A Brief Introduction to Model-Based Inference of Phylogeny

Brian R. Moore & Sebastian Höhna
Department of Evolution and Ecology
University of California, Bervis
Bodega Workshop, 2017

Outline

I. Phylogenetic inference as a statistical problem

Pose a question, build a model, collect some data, estimate parameters

II. Phylogenetic data

What are the relevant observations?

III. Anatomy of a phylogenetic model

All models include three main components

IV. Introduction to basic probability theory

Making friends with some useful math

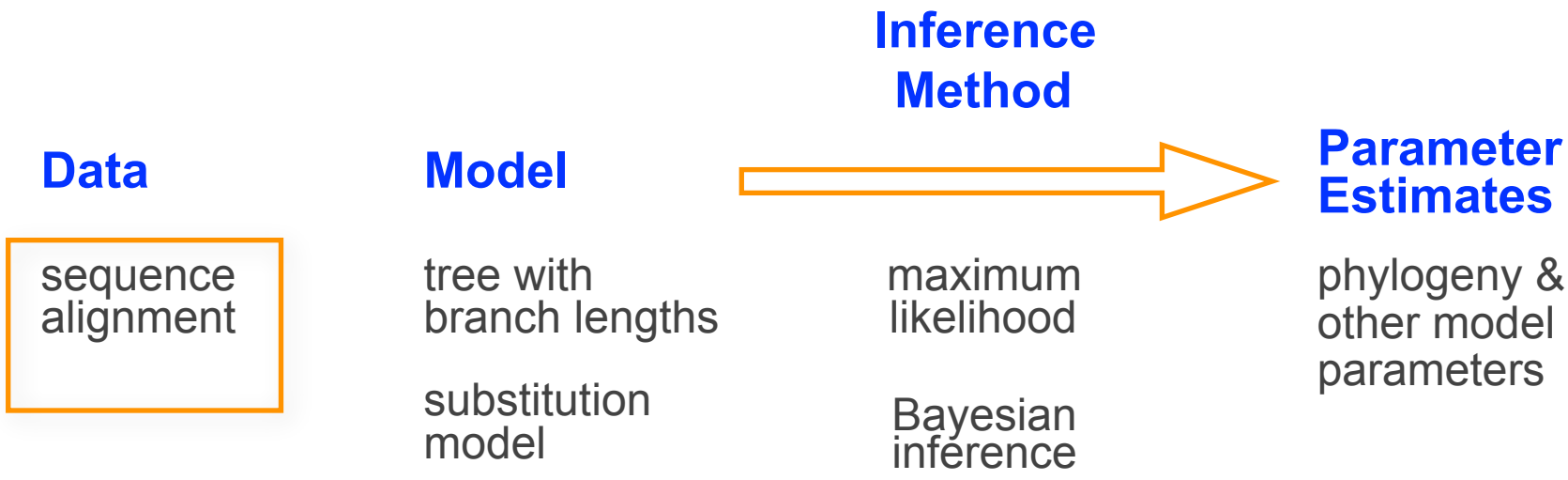
V. Models of discrete character change

Continuous-time Markov what now?

Statistical Estimation of Phylogeny: An Outline

Generic statistical paradigm	Statistical phylogenetic paradigm
pose a substantive question	what is the phylogeny of my study group?
develop a stochastic model with parameters that, if know, would answer the question	develop a phylogenetic model with a tree (and branch lengths) and a Markov model describing how traits change over the tree
collect observations that are informative about model parameters	assemble a data matrix (*e.g.*, of DNA sequences) sampled from members of your study group
find the best estimate of model parameters (by some means) conditioned on (*i.e.*, given) the data at hand	find the best estimate of phylogeny (and other model parameters) using a likelihood-based method (maximum-likelihood or Bayesian inference)

Statistical Estimation of Phylogeny: An Outline



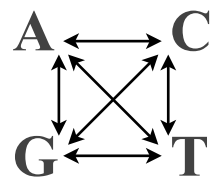
The Data: Nucleotide Sequences

I. The state space is comprised of the four nucleic acids

Purines

A (adenine)

G (guanine)



Pyrimidines

C (cytosine)

T (thymine)

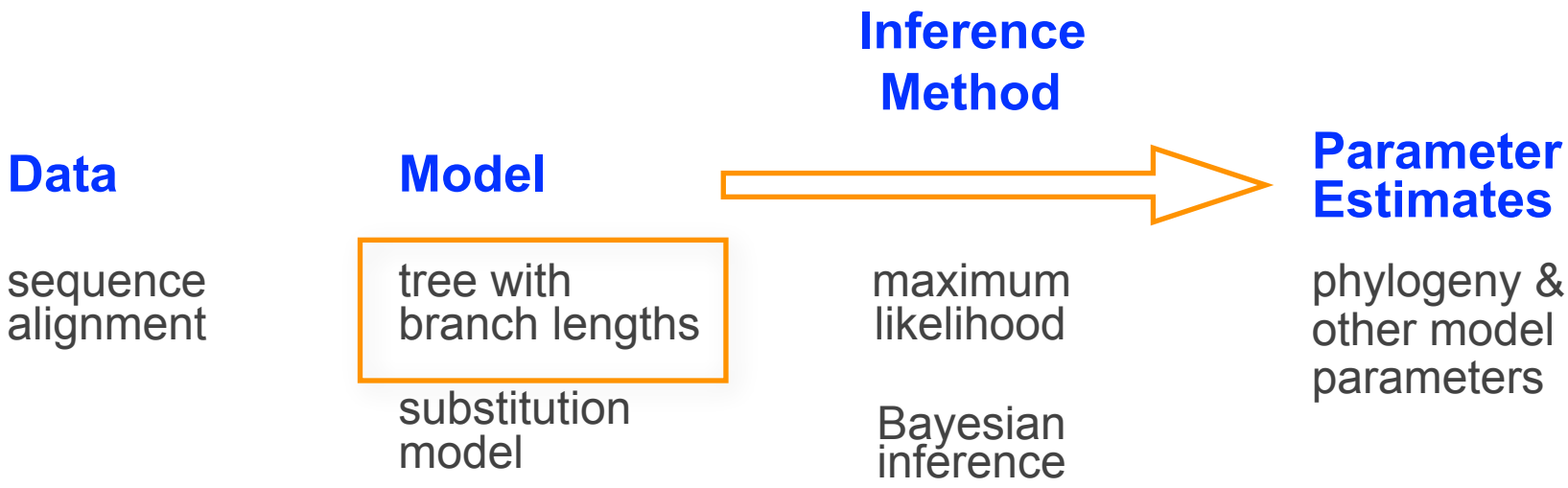
II. The sequences data are arranged in tables called data matrices

```
species 1     ACGCACCGGCGCAGTCA....
species 2     ACGTTCAGGCGGTCA....
species 3     ACGTTCACCGGCGCAGTCA....
species 4     ACGTTCACCCGCAGTCA....
```

III. The process of establishing homology is referred to as alignment

```
species 1     ACG--CACCGGCGCAGTCA....
species 2     ACGTTCA--GGCG--GTCA....
species 3     ACGTTCACCGGCGCAGTCA....
species 4     ACGTTCACC--CGCAGTCA....
```

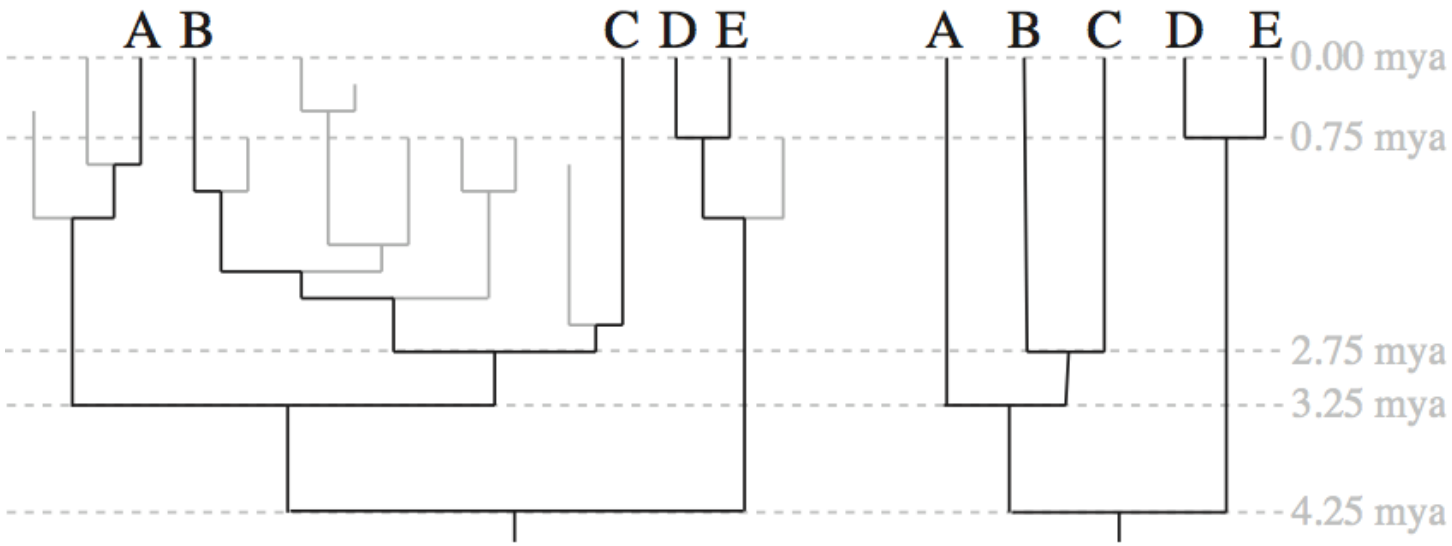
Statistical Estimation of Phylogeny: An Outline



The Model: Phylogeny Parameter

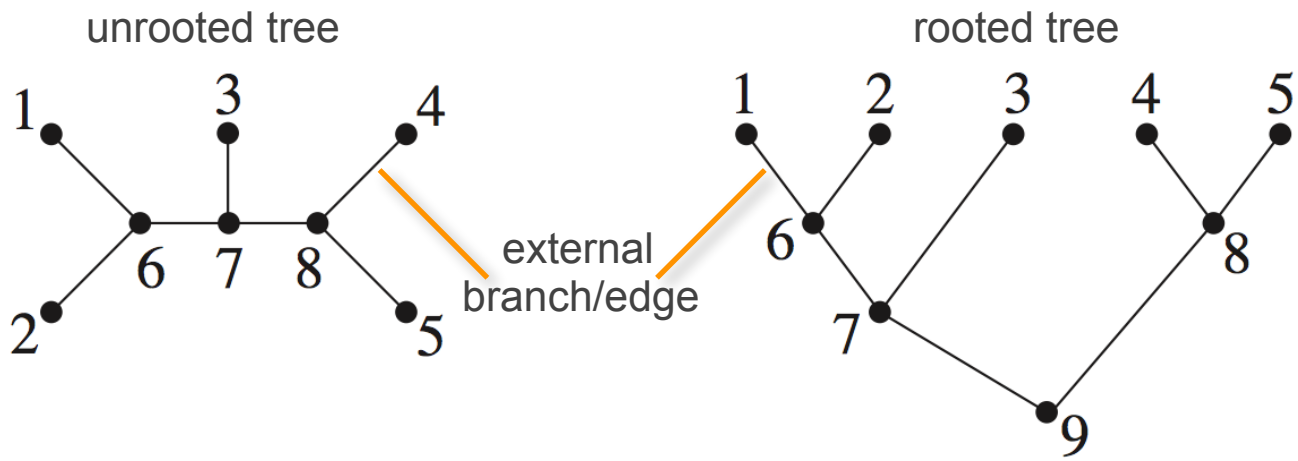
I. Phylogenies are estimates of genealogical (evolutionary) relationships

The topology specifies a nested set of common-ancestry relationships



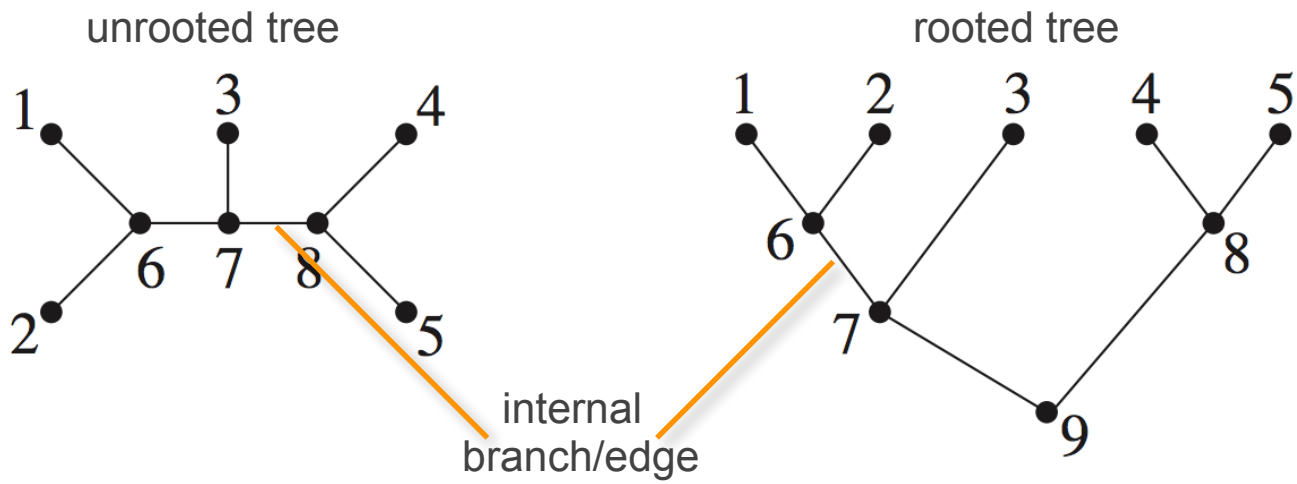
The Model: Phylogeny Parameter

Tree terms & concepts



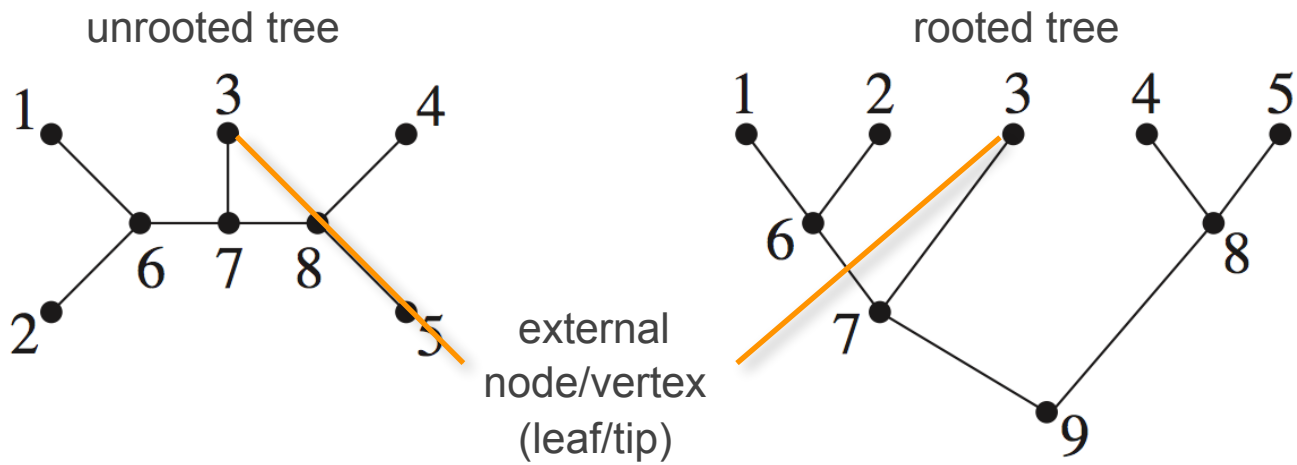
The Model: Phylogeny Parameter

Tree terms & concepts



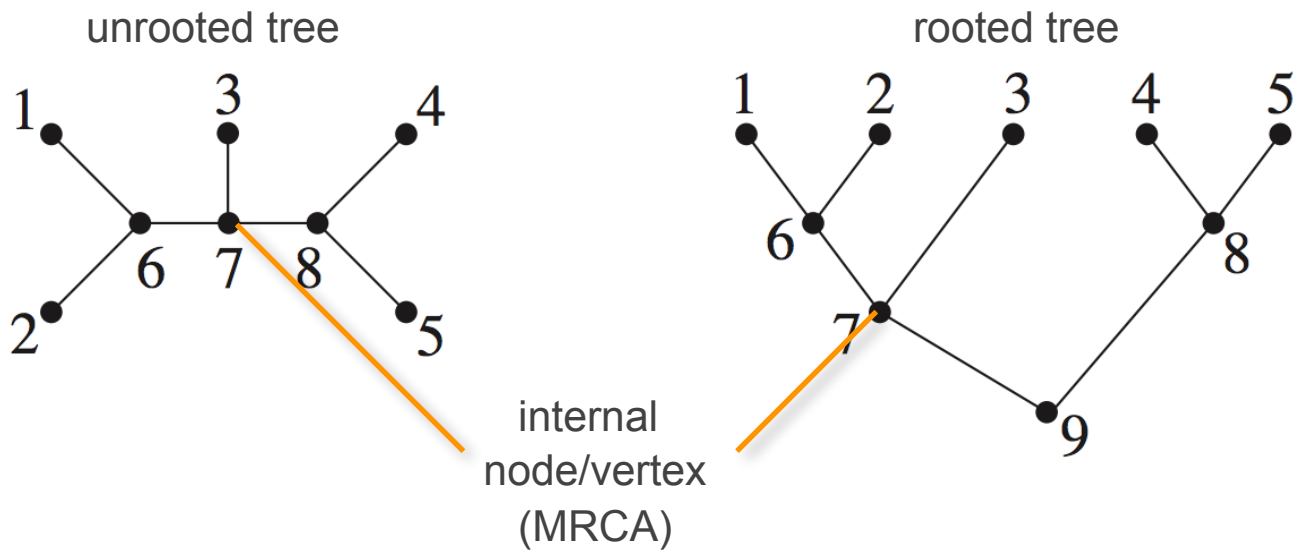
The Model: Phylogeny Parameter

Tree terms & concepts



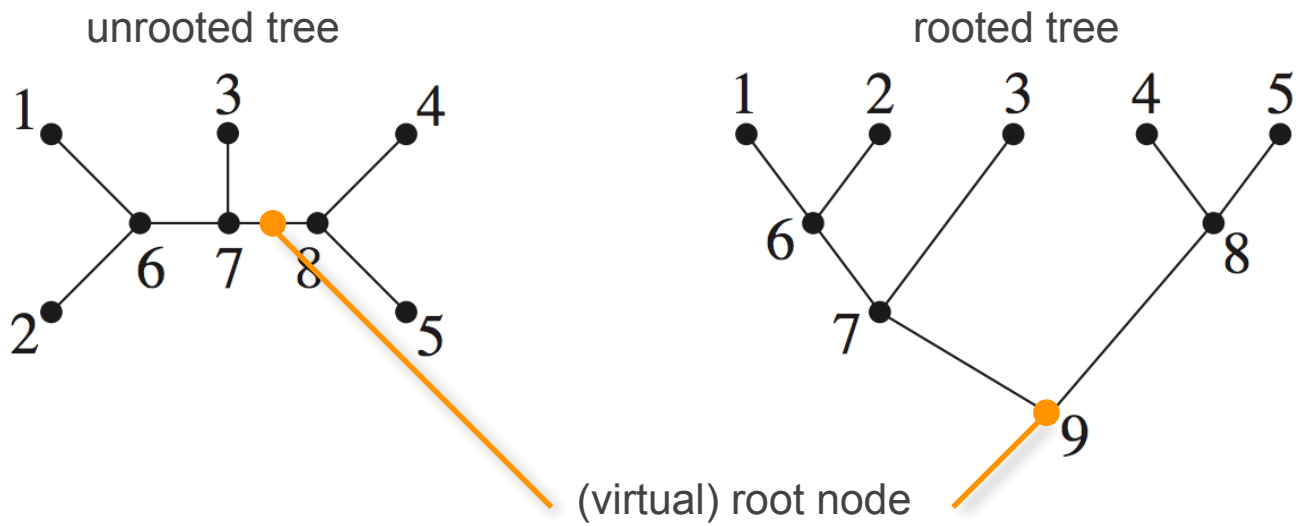
The Model: Phylogeny Parameter

Tree terms & concepts



The Model: Phylogeny Parameter

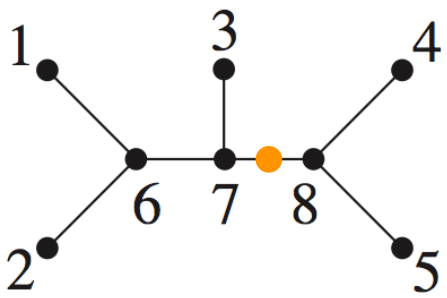
Tree terms & concepts



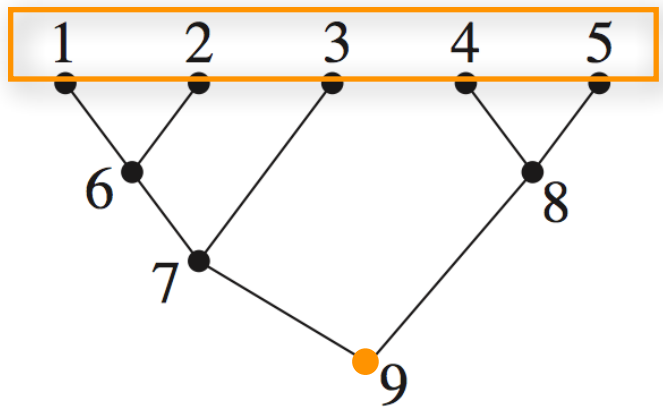
The Model: Phylogeny Parameter

Tree terms & concepts

unrooted tree



rooted tree



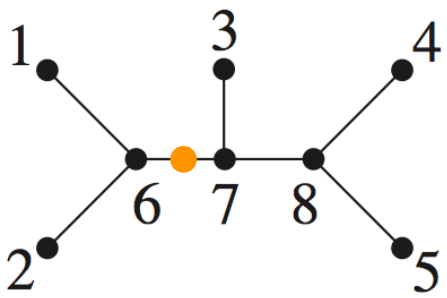
Phylogenies specify a temporal dimension...unrooted trees do not

Q. Are unrooted trees phylogenies?

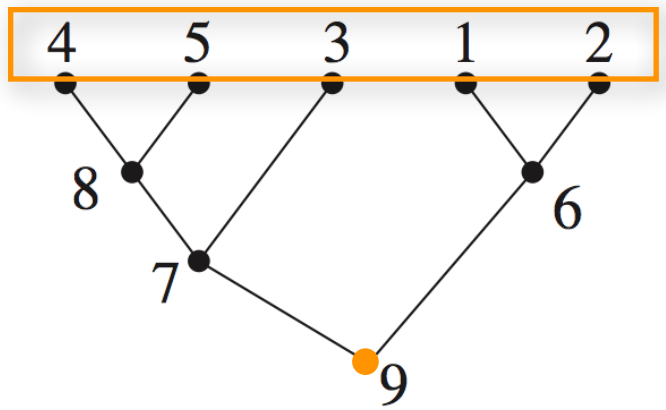
The Model: Phylogeny Parameter

Tree terms & concepts

unrooted tree



rooted tree



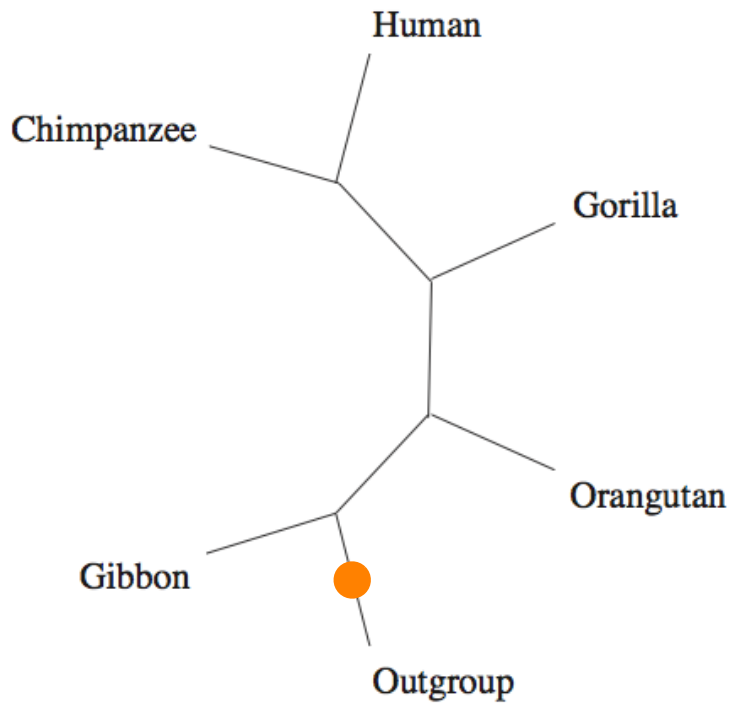
Phylogenies specify a temporal dimension...unrooted trees do not

Q. Are unrooted trees phylogenies?

A. No, because they constrain but do not specify relationships

The Model: Phylogeny Parameter

Tree terms & concepts



Typically, we root trees using the outgroup method

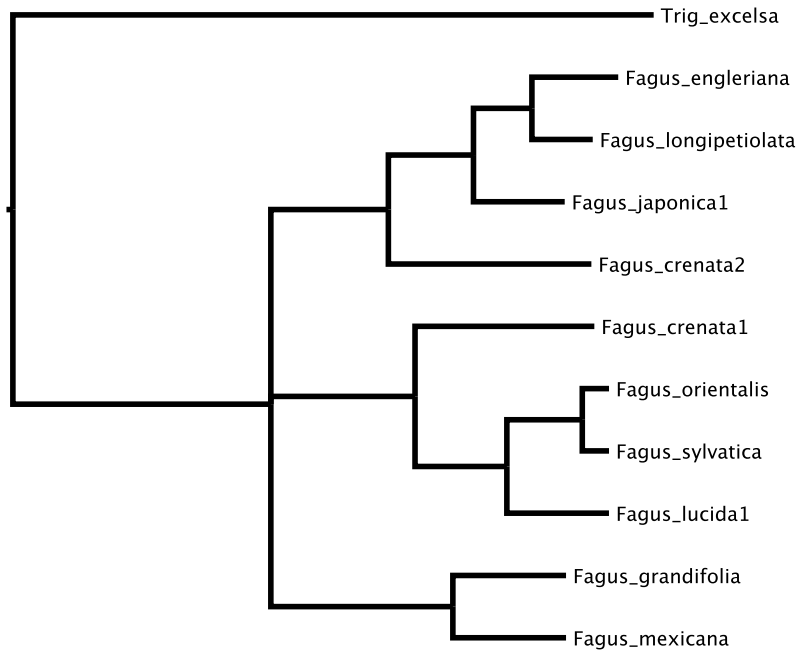
The Model: Phylogeny Parameter

I. Phylogenies are estimates of genealogical (evolutionary) relationships

The topology specifies a nested set of common-ancestry relationships

II. The branches of trees may reflect different quantities

The type of phylogeny depends on the nature of the branch lengths it depicts



Phylogram: expected number of substitutions/site

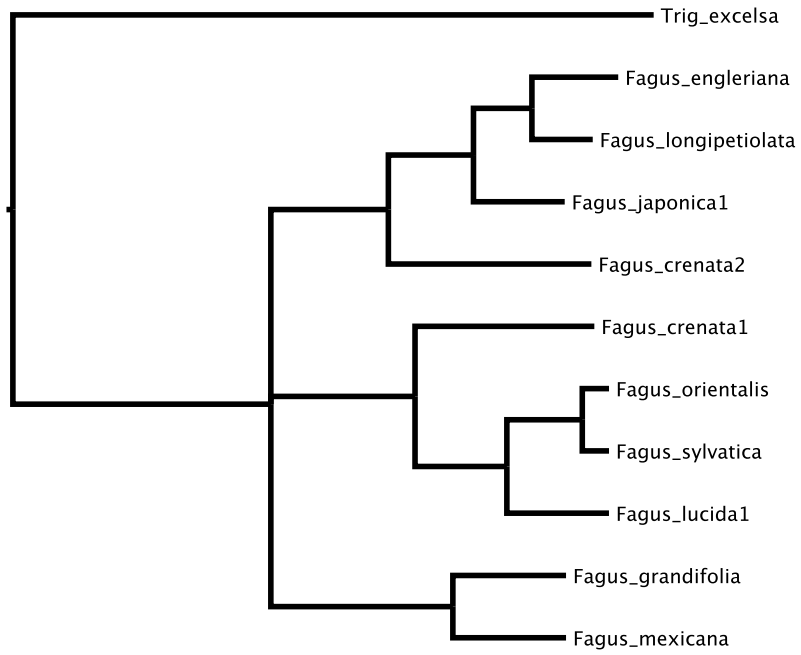
The Model: Phylogeny Parameter

I. Phylogenies are estimates of genealogical (evolutionary) relationships

The topology specifies a nested set of common-ancestry relationships

II. The branches of trees may reflect different quantities

The type of phylogeny depends on the nature of the branch lengths it depicts



Chronogram: relative/absolute branch durations and node ages

The Model: Phylogeny Parameter

I. Phylogenies are estimates of genealogical (evolutionary) relationships

The topology specifies a nested set of common-ancestry relationships

II. The branches of trees may reflect different quantities

The type of phylogeny depends on the nature of the branch lengths it depicts

III. Phylogenies are unusual and difficult parameters to model

They are discrete in nature (directed acyclic binary graphs; aka ‘DAGS’)

The Model: Phylogeny Parameter

I. Phylogenies are estimates of genealogical (evolutionary) relationships

The topology specifies a nested set of common-ancestry relationships

II. The branches of trees may reflect different quantities

The type of phylogeny depends on the nature of the branch lengths it depicts

III. Phylogenies are unusual and difficult parameters to model

They are discrete in nature (directed acyclic binary graphs; aka ‘DAGS’)

The combinatorics are such that the solution space can be vast

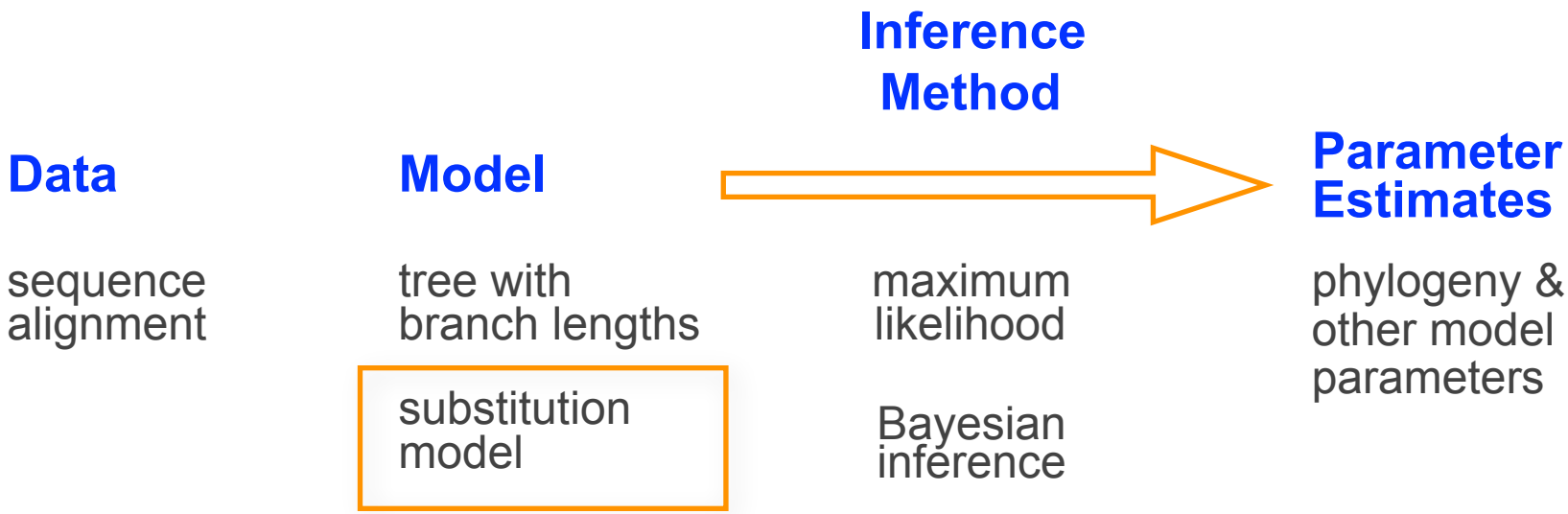
The Model: Phylogeny Parameter

How many trees are there?

s	$T_U(s)$	$T_R(s)$
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	2.216×10^{20}	8.201×10^{21}
50	2.838×10^{74}	2.753×10^{76}
100	1.700×10^{182}	3.350×10^{184}
500	1.012×10^{1277}	1.008×10^{1280}
1000	1.927×10^{2860}	3.848×10^{2863}

45 species

Statistical Estimation of Phylogeny: An Outline



The Model: Character Evolution Model

Continuous-time Markov Chains (CTMC)

Evolution of discrete traits (*e.g.*, substitution models, morphological models)

Diffusion model

Evolution of continuous traits (*e.g.*, Brownian, OU, Levey process models)

Outline

I. Phylogenetic inference as a statistical problem

Pose a question, build a model, collect some data, estimate parameters

II. Phylogenetic data

What are the relevant observations?

III. Anatomy of a phylogenetic model

All models include three main components

IV. Introduction to basic probability theory

Making friends with some useful math

V. Models of discrete character change

Continuous-time Markov what now?

A Brief Probabilistic Digression

Conditional Probability

The probability of observing A given that B has occurred, $\Pr(A \mid B)$, is just the fraction of cases in which B occurs, $\Pr(B)$, that A also occurs, $\Pr(A,B)$.

$$\Pr(CDN|male) = \frac{\Pr(CDN,male)}{\Pr(male)} = \frac{3/40}{15/40} = 0.2$$

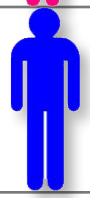
			Σ
	23/40	2/40	25/40
	12/40	3/40	15/40
Σ	35/40	5/40	40/40

What is the probability that a student is Canadian, given that he is male?

A Brief Probabilistic Digression

Joint Probability

The probability of observing A and B, $\Pr(A,B)$, can be determined by rearranging the expression for conditional likelihood.

			Σ
	23/40	2/40	25/40
	12/40	3/40	15/40
Σ	35/40	5/40	40/40

$$\Pr(A,B) = \Pr(A|B)\,\Pr(B)$$
$$= \Pr(B|A)\Pr(A)$$

$$\Pr(CDN,male) = \Pr(CDN|male)\,\Pr(male)$$
$$= 3/15 \text{ x } 15/40$$
$$= 3/40 = 0.075$$
$$= \Pr(male|CDN)\,\Pr(CDN)$$
$$= 3/5 \text{ x } 5/40$$
$$= 3/40 = 0.075$$

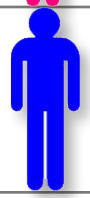
*What is the probability that a student is **both** male **and** Canadian?*

A Brief Probabilistic Digression

Marginal Probability

The unconditional probability of an observation A or B

$$\Pr(CDN) = 2/40 + 3/40$$
$$= 5/40 = 0.375$$

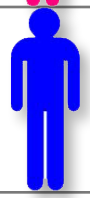
			Σ
	23/40	2/40	25/40
	12/40	3/40	15/40
Σ	35/40	5/40	40/40

What is the marginal probability of being Canadian?

A Brief Probabilistic Digression

The Law of Total Probability

The unconditional probabilities sum to unity

			Σ
	23/40	2/40	25/40
	12/40	3/40	15/40
Σ	35/40	5/40	40/40

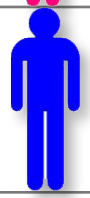
$$\begin{aligned}\Pr(total) &= 25/40 + 15/40 \\ &= 35/40 + 5/40 \\ &= 40/40 = 1.0\end{aligned}$$

What is the probability of all possible combinations of gender/nationality?

A Brief Probabilistic Digression

Statistical Independence

The two events are independent if the product of their marginal probabilities equals the joint probability of those two events.

			Σ
	23/40	2/40	25/40
	12/40	3/40	15/40
Σ	35/40	5/40	40/40

$$\Pr(CDN, male) \stackrel{?}{=} \Pr(CDN)\,\Pr(male)$$

$$3/40 \stackrel{?}{=} 5/40 \times 15/40$$

$$0.075 \neq 0.047$$

Are the probabilities of being male and Canadian independent?

Outline

I. Phylogenetic inference as a statistical problem

Pose a question, build a model, collect some data, estimate parameters

II. Phylogenetic data

What are the relevant observations?

III. Anatomy of a phylogenetic model

All models include three main components

IV. Introduction to basic probability theory

Making friends with some useful math

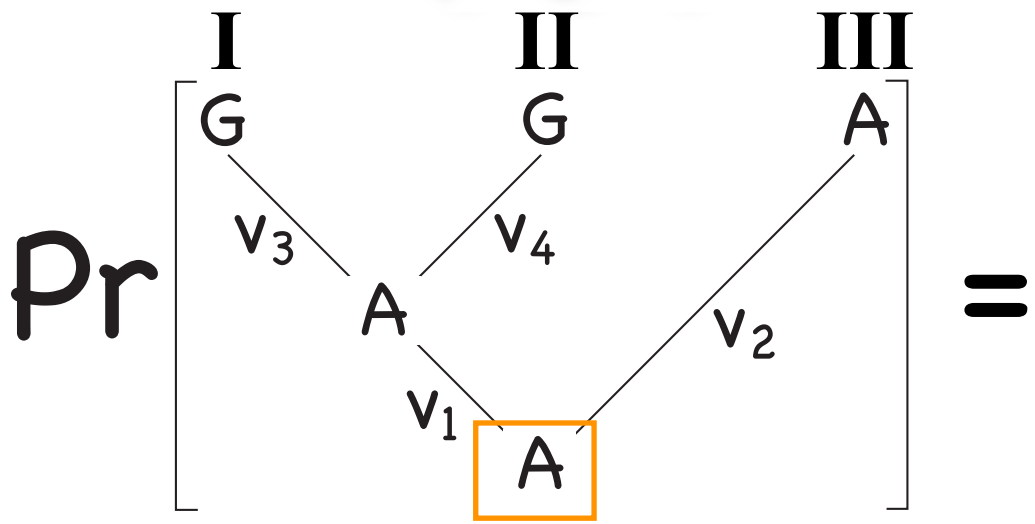
V. Models of discrete character change

Continuous-time Markov what now?

An Overview of Phylogenetic Inference

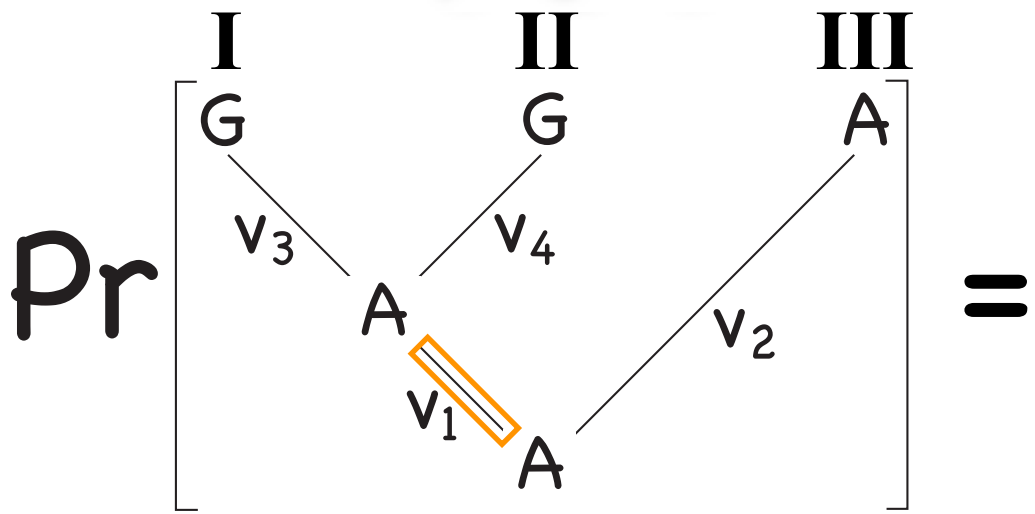
Species	Sequence data
Species I	GCG--CACCGGCGCAGTCA....
Species II	GCGTTCA--GGCG--GTCA....
Species III	ACGTTCACCGGCGCAGTCA....

An Overview of Phylogenetic Inference



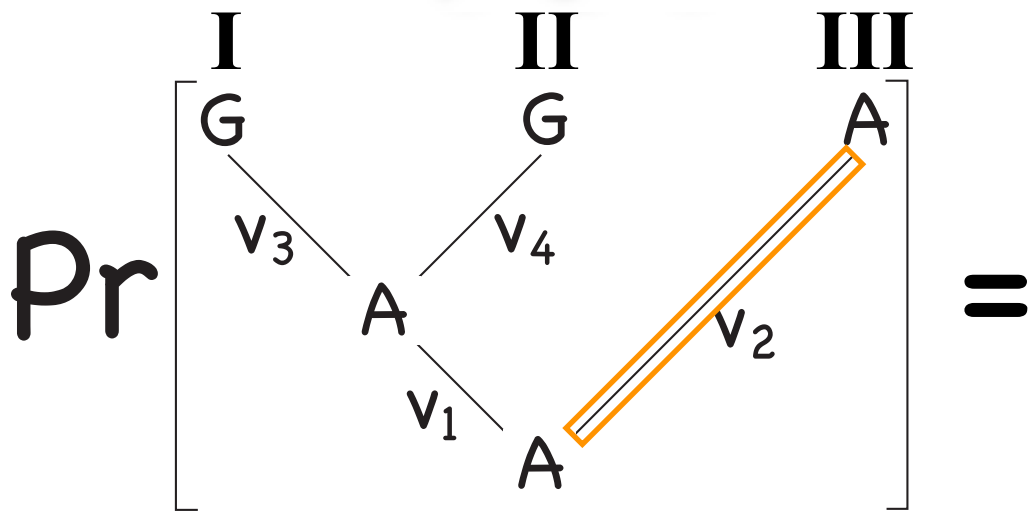
π_A

An Overview of Phylogenetic Inference



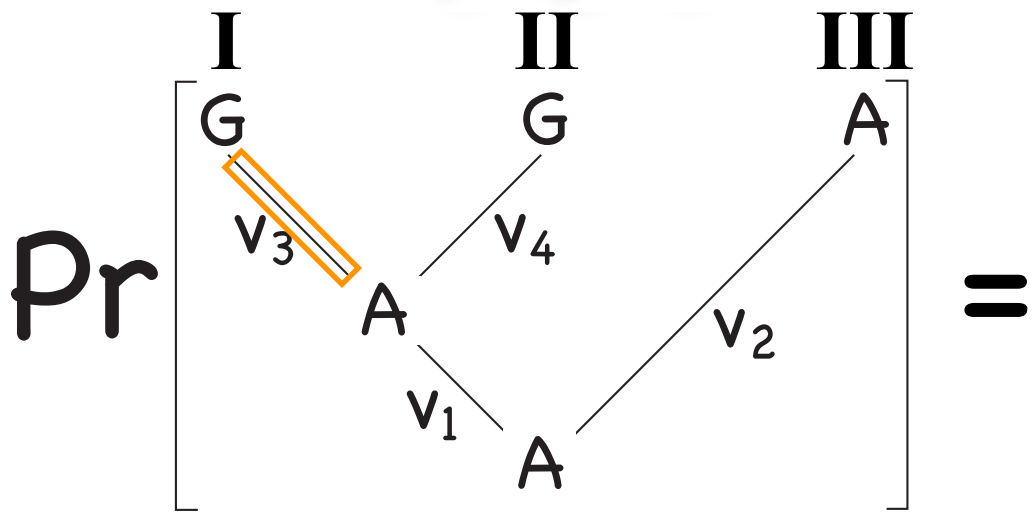
$$\pi_A \times p_{AA}(v_1)$$

An Overview of Phylogenetic Inference



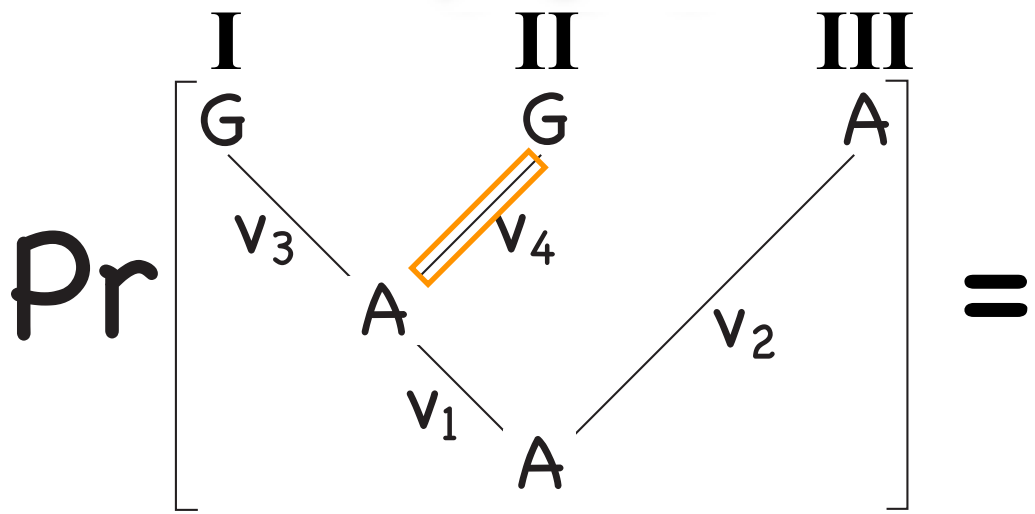
$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2)$$

An Overview of Phylogenetic Inference



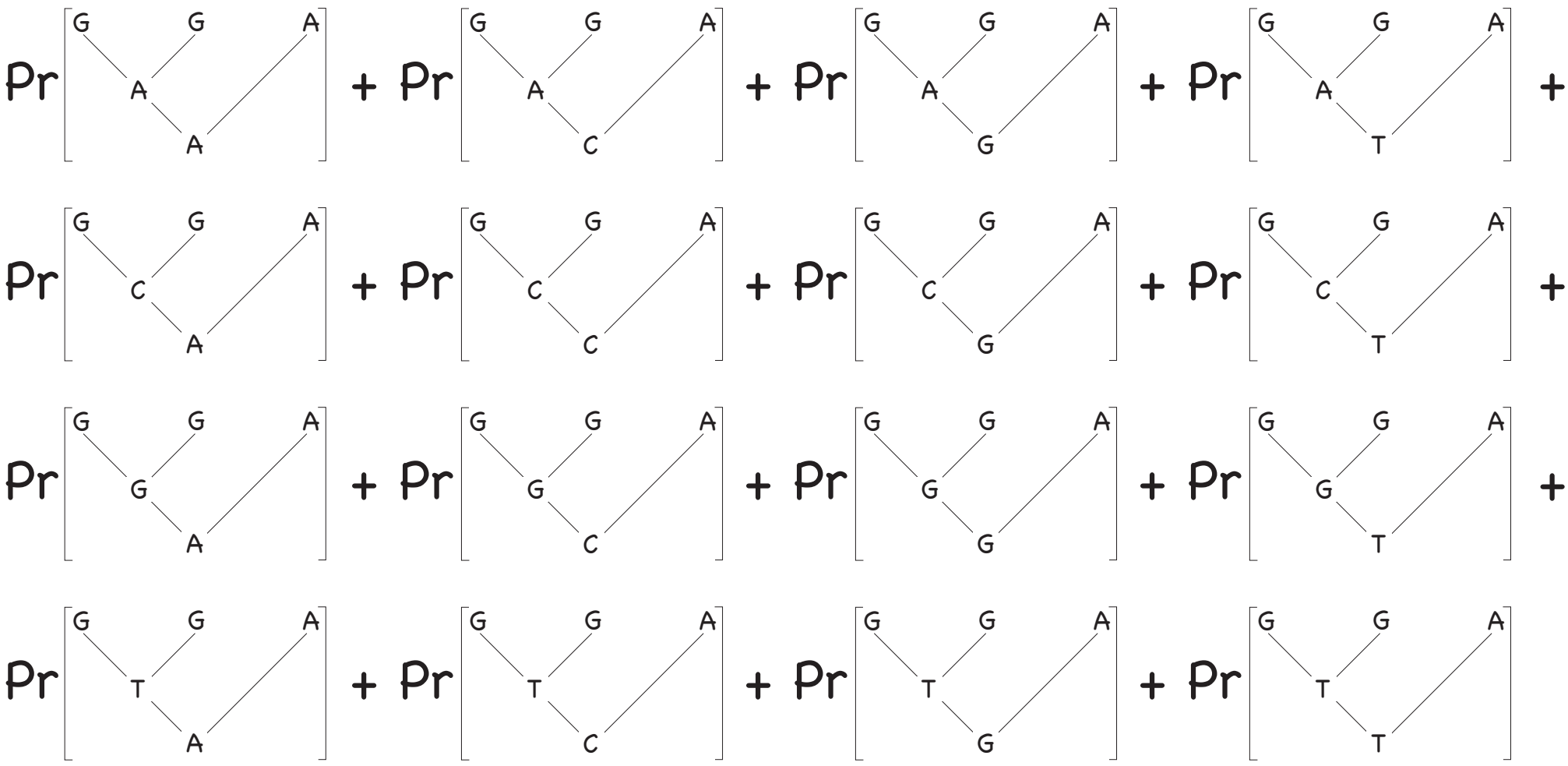
$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3)$$

An Overview of Phylogenetic Inference

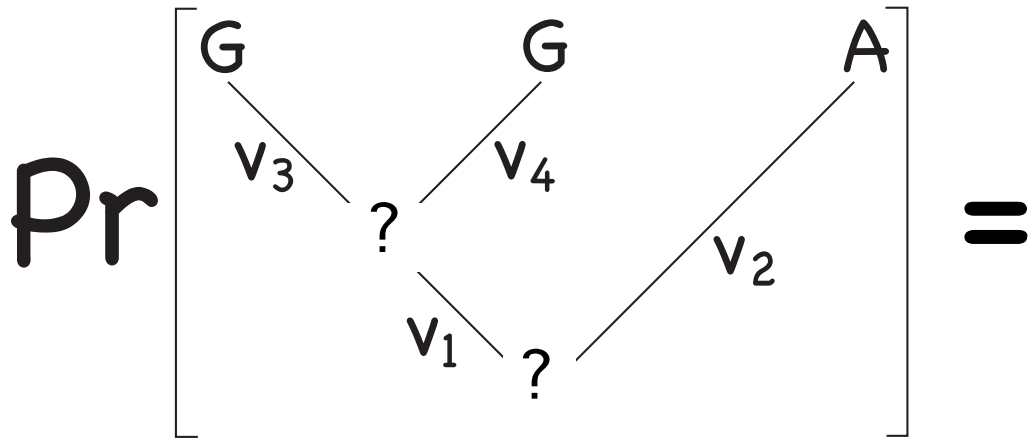


$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3) \times p_{AG}(v_4)$$

An Overview of Phylogenetic Inference



An Overview of Phylogenetic Inference



$$\pi_i \times p_{ij}(v_1) \times p_{iA}(v_2) \times p_{jG}(v_3) \times p_{jG}(v_4)$$

π_i	Stationary frequencies
$p_{ij}(v)$	Transition probabilities

Stochastic Models of Nucleotide Substitution

Continuous-time Markov Models

- Character change (nucleotide substitution) is modeled as a continuous-time Markov chain (CTMC)
- Stochastic model in which the next state of the chain depends only on the current state

The model is central to model-based inference

- Even if the parameters of the substitution model are not of direct interest, they are nevertheless critical to estimation of the focal model parameters

The Instantaneous-Rate Matrix

A Continuous-time Markov model is defined by a matrix of substitution rates

A table that specifies the rates of all possible changes between states.

The rate matrix allows us to calculate important quantities:

The probability of observing a substitution over a specified interval

The probability of observing the process in a specified state

The Instantaneous-Rate Matrix

A hypothetical instantaneous-rate matrix

		To			
		A	C	G	T
From	A	−1.916	0.541	0.787	0.588
	C	0.148	−1.069	0.415	0.506
	G	0.286	0.170	−0.591	0.135
	T	0.525	0.236	0.594	−1.355

This table of rates specifies the instantaneous rate of change between sates.

The rates are in terms of the expected number of substitutions per site.

The rates are scaled so that the average rate of substitution is one.

A Mechanistic Interpretation of the Rate Matrix

A hypothetical instantaneous-rate matrix

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

If the current state of the Markov chain is i, the next substitution will occur after an exponentially distributed waiting time with rate parameter -q_{ii}

A Mechanistic Interpretation of the Rate Matrix

A hypothetical instantaneous-rate matrix

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

If the current state of the Markov chain is i, the next substitution will occur after an exponentially distributed waiting time with rate parameter $-q_{ii}$

We can read this rate parameter directly from the instantaneous-rate matrix: e.g., if the current state is C, the rate parameter is $-q_{CC} = -(-1.069) = 1.069$, or equivalently $q_{CA} + q_{CG} + q_{CT} = 0.148 + 0.415 + 0.506 = 1.069$

When an event occurs, the rate matrix also specifies the probabilities of all possible substitutions: $P(i \to j) = q_{ij} / -q_{ii}$

$$P(C \to A) = q_{CA} / -q_{CC} = 0.148 \div 1.069 = 0.138$$
$$P(C \to G) = q_{CG} / -q_{CC} = 0.415 \div 1.069 = 0.388$$
$$P(C \to T) = q_{CT} / -q_{CC} = 0.506 \div 1.069 = 0.474$$
$$\sum P_{ij} = 1.0$$

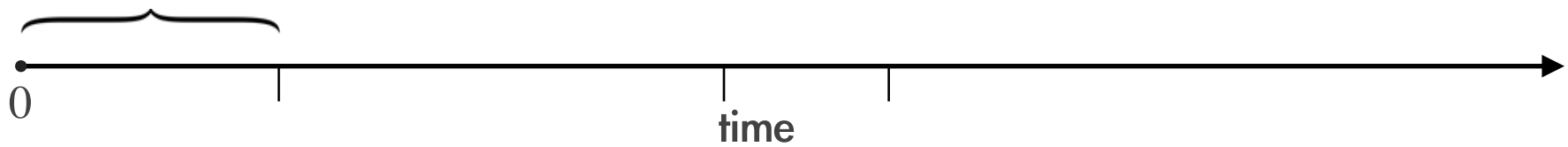
Another Probabilistic Digression

Waiting times in a Poisson process

Some rare, discrete event that occurs at a constant rate in continuous time is described by a Poisson process

These events occur with a rate λ

The waiting (sojourn) time for the first event



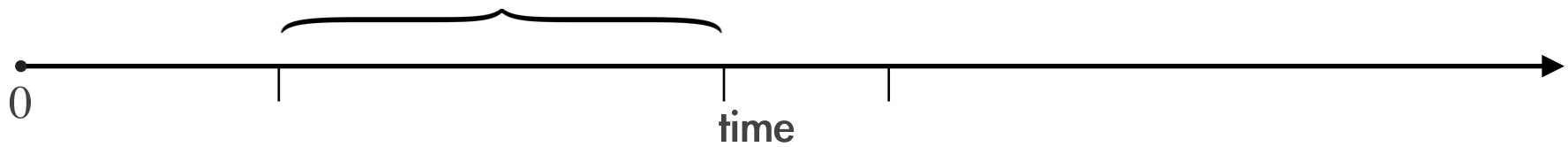
Another Probabilistic Digression

Waiting times in a Poisson process

Some rare, discrete event that occurs at a constant rate in continuous time is described by a Poisson process

These events occur with a rate λ

The waiting (sojourn) time for the second event



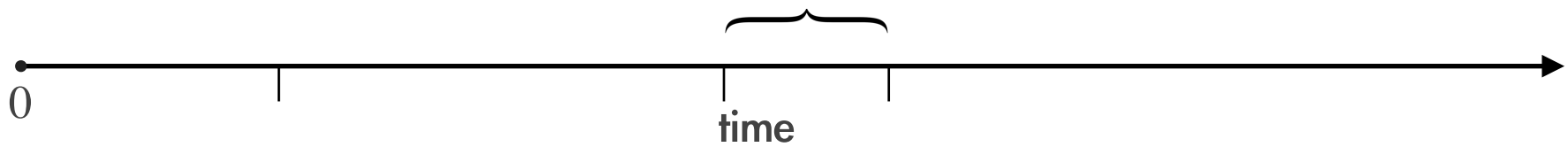
Another Probabilistic Digression

Waiting times in a Poisson process

Some rare, discrete event that occurs at a constant rate in continuous time is described by a Poisson process

These events occur with a rate λ

The waiting (sojourn) time for the third event



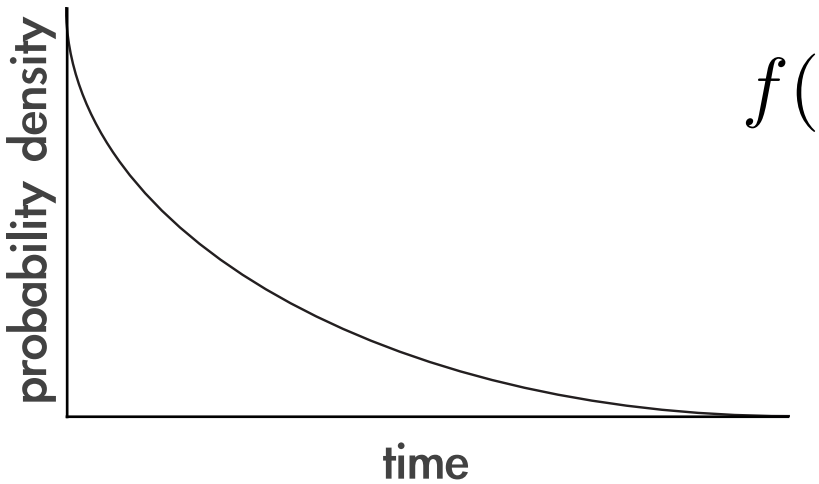
Another Probabilistic Digression

Waiting times in a Poisson process

Some rare, discrete event that occurs at a constant rate in continuous time is described by a Poisson process

These events occur with a rate λ

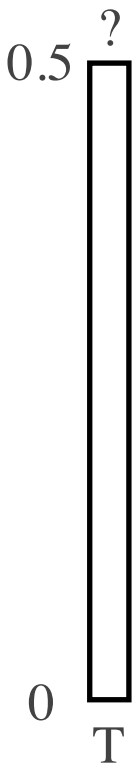
The waiting (sojourn) times are exponentially distributed random variables



$$f(t) = \lambda e^{-\lambda t}$$

A Mechanistic Interpretation of the Rate Matrix

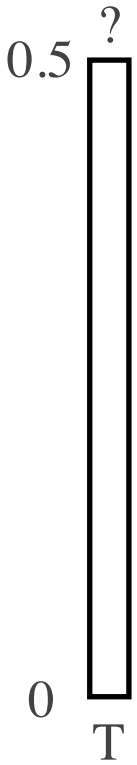
A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation

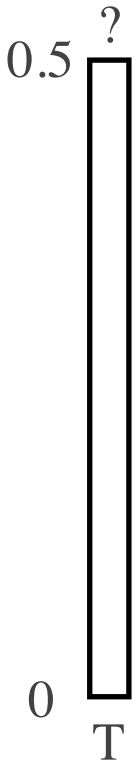


$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Rate of leaving the current state, $\mathrm{T} = 1.355$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



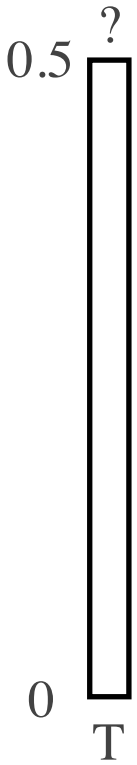
$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Probability of changing to A:

$$P(T \rightarrow A) = q_{TA} / -q_{TT} = 0.525 \div 1.355 = 0.387$$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

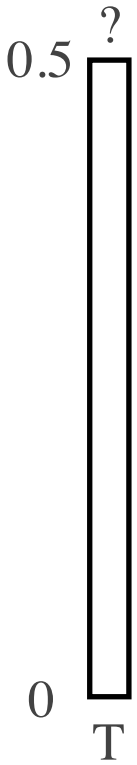
Probability of changing to C:

$$P(T \rightarrow A) = q_{TA} / -q_{TT} = 0.525 \div 1.355 = 0.387$$

$$P(T \rightarrow C) = q_{TC} / -q_{TT} = 0.236 \div 1.355 = 0.174$$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Probability of changing to G:

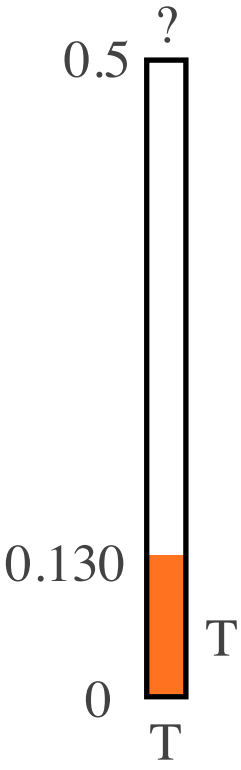
$$P(T \rightarrow A) = q_{TA} / -q_{TT} = 0.525 \div 1.355 = 0.387$$

$$P(T \rightarrow C) = q_{TC} / -q_{TT} = 0.236 \div 1.355 = 0.174$$

$$P(T \rightarrow G) = q_{TG} / -q_{TT} = 0.594 \div 1.355 = 0.438$$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate an exponentially distributed waiting time, x:

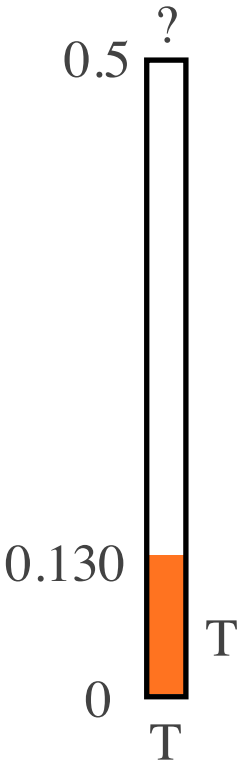
rate when process is in T: $\lambda = 1.355$

generate uniformly distributed random number with the die: $u = 0.839$

$x = -1/\lambda \ln(u) = 0.130$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate an exponentially distributed waiting time, x:

rate when process is in T: $\lambda = 1.355$

generate uniformly distributed random number with the die: $u = 0.839$

$x = -1/\lambda \ln(u) = 0.130$

Probabilities of substitution events in state T:

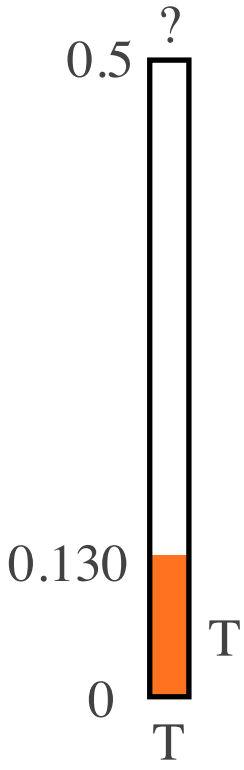
$P(T \rightarrow A) = q_{TA}/-q_{TT} = 0.525 \div 1.355 = 0.387$

$P(T \rightarrow C) = q_{TC}/-q_{TT} = 0.236 \div 1.355 = 0.174$

$P(T \rightarrow G) = q_{TG}/-q_{TT} = 0.594 \div 1.355 = 0.438$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate an exponentially distributed waiting time, x:

rate when process is in T: $\lambda = 1.355$

generate uniformly distributed random number with the die: $u = 0.839$

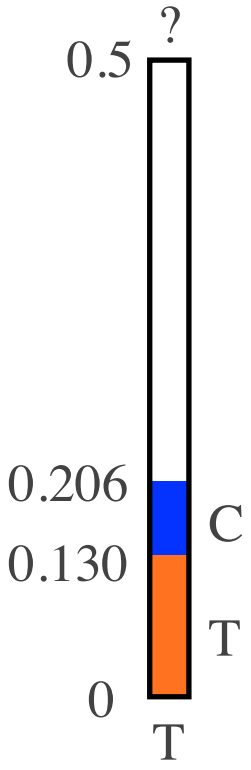
$x = -1/\lambda \ln(u) = 0.130$

Specify a set of intervals:	intervals	
$P(T \rightarrow A) = 0.387$	$0 - 0.387$	(choose A)
$P(T \rightarrow C) = 0.174$	$0.387 - 0.561$	(choose C)
$P(T \rightarrow G) = 0.438$	$0.561 - 1$	(choose T)

Generate a new uniformly distributed number, u, to select substitution event

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate waiting time to next event:

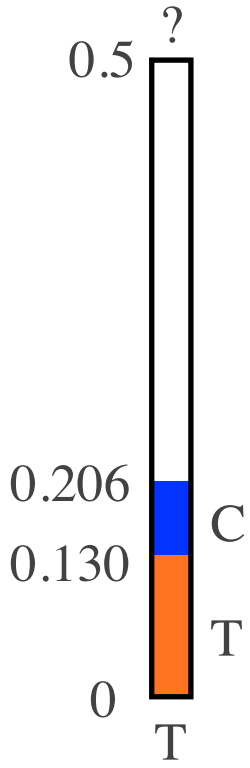
in state C: $-q_{CC} = \lambda = 1.069$

$u = \text{uniform}(0,1) = 0.922$

$x = -1/\lambda \ln(u) = 0.076$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate waiting time to next event:

in state C: $-q_{CC} = \lambda = 1.069$

$u = \text{uniform}(0,1) = 0.922$

$x = -1/\lambda \ln(u) = 0.076$

Substitution probabilities in state C:

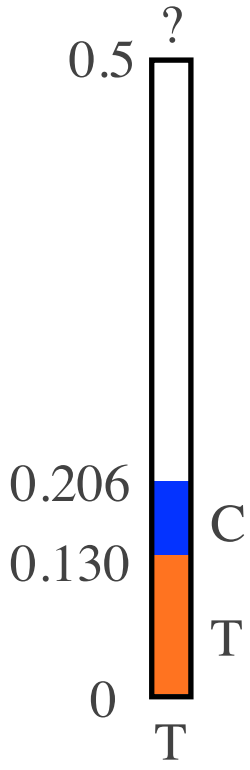
$P(C \rightarrow A) = q_{CA}/-q_{CC} = 0.148 \div 1.069 = 0.138$

$P(C \rightarrow G) = q_{CG}/-q_{CC} = 0.415 \div 1.069 = 0.388$

$P(C \rightarrow T) = q_{CT}/-q_{CC} = 0.506 \div 1.069 = 0.474$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate waiting time to next event:

in state C: $-q_{CC} = \lambda = 1.069$

$u = \text{uniform}(0,1) = 0.922$

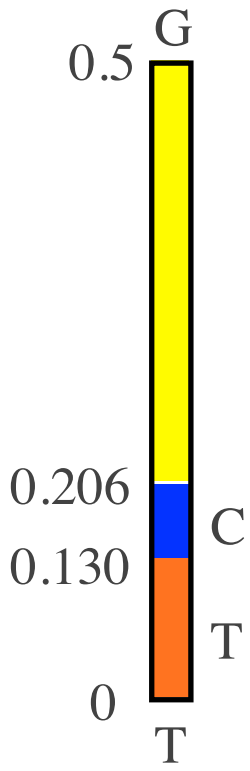
$x = -1/\lambda \ln(u) = 0.076$

Specify a set of intervals: **intervals**

$P(C \rightarrow A) = 0.138$	$0 - 0.138$	(choose A)
$P(C \rightarrow G) = 0.388$	$0.138 - 0.526$	(choose G)
$P(C \rightarrow T) = 0.474$	$0.526 - 1$	(choose T)

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate waiting time to next event:

in state G: $-q_{GG} = \lambda = 0.591$

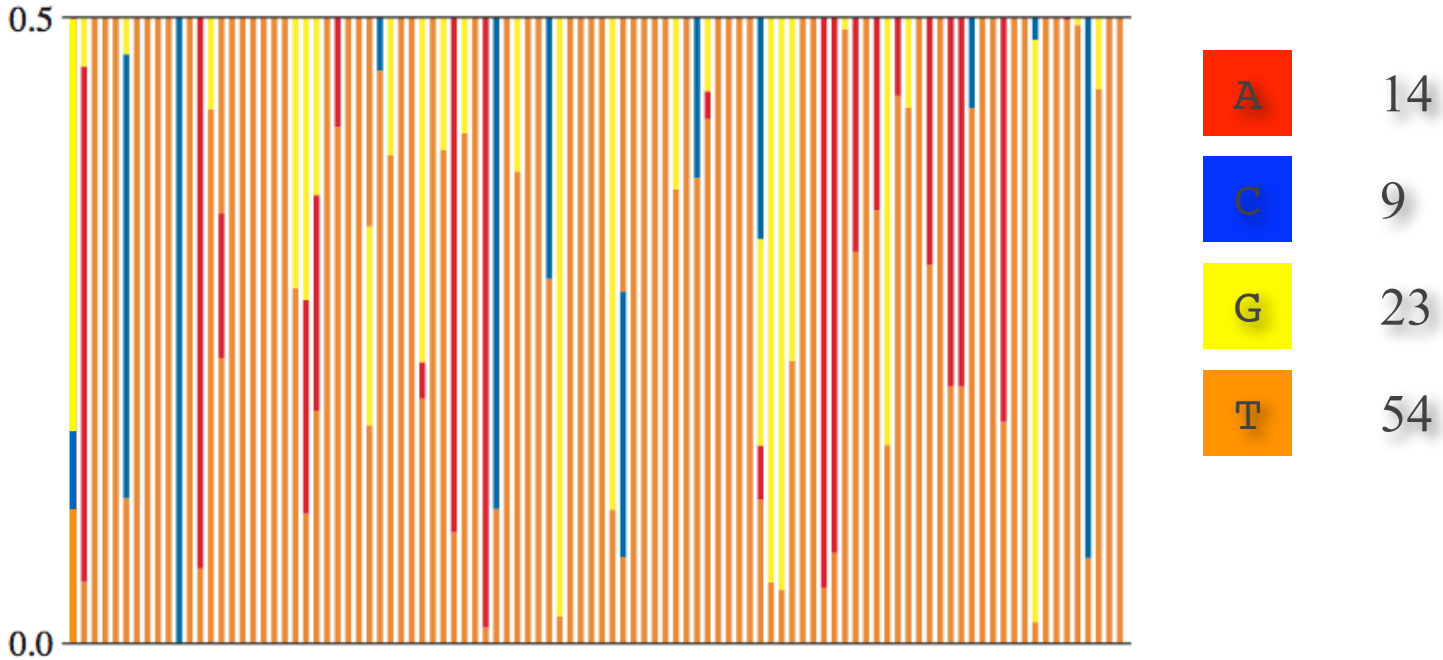
$u = \text{uniform}(0,1) = 0.341$

$x = -1/\lambda \ln(u) = 1.820$

Terminate simulation (in state G)

Transition Probabilities of a CTMC

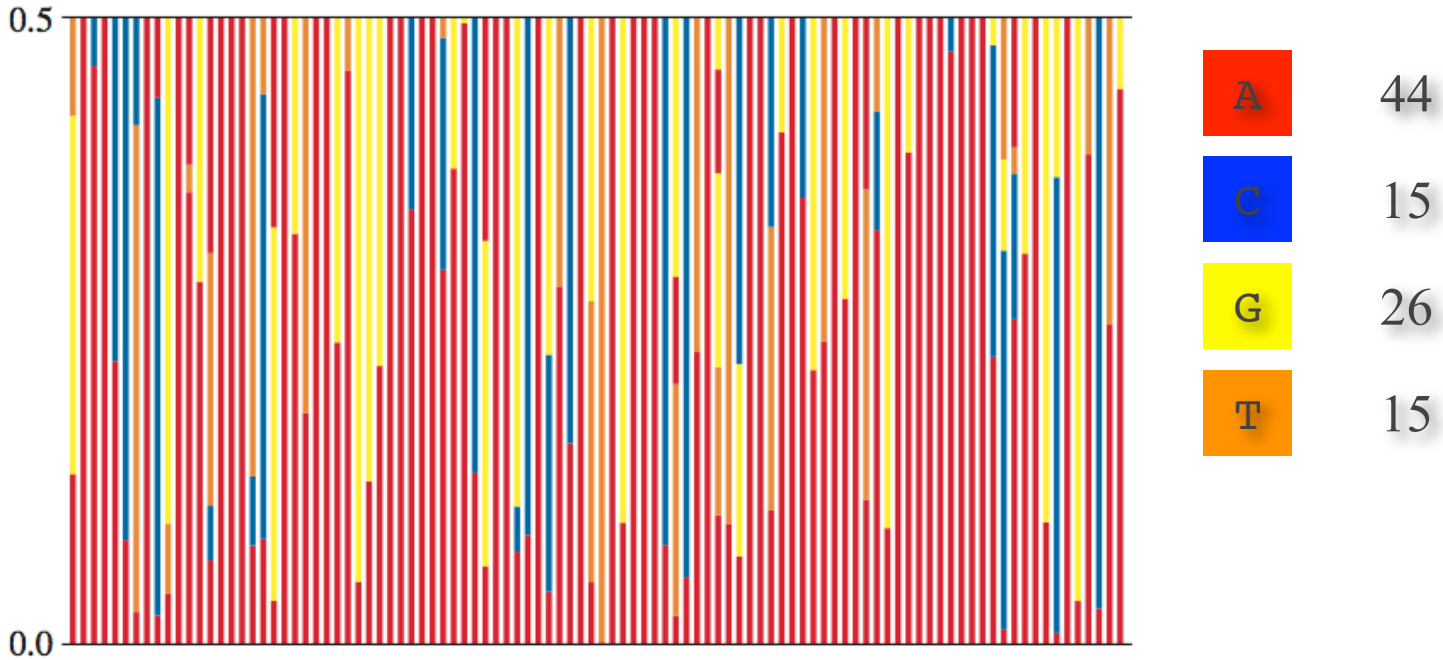
Estimating transition probabilities using Monte Carlo simulation



		To A	C	G	T
From	A				
	C				
	G				
	T	0.14	0.09	0.23	0.54

Transition Probabilities of a CTMC

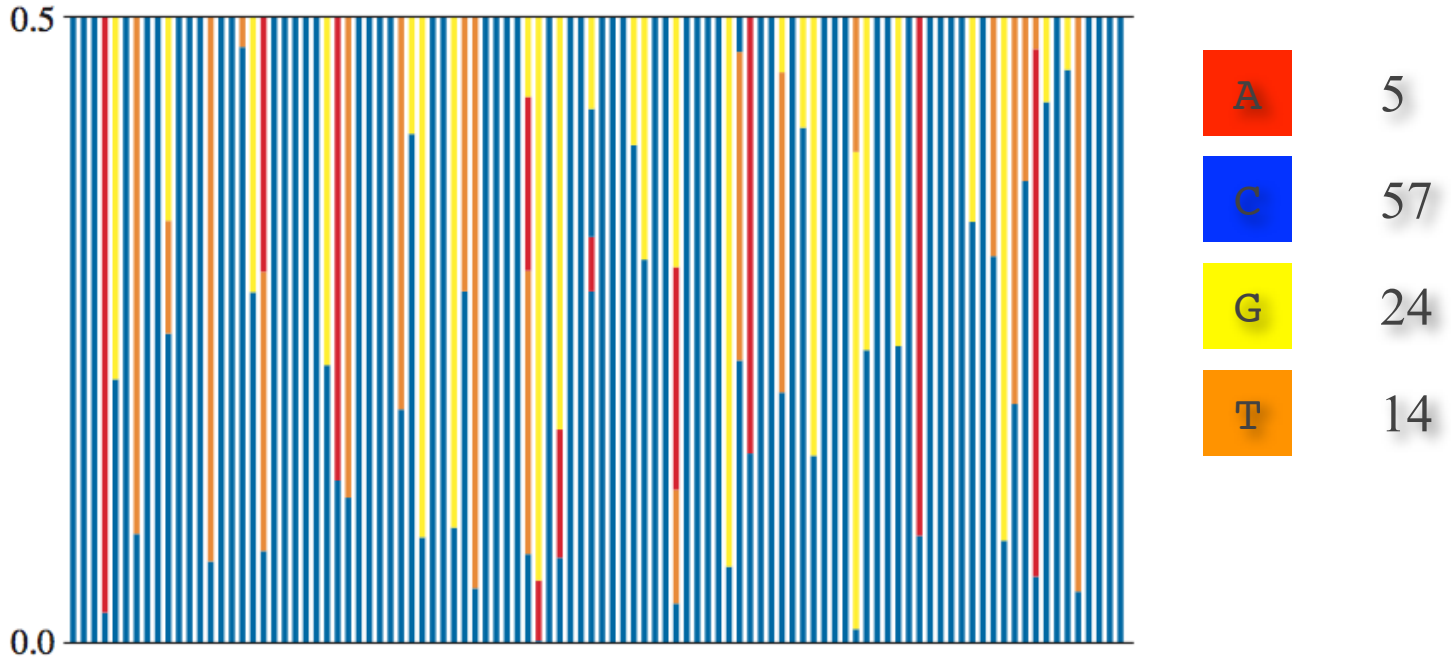
Realizations of 100 replicate simulations starting in state A



		To A	C	G	T
From	A	0.44	0.15	0.26	0.15
	C				
	G				
	T	0.14	0.09	0.23	0.54

Transition Probabilities of a CTMC

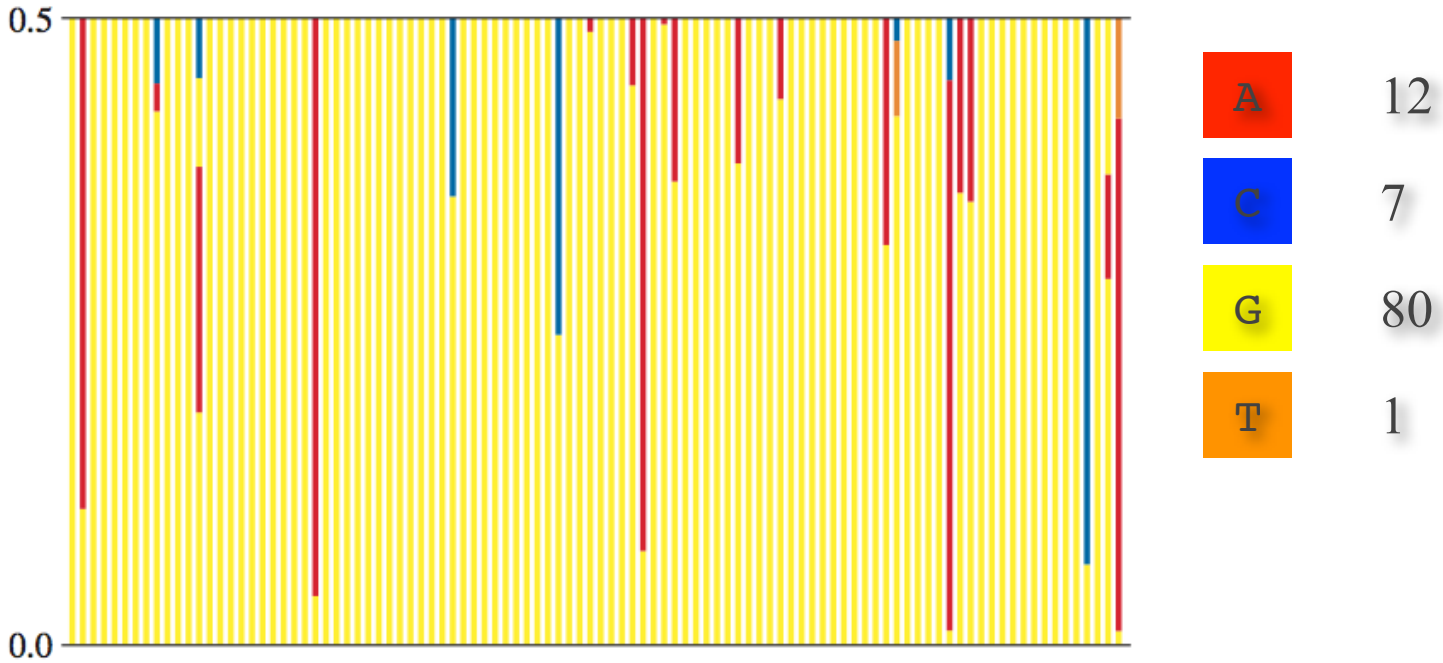
Realizations of 100 replicate simulations starting in state C



		To A	C	G	T
From	A	0.44	0.15	0.26	0.15
	C	0.05	0.57	0.24	0.14
	G				
	T	0.14	0.09	0.23	0.54

Transition Probabilities of a CTMC

Realizations of 100 replicate simulations starting in state G



		To A	C	G	T
From	A	0.44	0.15	0.26	0.15
	C	0.05	0.57	0.24	0.14
	G	0.12	0.07	0.80	0.01
	T	0.14	0.09	0.23	0.54

Transition Probabilities of a CTMC

Accuracy of Monte Carlo approximation depends on the number of replicates

100 replicates

		To A	C	G	T
From	A	0.44	0.15	0.26	0.15
	C	0.05	0.57	0.24	0.14
	G	0.12	0.07	0.80	0.01
	T	0.14	0.09	0.23	0.54

100,000 replicates

		To A	C	G	T
From	A	0.42119	0.15365	0.26361	0.16155
	C	0.06209	0.60811	0.17602	0.15378
	G	0.08834	0.07241	0.77796	0.06129
	T	0.13534	0.09411	0.22724	0.54331

Transition Probabilities of a CTMC

Exact solutions for the transition probabilities: matrix exponentiation

Monte Carlo simulation is computationally expensive and unnecessary, as the transition probabilities can be solved 'exactly'

The transition probability matrix, $\mathbf{P}$, can be solved by exponentiating the product of the instantaneous-rate matrix, $\mathbf{Q}$, and the branch length, v: $\mathbf{P}(v) = e^{\mathbf{Q}v}$

The exact solution for the transition probability matrix for our instantaneous-rate matrix and branch length (0.5) is:

$$\mathbf{P}(\nu) = \{p_{ij}(\nu)\} = \begin{pmatrix} 0.422927 & 0.153118 & 0.263330 & 0.160625 \\ 0.062896 & 0.609068 & 0.175153 & 0.152883 \\ 0.087566 & 0.071950 & 0.778271 & 0.062212 \\ 0.134967 & 0.093601 & 0.226962 & 0.544470 \end{pmatrix}$$

Compare with approximate solution (based on 100,000 replicates)

		To			
		A	C	G	T
From	A	0.42119	0.15365	0.26361	0.16155
	C	0.06209	0.60811	0.17602	0.15378
	G	0.08834	0.07241	0.77796	0.06129
	T	0.13534	0.09411	0.22724	0.54331

Transition Probabilities of a CTMC

An aside about transition probabilities

Transition probabilities account for all possible histories that a CTMC can end in a particular state, given a particular starting state (and fully specified model)

Transition probabilities play a key role in computing the likelihood, as they avoid the need to condition on a particular history of character change (nucleotide substitution)

Stationary Frequencies of a CTMC

Transition probabilities

The probability of observing state j conditioned on starting in state i and running the process over a branch of length v; i.e., $p_{ij}(v)$,

Can be estimated by Monte Carlo simulation or matrix exponentiation, $P(v) = e^{Qv}$

Stationary frequencies

The long-term probability of observing the process in state j

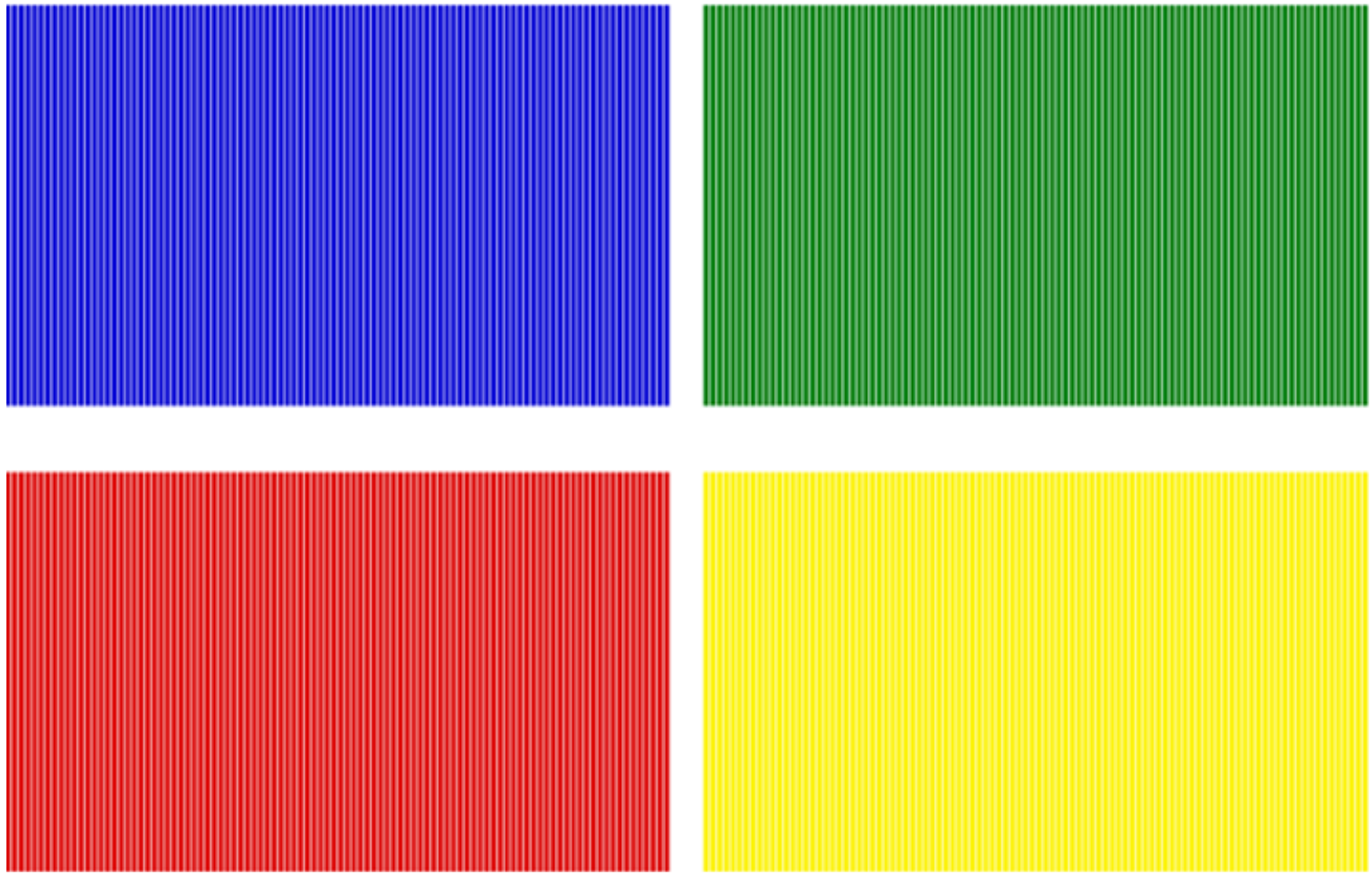
Our hypothetical rate matrix:

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Transition probabilities over a branch of length $v = 0.0$:

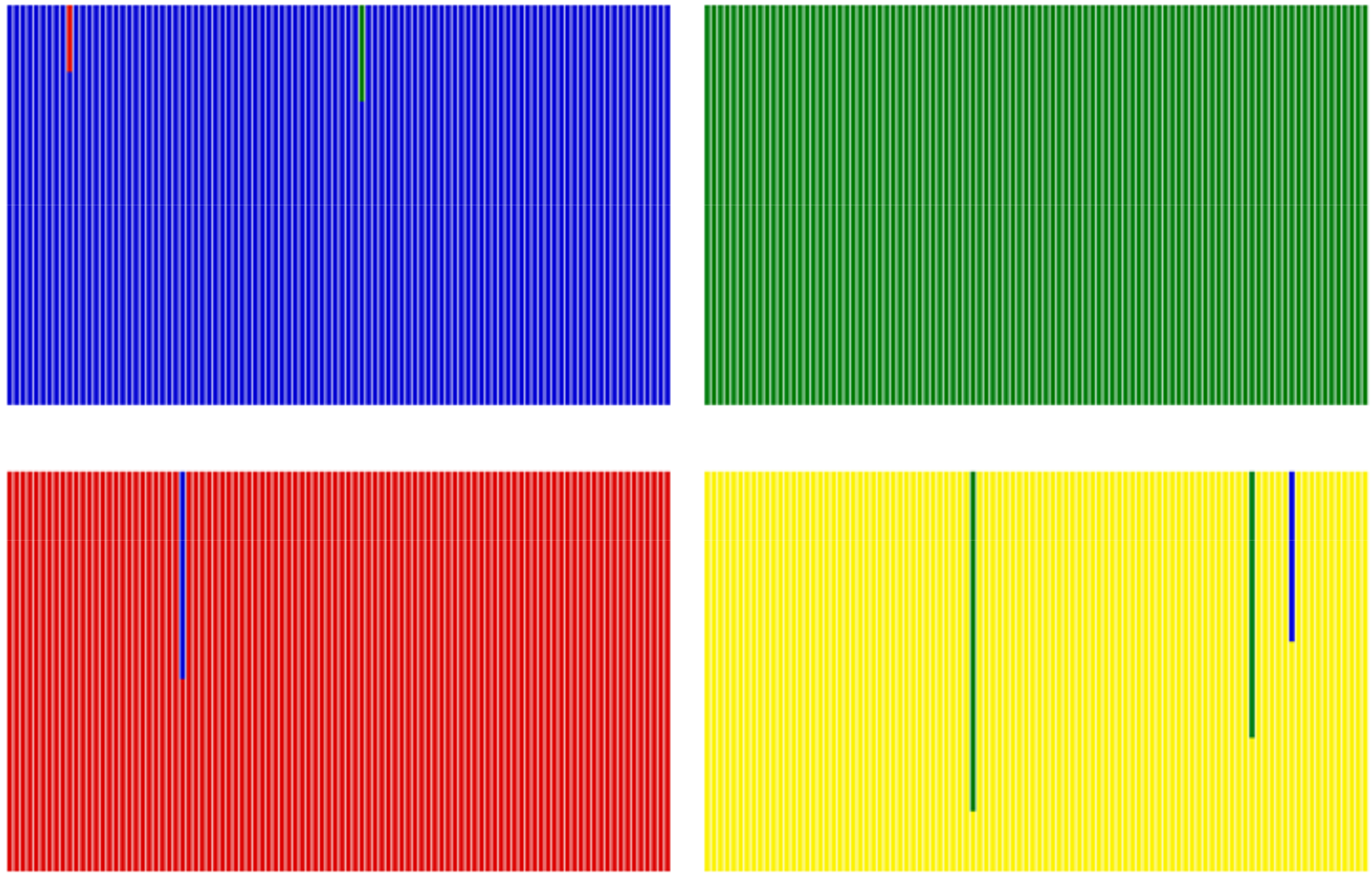
$$\mathbf{P}(0.0) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

Stationary Frequencies of a CTMC



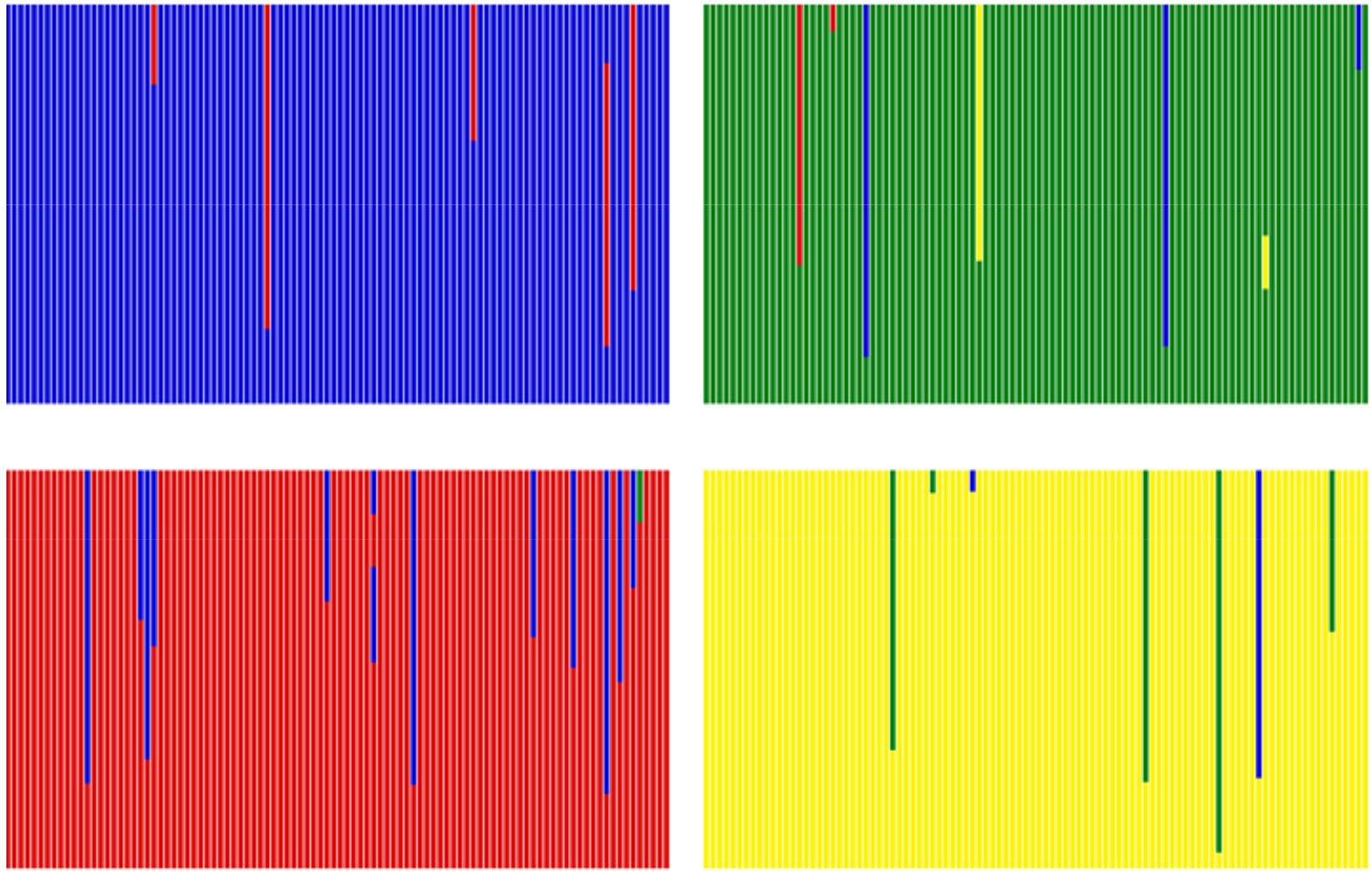
$$\mathbf{P}(0.0) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

Stationary Frequencies of a CTMC



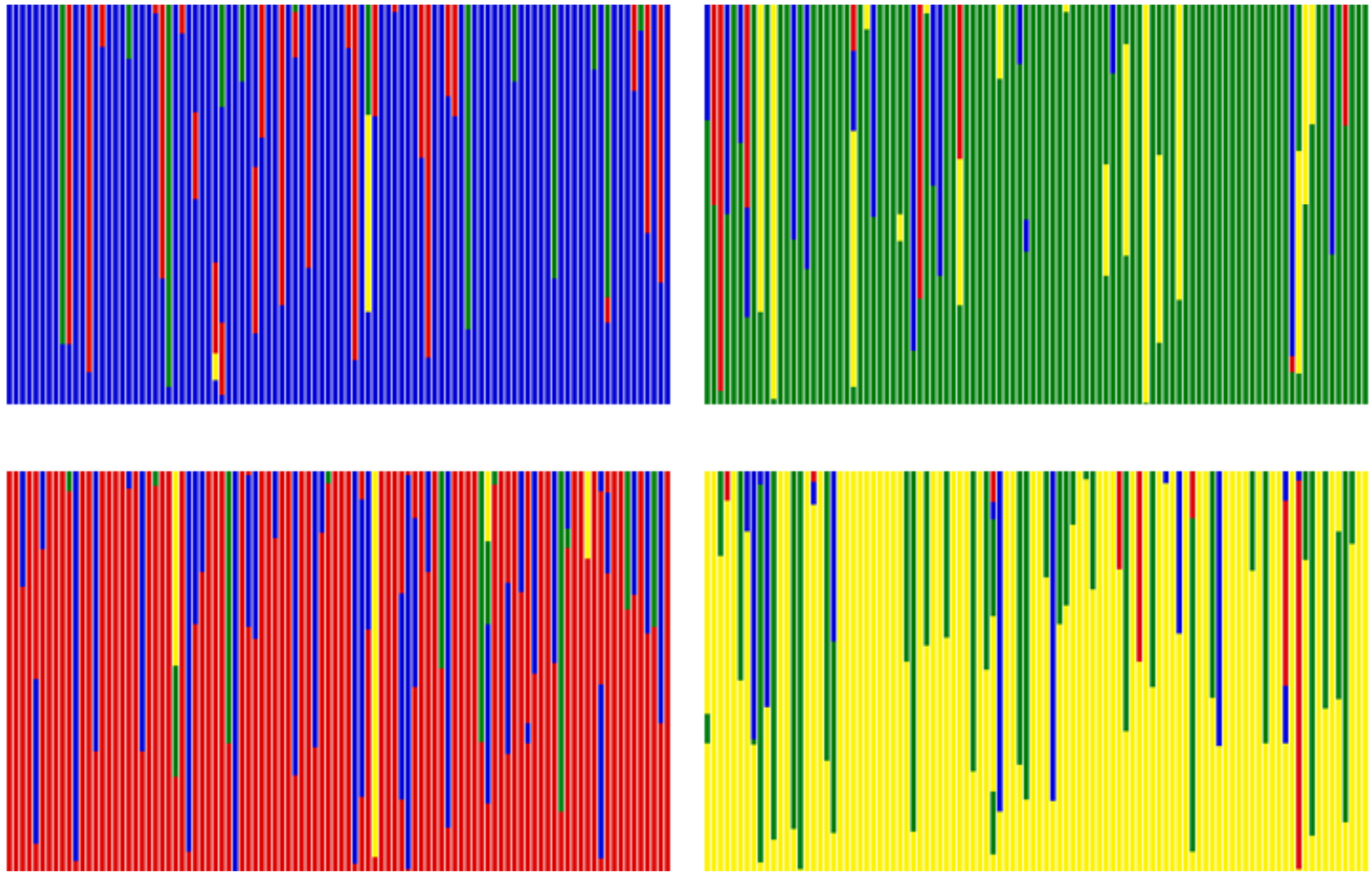
$$\mathbf{P}(0.01) = \begin{pmatrix} 0.981 & 0.005 & 0.008 & 0.006 \\ 0.001 & 0.989 & 0.004 & 0.005 \\ 0.003 & 0.002 & 0.994 & 0.001 \\ 0.005 & 0.002 & 0.006 & 0.986 \end{pmatrix}$$

Stationary Frequencies of a CTMC



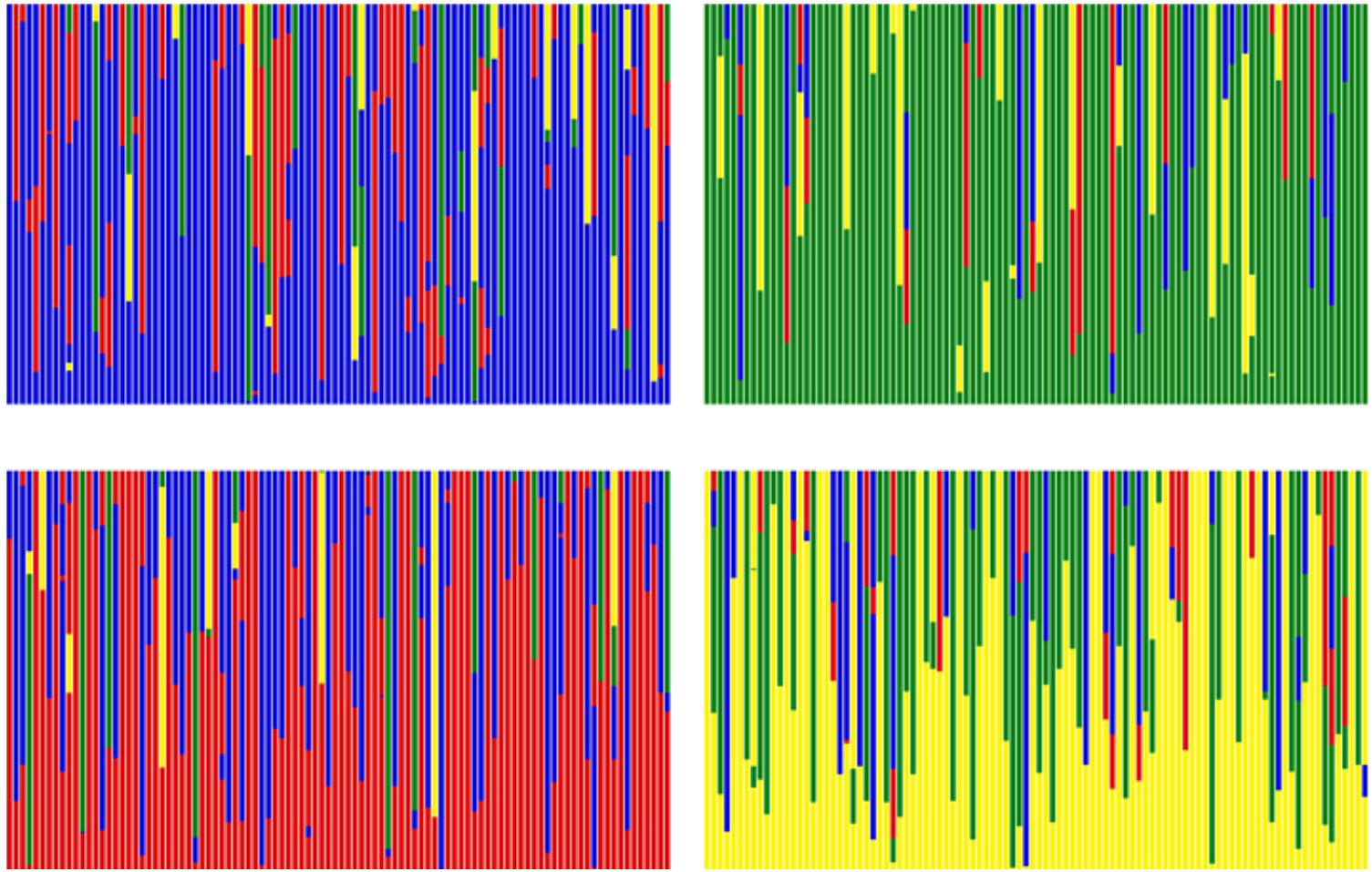
$$\mathbf{P}(0.10) = \begin{pmatrix} 0.828 & 0.048 & 0.072 & 0.052 \\ 0.014 & 0.900 & 0.040 & 0.046 \\ 0.026 & 0.017 & 0.944 & 0.013 \\ 0.046 & 0.023 & 0.056 & 0.876 \end{pmatrix}$$

Stationary Frequencies of a CTMC



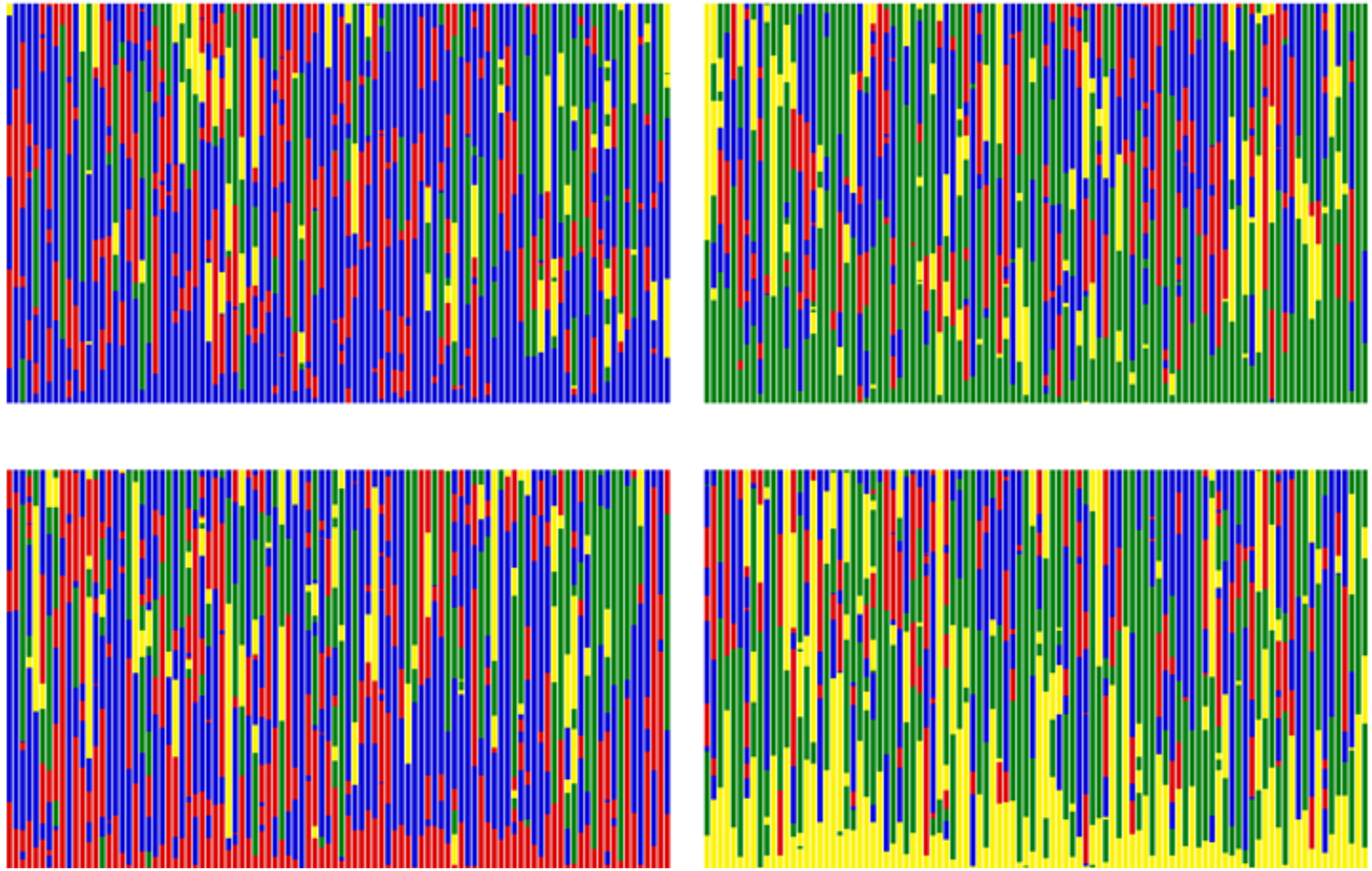
$$\mathbf{P}(0.50) = \begin{pmatrix} 0.423 & 0.153 & 0.263 & 0.161 \\ 0.063 & 0.609 & 0.175 & 0.153 \\ 0.088 & 0.072 & 0.778 & 0.062 \\ 0.135 & 0.094 & 0.227 & 0.544 \end{pmatrix}$$

Stationary Frequencies of a CTMC



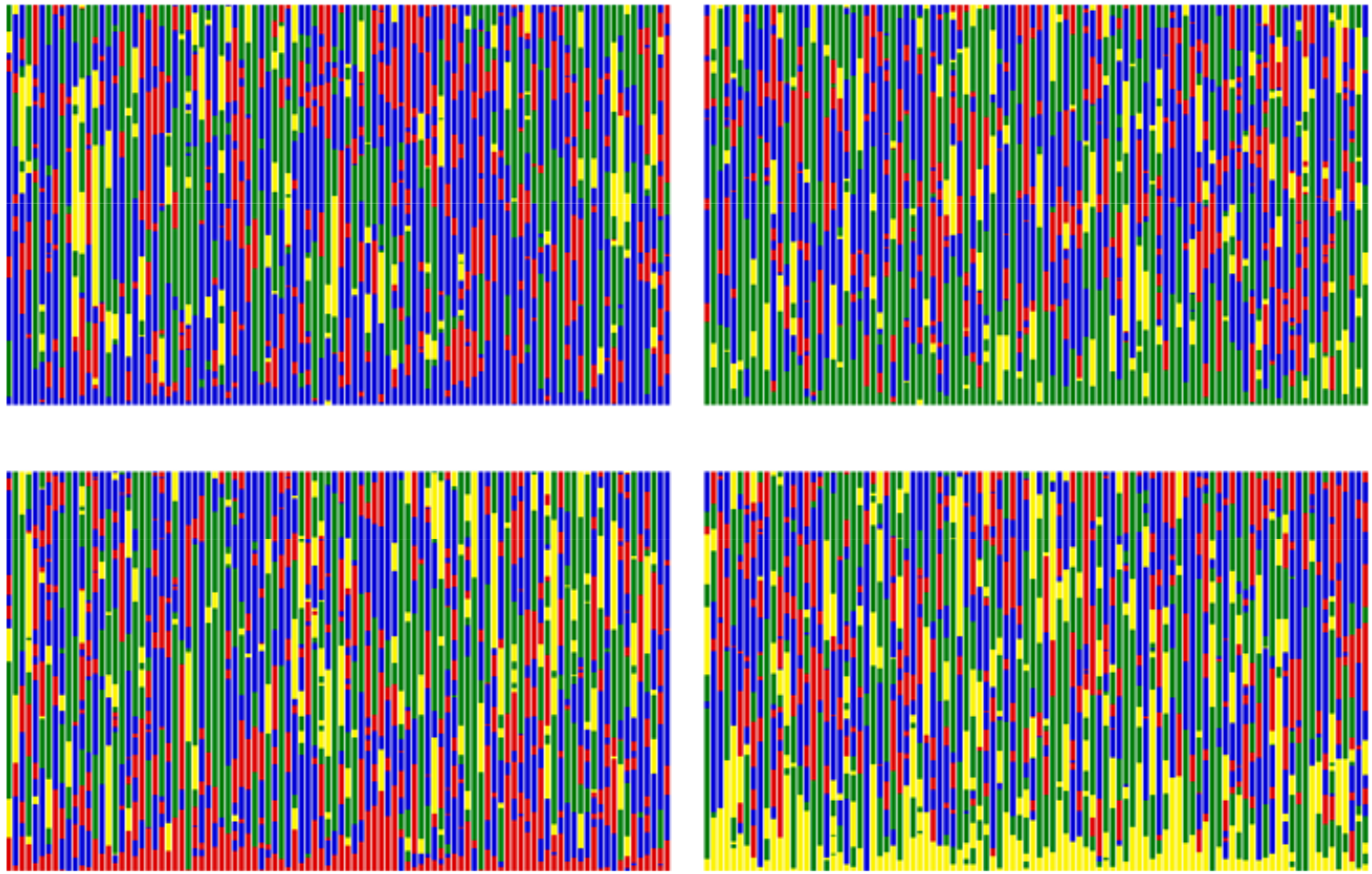
$$\mathbf{P}(1.0) = \begin{pmatrix} 0.233 & 0.192 & 0.379 & 0.195 \\ 0.101 & 0.408 & 0.295 & 0.197 \\ 0.118 & 0.119 & 0.655 & 0.107 \\ 0.156 & 0.145 & 0.352 & 0.347 \end{pmatrix}$$

Stationary Frequencies of a CTMC



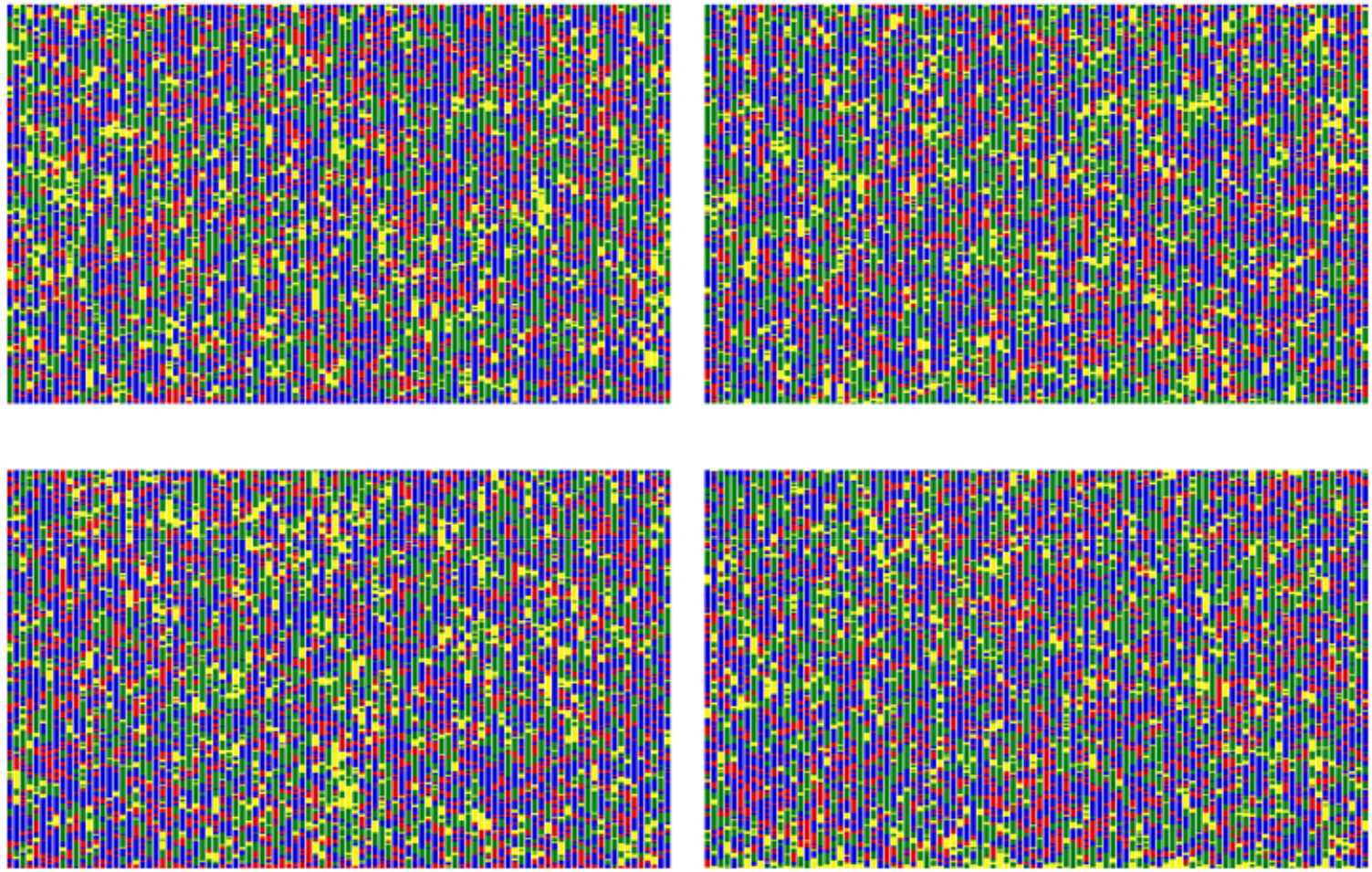
$$\mathbf{P}(5.0) = \begin{pmatrix} 0.138 & 0.188 & 0.494 & 0.180 \\ 0.138 & 0.190 & 0.492 & 0.181 \\ 0.137 & 0.187 & 0.497 & 0.178 \\ 0.138 & 0.188 & 0.494 & 0.180 \end{pmatrix}$$

Stationary Frequencies of a CTMC



$$\mathbf{P}(10.0) = \begin{pmatrix} 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \end{pmatrix}$$

Stationary Frequencies of a CTMC



$$\mathbf{P}(100.0) = \begin{pmatrix} 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \end{pmatrix}$$

Stationary Frequencies of a CTMC

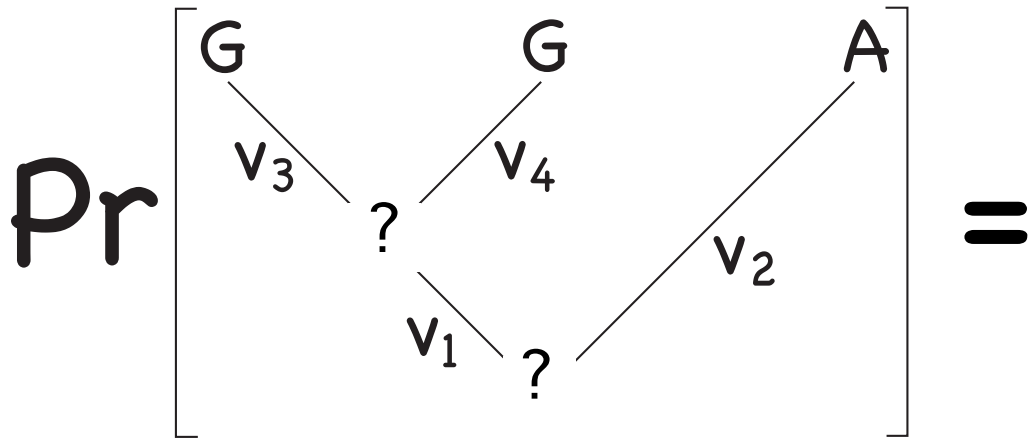
Stationary frequencies

The probability of observing the process in a particular state j after a long (infinite) period of time

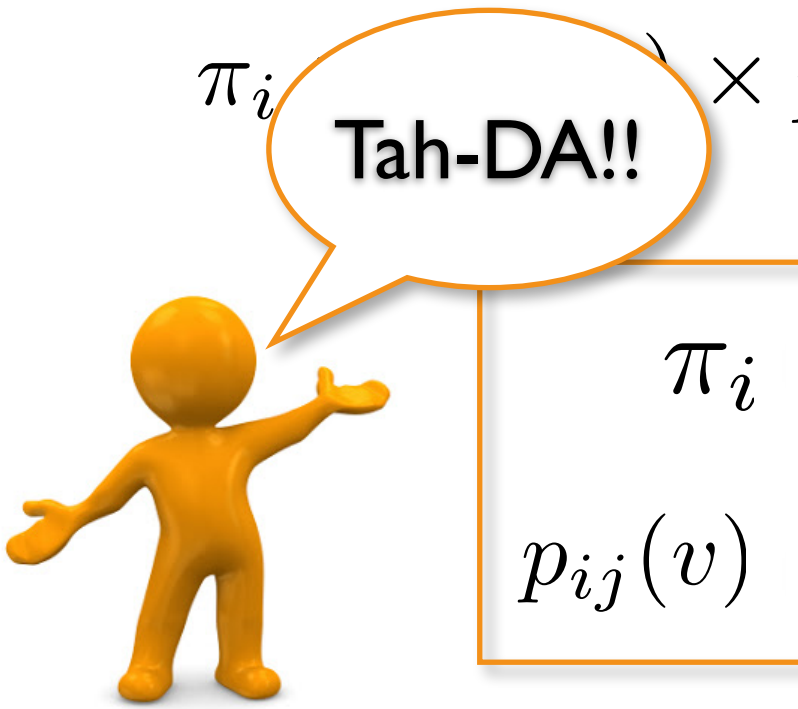
Also referred to as the 'invariant distribution' , 'stationary distribution' , 'stationary probabilities' , 'prior probabilities'

When the continuous time Markov chain is at stationarity, the stochastic process has 'forgotten' the starting state: the process ends in a given state with the same probability, regardless of the starting state

Putting It All Together…



$$\pi_i \times p_{jA}(v_2) \times p_{jG}(v_3) \times p_{iG}(v_4)$$



π_i	Stationary frequencies
$p_{ij}(v)$	Transition probabilities

Stochastic Mechanisms of Character Change

Continuous-time Markov models

describe the stochastic process by which traits (nucleotides) evolve over the tree

Instantaneous-rate matrix, $\mathbf{Q}$

completely describes the stochastic process by specifying:

Transition probabilities: $p_{ij}(v)$, the probability of observing state j conditioned on starting in state i and running the process over a branch of length v can be estimated by Monte Carlo simulation or matrix exponentiation, $P(v) = e^{\mathbf{Q}v}$

Stationary frequencies: the long-term probability of observing the chain in state j

Stochastic Mechanisms of Character Change

Instantaneous-rate matrix, $\mathbf{Q}$, and the transition probability matrix, $\mathbf{P}$

The instantaneous rate matrix describes the probability of change between each state in an infinitesimal time interval, $q_{ij}(\partial t)$

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

The transition probability matrix, $\mathbf{P}(v) = \{p_{ij}(v)\}$, describes the probability of observing state j given that we started in state i and ran the process over a branch of length v

$$\mathbf{P}(v) = \{p_{ij}(v)\} = \begin{pmatrix} 0.422927 & 0.153118 & 0.263330 & 0.160625 \\ 0.062896 & 0.609068 & 0.175153 & 0.152883 \\ 0.087566 & 0.071950 & 0.778271 & 0.062212 \\ 0.134967 & 0.093601 & 0.226962 & 0.544470 \end{pmatrix}$$

The relationship between $\mathbf{Q}$ and $\mathbf{P}$ is $P(v) = e^{\mathbf{Q}v}$

the transition probabilities integrate over all possible histories by which an initial state i can give rise to an end state j over branch length v

Statistical Estimation of Phylogeny: An Outline

Data

sequence alignment

Model

tree with branch lengths

substitution model

maximum likelihood

Bayesian inference

Parameter Estimates

phylogeny & other model parameters

After Coffee!!