Introduction to modelling sequence evolution

Bastien Boussau
Bastien.boussau@univ-lyon1.fr
@bastounette

Who am I?

- CNRS researcher at LBBE in Lyon, France
- Interested in methods for sequence and genome evolution, and in their application
- Keywords: gene tree-species tree, phylogenetic reconstruction (neural networks recently), site- and branch-heterogeneous models of sequence evolution, genome-phenotype associations

This course

- A lot of the good stuff was borrowed from Brian Moore's slides

(http://phylolab.org/)

- The bad stuff is mine

Aims

Understand the main ideas underlying models of sequence evolution

- To do so, we will:
  - Introduce important probability notions
  - Play with models of character evolution through simulations
- Briefly present some of the main models of nucleotide evolution

Why modelling sequence evolution?

Generic statistical paradigm

- Question about some part of the world
- Model of how this part of the world works
- Collect data
- Estimate parameters of the model that allow answering the question

Why modelling sequence evolution?

Generic statistical paradigm

- Question about some part of the world
- Model of how this part of the world works
- Collect data
- Estimate parameters of the model that allow answering the question

Example

- Is my coin fair?
- Repeated throws=independent identically distributed *Bernoulli* draws
- Throw coin *N* times
- Estimate probability of heads

Why modelling sequence evolution?

Generic statistical paradigm

- Question about some part of the world
- Model of how this part of the world works
- Collect data
- Estimate parameters of the model that allow answering the question

Phylogeny example

- Are transitions as probable as transversions in rodents?
- Sites of alignment=independent identically distributed Markov chains running along a phylogeny
- Sequence rodents
- Estimate transition/transversion ratio

Why modelling sequence evolution?

Generic statistical paradigm

- Question about some part of the world
- Model of how this part of the world works
- Collect data
- Estimate parameters of the model that allow answering the question

Phylogeny example

- Are transitions as probable as transversions in rodents?
- *Sites of alignment=independent identically distributed Markov chains running along a phylogeny*
- Sequence rodents
- Estimate transition/transversion ratio

Why are we interested in simulations?

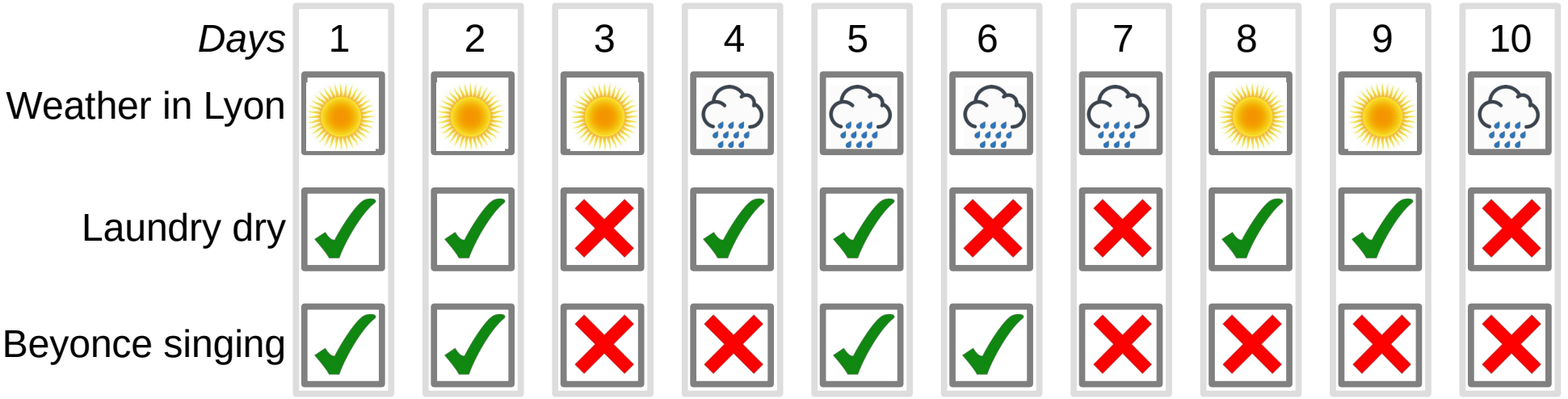
- Simulating data forces us to think in terms of a generating process
- By comparing true to simulated data, we can get a sense of how realistic our model is
- Simulations are also central to a lot of inferential problems:
  - Validation of inference methods
  - Posterior predictive tests
  - Approximate Bayesian Computation (ABC)
  - ...

Useful probability concepts

- Conditional probabilities
- Independence/intersection
- Union
- Bayes theorem
- Common distributions that will be useful in this talk:
  - Bernoulli
  - Binomial
  - Poisson
  - Exponential

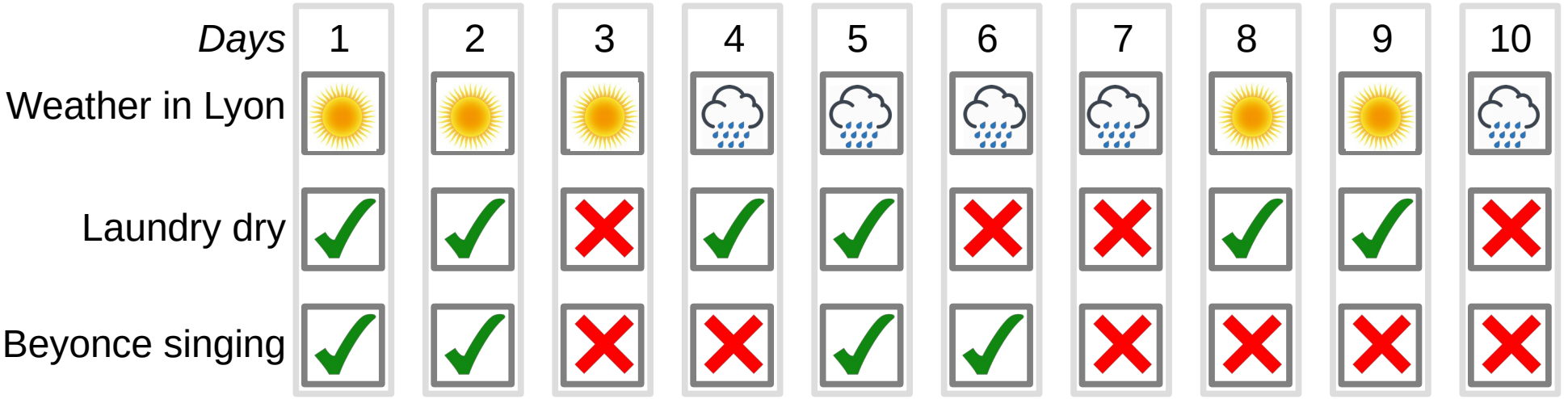
Crash course in probability

Record of various events during 10 days



Crash course in probability

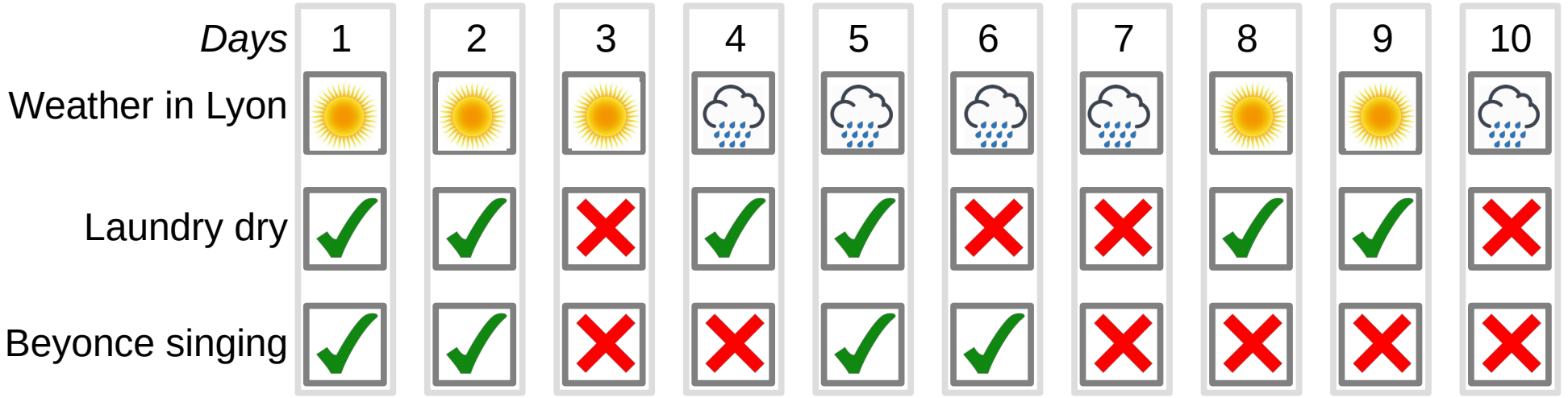
Record of various events during 10 days



$P(rainy) = ?$

Crash course in probability

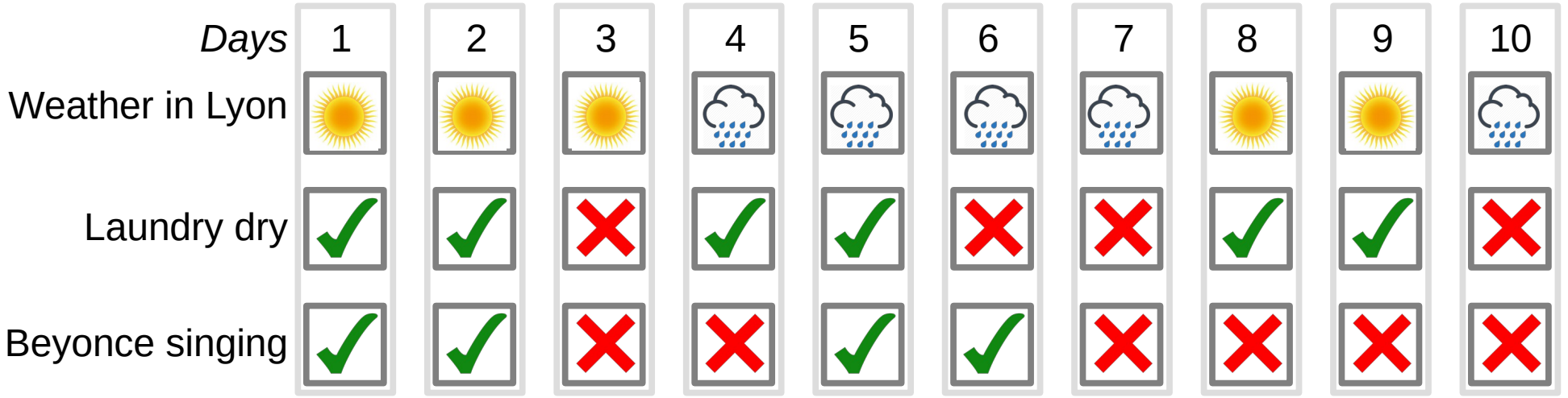
Record of various events during 10 days



$P(rainy) = 0.5$ $P(sunny) = 1 - P(rainy) = 0.5$

Crash course in probability

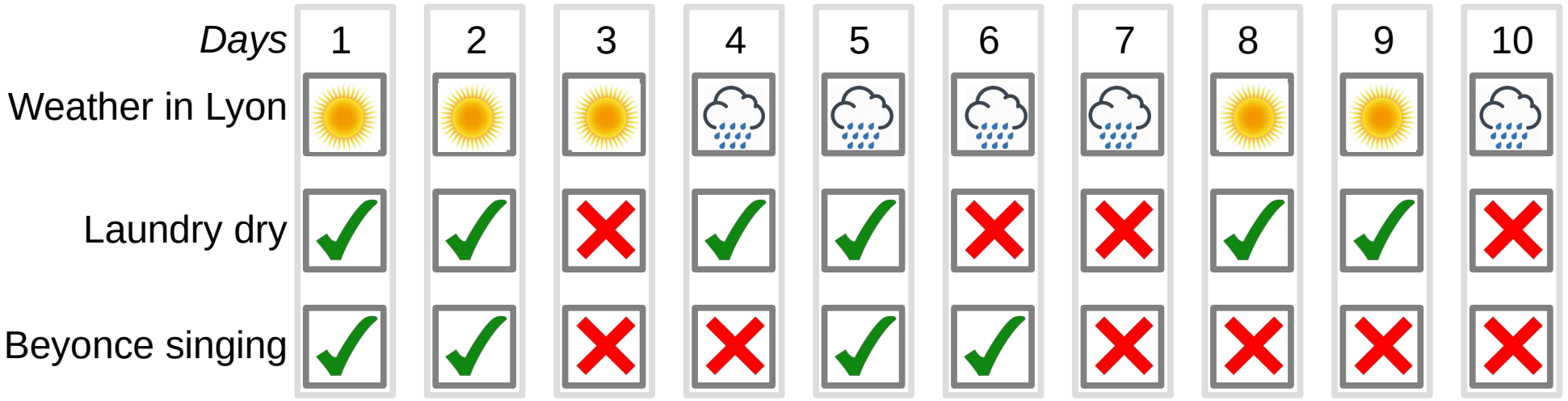
Record of various events during 10 days



$P(rainy) = 0.5$ $P(sunny) = 1 - P(rainy) = 0.5$ $P(dry\ laundry) = 0.6$

Crash course in probability

Record of various events during 10 days



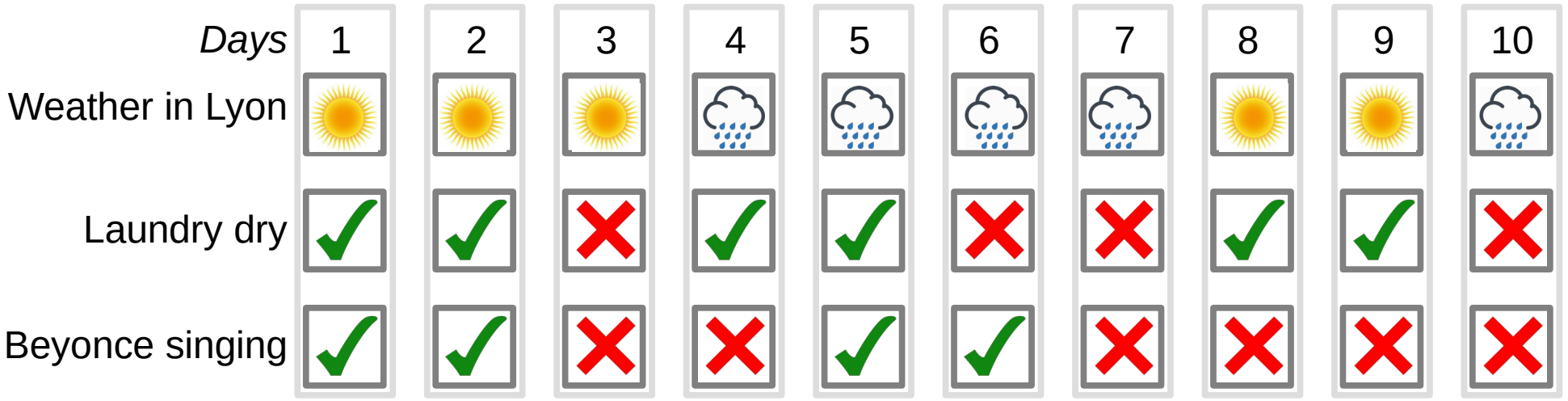
$P(rainy) = 0.5$ $P(sunny) = 1 - P(rainy) = 0.5$ $P(dry\ laundry) = 0.6$

$P(dry\ laundry | sunny) = ?$

$P(dry\ laundry | rainy) = ?$

Crash course in probability

Record of various events during 10 days



$P(rainy)=0.5$ $P(sunny)=1-P(rainy)=0.5$ $P(dry\,laundry)=0.6$

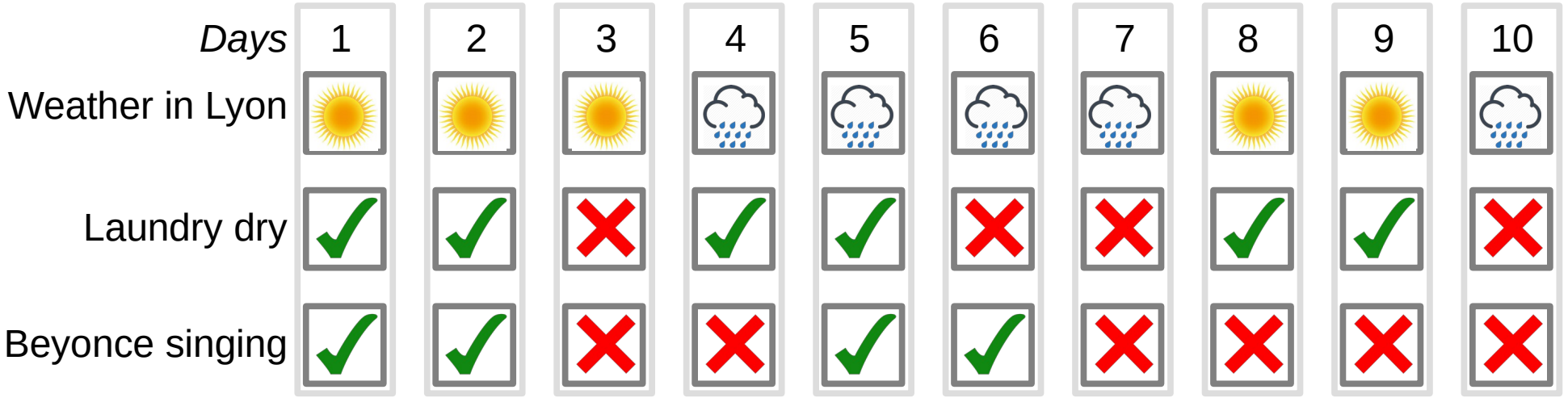
$P(dry\,laundry|sunny)=0.8$

$P(dry\,laundry|rainy)=0.4$

Conditional probability: P(A\|B)

Crash course in probability

Record of various events during 10 days



$P(rainy) = 0.5$ $P(sunny) = 1 - P(rainy) = 0.5$ $P(dry\ laundry) = 0.6$

$P(dry\ laundry | sunny) = 0.8$

$P(dry\ laundry | rainy) = 0.4$

$P(dry\ laundry \wedge sunny) = 0.4$

Crash course in probability

Record of various events during 10 days

Days	1	2	3	4	5	6	7	8	9	10
Weather in Lyon										
Laundry dry										
Beyonce singing										

$P(rainy)=0.5$ $P(sunny)=1-P(rainy)=0.5$ $P(dry\,laundry)=0.6$

$P(dry\,laundry|sunny)=0.8$

$P(dry\,laundry \wedge sunny)=0.4$

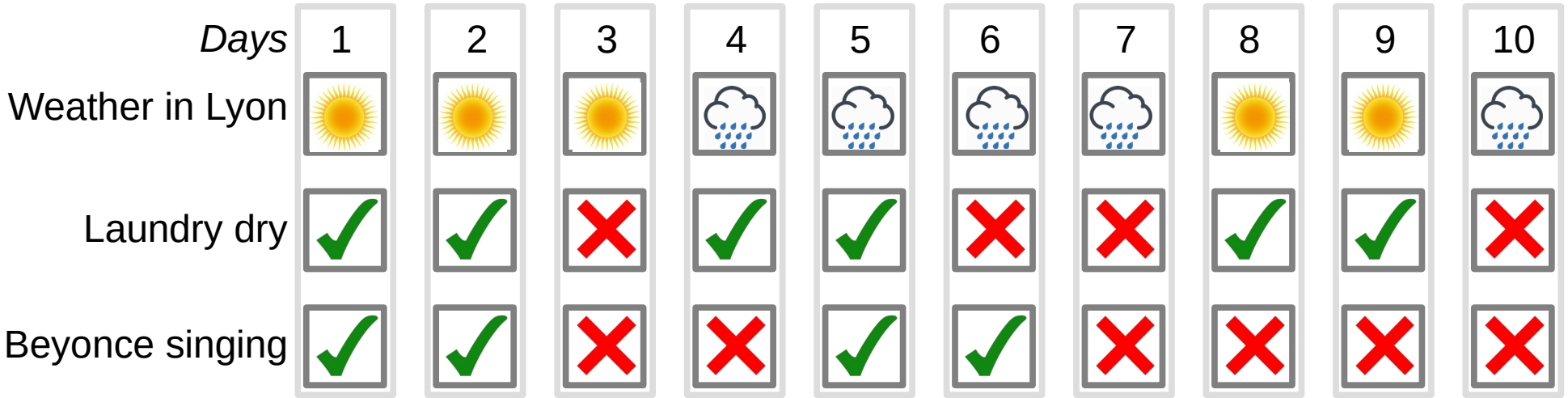
$P(dry\,laundry|rainy)=0.4$

$P(Beyonce\,singing)=0.4$

$P(Beyonce\,singing)=P(Beyonce\,singing|rainy)=P(Beyonce\,singing|sunny)=0.4$

Crash course in probability

Record of various events during 10 days



$P(rainy) = 0.5 \qquad P(sunny) = 1 - P(rainy) = 0.5$

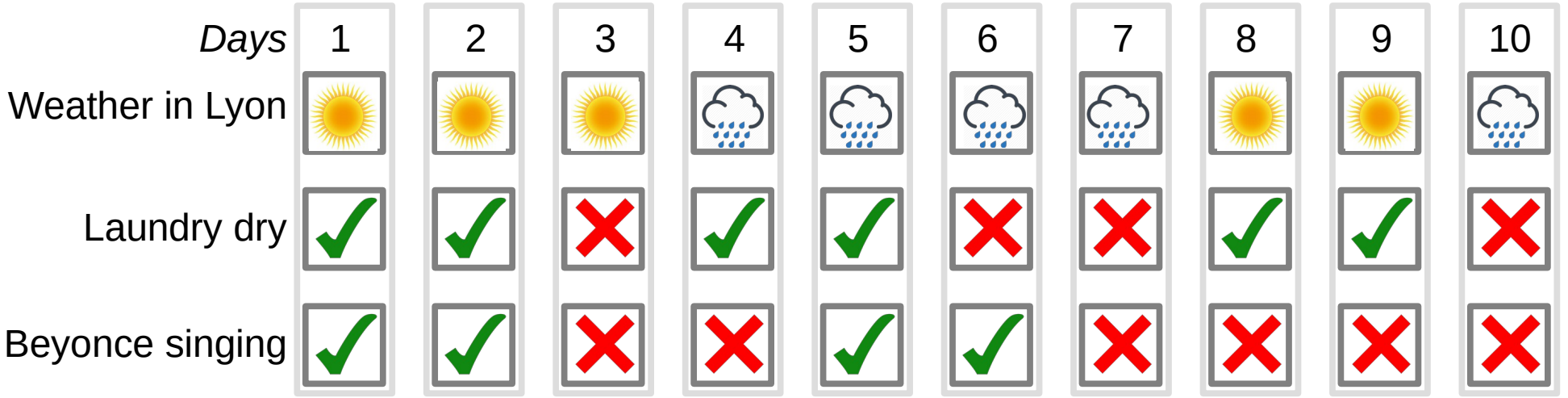
The events "Beyonce singing" and "sunny" are independent

$P(Beyonce\ singing) = 0.4$

$P(Beyonce\ singing) = P(Beyonce\ singing | rainy) = P(Beyonce\ singing | sunny) \overset{20}{=} 0.4$

Crash course in probability

Record of various events during 10 days



$P(rainy)=0.5$ $P(sunny)=1-P(rainy)=0.5$ $P(dry\,laundry)=0.6$

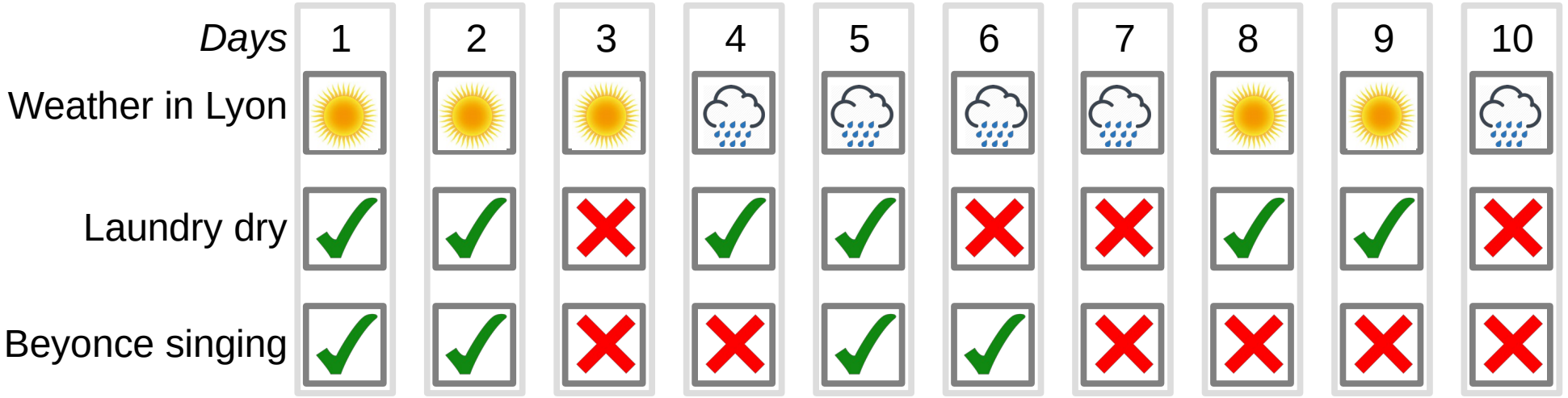
$P(dry\,laundry|sunny)=0.8$

$P(dry\,laundry|rainy)=0.4$

The events "dry laundry" and "sunny" are NOT independent

Crash course in probability

Record of various events during 10 days



$$P(rainy) = 0.5 \qquad P(sunny) = 1 - P(rainy) = 0.5 \qquad P(dry\ laundry) = 0.6$$

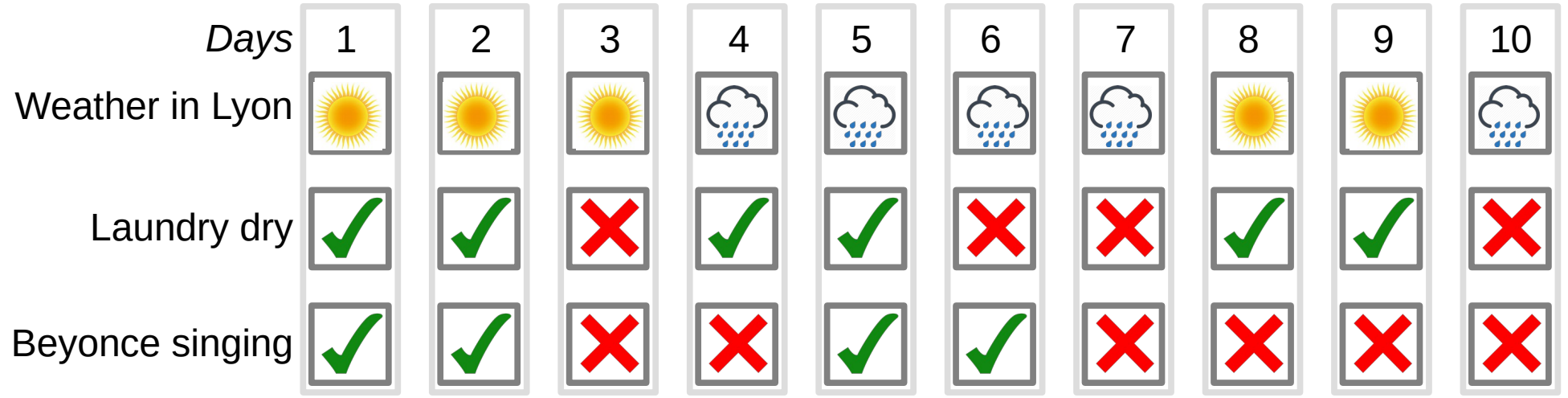
$$P(dry\ laundry \mid sunny) = 0.8 \qquad P(dry\ laundry \mid rainy) = 0.4$$

$$\begin{aligned} P(dry\ laundry) &= P(dry\ laundry \mid sunny) \times P(sunny) \\ &\quad + P(dry\ laundry \mid rainy) \times P(rainy) \\ &= 0.8 \times 0.5 + 0.4 \times 0.5 = 0.6 \end{aligned}$$

Bayes formula

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(B \wedge A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



$P(rainy) = 0.5$ $P(sunny) = 1 - P(rainy) = 0.5$ $P(dry\ laundry) = 0.6$

$P(dry\ laundry | sunny) = 0.8$

$P(dry\ laundry | rainy) = 0.4$

$P(dry\ laundry \wedge sunny) = 0.4$

$$P(sunny|dry\ laundry) = \frac{P(sunny \wedge dry\ laundry)}{P(dry\ laundry)} = \frac{P(dry\ laundry \wedge sunny)}{P(dry\ laundry)}$$

$$P(sunny|dry\ laundry) = \frac{P(dry\ laundry|sunny)\,P(sunny)}{P(dry\ laundry)}$$

Useful distributions

- *Discrete distributions (values in {0,1}, {0,1,2...}):*
  - **Bernoulli**: coin flip: $P(X=1)=p\,;P(X=0)=1-p$

Useful distributions

- *Discrete distributions (values in {0,1}, {0,1,2...}):*
  - **Bernoulli**: coin flip: $P(X=1)=p\,;P(X=0)=1-p$
  - **Binomial**: how many heads in several coin flips:

$$Pr(k;n,p)=\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

Useful distributions

- *Discrete distributions (values in {0,1}, {0,1,2...}):*
  - **Bernoulli**: coin flip: $P(X=1)=p\,;P(X=0)=1-p$
  - **Binomial**: how many heads in several coin flips:

$$Pr(k;n,p)=\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

  - **Poisson**: how many events of a type over a continuous time: how many meteorites with diameter > 1m in a year:

$$P(k\text{ events in interval})=e^{-\lambda}\frac{\lambda^k}{k!}$$

Useful distributions

- *Discrete distributions (values in {0,1}, {0,1,2...}):*
  - **Bernoulli**: coin flip: $P(X=1)=p\,;P(X=0)=1-p$
  - **Binomial**: how many heads in several coin flips:

$$Pr(k;n,p)=\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

  - **Poisson**: how many events of a type over a continuous time: how many meteorites with diameter > 1m in a year:

$$P(k\text{ events in interval})=e^{-\lambda}\frac{\lambda^k}{k!}$$

- *Continuous distributions (values in $\mathbb{R}$, [0,1]...):*
  - **Exponential**: Time between events in a Poisson process: how much time between two meteorites with diameter >1m:

$$f(x;\lambda)=\begin{cases}\lambda e^{-\lambda x} & x\geq 0,\\ 0 & x<0.\end{cases}$$

More on waiting times in Poisson processes

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

Siméon Poisson (1821)

More on waiting times in Poisson processes

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

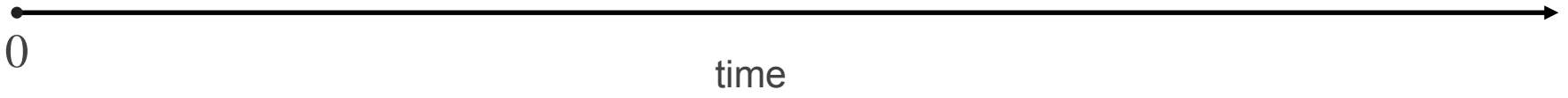
Siméon Poisson (1821)

More on waiting times in Poisson processes

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



More on waiting times in Poisson processes

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

More on waiting times in Poisson processes

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

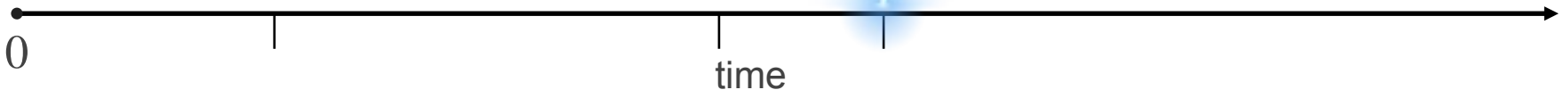
These events occur with a rate λ

More on waiting times in Poisson processes

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



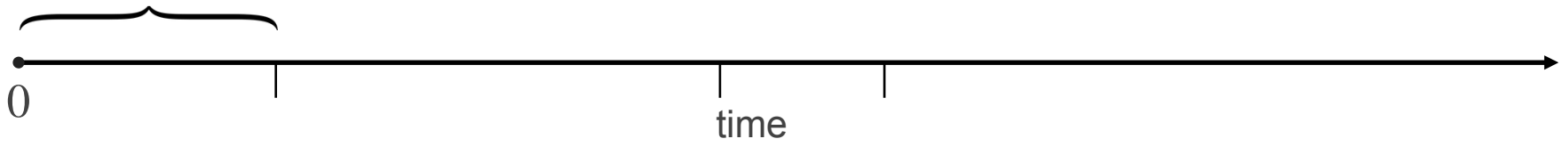
More on waiting times in Poisson processes

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



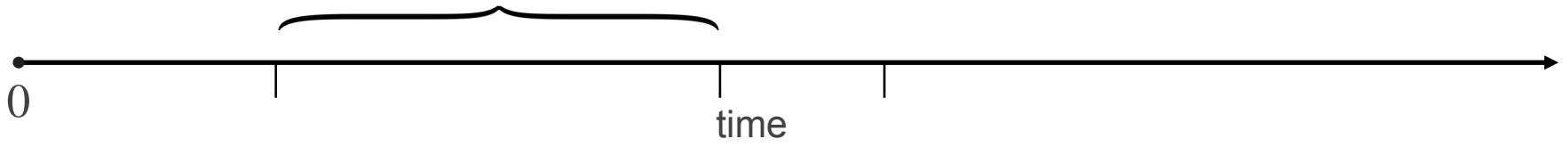
More on waiting times in Poisson processes

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



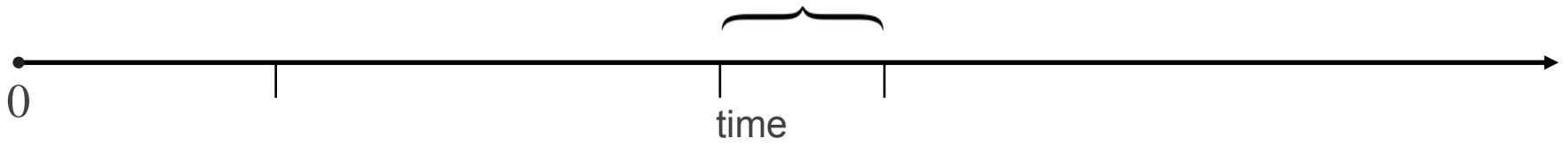
More on waiting times in Poisson processes

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



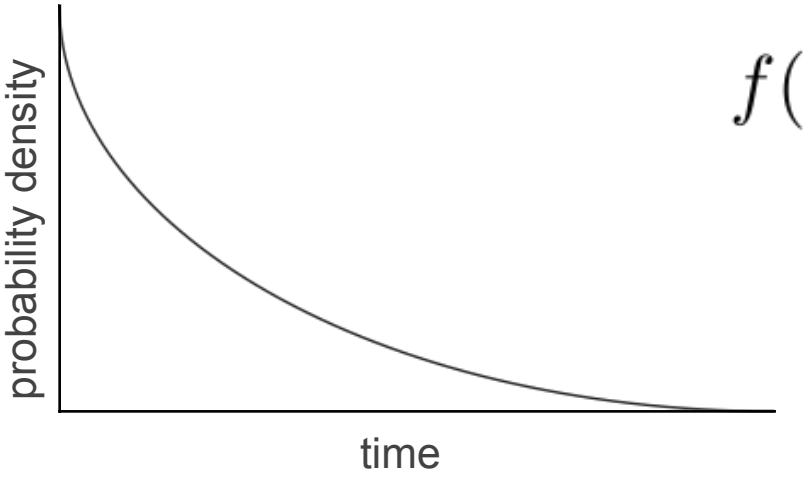
More on waiting times in Poisson processes

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



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

Aims

Understand the main ideas underlying models of sequence evolution

- To do so, we will:
  - Introduce important probability notions
  - Play with models of character evolution through simulations
- Briefly present some of the main models of nucleotide evolution

Stochastic Models of Nucleotide Substitution

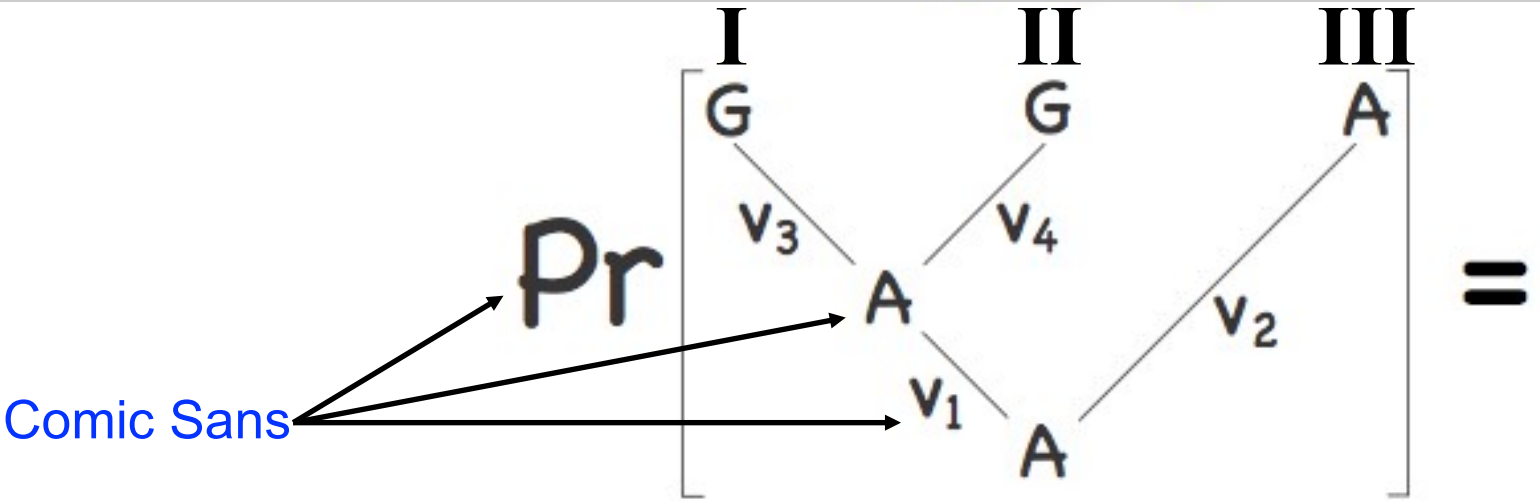
Species	Sequence data
Species I	GCG--CACCGGCGCAGTCA....
Species II	GCGTTCA--GGCG--GTCA....
Species III	ACGTTCACCGGCGCAGTCA....

How do we compute the likelihood of a site pattern?

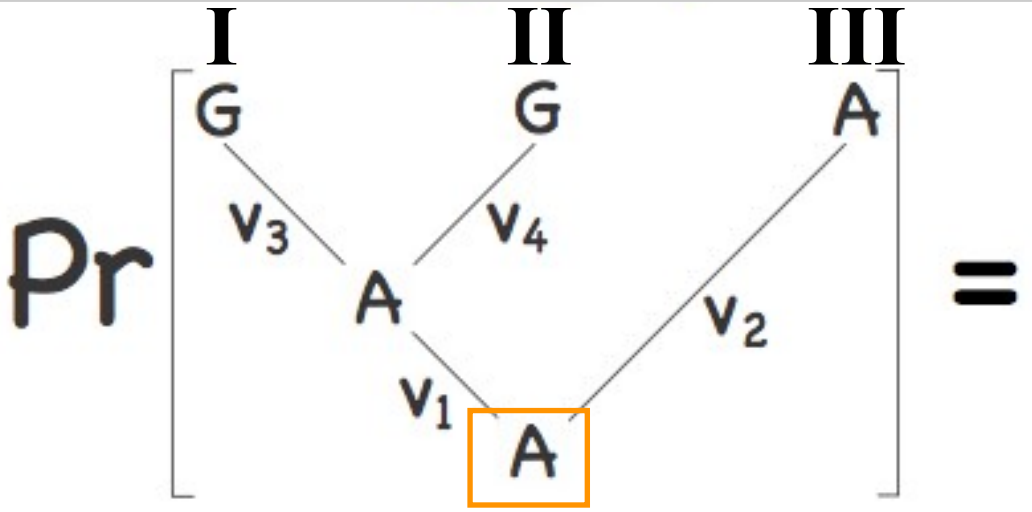
First, we'll see how we compute the likelihood of a site history.

To do that, we'll use simulations.

Stochastic Models of Nucleotide Substitution

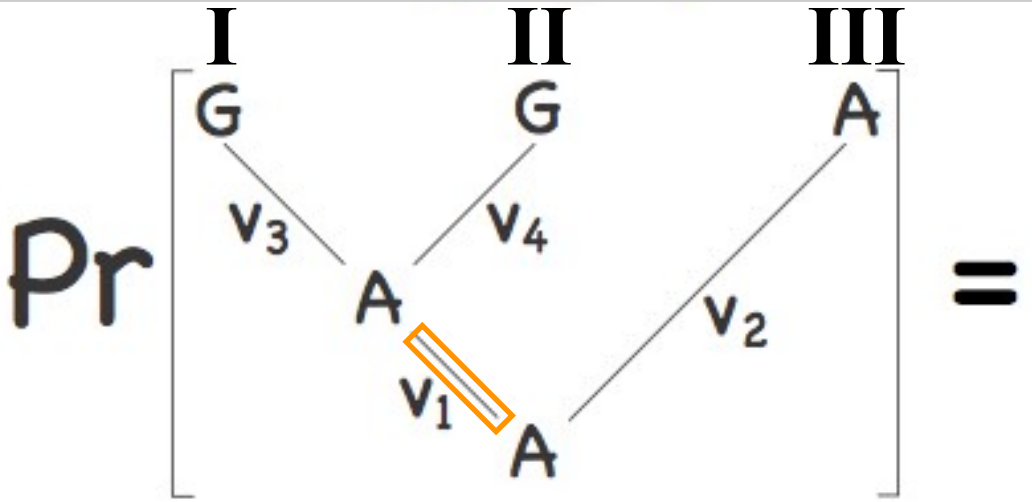


Stochastic Models of Nucleotide Substitution



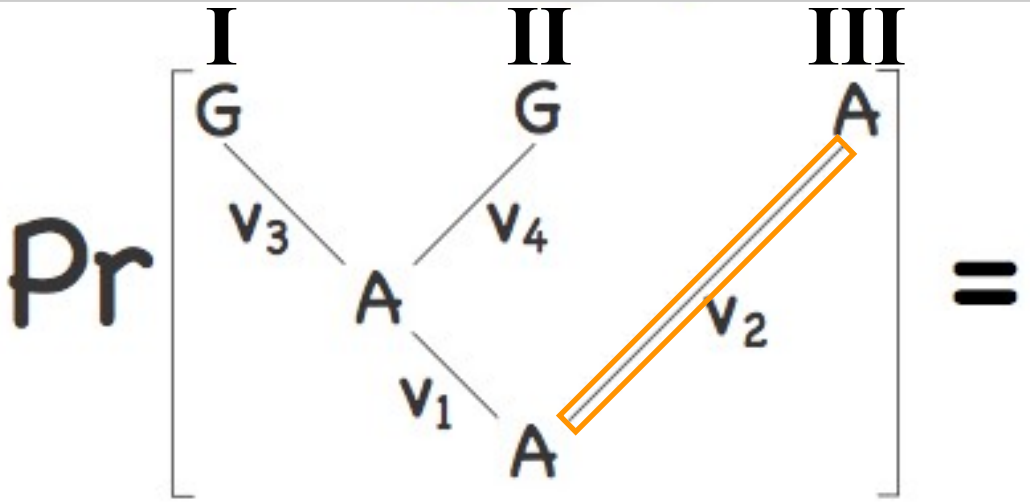
π_A

Stochastic Models of Nucleotide Substitution



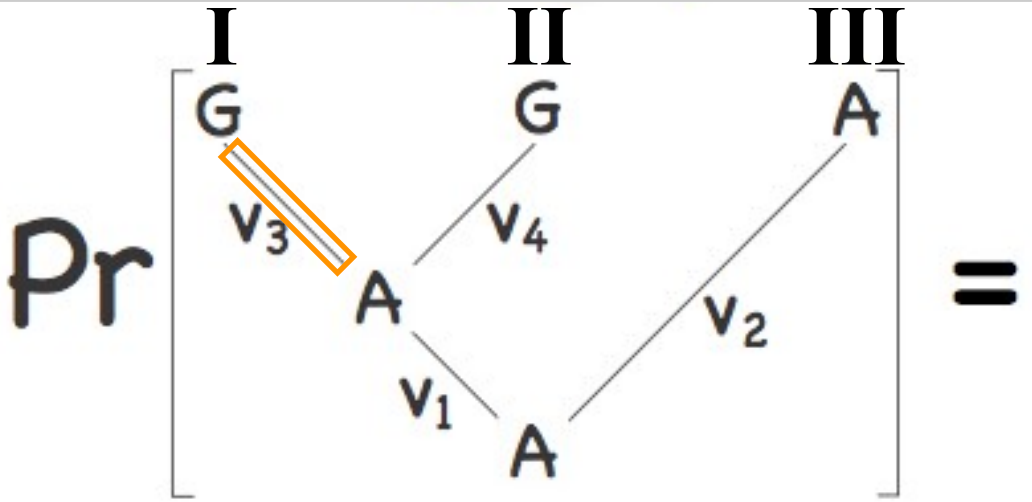
$$\pi_A \times p_{AA}(v_1)$$

Stochastic Models of Nucleotide Substitution



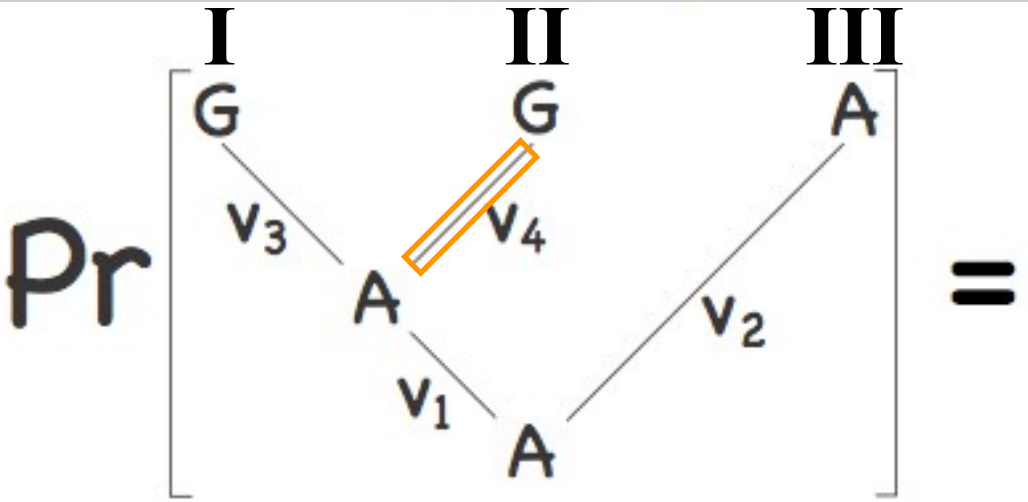
$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2)$$

Stochastic Models of Nucleotide Substitution



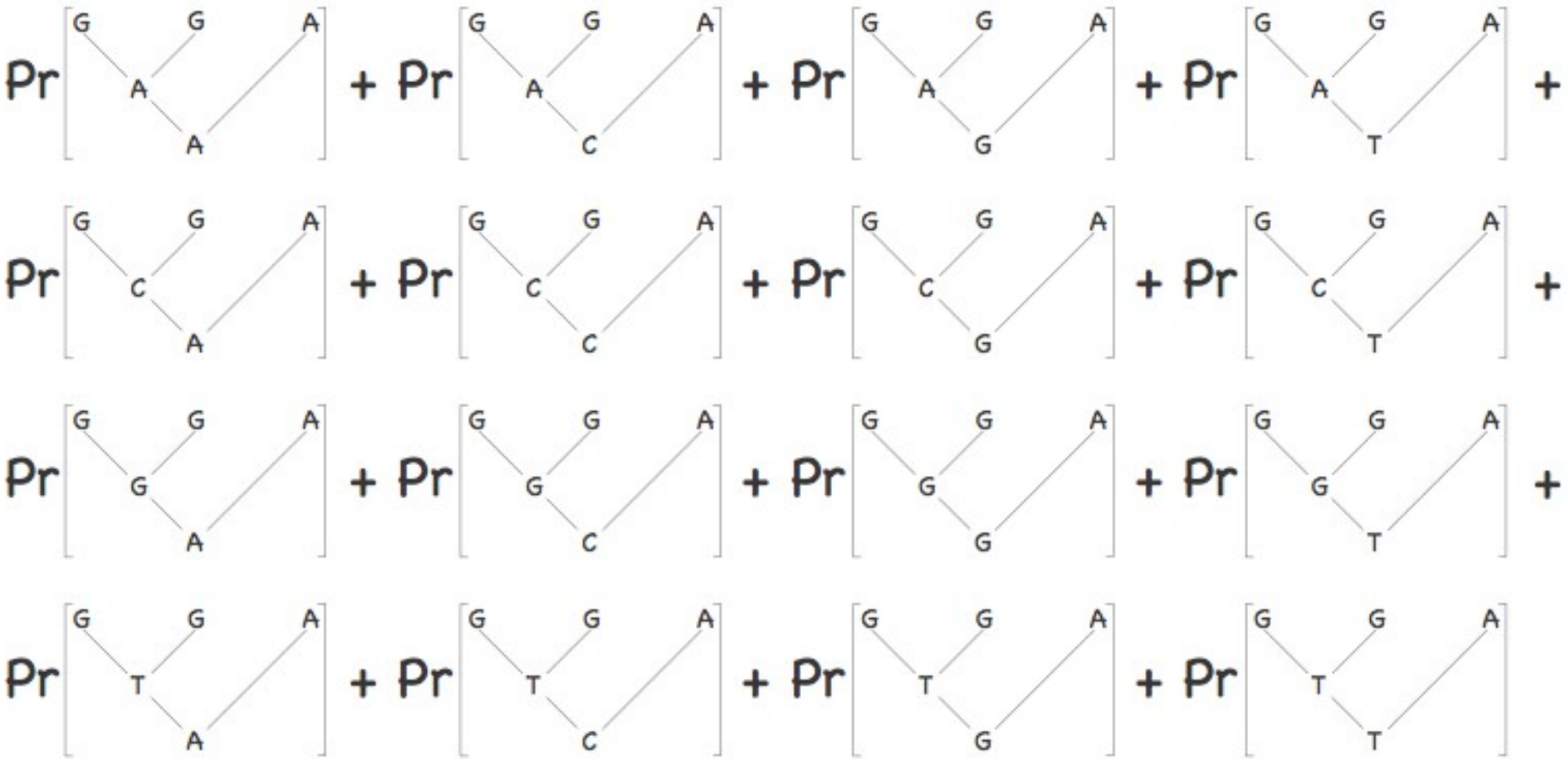
$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3)$$

Stochastic Models of Nucleotide Substitution

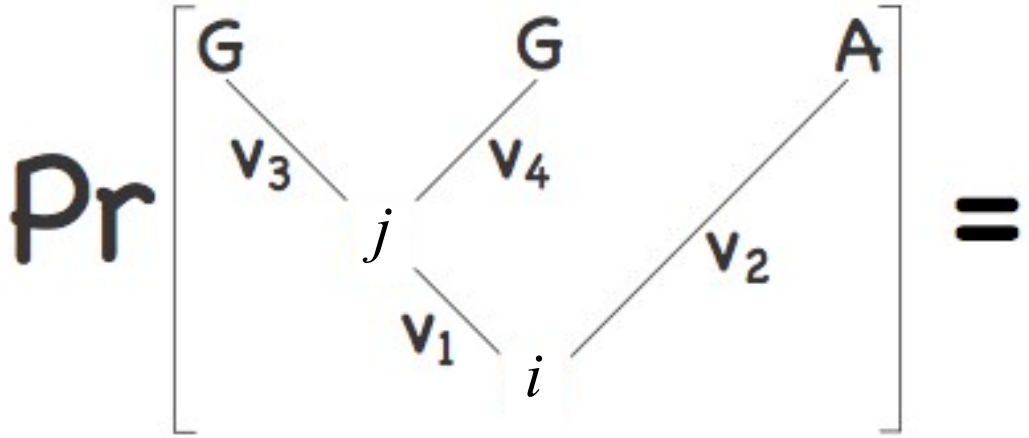


$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3) \times p_{AG}(v_4)$$

Stochastic Models of Nucleotide Substitution



Stochastic Models of Nucleotide Substitution



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

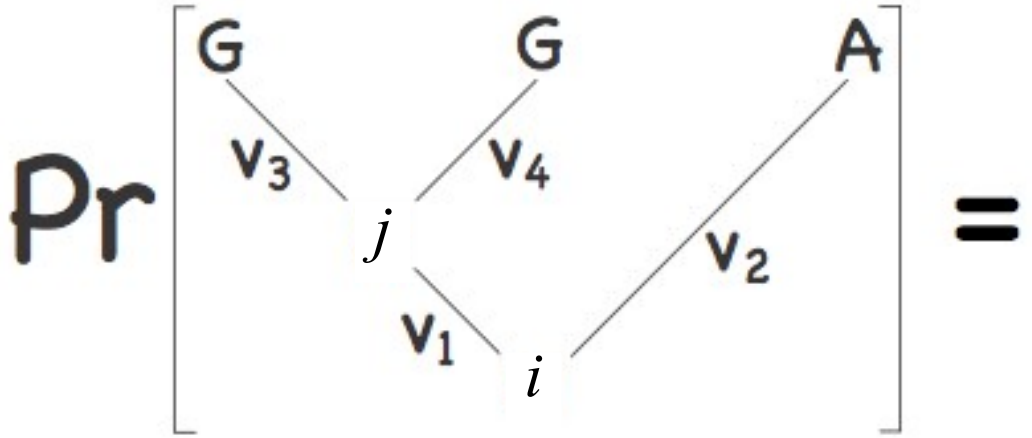
Stochastic Models of Nucleotide Substitution

$$\Pr\left[\begin{array}{ccccc} \text{G} & & \text{G} & & \text{A} \\ & v_3 \searrow & & \swarrow v_4 & \\ & & j & & \\ & & & \nwarrow v_2 & \\ & v_1 \searrow & & & \\ & & i & & \end{array}\right] =$$

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

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

Stochastic Models of Nucleotide Substitution



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

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

$p_{ij}(v)$ Stochastic Models of Nucleotide Substitution

Continuous-time Markov Chains (CTMC)

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

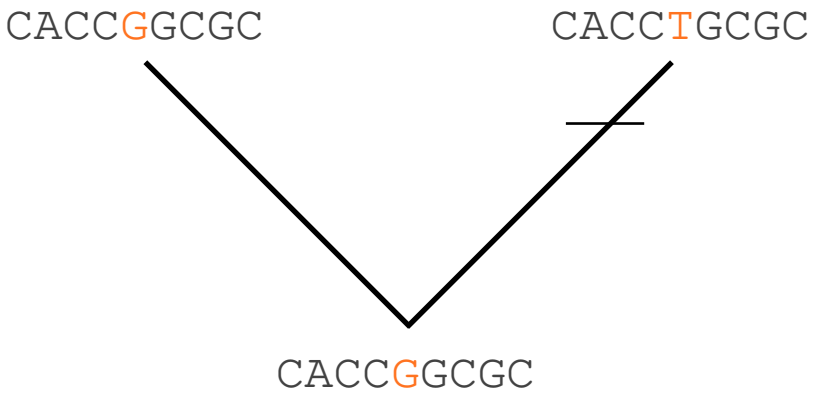
We will introduce:

- Substitution rates
- Substitution probabilities
- Stationary frequencies
- Relative substitution rates

$p_{ij}(v)$ Stochastic Models of Nucleotide Substitution

Models describe changes in the nucleotide sites at the species level

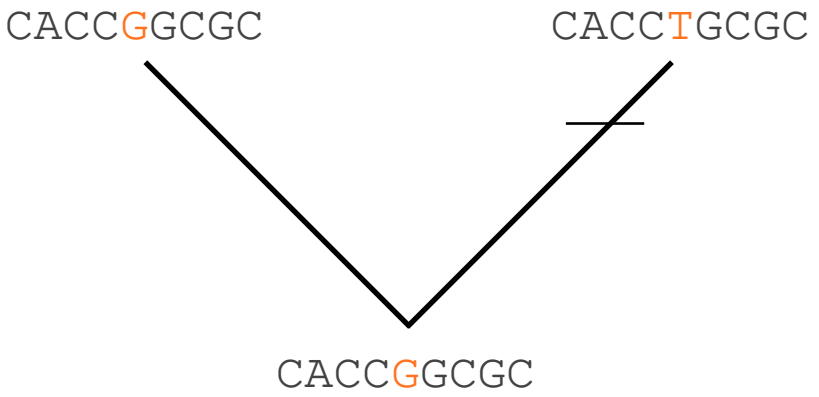
We are (generally) not trying to model the dynamics of allele frequencies in populations



$p_{ij}(v)$ Stochastic Models of Nucleotide Substitution

Models describe changes in the nucleotide sites at the species level

We are (generally) not trying to model the dynamics of allele frequencies in populations



We are modeling the process of **nucleotide substitution**, which describes the outcome of the mutation and fixation processes within populations

$p_{ij}(v)$ Stochastic Models of Nucleotide Substitution

Continuous-time Markov Models

Character change (nucleotide substitution) is modeled as a continuous-time Markov chain (CTMC)

$p_{ij}(v)$ 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

$p_{ij}(v)$ 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

$p_{ij}(v)$

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.

$p_{ij}(v)$

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.

A hypothetical instantaneous-rate matrix

		To A	To C	To G	To 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.

$p_{ij}(v)$

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.

A hypothetical instantaneous-rate matrix

		To A	To C	To G	To 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.

$p_{ij}(v)$

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.

A hypothetical instantaneous-rate matrix

		To A	To C	To G	To 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.

$p_{ij}(v)$

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.

A hypothetical instantaneous-rate matrix

		To A	To C	To G	To 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.

$p_{ij}(v)$

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.

A hypothetical instantaneous-rate matrix

		To A	To C	To G	To 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

$-1.916 = 0.541 + 0.787 + 0.588$

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.

The rows of the table must sum to zero.

$p_{ij}(v)$

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 $-q_{ii}$

$p_{ij}(v)$ 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 $-q_{cc} = -(-1.069) = 1.069$

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: $q_{CA} + q_{CG} + q_{CT} = 0.148 + 0.415 + 0.506 = 1.069$

$p_{ij}(v)$

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: $\mathrm{P}(i\text{->}j) = q_{ij} \div -q_{ii}$

$p_{ij}(v)$ 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: $\mathrm{P}(i\text{->}j) = q_{ij} \div -q_{ii}$

$$\mathrm{P(C->A)} = q_{CA} / -q_{CC} = 0.148 \mid 1.069 = 0.138$$

$p_{ij}(v)$ 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: $\mathrm{P}(i\text{->}j) = q_{ij} \div -q_{ii}$

$$\mathrm{P(C->A)} = q_{CA} / -q_{CC} = 0.148 \mid 1.069 = 0.138$$

$p_{ij}(v)$ 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: $\mathrm{P}(i\text{->}j) = q_{ij} \div -q_{ii}$

$$\mathrm{P(C->A)} = q_{CA} / -q_{CC} = 0.148 \mid 1.069 = 0.138$$
$$\mathrm{P(C->G)} = q_{CG} / -q_{CC} = 0.415 \mid 1.069 = 0.388$$

$p_{ij}(v)$ 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: $\mathrm{P}(i\text{->}j) = q_{ij} \div -q_{ii}$

$$\mathrm{P(C->A)} = q_{CA} / -q_{CC} = 0.148 \mid 1.069 = 0.138$$
$$\mathrm{P(C->G)} = q_{CG} / -q_{CC} = 0.415 \mid 1.069 = 0.388$$
$$\mathrm{P(C->T)} = q_{CT} / -q_{CC} = 0.506 \mid 1.069 = 0.474$$

$p_{ij}(v)$ 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: $P(i\text{->}j) = q_{ij} \div -q_{ii}$

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

Developing Intuition for CTMCs: A Monte Carlo Simulation Experiment

What the heck is Monte Carlo Simulation?

We generate a number of **replicate outcomes** (we will perform multiple trials)

Developing Intuition for CTMCs: A Monte Carlo Simulation Experiment

What the heck is Monte Carlo Simulation?

We generate a number of **replicate outcomes** (we will perform multiple trials) of a **stochastic process** (our replicate experiments involve randomness)

Developing Intuition for CTMCs: A Monte Carlo Simulation Experiment

What the heck is Monte Carlo Simulation?

We generate a number of **replicate outcomes** (we will perform multiple trials)

of a **stochastic process** (our replicate experiments involve randomness)

under a **fully specified model** (with specific values for all model parameters)

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation

We will assume that we have a **fully specified phylogenetic model**:

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation

We will assume that we have a **fully specified phylogenetic model**:

there is a single branch (the tree ‘topology')

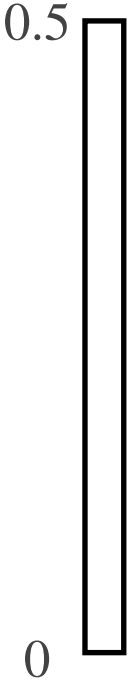
A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation

We will assume that we have a **fully specified phylogenetic model**:

there is a single branch (the tree 'topology')

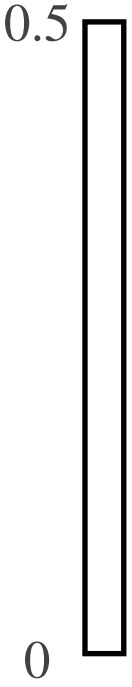
the branch has a length of 0.5 (expected substitutions/site)



A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation

We will assume that we have a **fully specified phylogenetic model**:



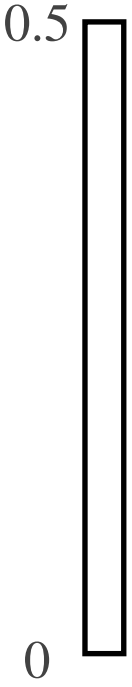
there is a single branch (the tree 'topology')
the branch has a length of 0.5 (expected substitutions/site)
the instantaneous-rate matrix is known

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

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation

We will assume that we have a **fully specified phylogenetic model**:



there is a single branch (the tree 'topology')
the branch has a length of 0.5 (expected substitutions/site)
the instantaneous-rate matrix is known

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

What is the matrix of substitution probabilities over our branch of length 0.5?

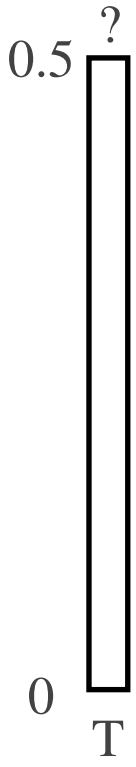
A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation

We will use a random-number generator to mimic the **stochastic process**

A Mechanistic Interpretation of the Rate Matrix

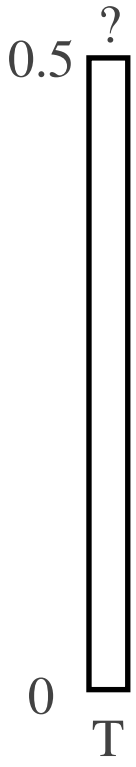
A simple Monte Carlo Simulation



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

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation

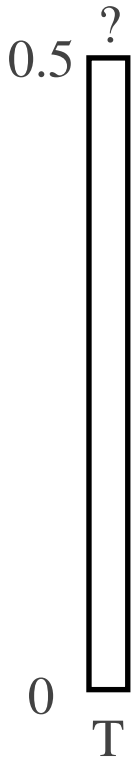


$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Rate of leaving the current state, T = 1.355

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



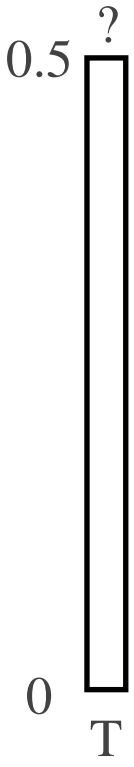
$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Probability of changing to A:

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

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



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

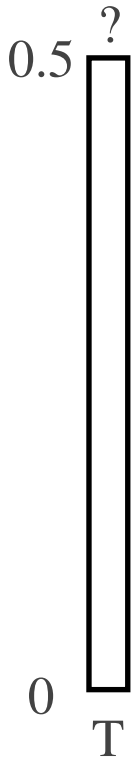
Probability of changing to C:

$$\text{P(T->A)} = q_{TA} / -q_{TT} = 0.525 \mid 1.355 = 0.387$$

$$\text{P(T->C)} = q_{TC} / -q_{TT} = 0.236 \mid 1.355 = 0.174$$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



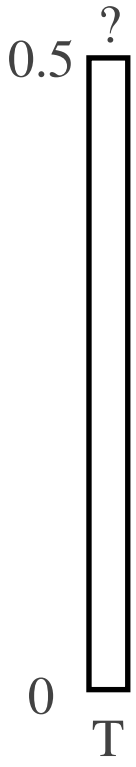
$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Probability of changing to G:

$$\text{P(T->A)} = q_{TA} / -q_{TT} = 0.525 \,|\, 1.355 = 0.387$$
$$\text{P(T->C)} = q_{TC} / -q_{TT} = 0.236 \,|\, 1.355 = 0.174$$
$$\text{P(T->G)} = q_{TG} / -q_{TT} = 0.594 \,|\, 1.355 = 0.438$$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation

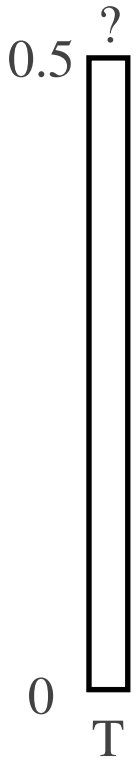


$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate an exponentially distributed waiting time, x:

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



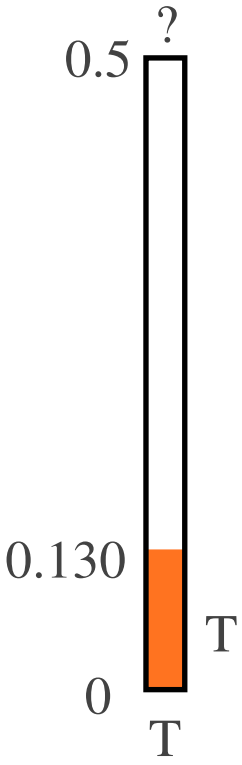
$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate an exponentially distributed waiting time, x:

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

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



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

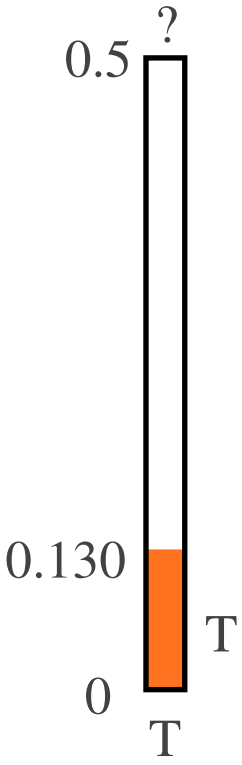
Generate an exponentially distributed waiting time, x:

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

Draw x: `x~dnExponential(1.355)` : x = 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: $\lambda = 1.355$

Draw x: `x~dnExponential(1.355)` : x = 0.130

Probabilities of substitution events in state T:

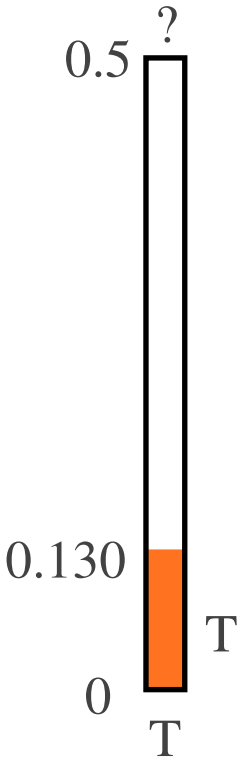
$$P(T\text{->}A) = q_{TA}/-q_{TT} = 0.525 \,|\, 1.355 = 0.387$$

$$P(T\text{->}C) = q_{TC}/-q_{TT} = 0.236 \,|\, 1.355 = 0.174$$

$$P(T\text{->}G) = q_{TG}/-q_{TT} = 0.594 \,|\, 1.355 = 0.438$$

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: $\lambda = 1.355$

Draw x: `x~dnExponential(1.355)` : x = 0.130

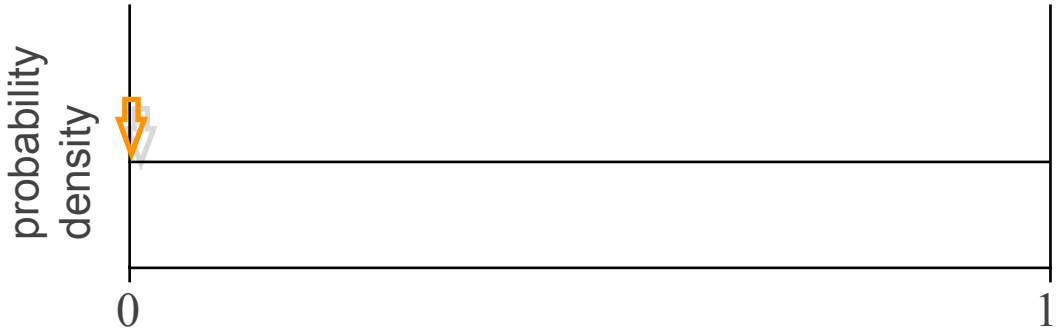
Specify a set of intervals:

	intervals	
P(T->A) = 0.387	0 – 0.387	(choose A)
P(T->C) = 0.174	0.387 – 0.561	(choose C)
P(T->G) = 0.438	0.561 – 1	(choose G)

Aside: Making Decisions within Random Numbers

Only the width of the bins matters, not their order

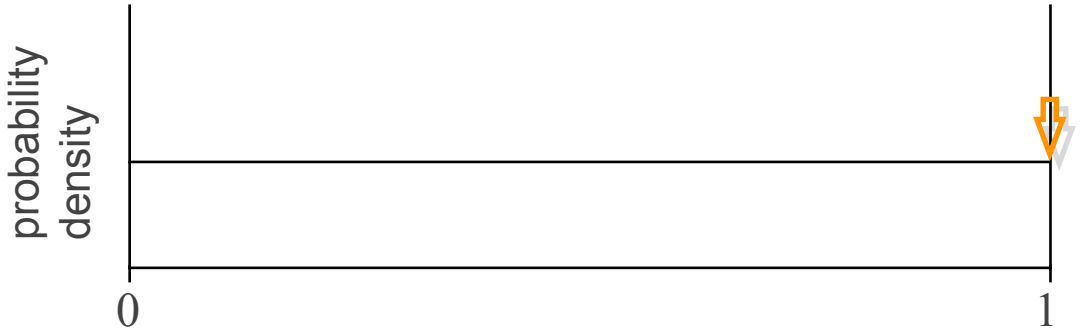
Our uniform$(0,1)$ random number will take any value between 0 and 1 with equal probability (by definition)



Aside: Making Decisions within Random Numbers

Only the width of the bins matters, not their order

Our uniform$(0,1)$ random number will take any value between 0 and 1 with equal probability (by definition)



Aside: Making Decisions within Random Numbers

Only the width of the bins matters, not their order

Our uniform$(0,1)$ random number will take any value between 0 and 1 with equal probability (by definition)

Imagine that there are two possible outcomes, A and B, which occur with probabilities:

Probability of outcomes:

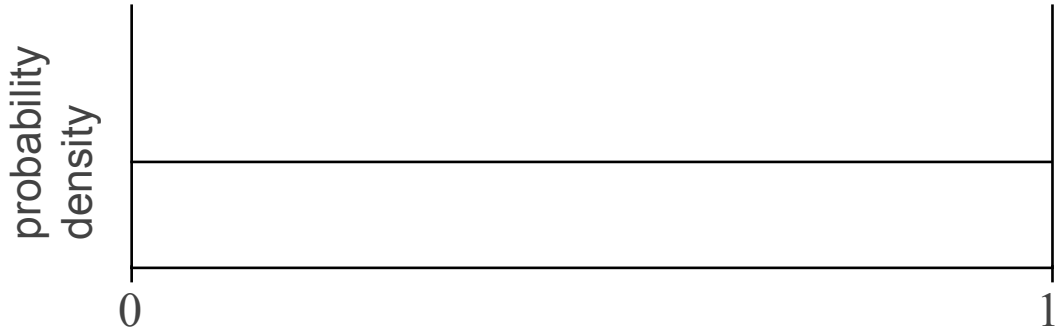
$$\mathrm{P}(\text{option A}) = 0.6$$

$$\mathrm{P}(\text{option B}) = 0.4$$

Aside: Making Decisions within Random Numbers

Only the width of the bins matters, not their order

Our uniform$(0,1)$ random number will take any value between 0 and 1 with equal probability (by definition)



Imagine that there are two possible outcomes, A and B, which occur with probabilities:

Probability of outcomes:

$$\mathrm{P}(\text{option A}) = 0.6$$

$$\mathrm{P}(\text{option B}) = 0.4$$

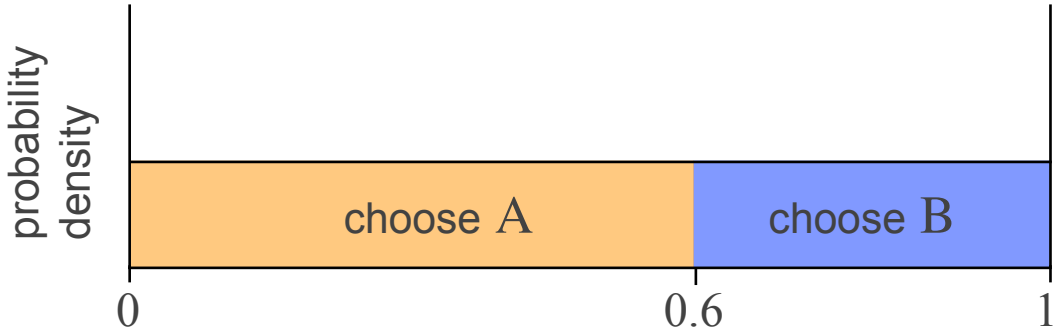
We can choose randomly (according to their probabilities) by specifying these intervals…

Probability of outcomes:	intervals	
$\mathrm{P}(\text{option A}) = 0.6$	$0.0 - 0.6$	(choose A)
$\mathrm{P}(\text{option B}) = 0.4$	$0.6 - 1.0$	(choose B)

Aside: Making Decisions within Random Numbers

Only the width of the bins matters, not their order

Our uniform$(0,1)$ random number will take any value between 0 and 1 with equal probability (by definition)



Imagine that there are two possible outcomes, A and B, which occur with probabilities:

Probability of outcomes:

$$P(\text{option A}) = 0.6$$

$$P(\text{option B}) = 0.4$$

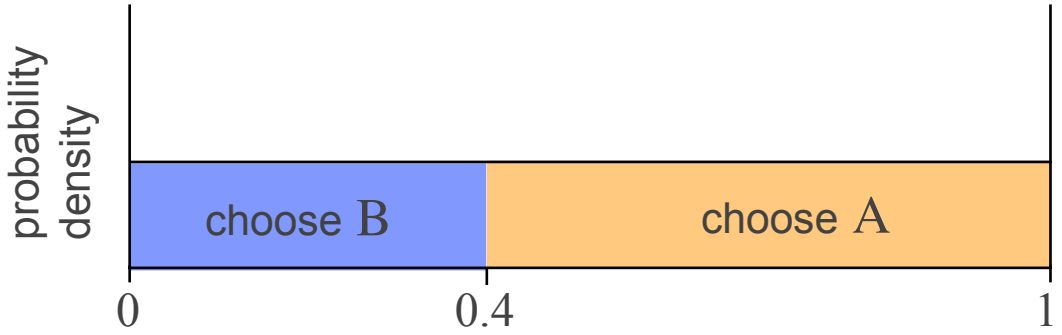
We can choose randomly (according to their probabilities) by specifying these intervals…

Probability of outcomes:	intervals	
$P(\text{option A}) = 0.6$	$0.0 - 0.6$	(choose A)
$P(\text{option B}) = 0.4$	$0.6 - 1.0$	(choose B)

Aside: Making Decisions within Random Numbers

Only the width of the bins matters, not their order

Our uniform$(0,1)$ random number will take any value between 0 and 1 with equal probability (by definition)



Imagine that there are two possible outcomes, A and B, which occur with probabilities:

Probability of outcomes:

$$P(\text{option A}) = 0.6$$

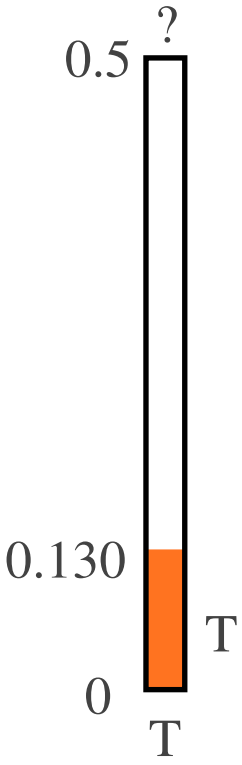
$$P(\text{option B}) = 0.4$$

Or equivalently by specifying these intervals…

Probability of outcomes:	intervals	
$P(\text{option A}) = 0.6$	$0.4 - 1.0$	(choose A)
$P(\text{option B}) = 0.4$	$0.0 - 0.6$	(choose B)

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: $\lambda = 1.355$

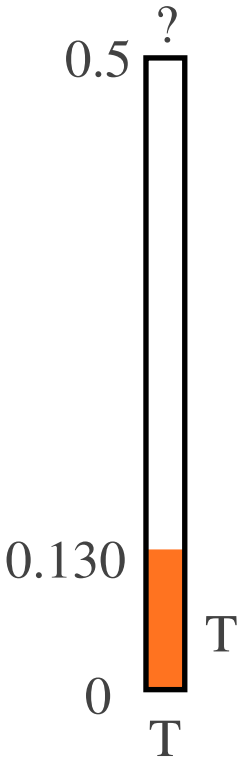
Draw x: `x~dnExponential(1.355)` : x = 0.130

Specify a set of intervals:

	intervals	
P(T->A) = 0.387	0 – 0.387	(choose A)
P(T->C) = 0.174	0.387 – 0.561	(choose C)
P(T->G) = 0.438	0.561 – 1	(choose G)

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: $\lambda = 1.355$

Draw x: `x~dnExponential(1.355)` : x = 0.130

Specify a set of intervals:

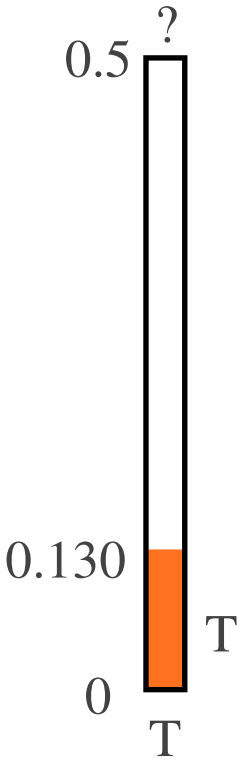
	intervals	
P(T->A) = 0.387	0 – 0.387	(choose A)
P(T->C) = 0.174	0.387 – 0.561	(choose C)
P(T->G) = 0.438	0.561 – 1	(choose G)

Draw a uniformly distributed number, u, to select substitution event:

`u~dnUniform(0,1)` : u=0.446

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: $\lambda = 1.355$

Draw x: `x~dnExponential(1.355)` : x = 0.130

Specify a set of intervals:

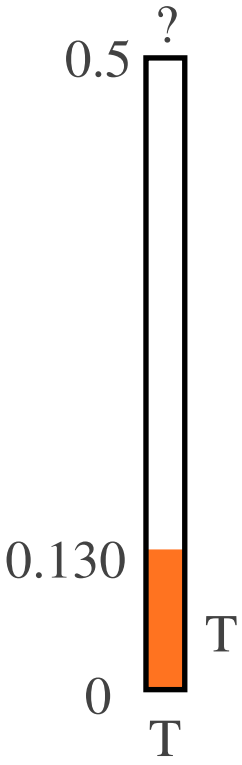
	intervals	
P(T->A) = 0.387	0 – 0.387	(choose A)
P(T->C) = 0.174	0.387 – 0.561	(choose C)
P(T->G) = 0.438	0.561 – 1	(choose G)

Draw a uniformly distributed number, u, to select substitution event:

`u~dnUniform(0,1)` : u=0.446

A Mechanistic Interpretation of the Rate Matrix

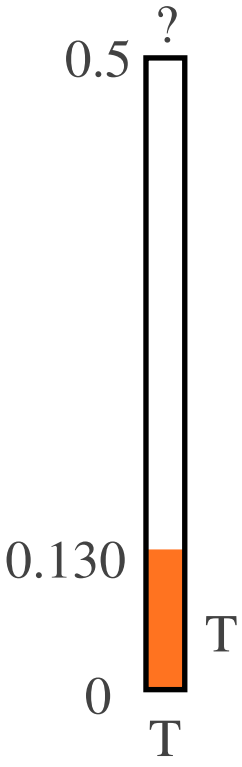
A simple Monte Carlo Simulation



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

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



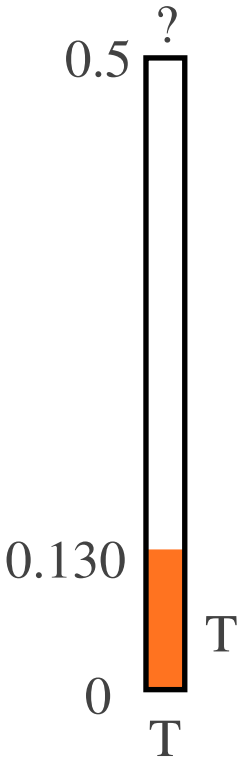
$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate waiting time to next event:

Rate when process in state C: $-q_{cc} = \lambda = 1.069$

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



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

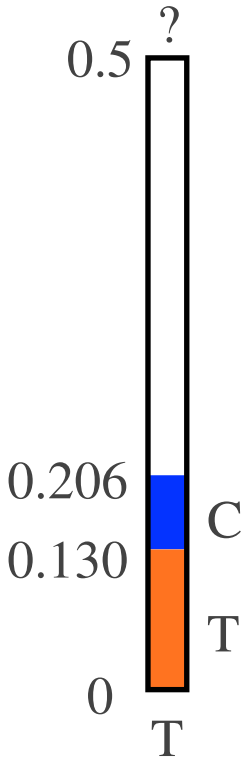
Generate waiting time to next event:

Rate when process in state C: $-q_{cc} = \lambda = 1.069$

Draw x: `x~dnExponential(1.069)` : x = 0.076

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



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

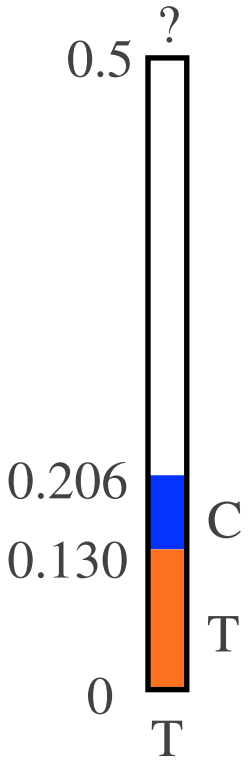
Generate waiting time to next event:

Rate when process in state C: $-q_{cc} = \lambda = 1.069$

Draw x: `x~dnExponential(1.069)` : x = 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:

Rate when process in state C: $-q_{cc} = \lambda = 1.069$

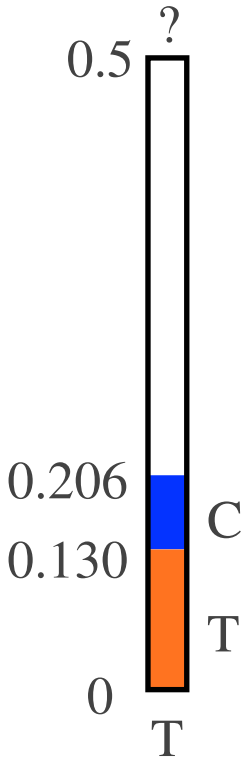
Draw x: `x~dnExponential(1.069)` : x = 0.076

Substitution probabilities in state C:

$$\text{P(C->A)} = q_{CA} / -q_{cc} = 0.148 \mid 1.069 = 0.138$$

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:

Rate when process in state C: $-q_{cc} = \lambda = 1.069$

Draw x: `x~dnExponential(1.069)` : x = 0.076

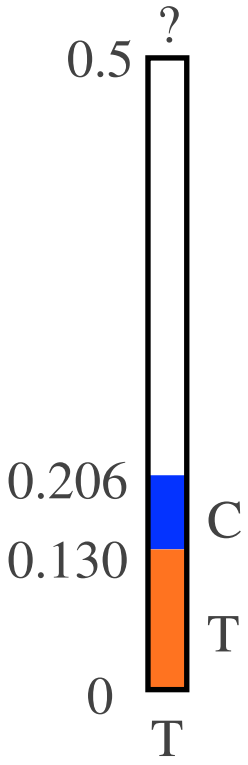
Substitution probabilities in state C:

$$P(C\text{->}A) = q_{CA}/-q_{cc} = 0.148 \mid 1.069 = 0.138$$

$$P(C\text{->}G) = q_{CG}/-q_{cc} = 0.415 \mid 1.069 = 0.388$$

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:

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

Draw x: `x~dnExponential(1.069)` : x = 0.076

Substitution probabilities in state C:

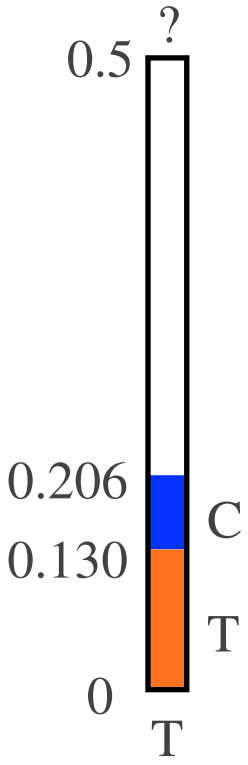
$$P(C\text{->}A) = q_{CA}/-q_{CC} = 0.148 \mid 1.069 = 0.138$$

$$P(C\text{->}G) = q_{CG}/-q_{CC} = 0.415 \mid 1.069 = 0.388$$

$$P(C\text{->}T) = q_{CT}/-q_{CC} = 0.506 \mid 1.069 = 0.474$$

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:

Rate when process in state C: $-q_{cc} = \lambda = 1.069$

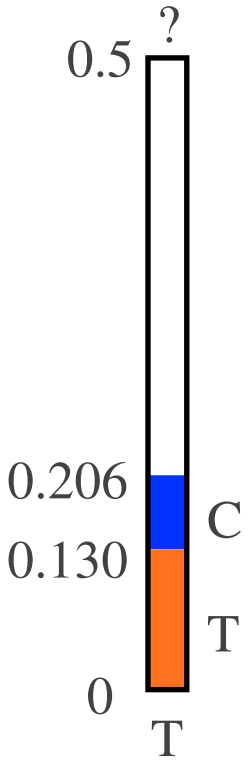
Draw x: `x~dnExponential(1.069)` : x = 0.076

Specify a set of intervals:

	intervals	
P(C->A) = 0.138	0 – 0.138	(choose A)
P(C->G) = 0.388	0.138 – 0.526	(choose G)
P(C->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:

Rate when process in state C: $-q_{cc} = \lambda = 1.069$

Draw x: `x~dnExponential(1.069)` : x = 0.076

Specify a set of intervals:

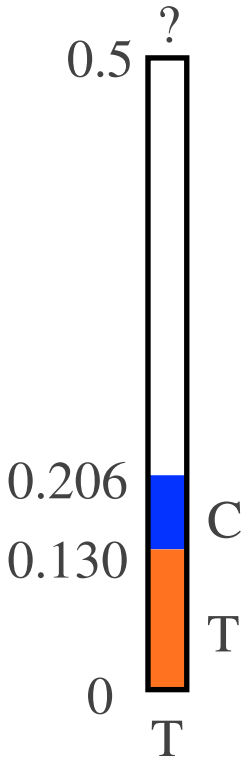
	intervals	
P(C->A) = 0.138	0 – 0.138	(choose A)
P(C->G) = 0.388	0.138 – 0.526	(choose G)
P(C->T) = 0.474	0.526 – 1	(choose T)

Draw a uniformly distributed number, u, to select substitution event:

`u~dnUniform(0,1)` : u=0.317

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:

Rate when process in state C: $-q_{cc} = \lambda = 1.069$

Draw x: `x~dnExponential(1.069)` : x = 0.076

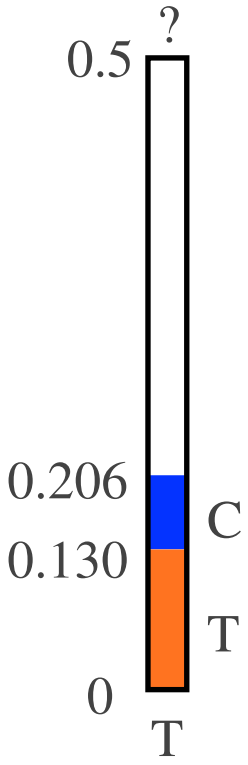
Specify a set of intervals:	intervals	
P(C->A) = 0.138	0 – 0.138	(choose A)
P(C->G) = 0.388	0.138 – 0.526	(choose G)
P(C->T) = 0.474	0.526 – 1	(choose T)

Draw a uniformly distributed number, u, to select substitution event:

`u~dnUniform(0,1)` : u=0.317

A Mechanistic Interpretation of the Rate Matrix

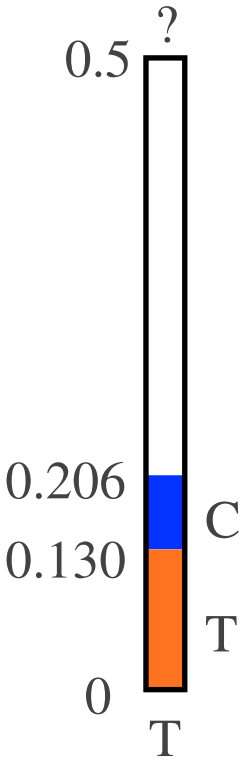
A simple Monte Carlo Simulation



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

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



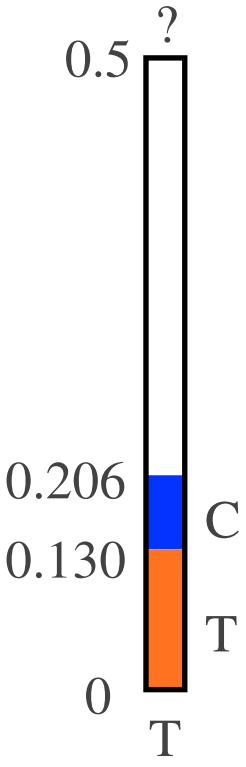
$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

Generate waiting time to next event:

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

A Mechanistic Interpretation of the Rate Matrix

A simple Monte Carlo Simulation



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

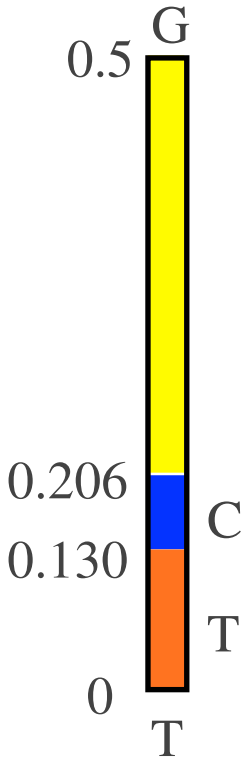
Generate waiting time to next event:

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

Draw x: `x~dnExponential(1.069)` : x = 1.820

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:

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

Draw x: `x~dnExponential(1.069)` : x = 1.820

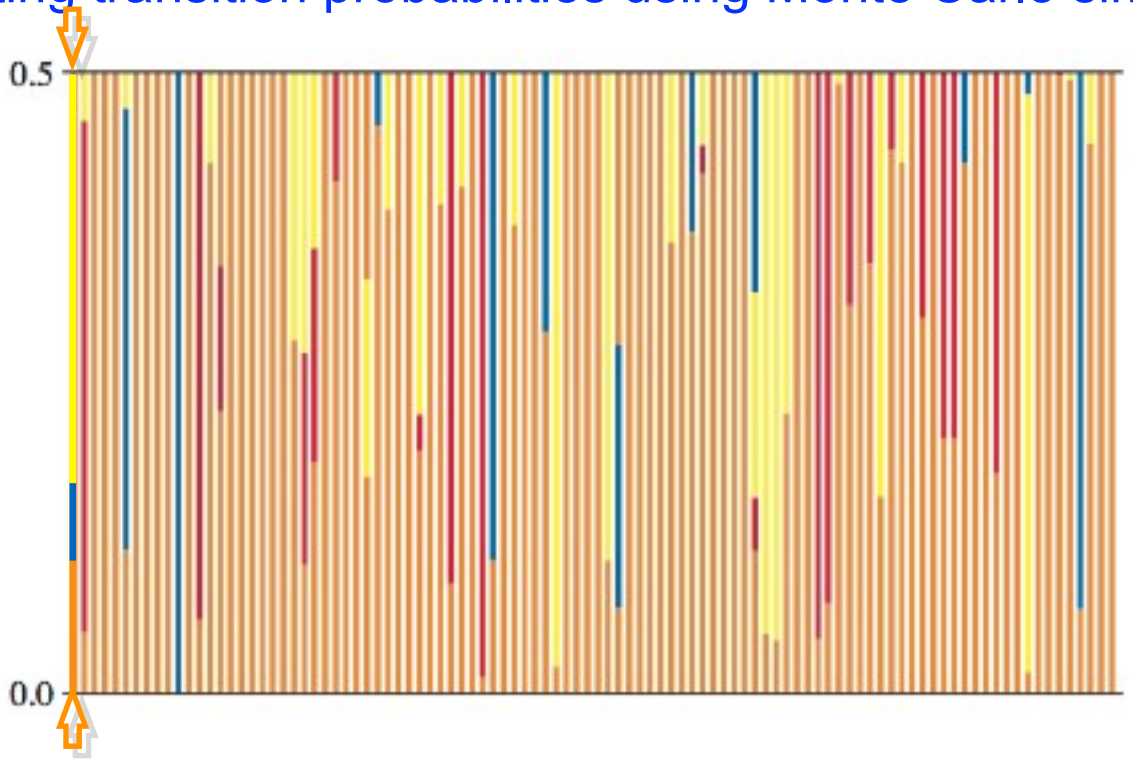
1.820 > 0.5-0.206 → *Terminate simulation (in state G)*

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation

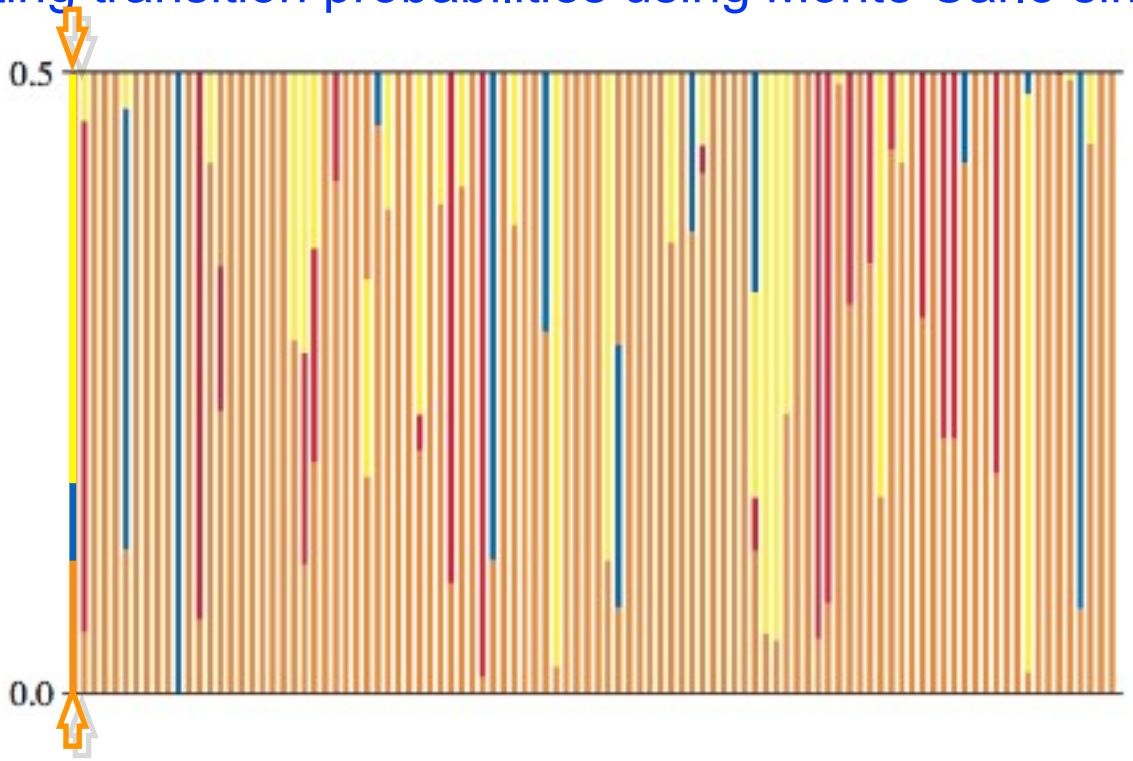


T→ C→ G

Ended in G with two changes

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation

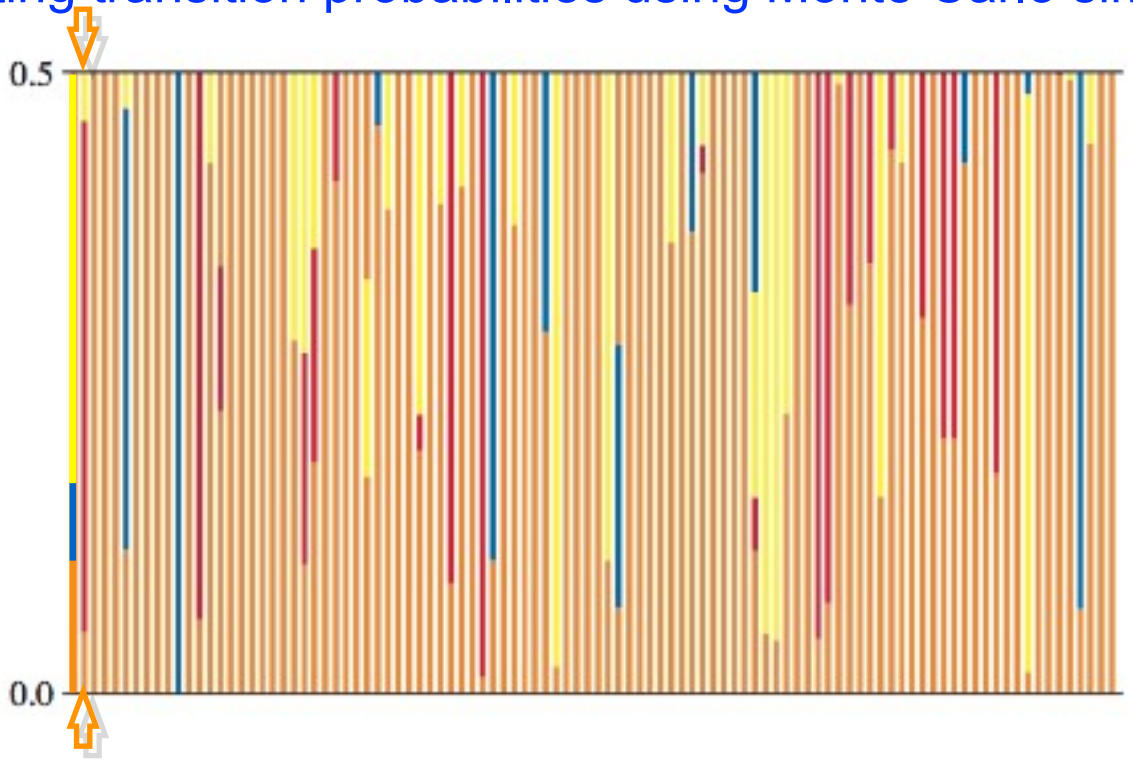


T→ C→ G

Ended in G with two changes

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



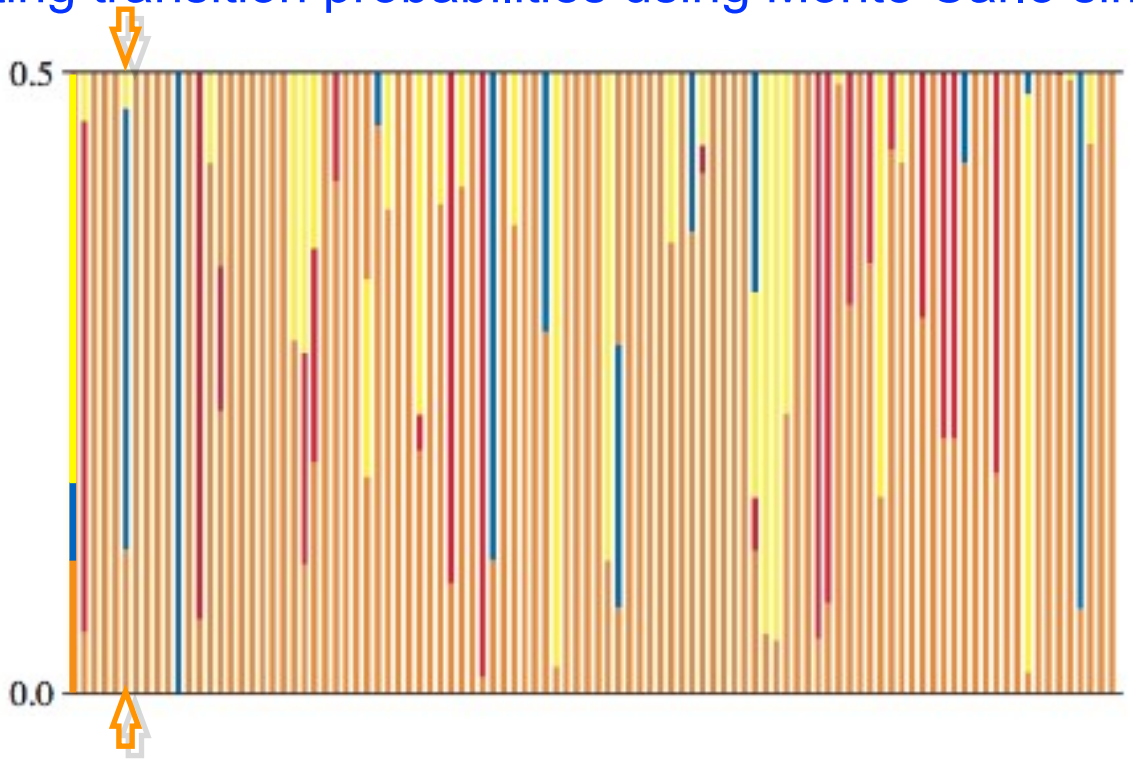
G 2

T → A → G

Ended in G with two changes

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation

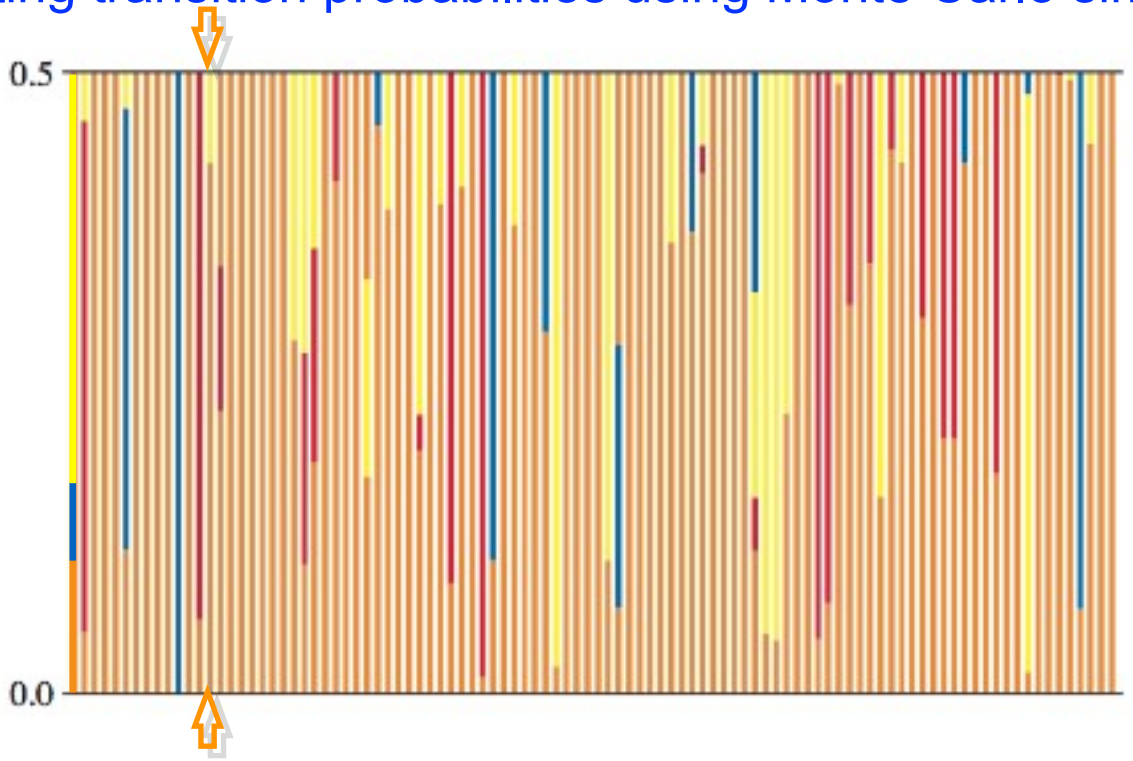


T → C → G

Ended in G with two changes

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



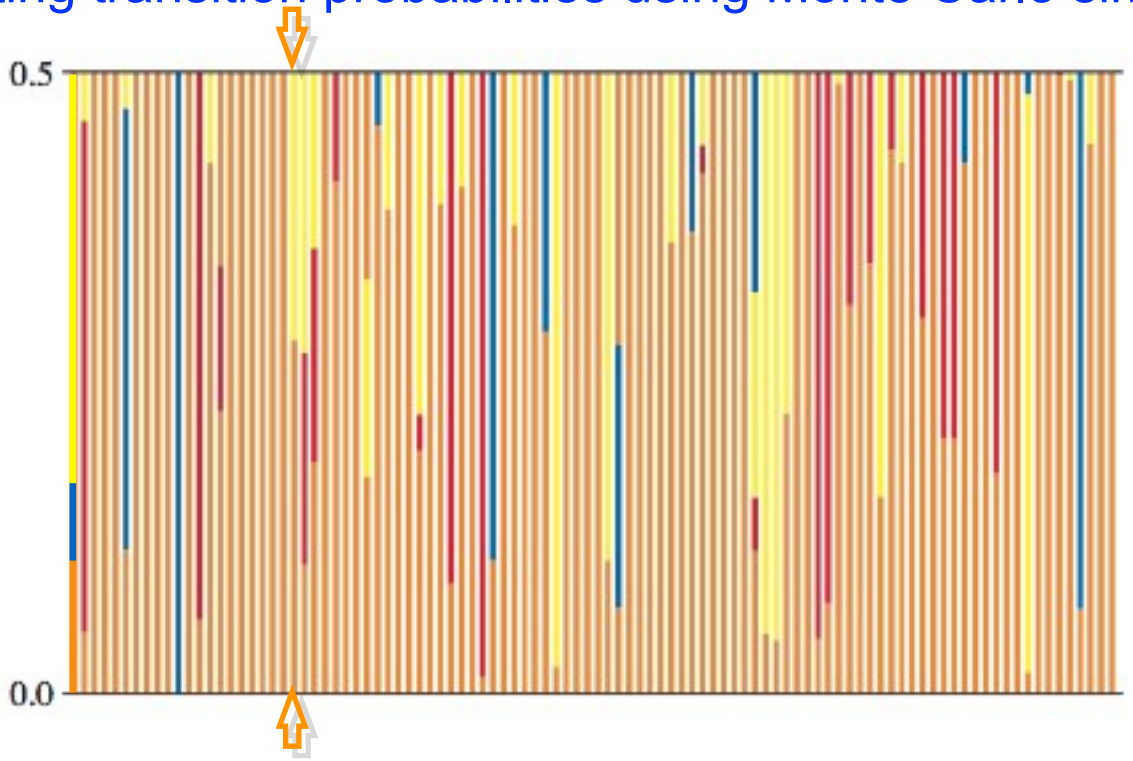
G 4

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



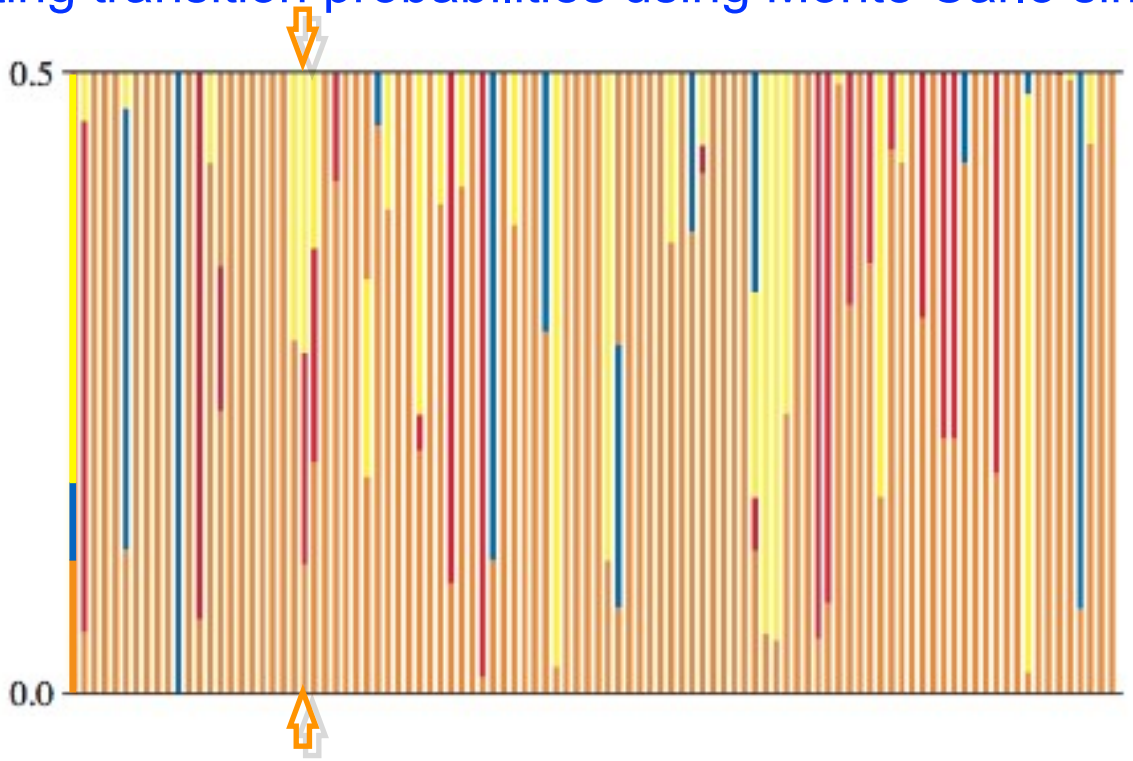
G 5

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

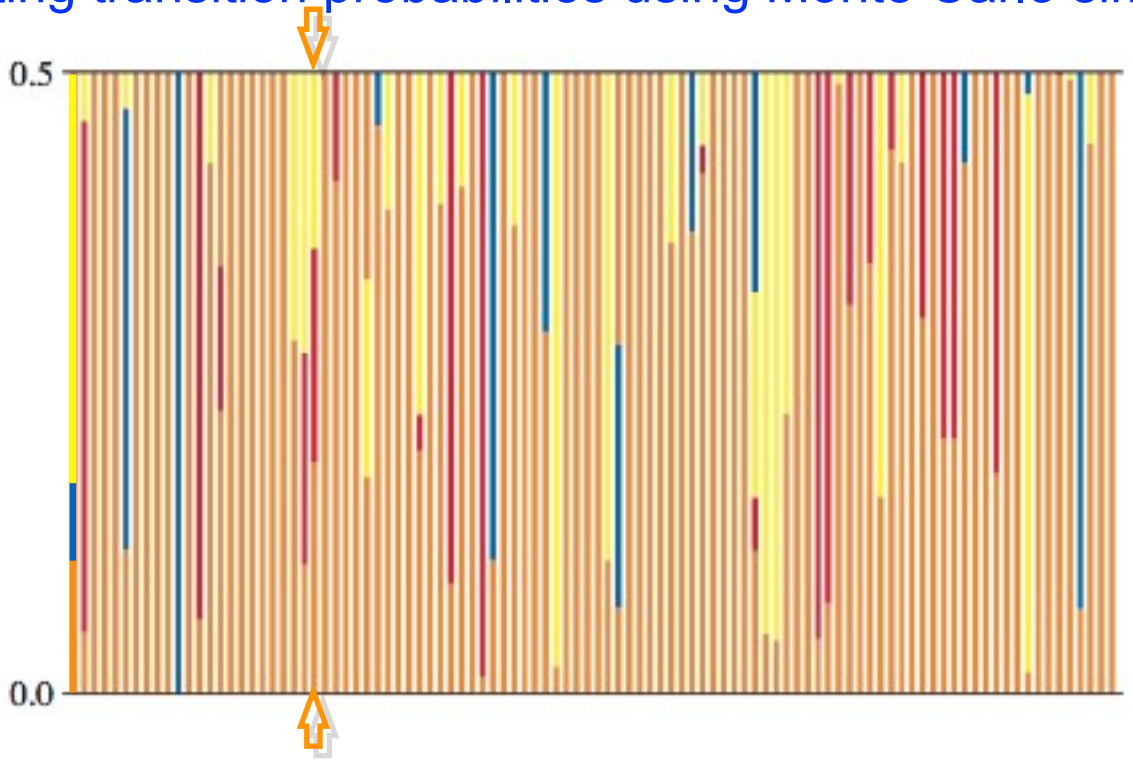
6

T → A → G

Ended in G with two changes

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



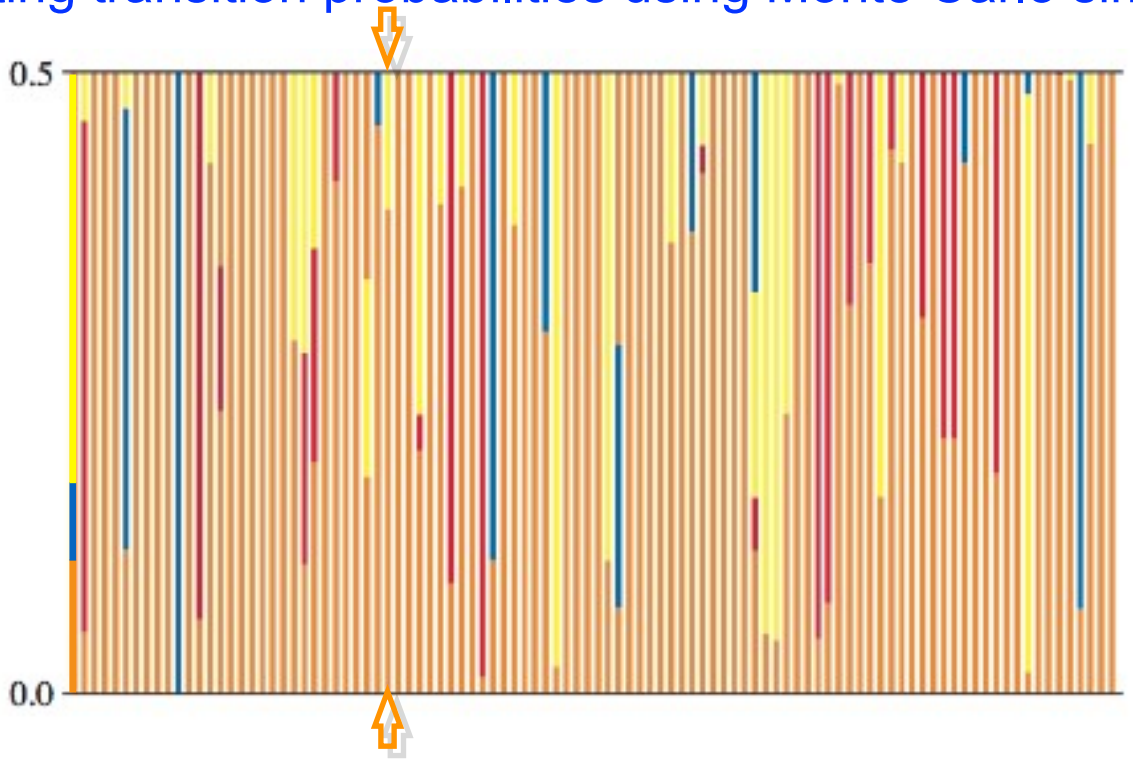
G 7

T → A → G

Ended in G with two changes

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

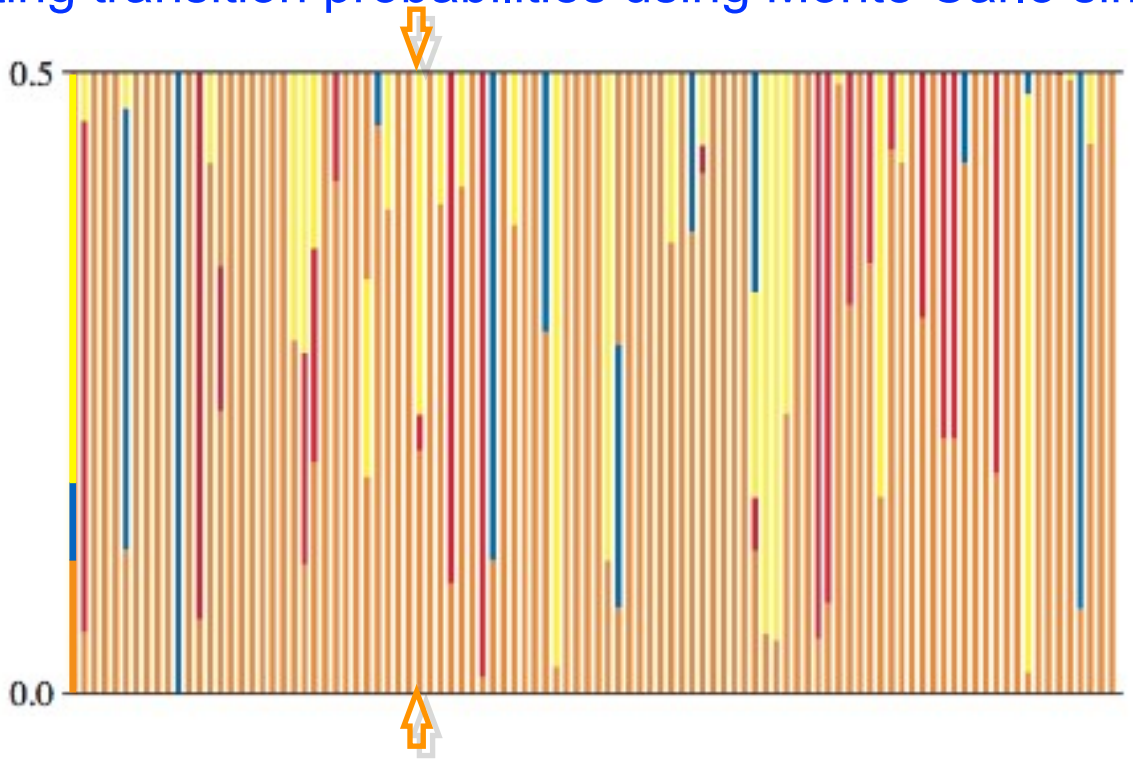
8

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

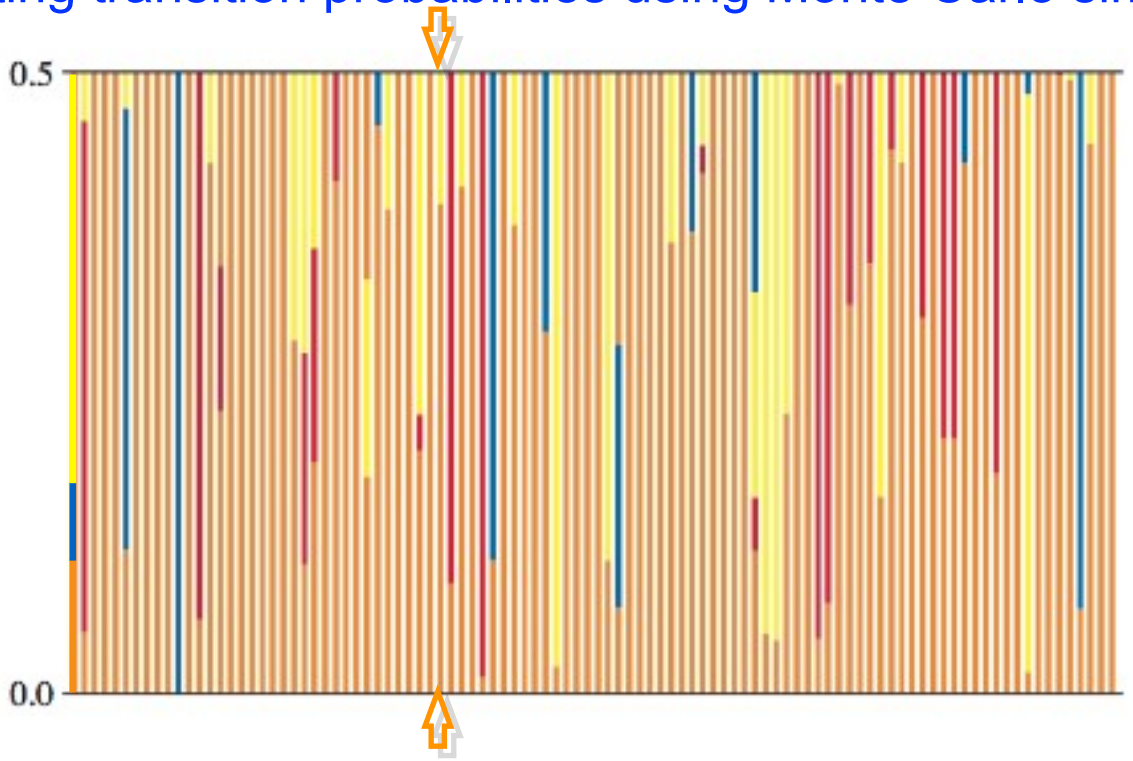
9

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



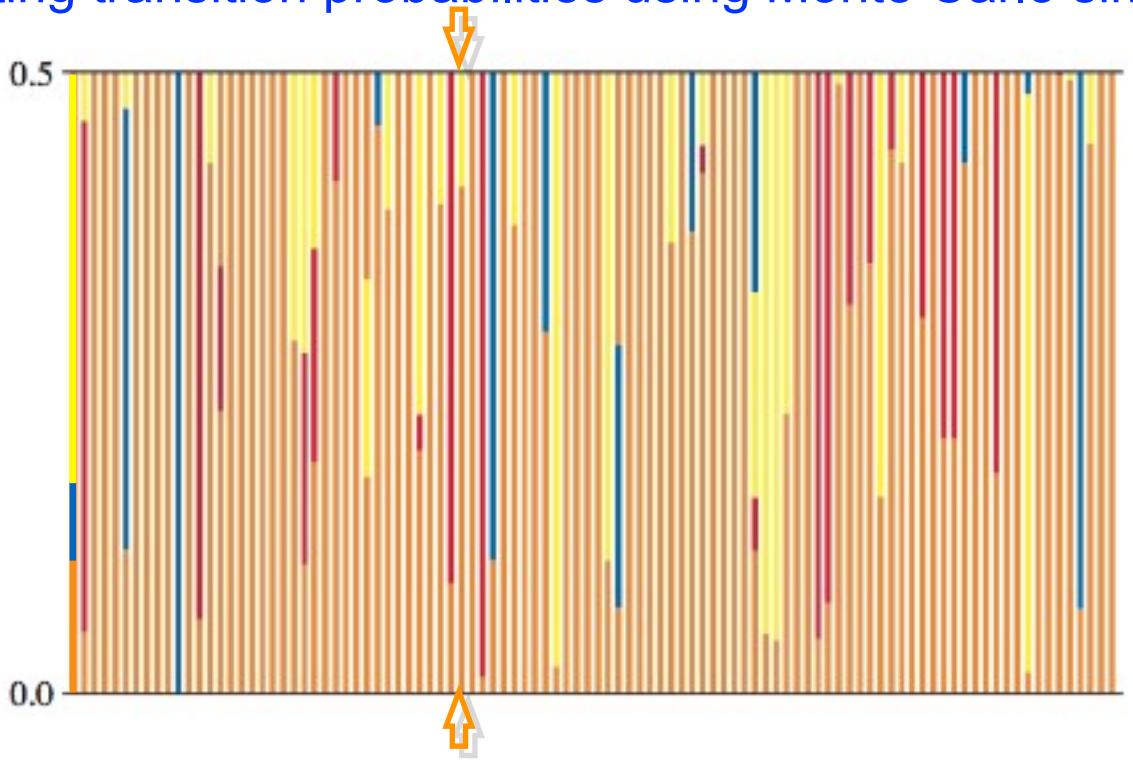
G 10

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

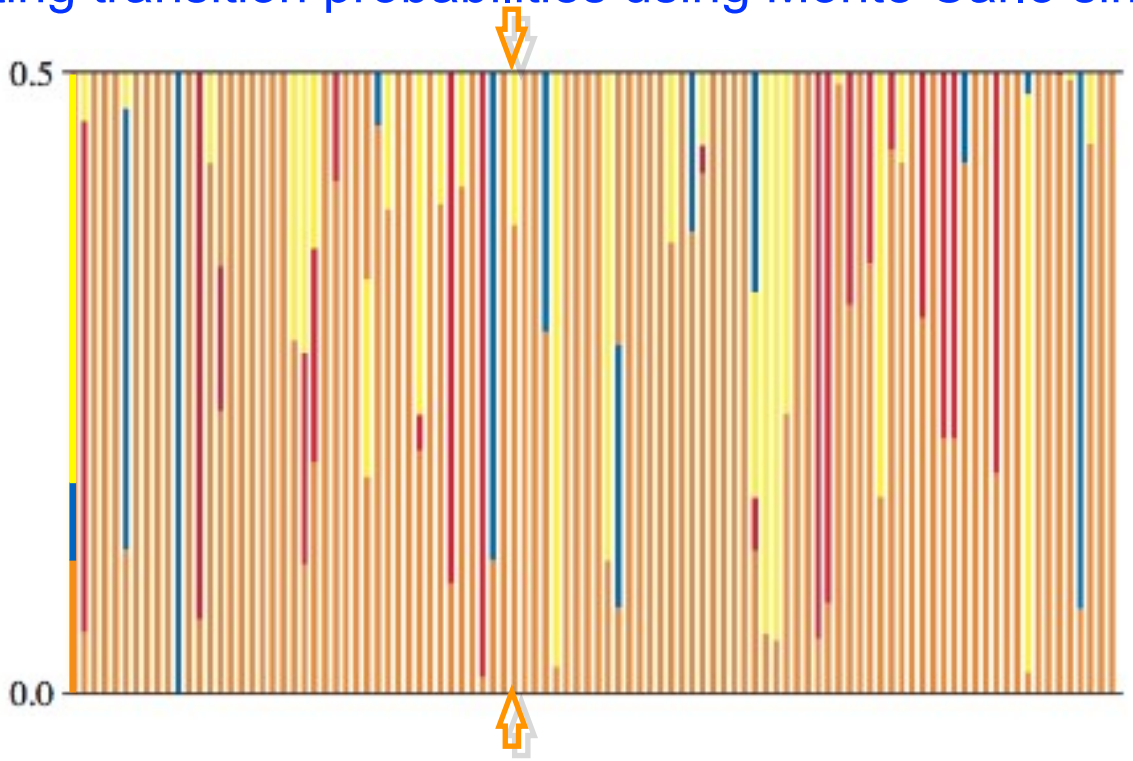
11

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



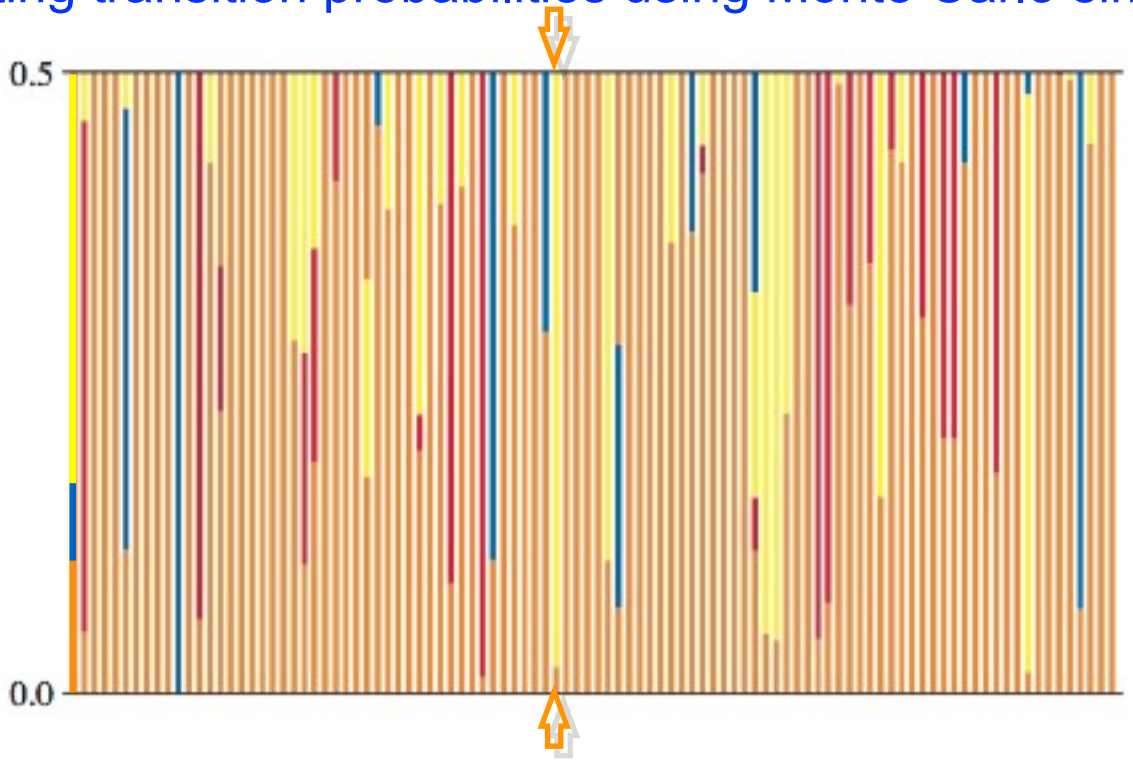
G 12

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

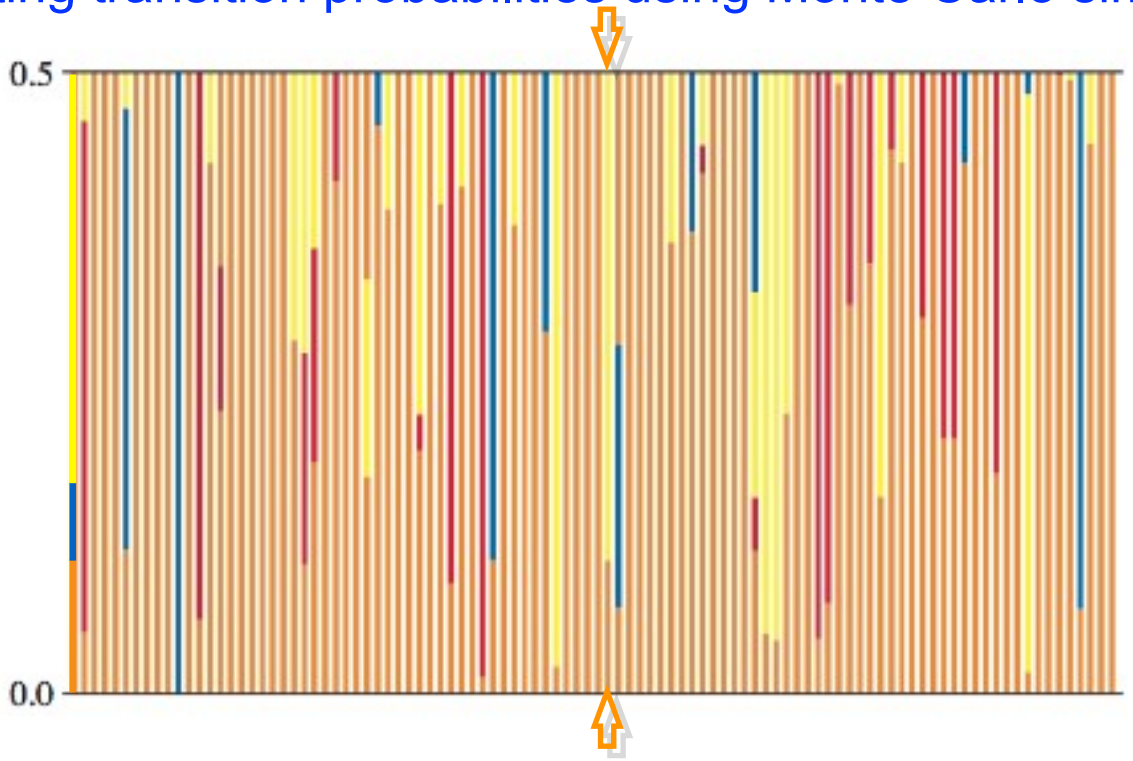
13

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



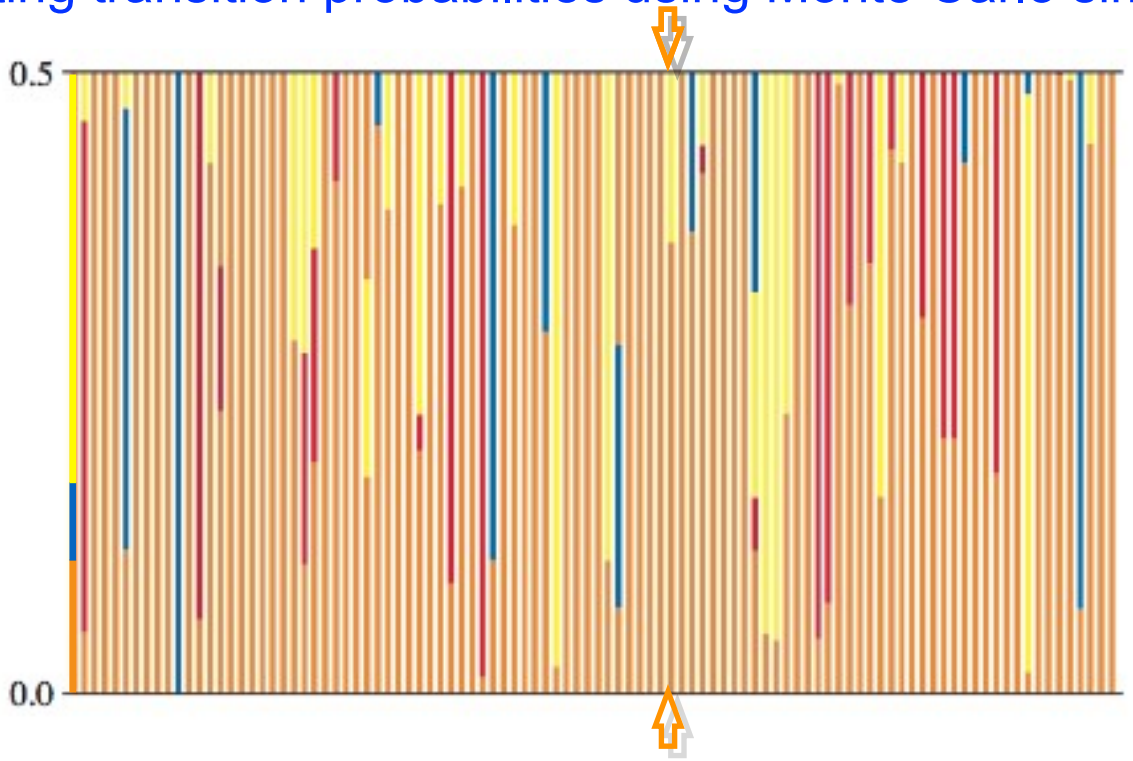
G

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



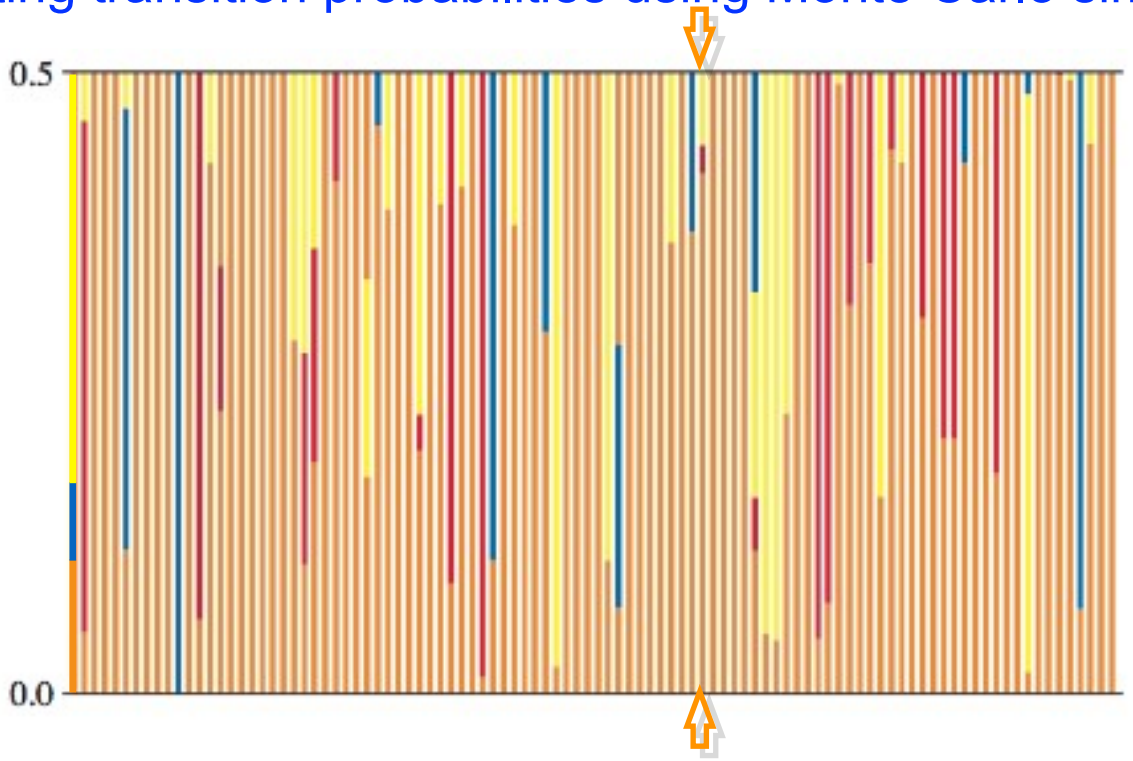
G

15

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

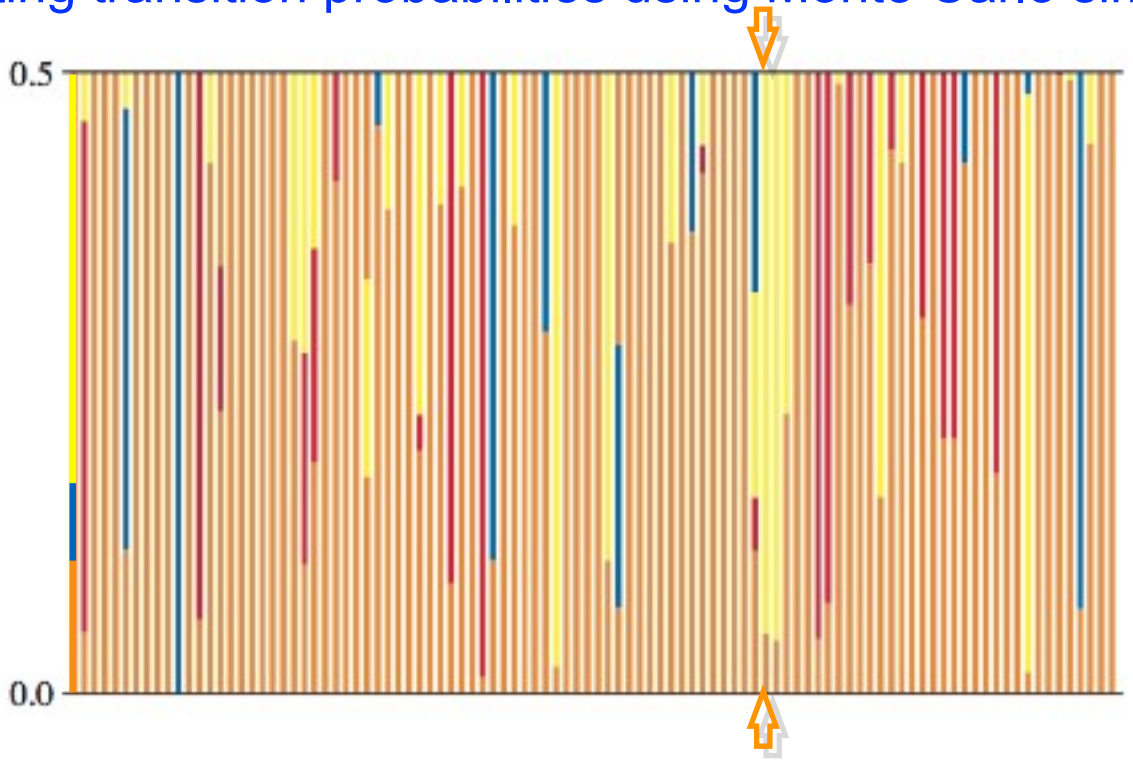
16

T → A → G

Ended in G with two changes

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

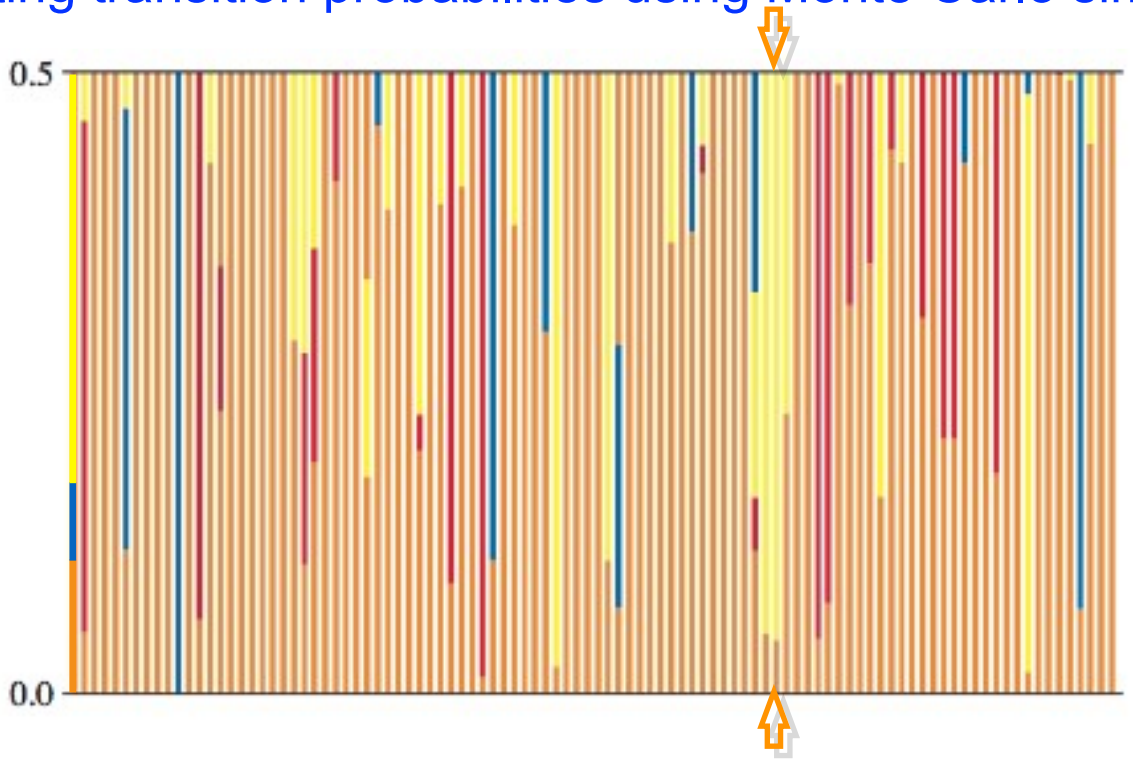
17

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

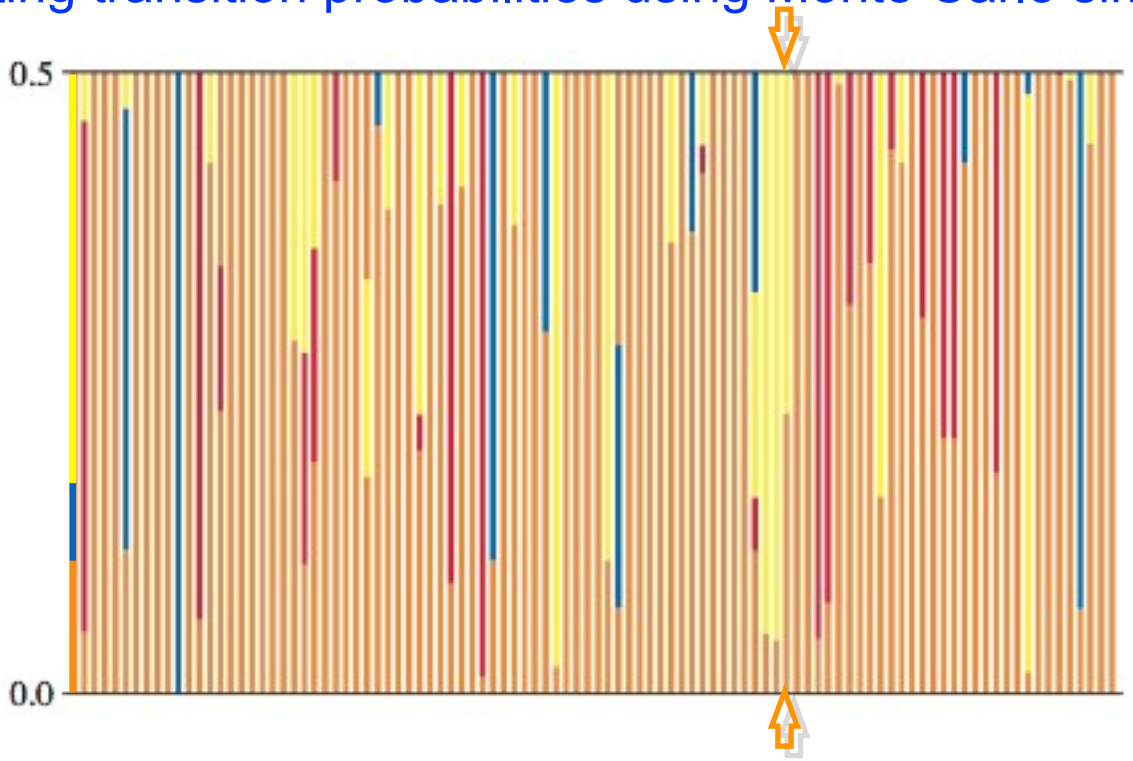
18

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

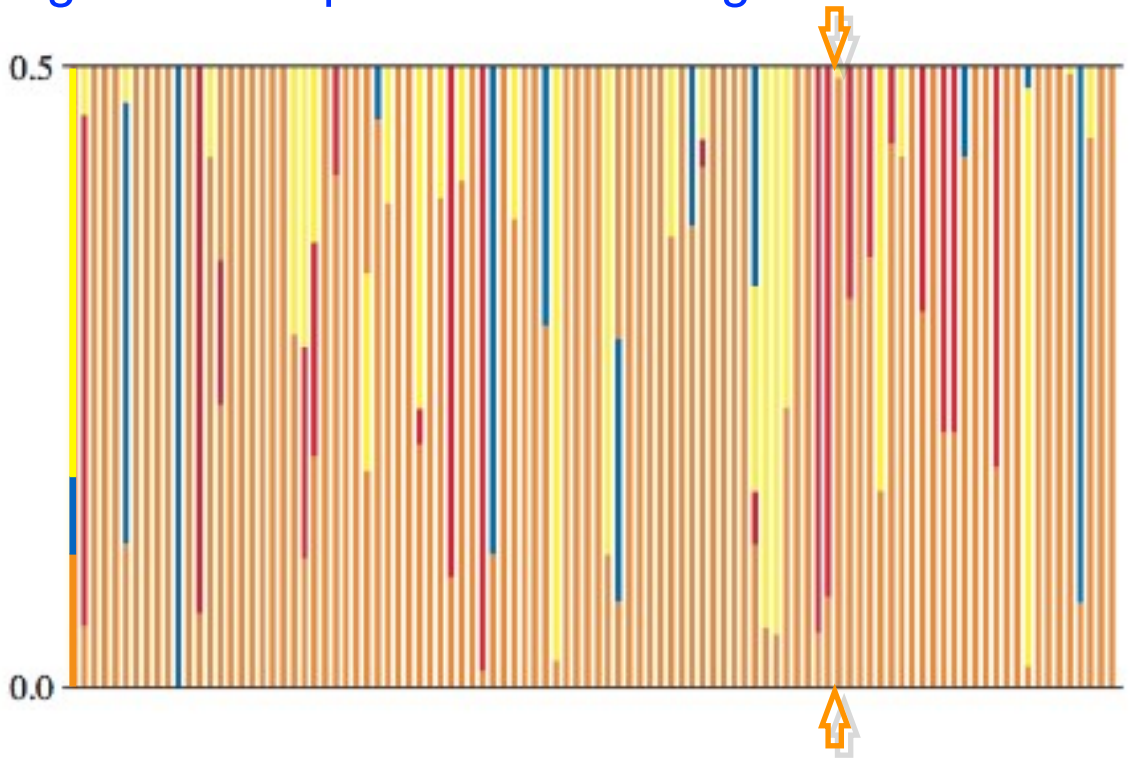
19

T → G

Ended in G with one change

Transition Probabilities of a CTMC

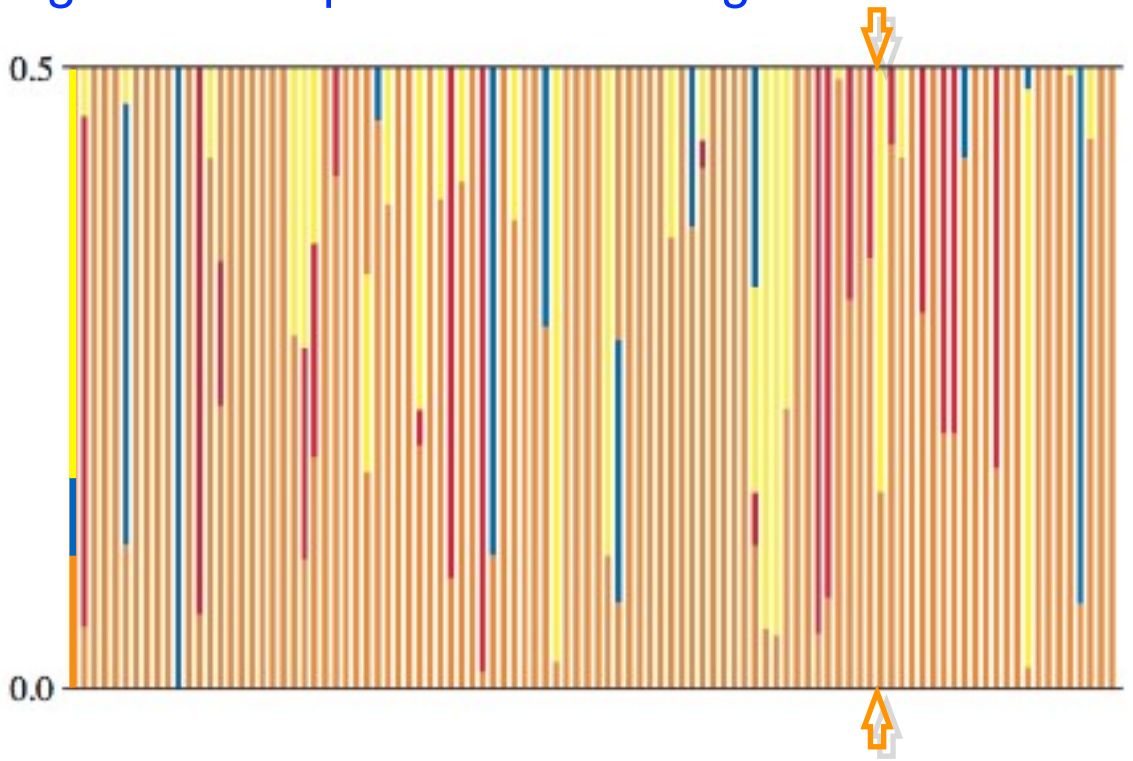
Estimating transition probabilities using Monte Carlo simulation



Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

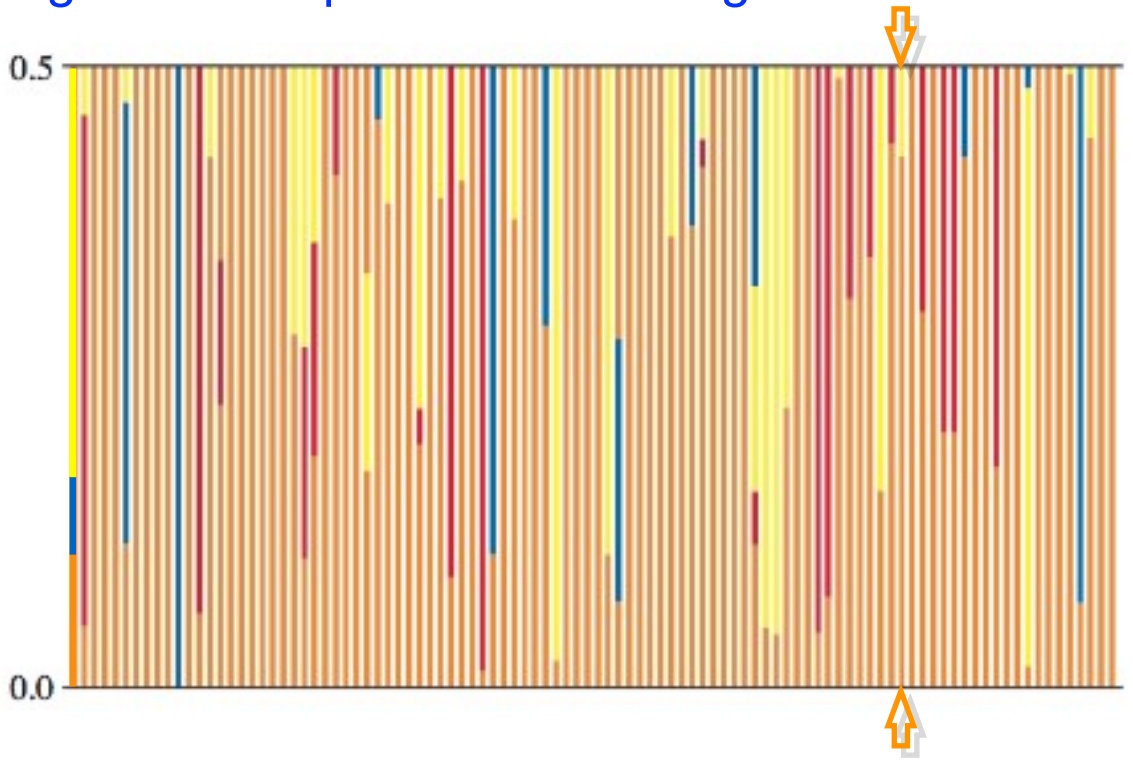
21

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



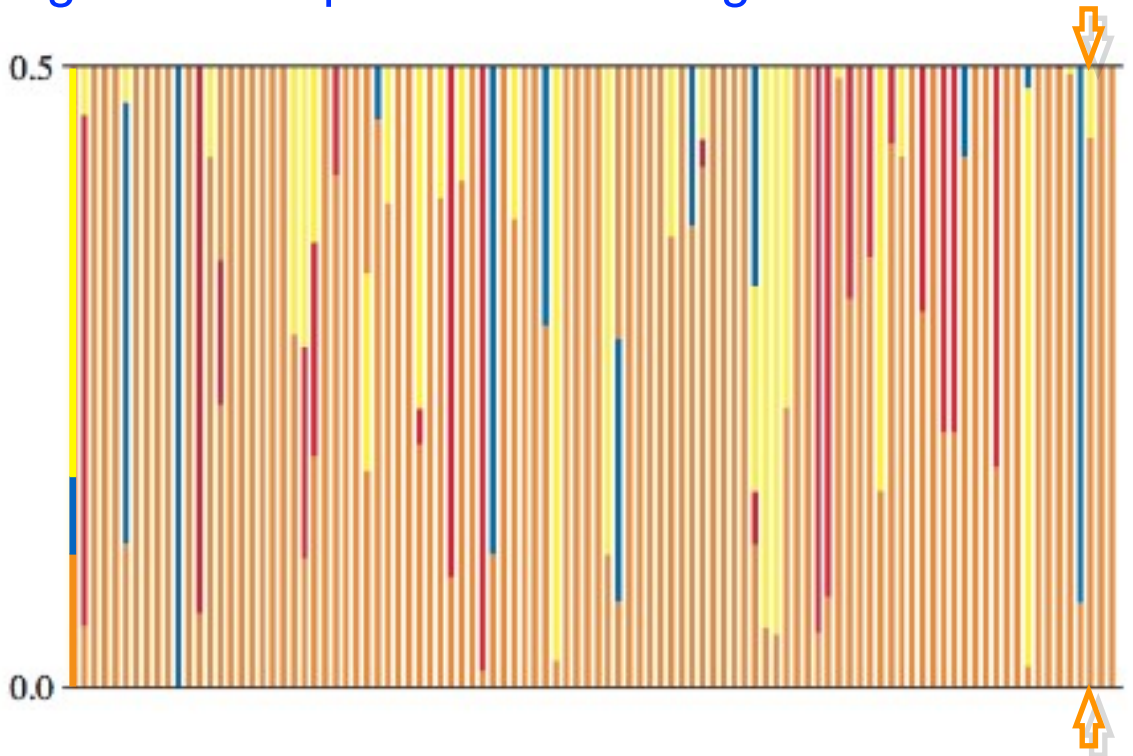
G 22

T → G

Ended in G with one change

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

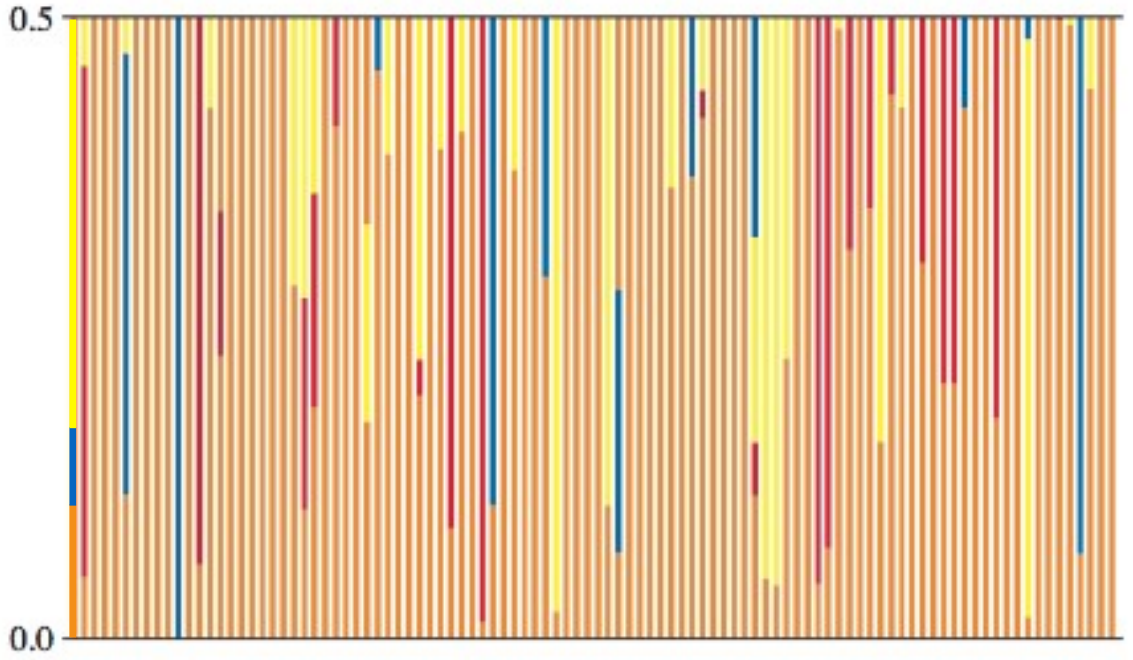
23

T → G

Ended in G with one change

Transition Probabilities of a CTMC

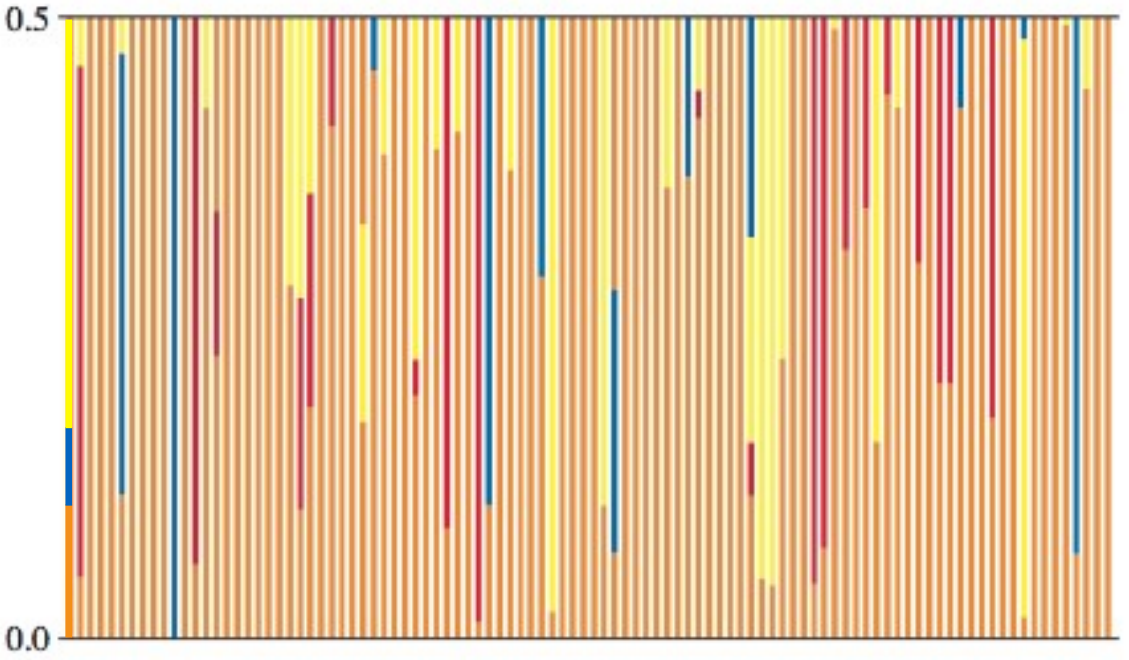
Estimating transition probabilities using Monte Carlo simulation



What is the probability that the process ends in G given that we started in T (and given the other parameters of the model)?

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



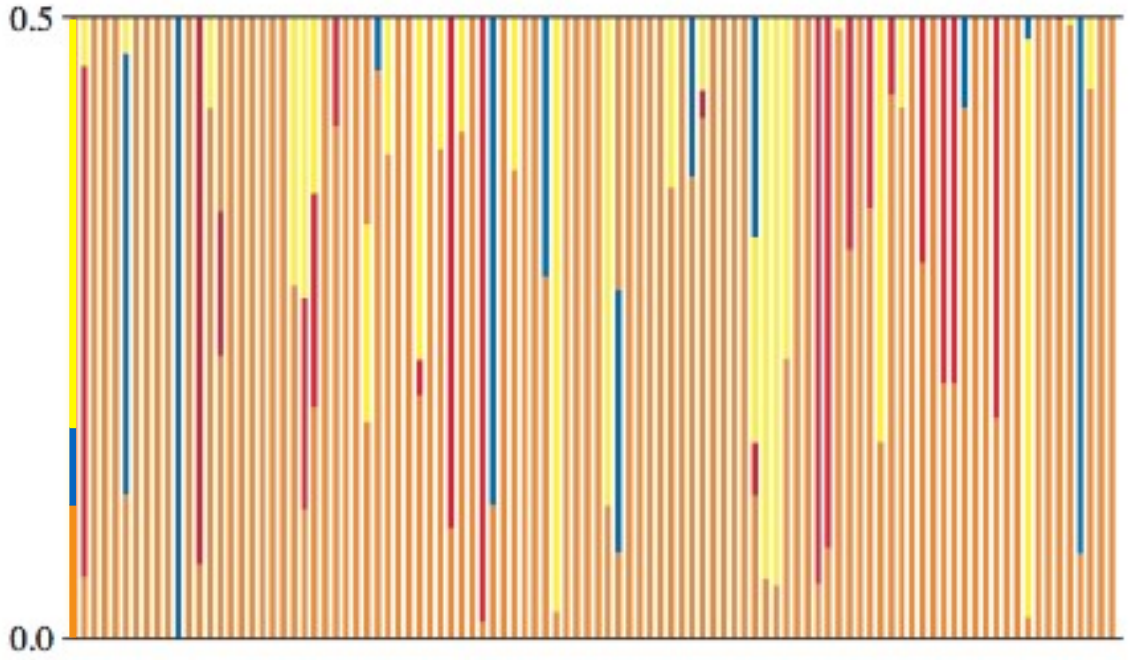
number of outcomes

G 23

What is the probability that the process ends in G given that we started in T (and given the other parameters of the model)?

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



G

number of outcomes

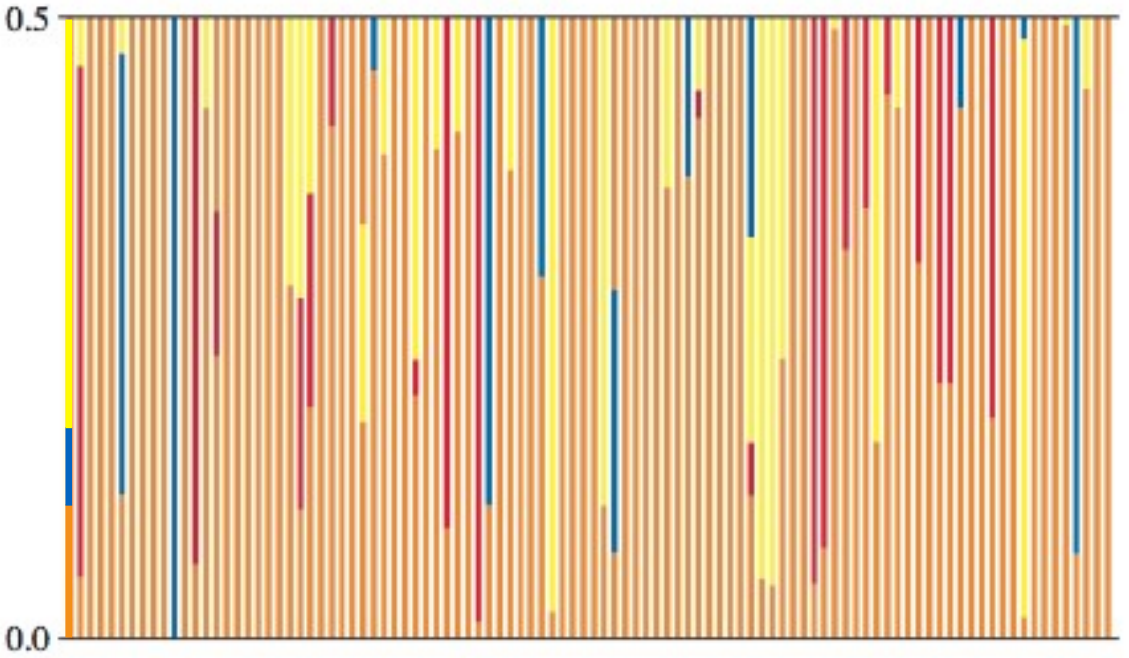
$23 \div 100$

number of simulations

What is the probability that the process ends in G given that we started in T (and given the other parameters of the model)?

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



number of outcomes

G

$23 \div 100$

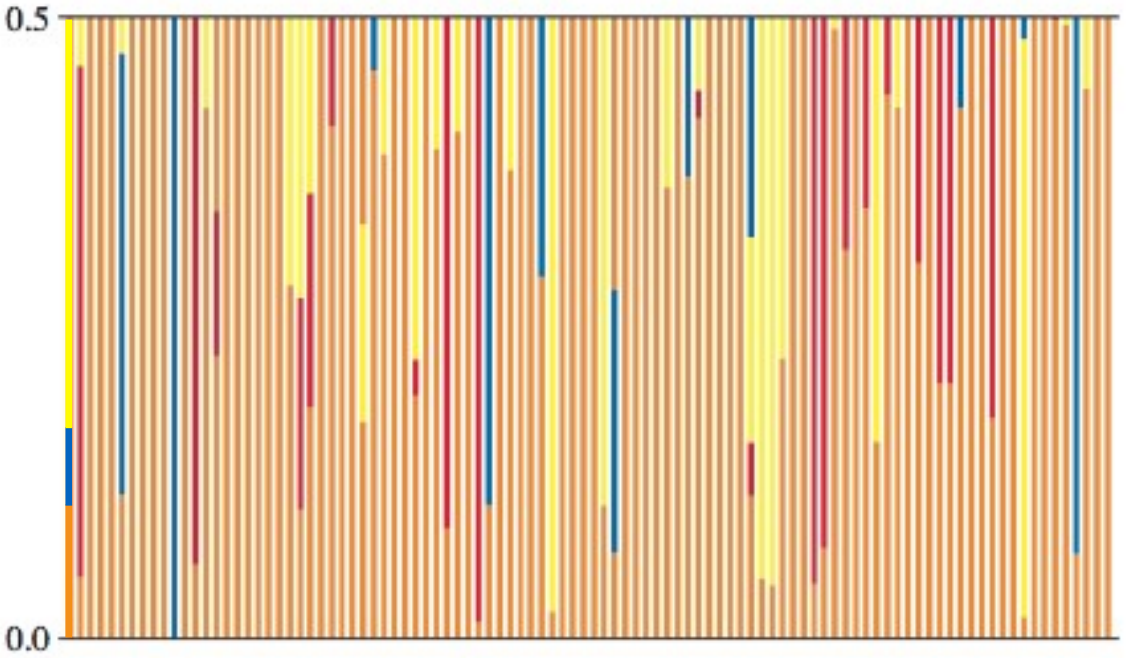
number of simulations

		To			
		A	C	G	T
From	A				
	C				
	G				
	T			0.23	

What is the probability that the process ends in G given that we started in T (and given the other parameters of the model)?

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation



number of outcomes

G

$23 \div 100$

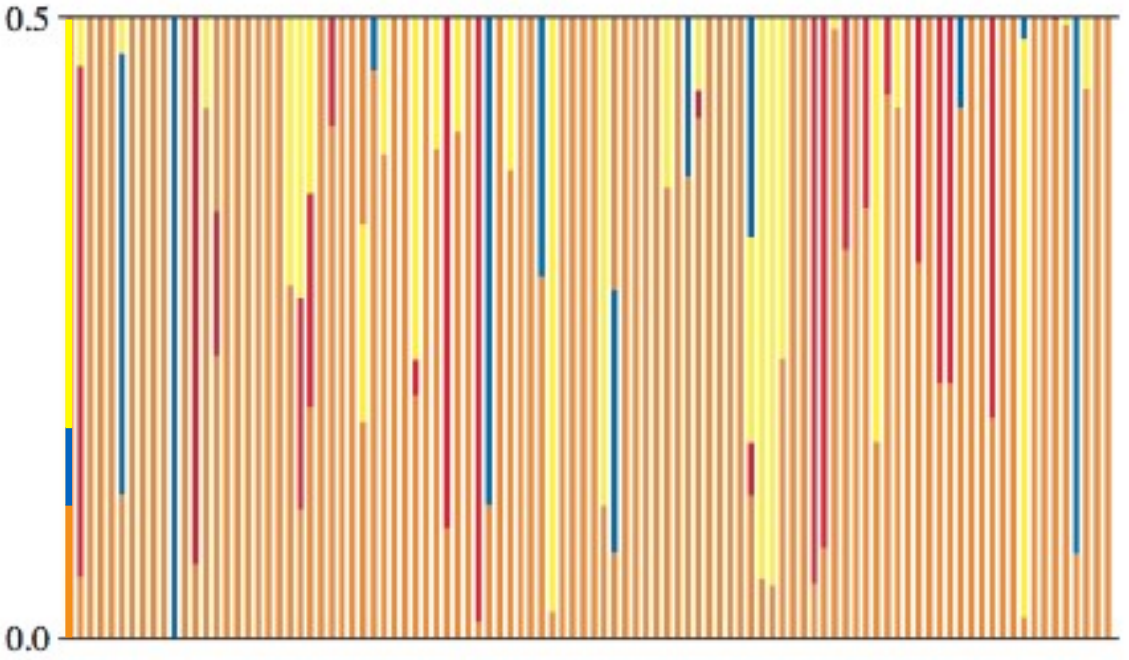
number of simulations

		To			
		A	C	G	T
From	A				
	C				
	G				
	T			0.23	

This ‘transition probability’ reflects all possible histories that start in T and end in G (*i.e.*, histories with different numbers and/or positions of changes)

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation

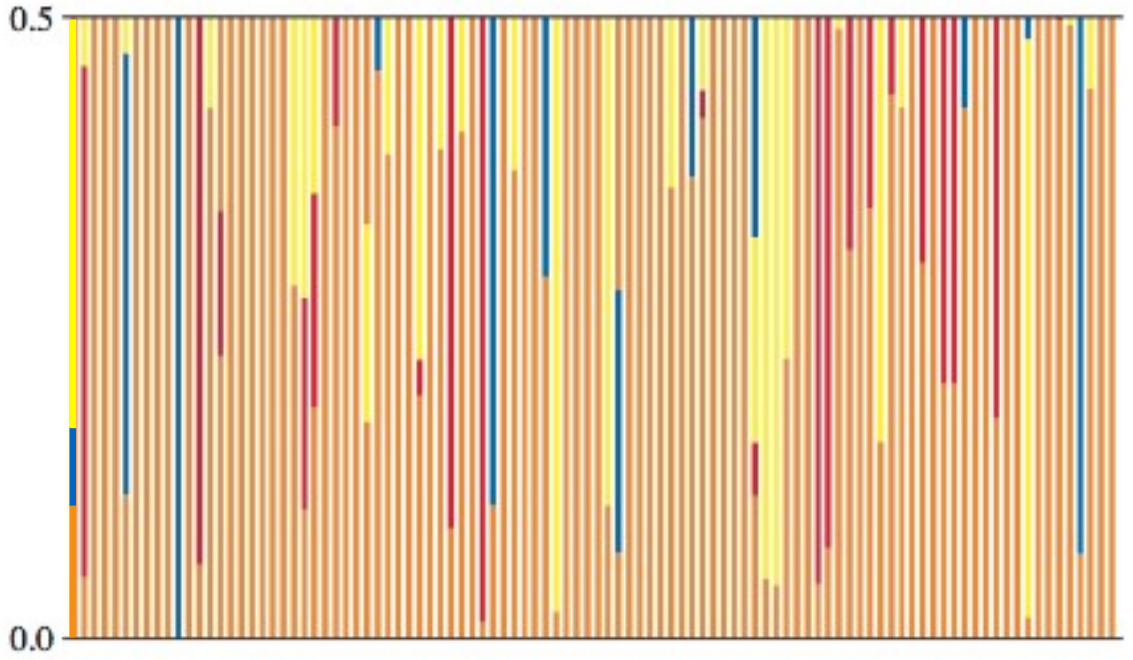


		To A	C	G	T
From	A				
	C				
	G				
	T	0.14		0.23	

We can repeat this process for the other end states…

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation

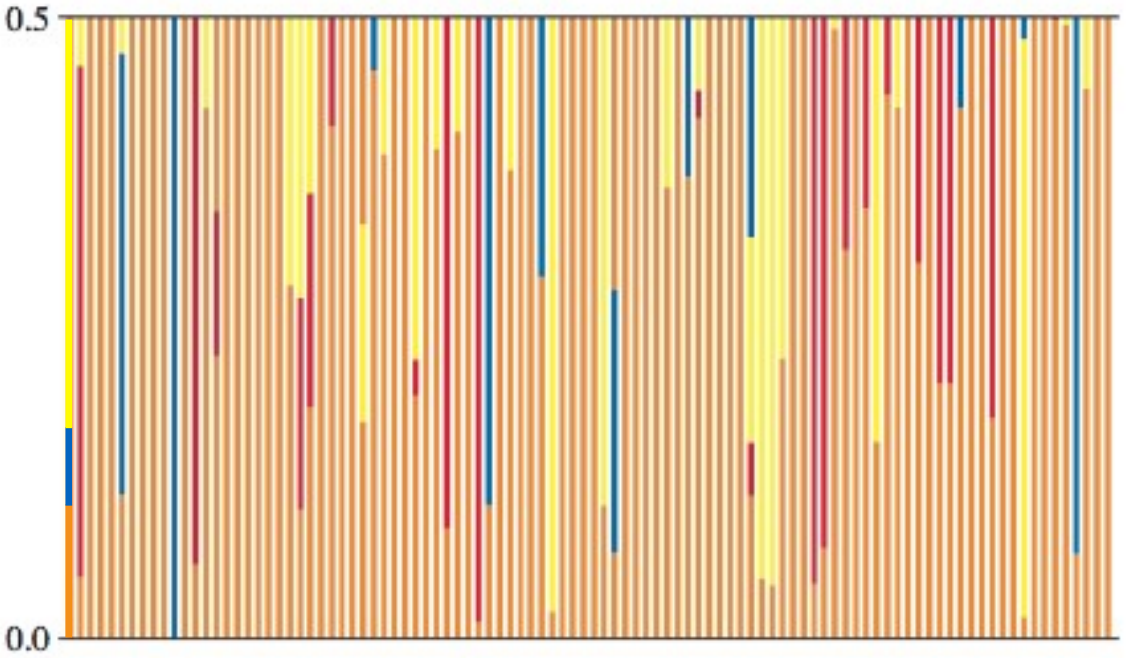


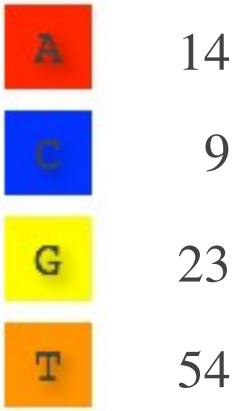
		To A	C	G	T
From	A				
	C				
	G				
	T	0.14	0.09	0.23	

We can repeat this process for the other end states…

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation





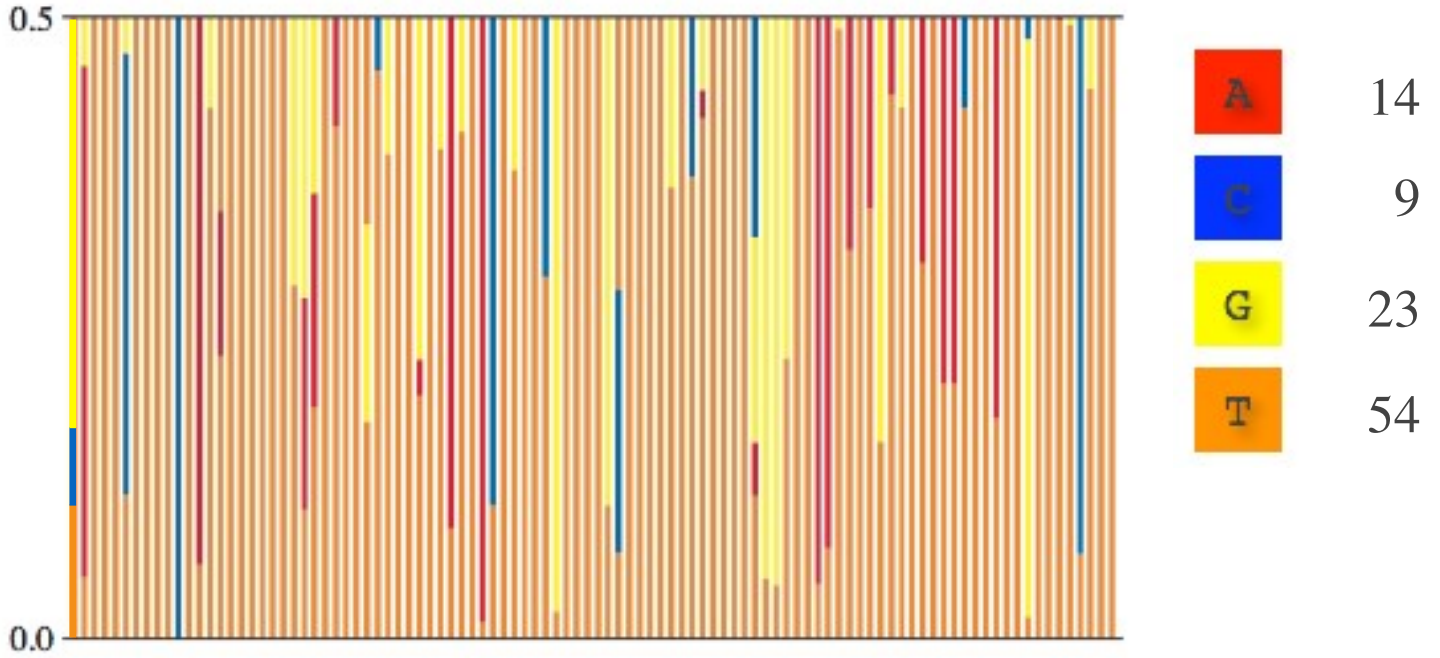
14
9
23
54

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

We can repeat this process for the other end states…

Transition Probabilities of a CTMC

Estimating transition probabilities using Monte Carlo simulation

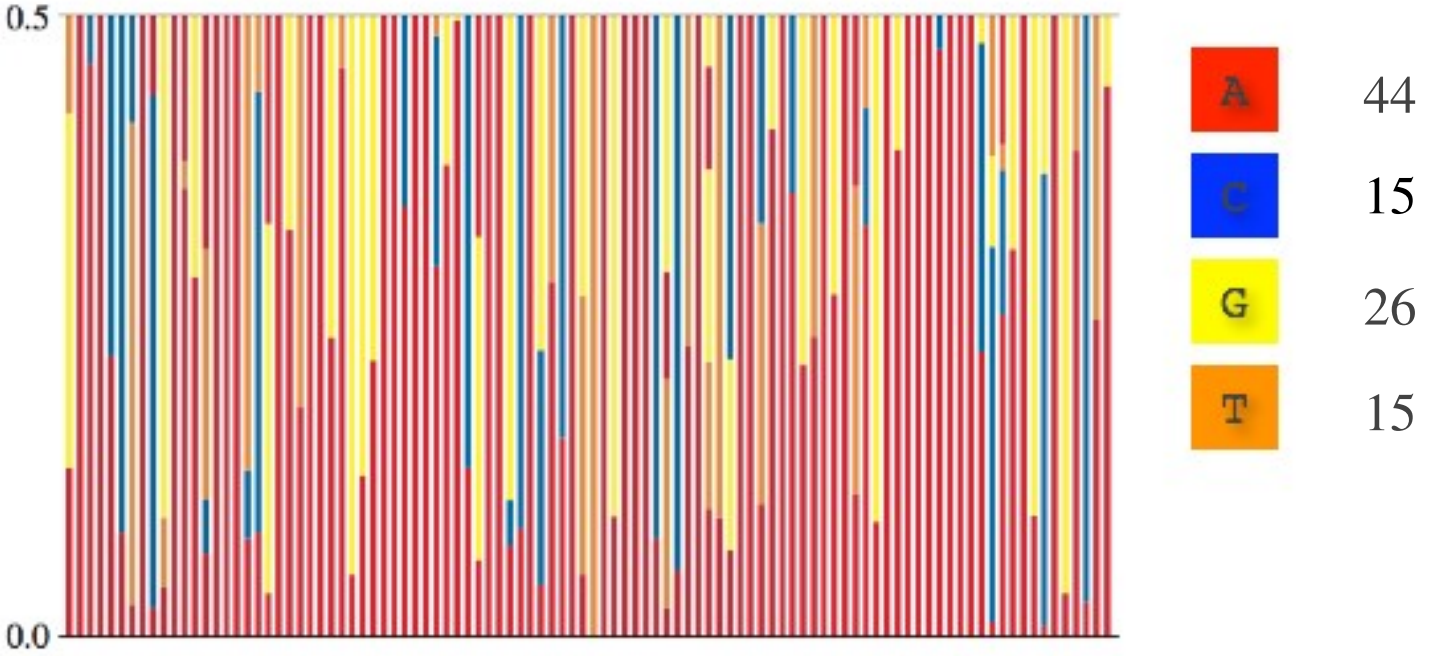


		To A	C	G	T
From	A				
	C				
	G				
	T	0.14	+ 0.09	+ 0.23	+ 0.54 = 1

Note that each row of the transition-probability matrix sums to 1 (*c.f.*, the Law of Total Probability).

Transition Probabilities of a CTMC

Realizations of 100 replicate simulations starting in state A

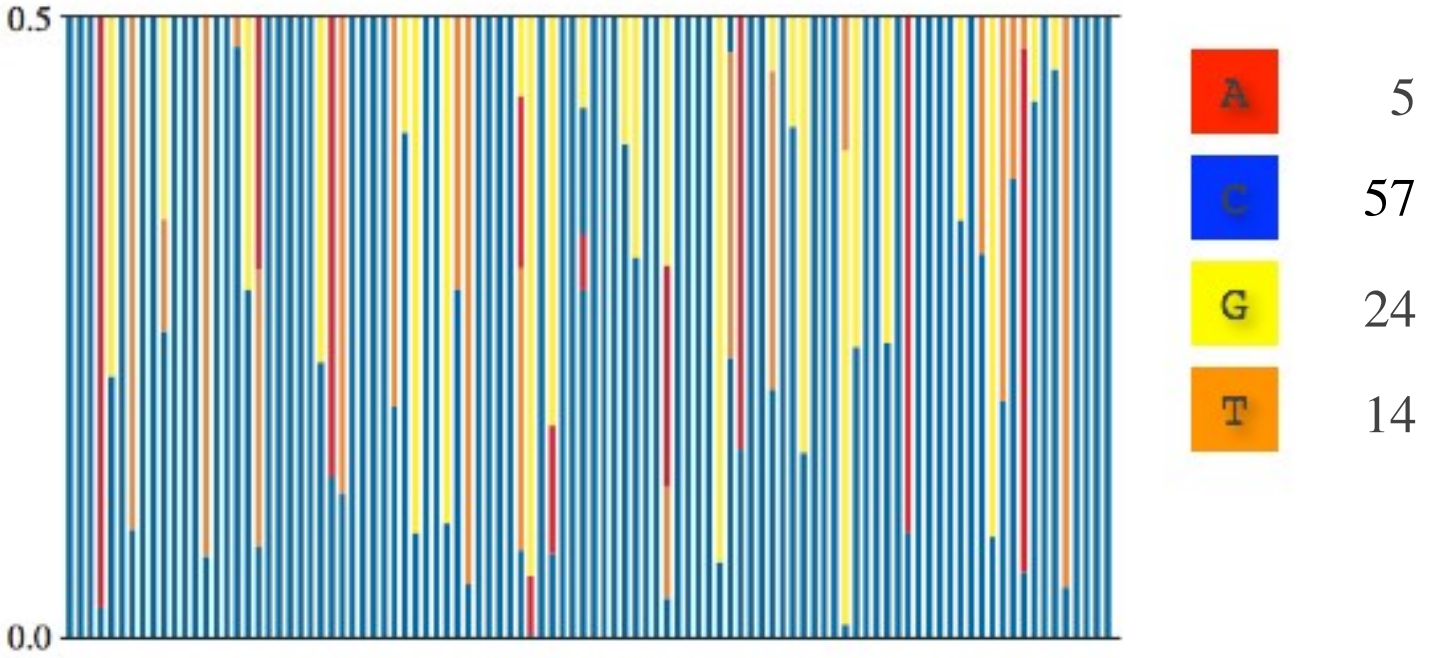


		To A	C	G	T
From	A	0.44	0.15	0.26	0.15
	C				
	G				
	T	0.14	0.09	0.23	0.54

We can perform new Monte Carlo simulations that start in A to fill out the corresponding row of the transition-probability matrix.

Transition Probabilities of a CTMC

Realizations of 100 replicate simulations starting in state C

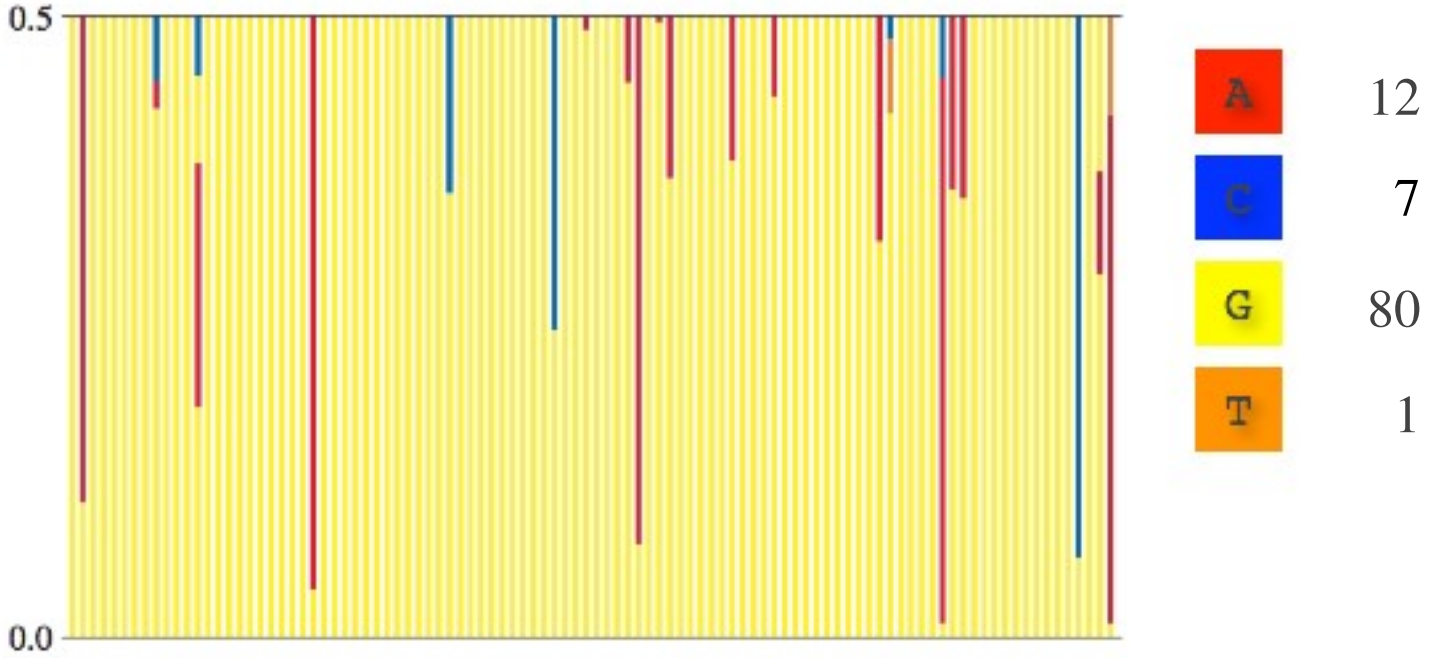


		To A	C	G	T
From	A	0.44	0.15	0.26	0.15
	C	0.05	0.57	0.24	0.14
	G				
	T	0.14	0.09	0.23	0.54

And then for simulations that start in C …

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

And finally for simulations that start in G.

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

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

Analytical solutions for the transition probabilities: matrix exponentiation

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

Transition Probabilities of a CTMC

Analytical solutions for the transition probabilities: matrix exponentiation

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

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: $\mathbf{P}(v) = e^{\mathbf{Q}v}$

Transition Probabilities of a CTMC

Analytical solutions for the transition probabilities: matrix exponentiation

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

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:

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

Transition Probabilities of a CTMC

Analytical solutions for the transition probabilities: matrix exponentiation

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

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

Analytical solutions for the transition probabilities: matrix exponentiation

Monte Carlo simulation is computationally expensive and unnecessary, as the

SIAM REVIEW
Vol. 45, No. 1, pp. 3–000

© 2003 Society for Industrial and Applied Mathematics

Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*

Cleve Moler†
Charles Van Loan‡

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

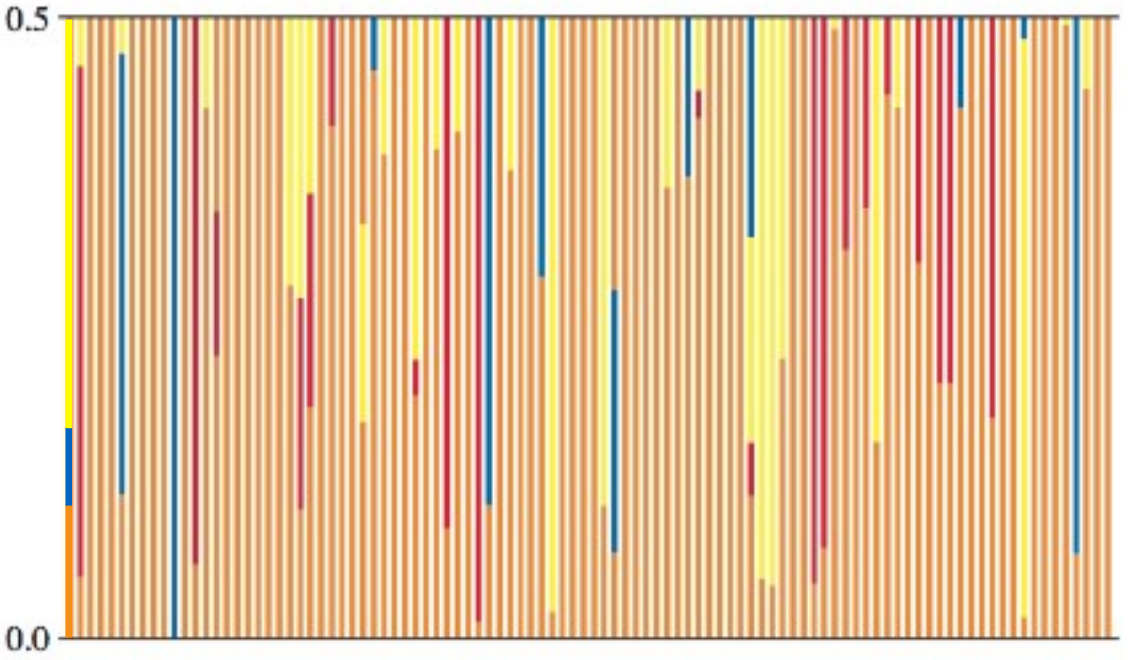
Transition Probabilities of a CTMC

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 of a CTMC

Reminder:



number of outcomes

G

$23 \div 100$

number of simulations

		To A	C	G	T
From	A				
	C				
	G				
	T			0.23	

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)

π_i

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, $P(v) = e^{Qv}$

π_i

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

π_i

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:

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

π_i

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:

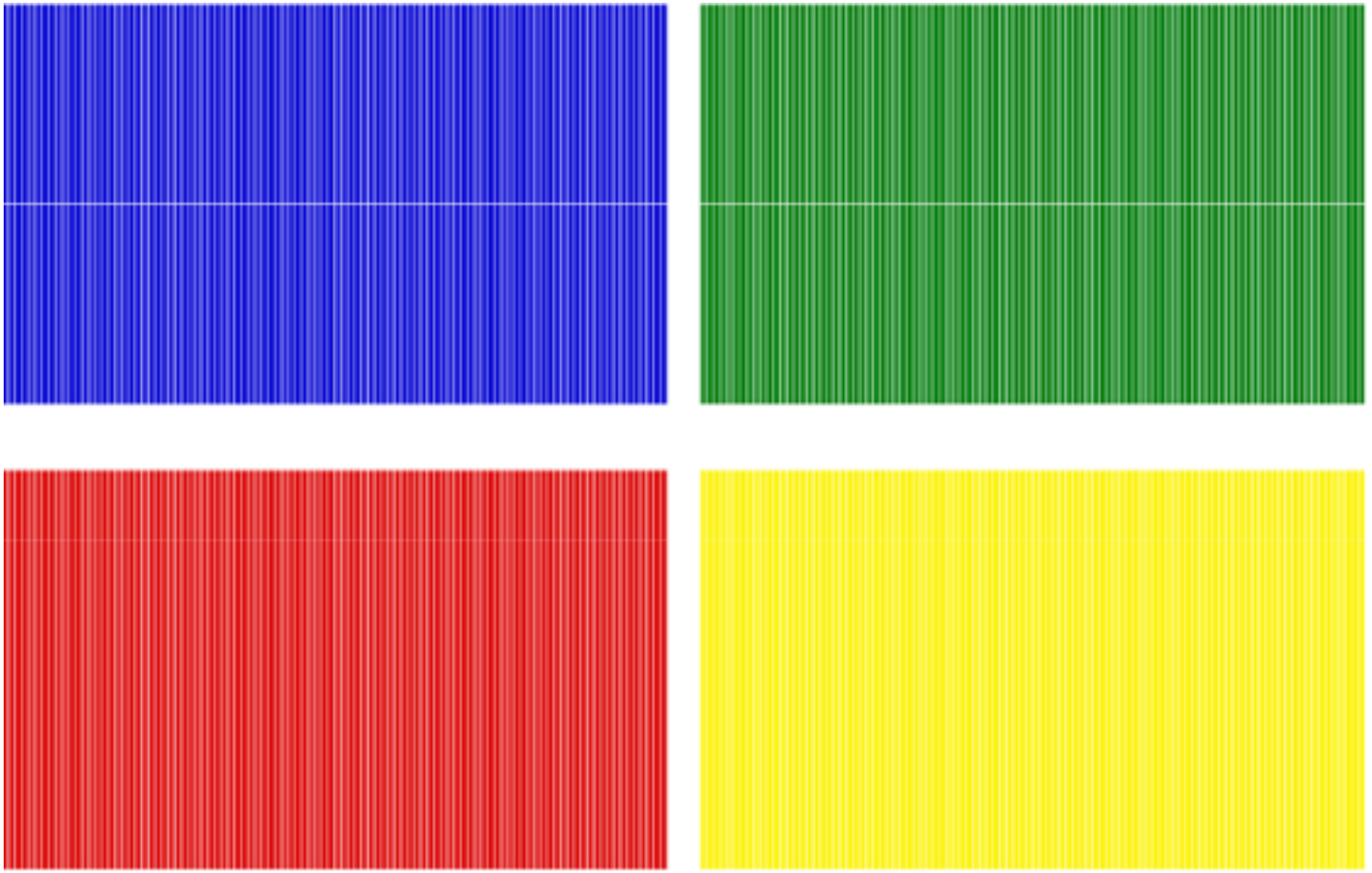
$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

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

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

π_i

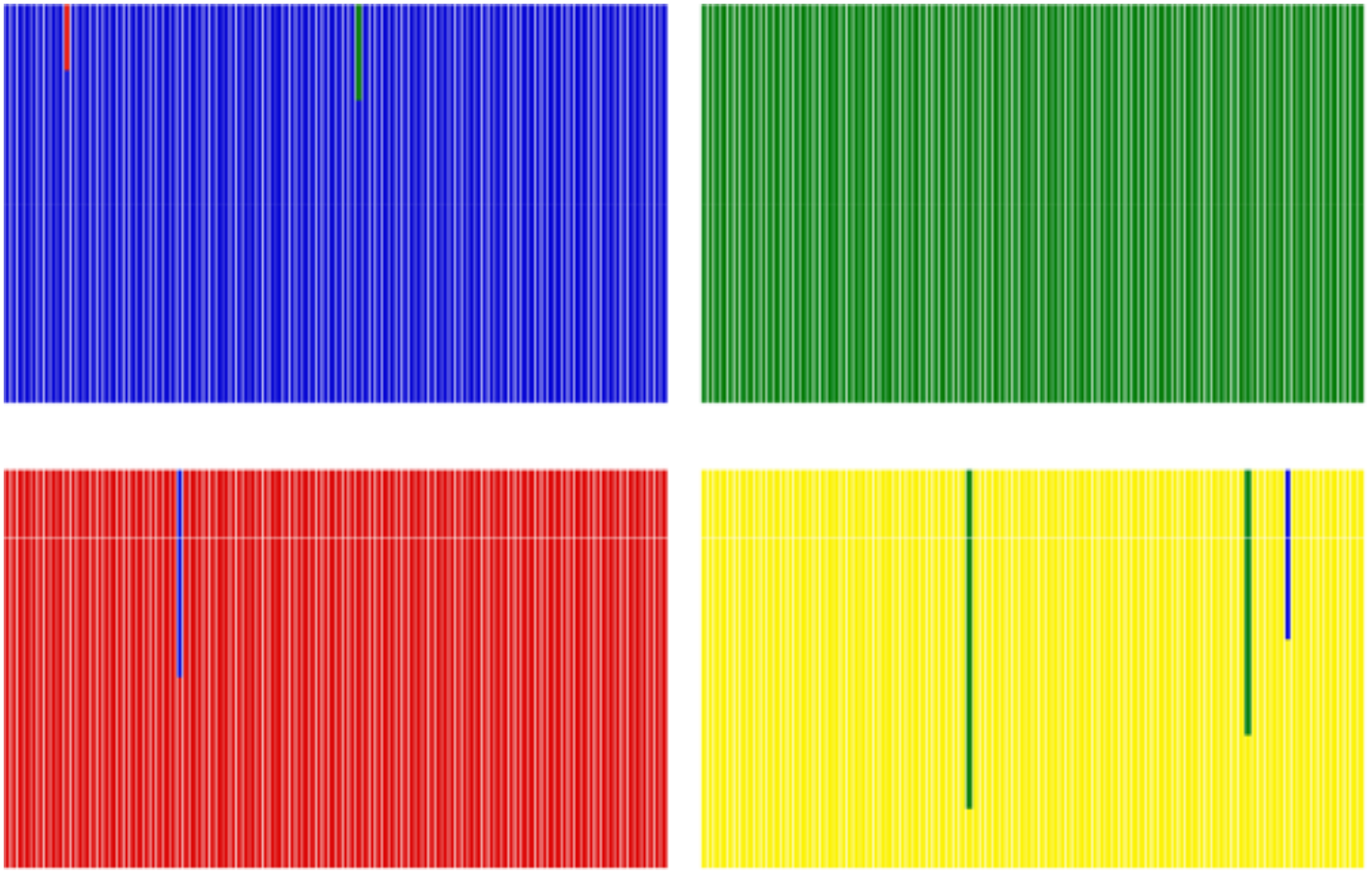
Stationary Frequencies of a CTMC



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

π_i

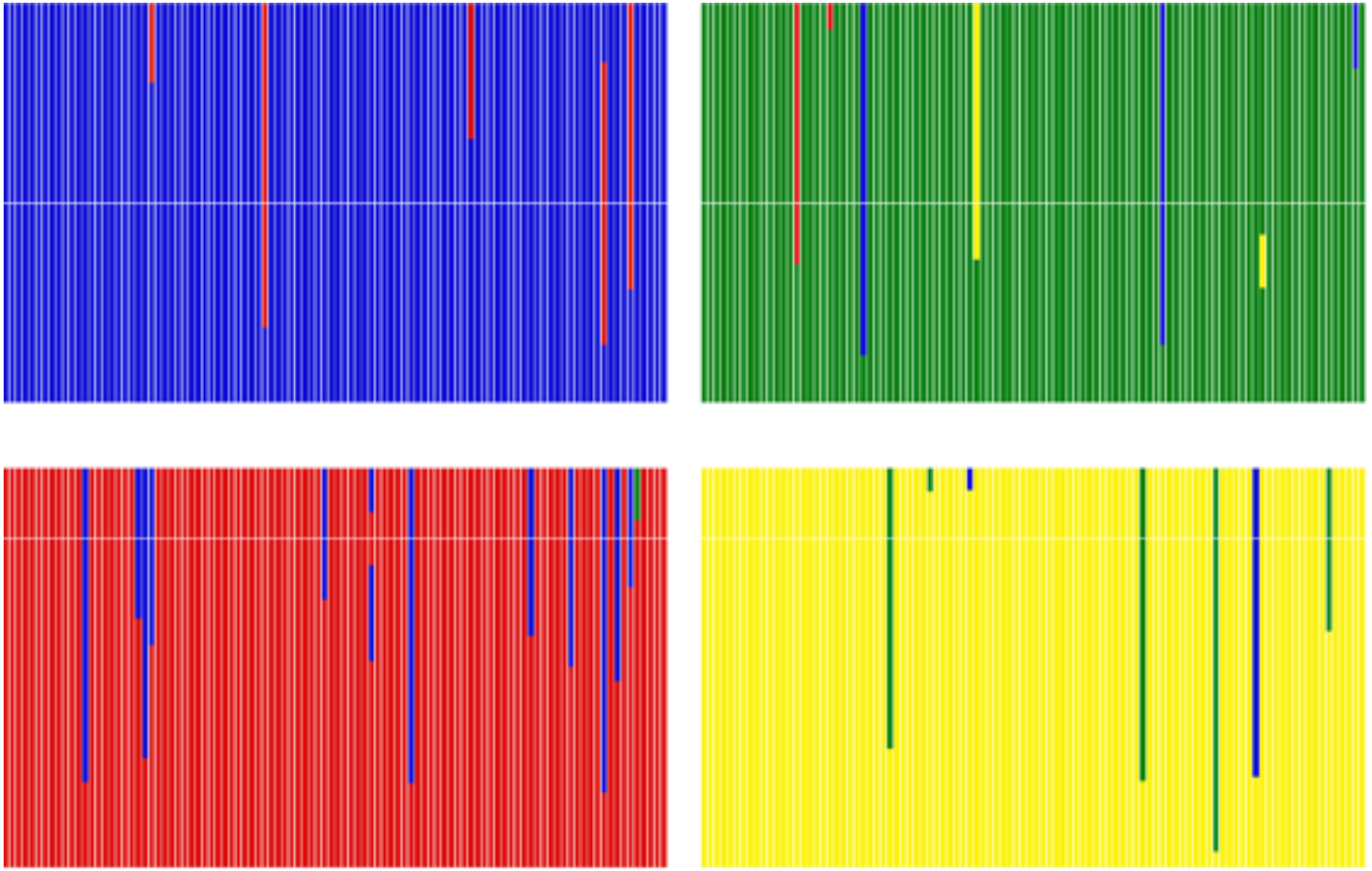
Stationary Frequencies of a CTMC



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

π_i

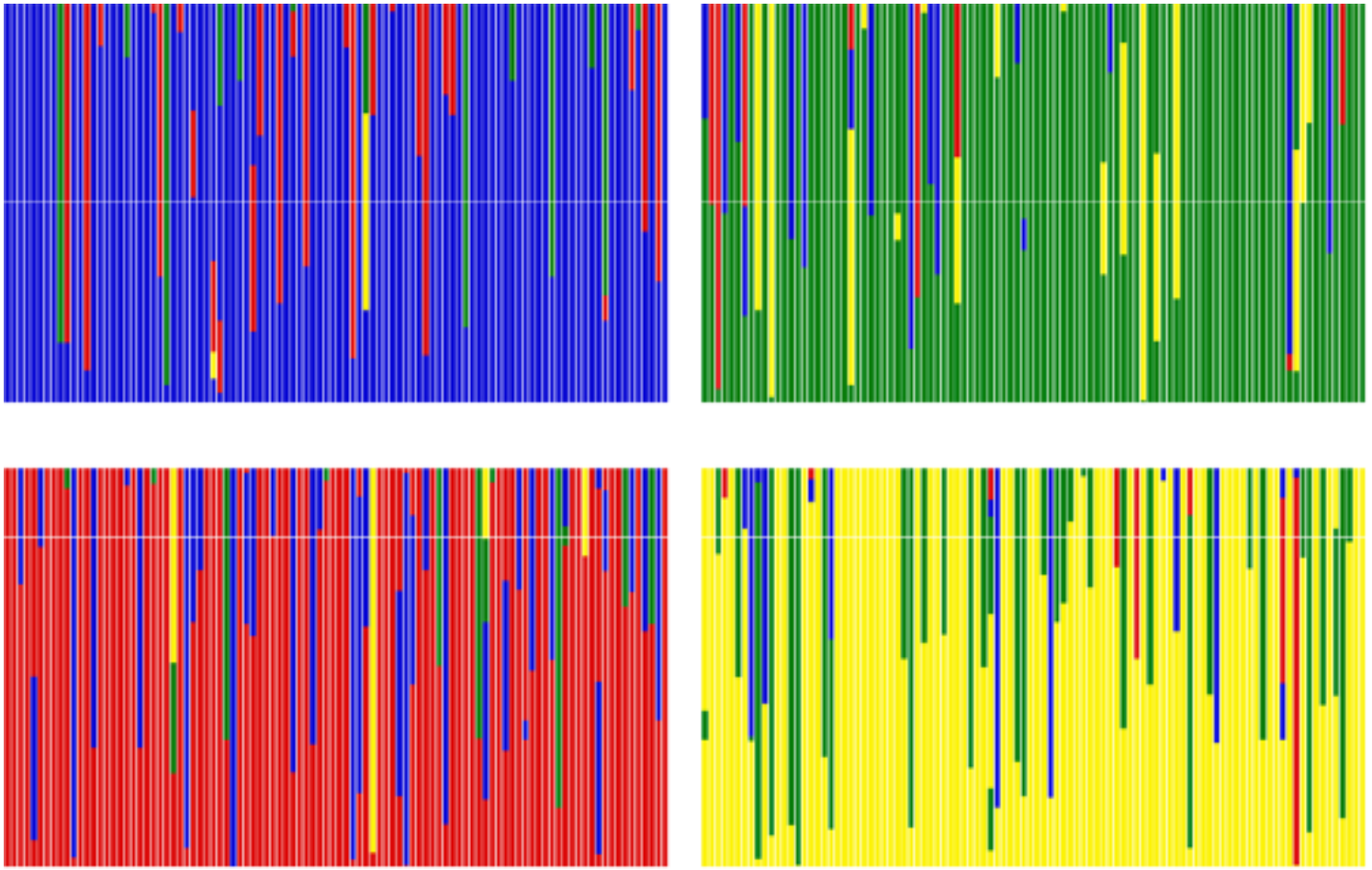
Stationary Frequencies of a CTMC



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

π_i

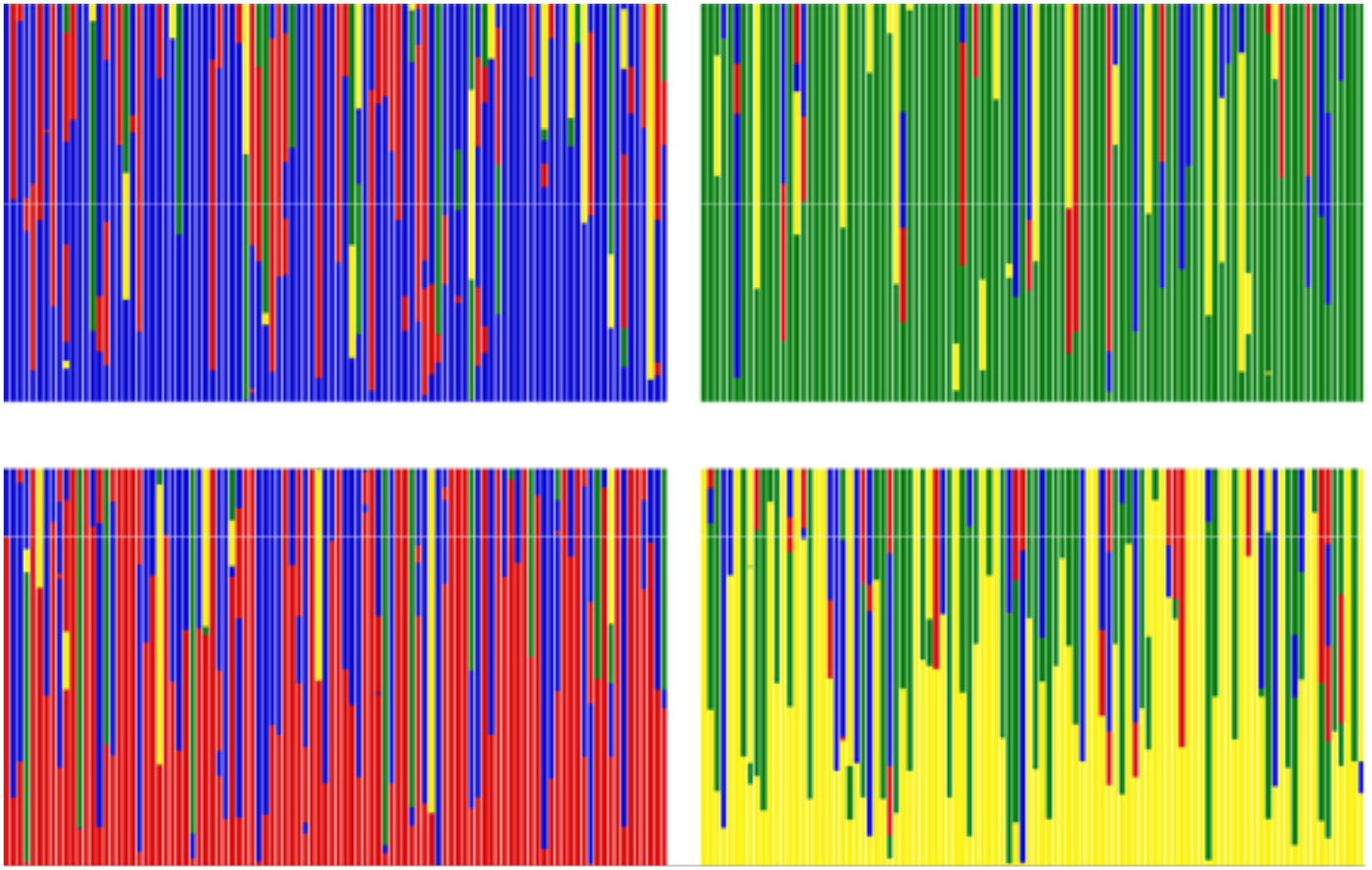
Stationary Frequencies of a CTMC



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

π_i

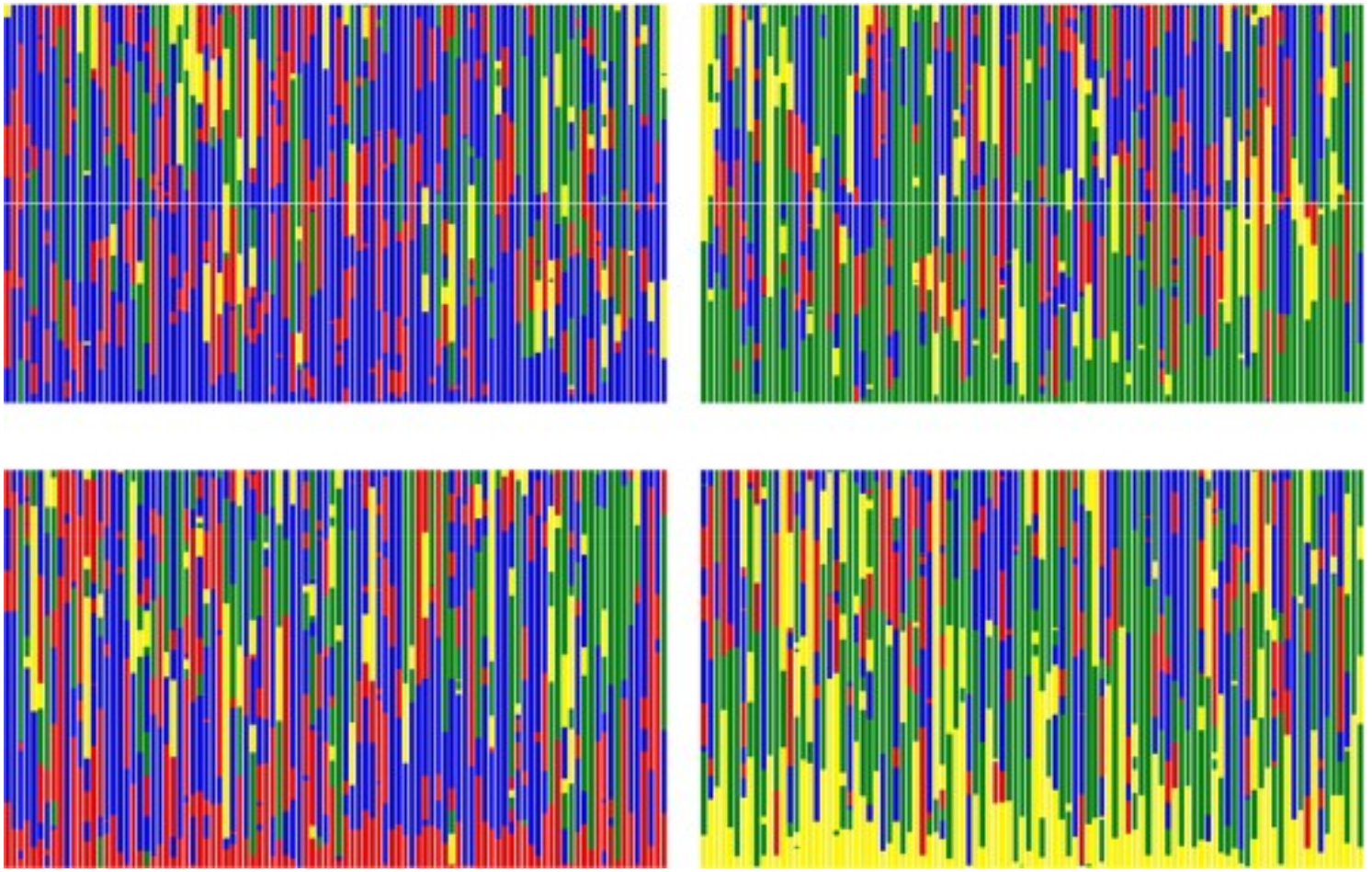
Stationary Frequencies of a CTMC



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

π_i

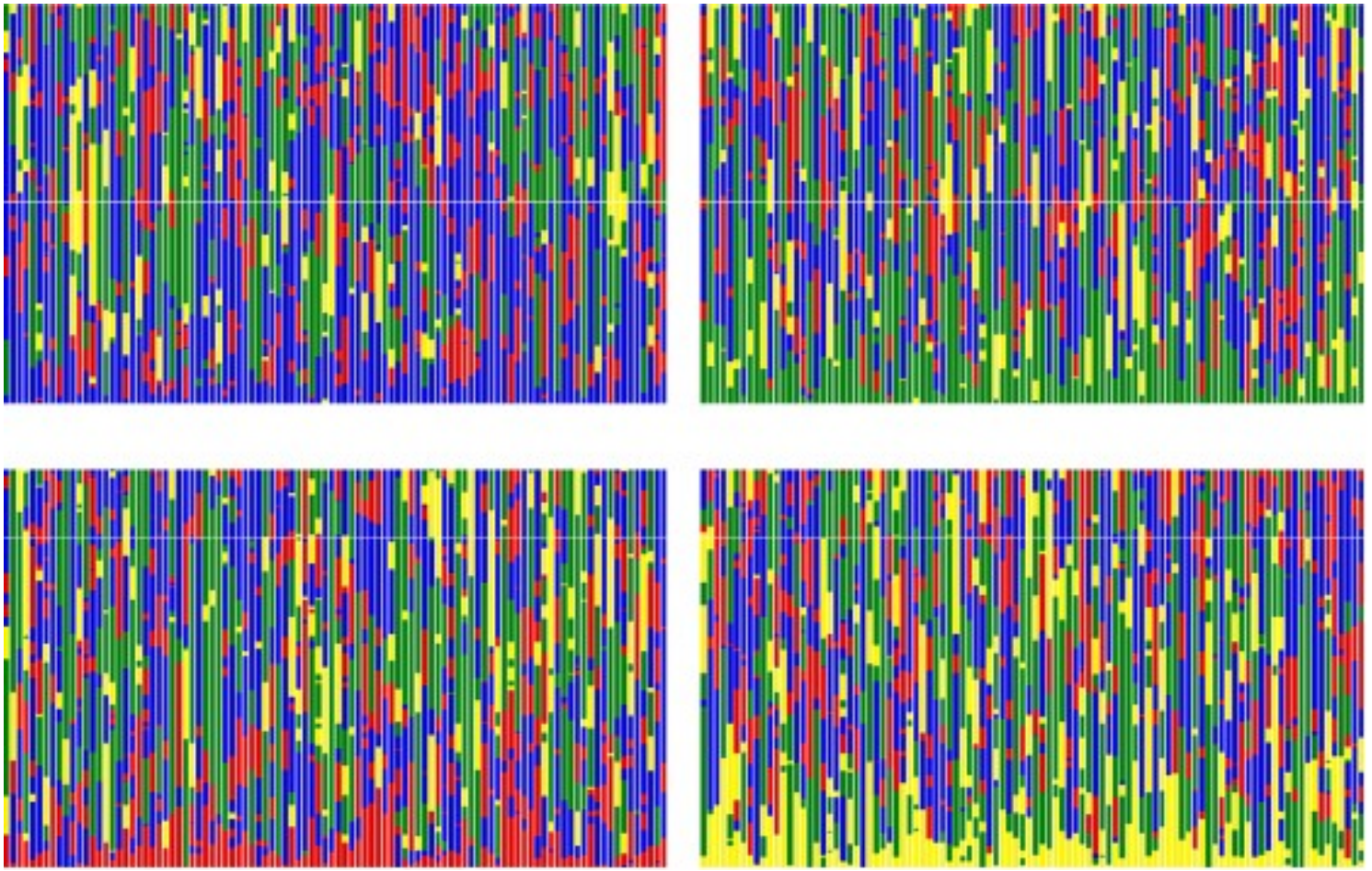
Stationary Frequencies of a CTMC



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

π_i

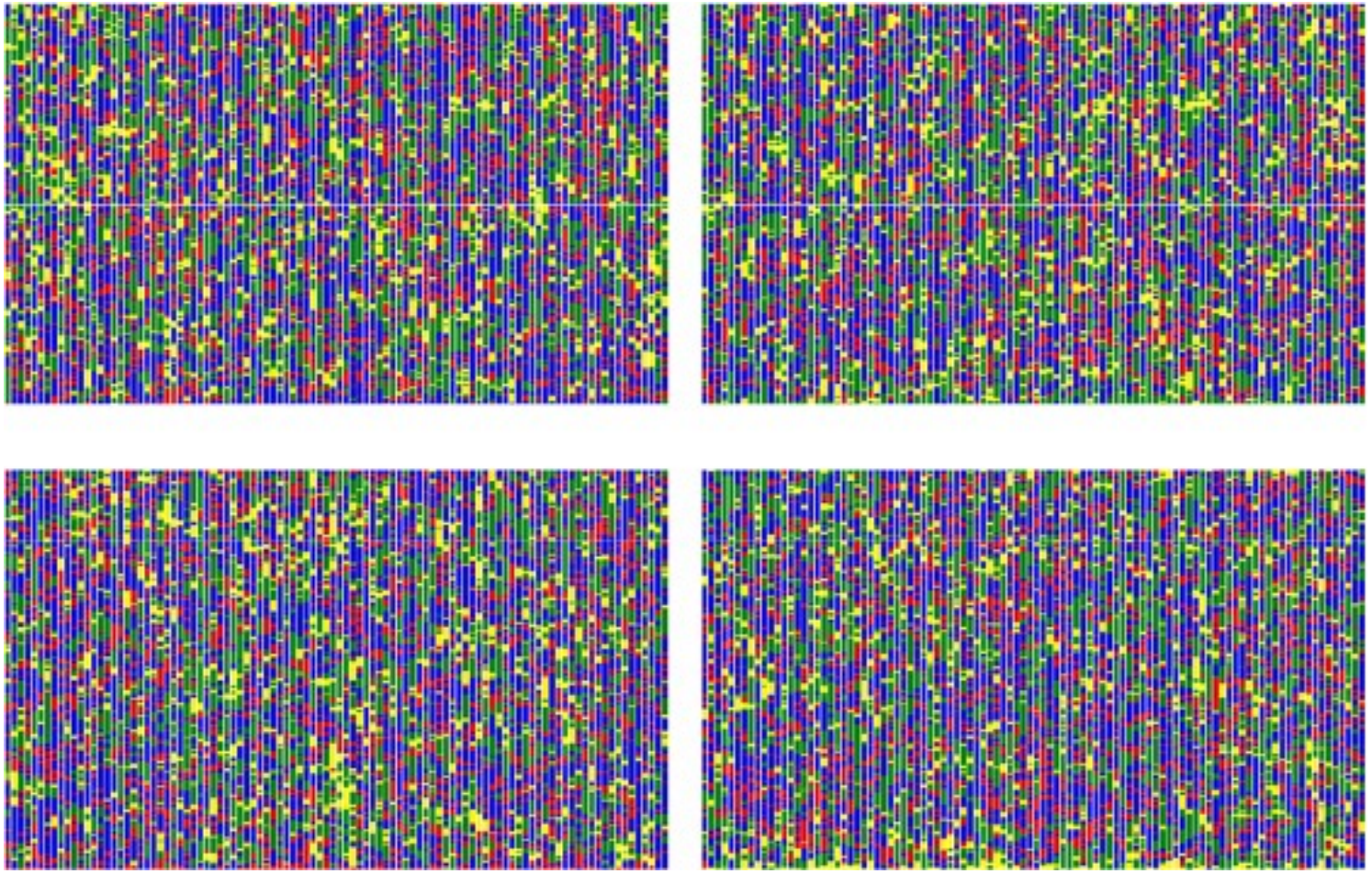
Stationary Frequencies of a CTMC



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

π_i

Stationary Frequencies of a CTMC



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

π_i

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

π_i

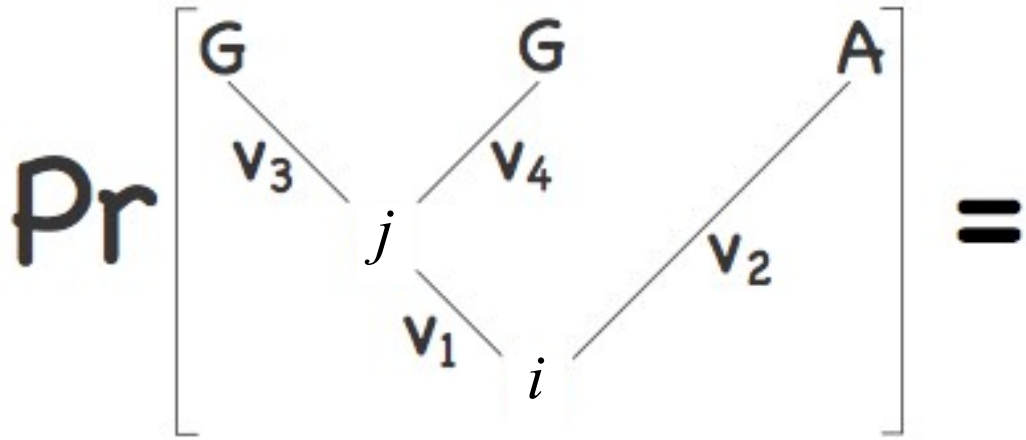
Stationary Frequencies of a CTMC

Stationary frequencies

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

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

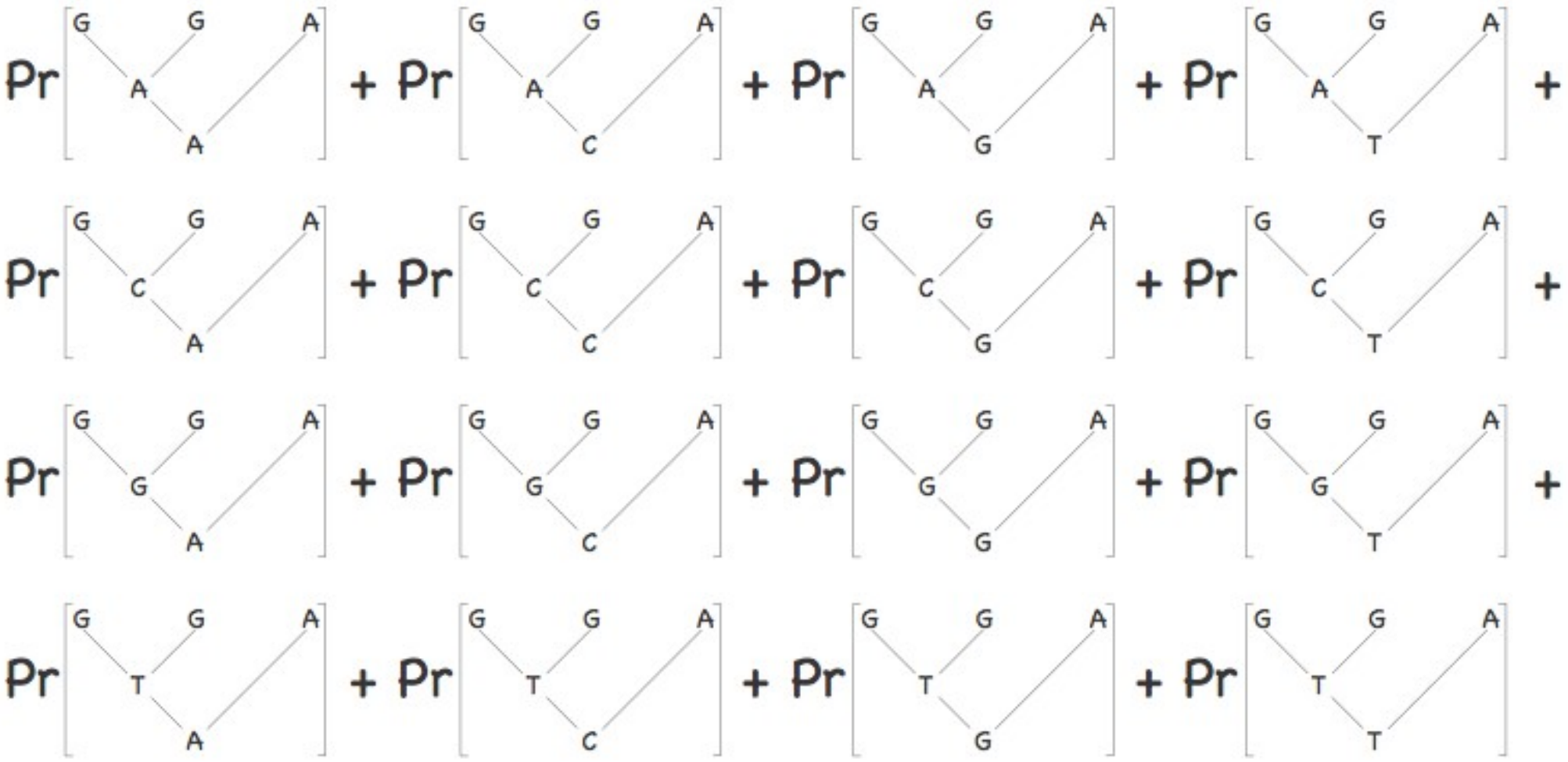
Now we can compute the likelihood of a site history



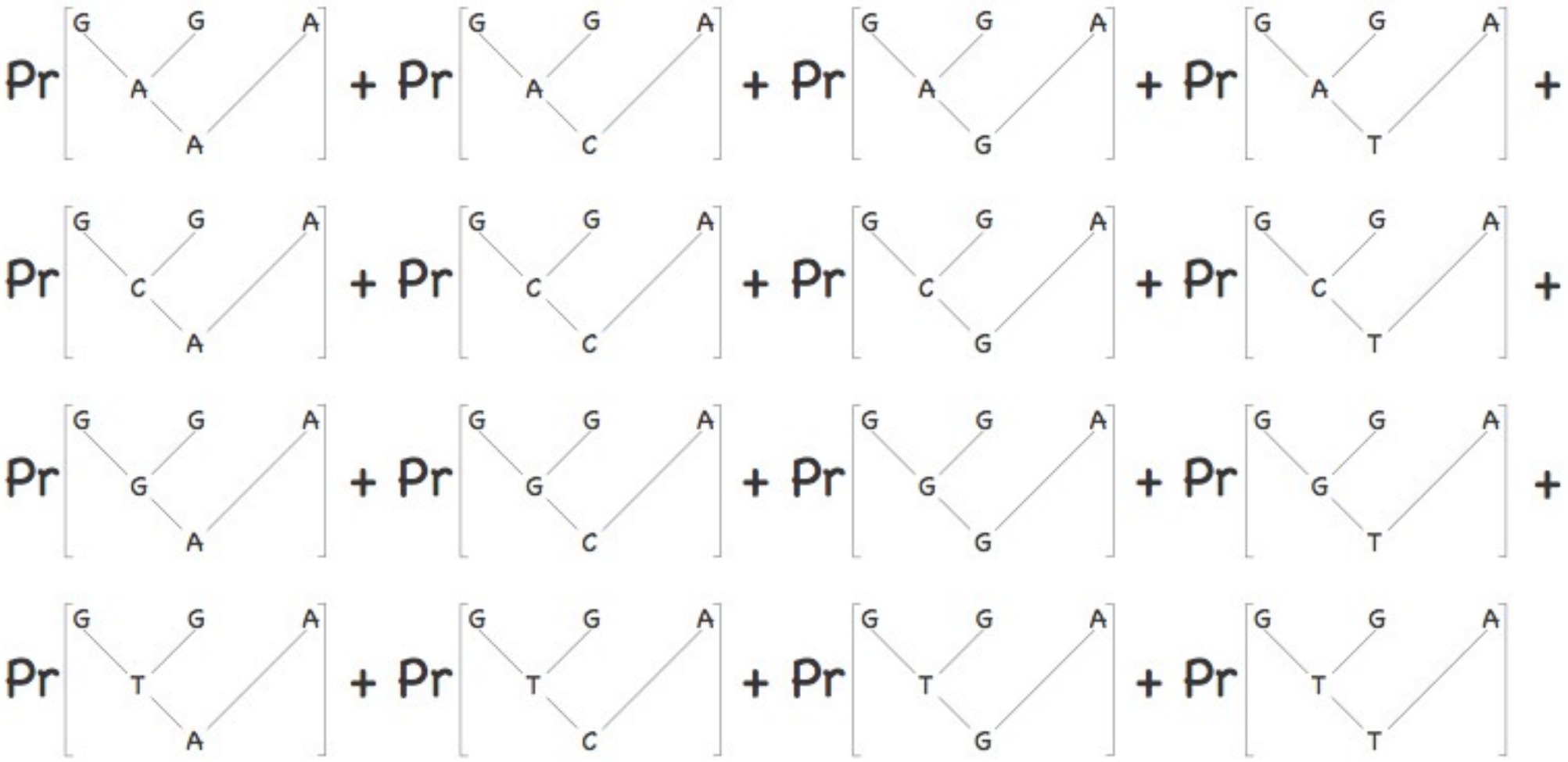
$$\pi_i \times p_{ij}(v_1) \times p_{iA}(v_2) \times p_{jG}(v_3) \times p_{jG}(v_4)$$

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

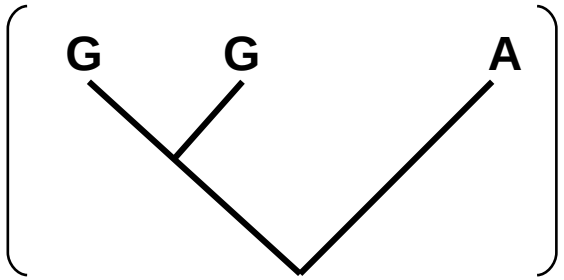
What do we get if we sum over all site histories?



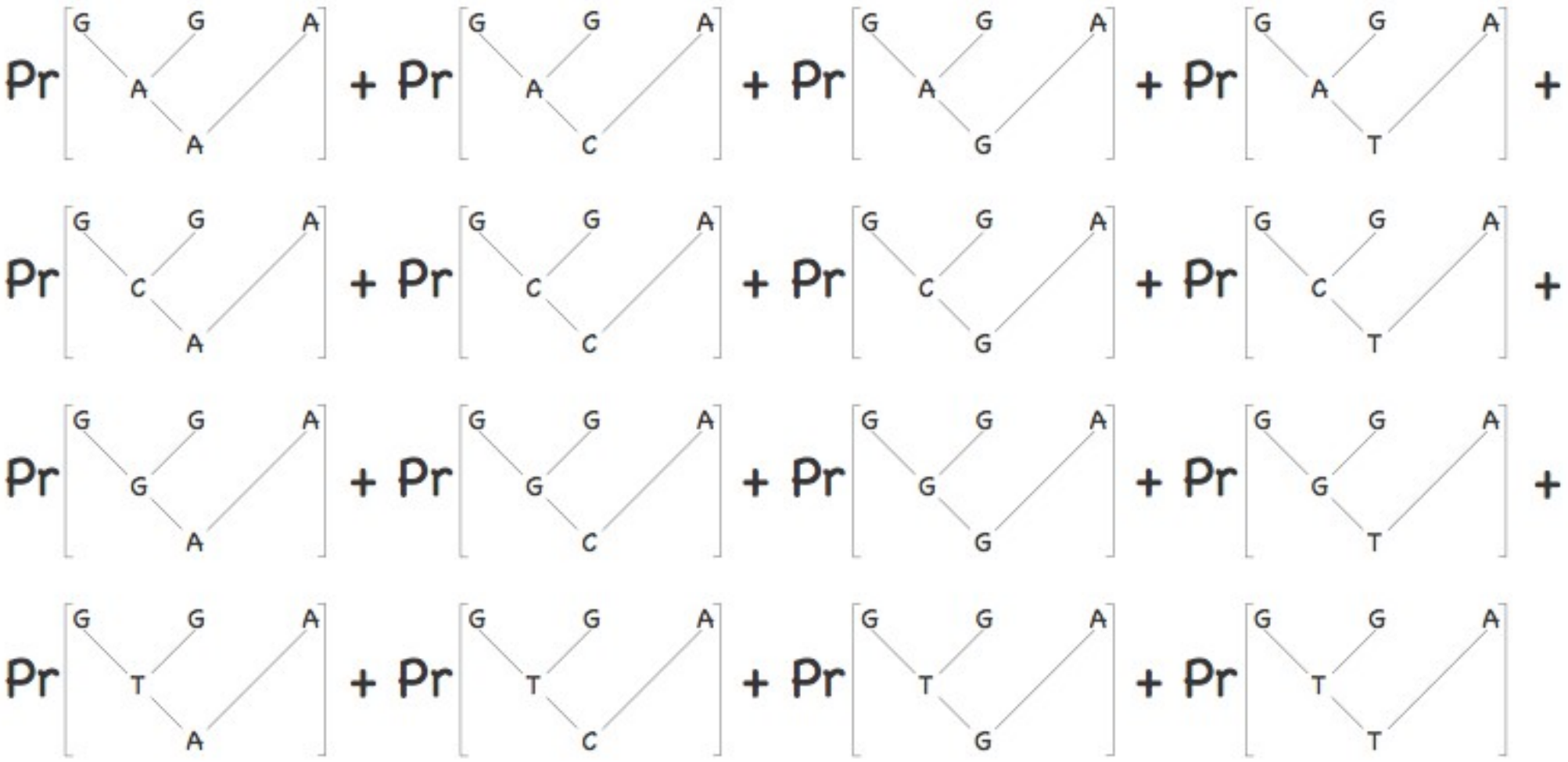
What do we get if we sum over all site histories?

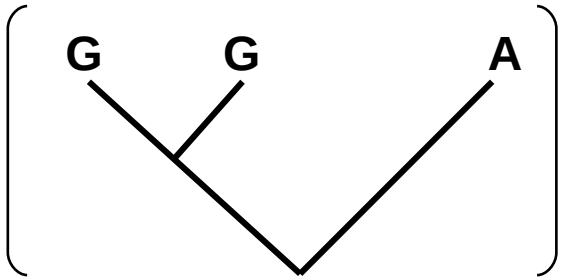


= Pr

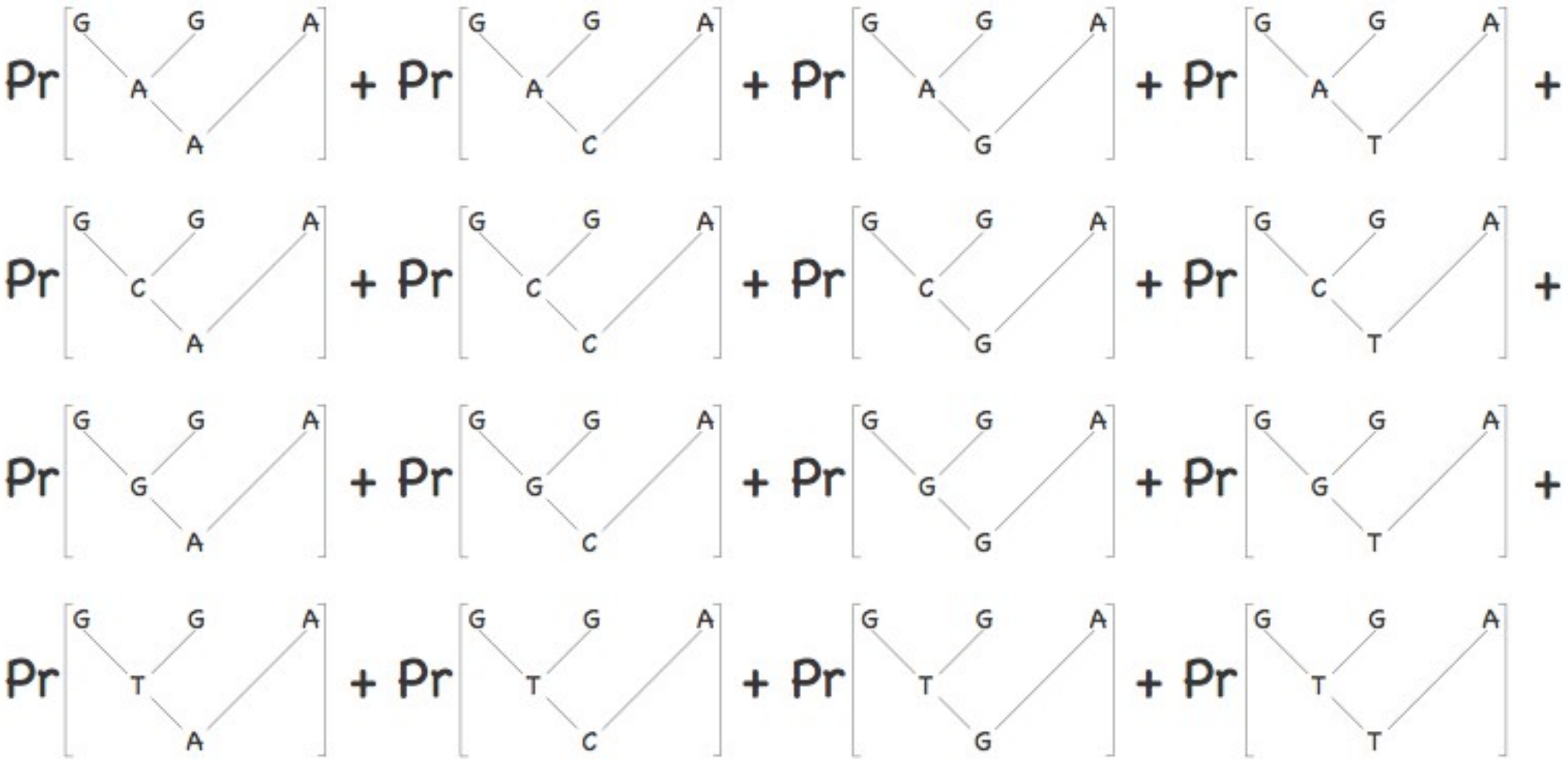


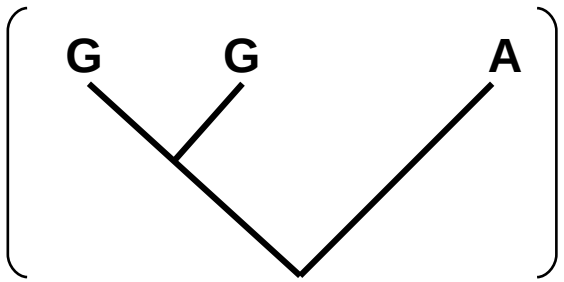
What do we get if we sum over all site histories?



= Pr  = P(GGA|tree, branch lengths, Q)

What do we get if we sum over all site histories?



= Pr 

= P(GGA|tree, branch lengths, Q)

This is the likelihood of site pattern (GGA)

Models of Character Change

Continuous-time Markov models

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

Models 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:

Models 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, $P(v) = e^{\mathbf{Q}v}$

Models 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

Models of Character Change

Instantaneous-rate matrix, **Q**, and the transition probability matrix, **P**

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

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

Models of Character Change

Instantaneous-rate matrix, **Q**, and the transition probability matrix, **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

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

Models of Character Change

Instantaneous-rate matrix, **Q**, and the transition probability matrix, **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

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

The relationship between **Q** and **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

Aims

Understand the main ideas underlying models of sequence evolution

- To do so, we will:
  - Introduce important probability notions
  - Play with models of character evolution through simulations
- Briefly present some of the main models of nucleotide evolution

Substitution models

Rate matrix

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a \pi_C & \mu b \pi_G & \mu c \pi_T \\ \mu a \pi_A & - & \mu d \pi_G & \mu e \pi_T \\ \mu b \pi_A & \mu d \pi_C & - & \mu f \pi_T \\ \mu c \pi_A & \mu e \pi_C & \mu f \pi_G & - \end{pmatrix}$$

Substitution models

Rate matrix

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a \pi_C & \mu b \pi_G & \mu c \pi_T \\ \mu a \pi_A & - & \mu d \pi_G & \mu e \pi_T \\ \mu b \pi_A & \mu d \pi_C & - & \mu f \pi_T \\ \mu c \pi_A & \mu e \pi_C & \mu f \pi_G & - \end{pmatrix}$$

Jukes and Cantor 1969

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu 1/4 & \mu 1/4 & \mu 1/4 \\ \mu 1/4 & - & \mu 1/4 & \mu 1/4 \\ \mu 1/4 & \mu 1/4 & - & \mu 1/4 \\ \mu 1/4 & \mu 1/4 & \mu 1/4 & - \end{pmatrix}$$

0 free parameter (1 if we do not impose one substitution per unit time)

Kimura 1980

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu 1/4 & \mu \kappa 1/4 & \mu 1/4 \\ \mu 1/4 & - & \mu 1/4 & \mu \kappa 1/4 \\ \mu \kappa 1/4 & \mu 1/4 & - & \mu 1/4 \\ \mu 1/4 & \mu \kappa 1/4 & \mu 1/4 & - \end{pmatrix}$$

1 transition/transversion ratio : 1 free parameter

Hasegawa, Kishino, Yano 1985

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu \pi_C & \mu \kappa \pi_G & \mu \pi_T \\ \mu \pi_A & - & \mu \pi_G & \mu \kappa \pi_T \\ \mu \kappa \pi_A & \mu \pi_C & - & \mu \pi_T \\ \mu \pi_A & \mu \kappa \pi_C & \mu \pi_G & - \end{pmatrix}$$

1 transition/transversion ratio
4 equilibrium frequencies:
4 free parameters

Substitution models

Rate matrix

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a \pi_C & \mu b \pi_G & \mu c \pi_T \\ \mu a \pi_A & - & \mu d \pi_G & \mu e \pi_T \\ \mu b \pi_A & \mu d \pi_C & - & \mu f \pi_T \\ \mu c \pi_A & \mu e \pi_C & \mu f \pi_G & - \end{pmatrix}$$

Jukes and Cantor 1969

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu 1/4 & \mu 1/4 & \mu 1/4 \\ \mu 1/4 & - & \mu 1/4 & \mu 1/4 \\ \mu 1/4 & \mu 1/4 & - & \mu 1/4 \\ \mu 1/4 & \mu 1/4 & \mu 1/4 & - \end{pmatrix}$$

0 free parameter (1 if we do not impose one substitution per unit time)

Kimura 1980

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu 1/4 & \mu \kappa 1/4 & \mu 1/4 \\ \mu 1/4 & - & \mu 1/4 & \mu \kappa 1/4 \\ \mu \kappa 1/4 & \mu 1/4 & - & \mu 1/4 \\ \mu 1/4 & \mu \kappa 1/4 & \mu 1/4 & - \end{pmatrix}$$

1 transition/transversion ratio : 1 free parameter

Hasegawa, Kishino, Yano 1985

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu \pi_C & \mu \kappa \pi_G & \mu \pi_T \\ \mu \pi_A & - & \mu \pi_G & \mu \kappa \pi_T \\ \mu \kappa \pi_A & \mu \pi_C & - & \mu \pi_T \\ \mu \pi_A & \mu \kappa \pi_C & \mu \pi_G & - \end{pmatrix}$$

1 transition/transversion ratio
4 equilibrium frequencies:
4 free parameters

All those are particular cases of the GTR model

General Time Reversible model of substitution

Rate matrix

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a \pi_C & \mu b \pi_G & \mu c \pi_T \\ \mu a \pi_A & - & \mu d \pi_G & \mu e \pi_T \\ \mu b \pi_A & \mu d \pi_C & - & \mu f \pi_T \\ \mu c \pi_A & \mu e \pi_C & \mu f \pi_G & - \end{pmatrix}$$

Lanave et al. 1984; Tavaré, 1986

4 ***stationary frequencies***: 3 parameters
6 ***exchangeability parameters***: 5 parameters (if we impose one substitution per unit time)

Summary on CTMCs

- We use CTMCs to model character evolution
- Given an instantaneous rate matrix, we can compute substitution probabilities on a branch with an arbitrary length
- We can combine these computations to compute the likelihood of a site history
- We can sum over site histories to get the likelihood of a site pattern
- The GTR family provides examples of instantaneous rate matrices