

# Lecture 4: A Brief Tour of Some Common Models of Nucleotide Substitution

Bastien Boussau & Brian Moore  
Laboratoire de Biométrie et Biologie Évolutive  
Université de Lyon  
CoME, 2019

# Outline

## I. A brief review of Continuous-Time Markov models

# Outline

I. A brief review of Continuous-Time Markov models

II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# Stochastic Mechanisms of Character Change

## Continuous-time Markov models

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

# Stochastic Mechanisms of Character Change

## Continuous-time Markov models

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

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

completely describes the stochastic process by specifying:

# Stochastic Mechanisms of Character Change

## Continuous-time Markov models

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

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

completely describes the stochastic process by specifying:

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

# Stochastic Mechanisms of Character Change

## Continuous-time Markov models

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

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

completely describes the stochastic process by specifying:

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

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

# Scaling the Instantaneous Rate Matrix

What is the scale of the instantaneous rates?

Substitution rate and time are confounded in most CTMC models

# Scaling the Instantaneous Rate Matrix

What is the scale of the instantaneous rates?

Substitution rate and time are confounded in most CTMC models

The expected (mean) number of substitutions along a branch is a function of the duration of that branch,  $t$ , and the substitution rate,  $u$

# Scaling the Instantaneous Rate Matrix

What is the scale of the instantaneous rates?

Substitution rate and time are confounded in most CTMC models

The expected (mean) number of substitutions along a branch is a function of the duration of that branch,  $t$ , and the substitution rate,  $u$

However, the timescale of the tree is typically unknown, so the time,  $t$ , and rate,  $u$ , cannot be estimated separately, but only as their product:  $v = ut$

# Scaling the Instantaneous Rate Matrix

What is the scale of the instantaneous rates?

Substitution rate and time are confounded in most CTMC models

The expected (mean) number of substitutions along a branch is a function of the duration of that branch,  $t$ , and the substitution rate,  $u$

However, the timescale of the tree is typically unknown, so the time,  $t$ , and rate,  $u$ , cannot be estimated separately, but only as their product:  $v = ut$

In order for us to be able to interpret the branch length,  $v$ , as the expected number of substitutions per site, the average substitution rate must be one

# Scaling the Instantaneous Rate Matrix

What is the scale of the instantaneous rates?

Substitution rate and time are confounded in most CTMC models

The expected (mean) number of substitutions along a branch is a function of the duration of that branch,  $t$ , and the substitution rate,  $u$

However, the timescale of the tree is typically unknown, so the time,  $t$ , and rate,  $u$ , cannot be estimated separately, but only as their product:  $v = ut$

In order for us to be able to interpret the branch length,  $v$ , as the expected number of substitutions per site, the average substitution rate must be one

This is achieved by scaling the instantaneous-rate matrix by the weighted average of substitution rates:

$$\mu = \sum_i \sum_j \pi_i q_{ij}$$

# Scaling the Instantaneous Rate Matrix

What is the scale of the instantaneous rates?

Substitution rate and time are confounded in most CTMC models

The expected (mean) number of substitutions along a branch is a function of the duration of that branch,  $t$ , and the substitution rate,  $u$

However, the timescale of the tree is typically unknown, so the time,  $t$ , and rate,  $u$ , cannot be estimated separately, but only as their product:  $v = ut$

In order for us to be able to interpret the branch length,  $v$ , as the expected number of substitutions per site, the average substitution rate must be one

This is achieved by scaling the instantaneous-rate matrix by the weighted average of substitution rates:

$$\mu = \sum_i \sum_j \pi_i q_{ij}$$

or equivalently:

$$\mu = - \sum_i \pi_i q_{ii}$$

# Scaling the Instantaneous Rate Matrix

What is the scale of the instantaneous rates?

Substitution rate and time are confounded in most CTMC models

The expected (mean) number of substitutions along a branch is a function of the duration of that branch,  $t$ , and the substitution rate,  $u$

However, the timescale of the tree is typically unknown, so the time,  $t$ , and rate,  $u$ , cannot be estimated separately, but only as their product:  $v = ut$

In order for us to be able to interpret the branch length,  $v$ , as the expected number of substitutions per site, the average substitution rate must be one

This is achieved by scaling the instantaneous-rate matrix by the weighted average of substitution rates:

$$\mu = \sum_i \sum_j \pi_i q_{ij} \quad \text{weighted sum of off-diagonal rates}$$

or equivalently:

$$\mu = - \sum_i \pi_i q_{ii} \quad \text{negative weighted sum of diagonal rates}$$

# Scaling the Instantaneous Rate Matrix

What is the scale of the instantaneous rates?

Substitution rate and time are confounded in most CTMC models

The expected (mean) number of substitutions along a branch is a function of the duration of that branch,  $t$ , and the substitution rate,  $u$

However, the timescale of the tree is typically unknown, so the time,  $t$ , and rate,  $u$ , cannot be estimated separately, but only as their product:  $v = ut$

In order for us to be able to interpret the branch length,  $v$ , as the expected number of substitutions per site, the average substitution rate must be one

This is achieved by scaling the instantaneous-rate matrix by the weighted average of substitution rates:

$$\mu = \sum_i \sum_j \pi_i q_{ij}$$

or equivalently:

$$\mu = - \sum_i \pi_i q_{ii}$$

The rate of change from state  $i$  to state  $j$  is weighted by the probability of starting in state  $i$ , which is the stationary frequency of state  $i$ ,  $\pi_i$

# Scaling the Instantaneous Rate Matrix

What is the scale of the instantaneous rates?

Substitution rate and time are confounded in most CTMC models

The expected (mean) number of substitutions along a branch is a function of the duration of that branch,  $t$ , and the substitution rate,  $u$

However, the timescale of the tree is typically unknown, so the time,  $t$ , and rate,  $u$ , cannot be estimated separately, but only as their product:  $v = ut$

In order for us to be able to interpret the branch length,  $v$ , as the expected number of substitutions per site, the average substitution rate must be one

This is achieved by scaling the instantaneous-rate matrix by the weighted average of substitution rates:

$$\mu = \sum_i \sum_j \pi_i q_{ij}$$

or equivalently:

$$\mu = - \sum_i \pi_i q_{ii}$$

The rate of change from state  $i$  to state  $j$  is weighted by the probability of starting in state  $i$ , which is the stationary frequency of state  $i$ ,  $\pi_i$

We then multiply each element in the instantaneous-rate matrix by the scalar  $1/\mu$

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

Stationary frequencies...

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

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

Stationary frequencies...

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

The (weighted) average substitution rate:

$$\mu = - \sum_i \pi_i q_{ii}$$

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

Stationary frequencies...

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

The (weighted) average substitution rate:

$$\mu = - \sum_i \pi_i q_{ii}$$

$$-(0.138 \times -1.916 + 0.188 \times -1.069 + 0.495 \times -0.591 + 0.179 \times -1.355) = 1$$

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

Stationary frequencies...

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

The (weighted) average substitution rate:

$$\mu = - \sum_i \pi_i q_{ii}$$

$$-(0.138 \times -1.916 + 0.188 \times -1.069 + 0.495 \times -0.591 + 0.179 \times -1.355) = 1$$

↑  
probability of  
starting in A

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

Stationary frequencies...

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

The (weighted) average substitution rate:

$$\mu = - \sum_i \pi_i q_{ii}$$

$$-(0.138 \times -1.916 + 0.188 \times -1.069 + 0.495 \times -0.591 + 0.179 \times -1.355) = 1$$

↑  
probability of  
starting in A

↑  
overall rate of  
change away from A

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

Stationary frequencies...

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

The (weighted) average substitution rate:

$$\mu = - \sum_i \pi_i q_{ii}$$

$$-(0.138 \times -1.916 + 0.188 \times -1.069 + 0.495 \times -0.591 + 0.179 \times -1.355) = 1$$

probability of  
starting in C

overall rate of  
change away from C

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

Stationary frequencies...

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

The (weighted) average substitution rate:

$$\mu = - \sum_i \pi_i q_{ii}$$

$$-(0.138 \times -1.916 + 0.188 \times -1.069 + 0.495 \times -0.591 + 0.179 \times -1.355) = 1$$

probability of starting in G  
overall rate of change away from G

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

Stationary frequencies...

$$\mathbf{P}(100.0) = \begin{pmatrix} \pi_A & \pi_C & \pi_G & \pi_T \\ 0.138 & 0.188 & 0.495 & \boxed{0.179} \\ 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \end{pmatrix}$$

The (weighted) average substitution rate:

$$\mu = - \sum_i \pi_i q_{ii}$$

$$-(0.138 \times -1.916 + 0.188 \times -1.069 + 0.495 \times -0.591 + 0.179 \times -1.355) = 1$$

probability of starting in T  
overall rate of change away from T

# Scaling the Instantaneous Rate Matrix

Let's check the scaling of our hypothetical rate matrix

The instantaneous-rate matrix...

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

Stationary frequencies...

$$\mathbf{P}(100.0) = \begin{pmatrix} \pi_A & \pi_C & \pi_G & \pi_T \\ 0.138 & 0.188 & 0.495 & \boxed{0.179} \\ 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \\ 0.138 & 0.188 & 0.495 & 0.179 \end{pmatrix}$$

The (weighted) average substitution rate:

$$\mu = - \sum_i \pi_i q_{ii}$$

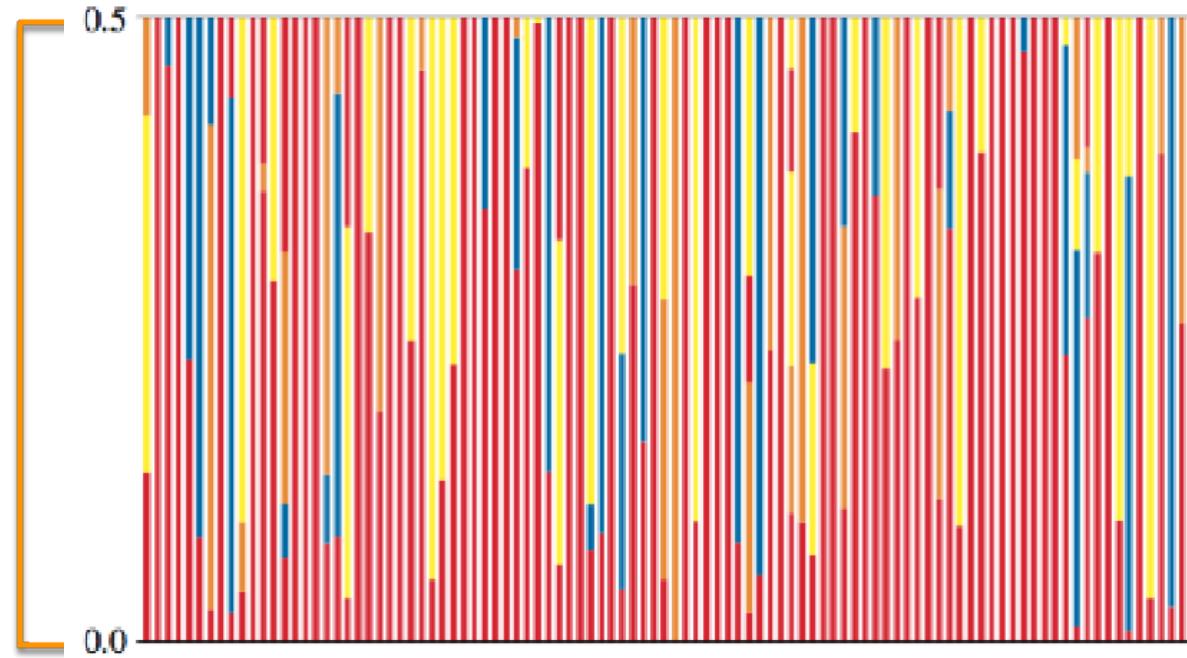
$$-(0.138 \times -1.916 + 0.188 \times -1.069 + 0.495 \times -0.591 + 0.179 \times -1.355) = 1$$

The average substitution rate in our instantaneous-rate matrix is therefore 1!

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The true branch length was  $\nu = 0.5$



For simulations starting in A, the average number of changes was  $z = 0.788$

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The true branch length was  $\nu = 0.5$

We compute the average number of changes for each starting state,  $z$



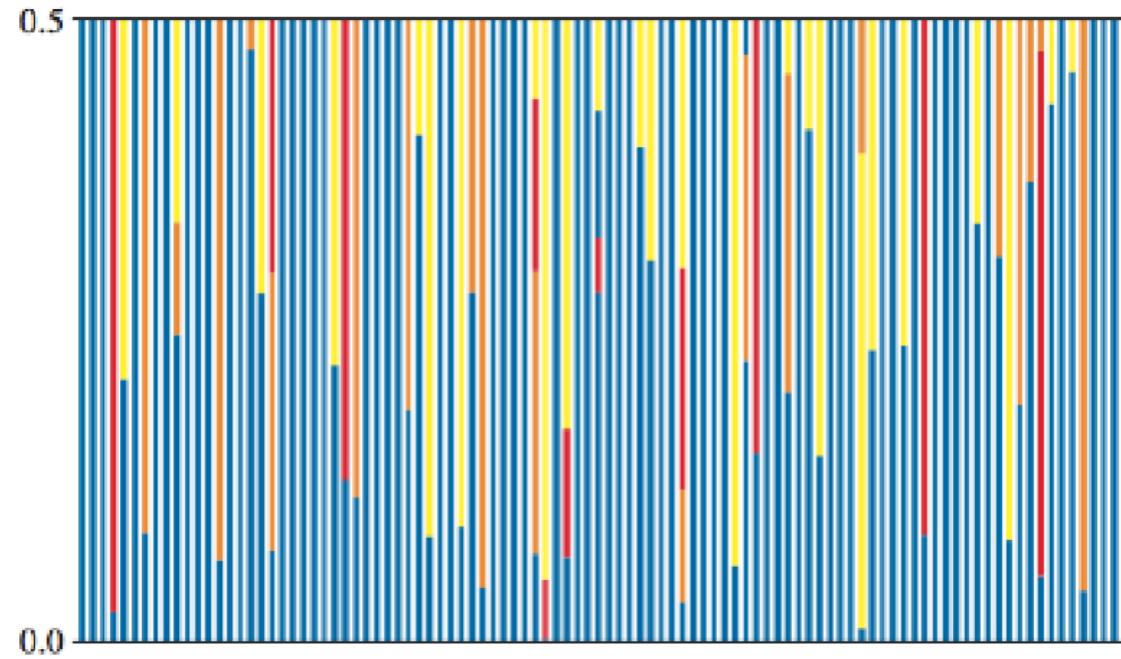
For simulations starting in A, the average number of changes was  $z = 0.788$

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The true branch length was  $\nu = 0.5$

We compute the average number of changes for each starting state,  $z$



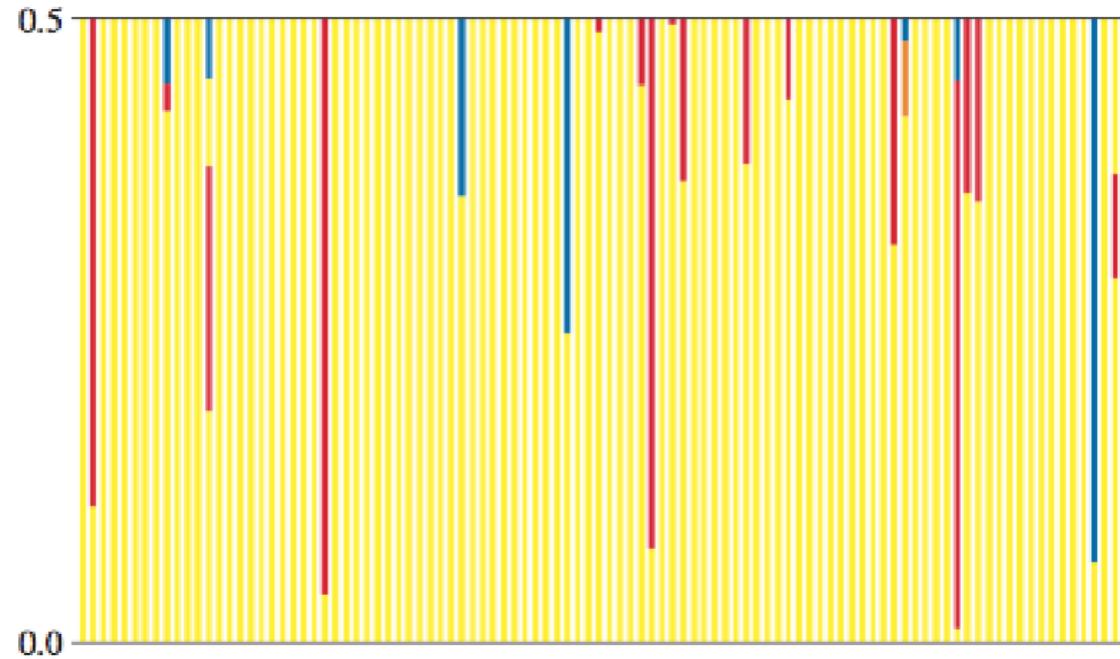
For simulations starting in C, the average number of changes was  $z = 0.541$

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The true branch length was  $\nu = 0.5$

We compute the average number of changes for each starting state,  $z$



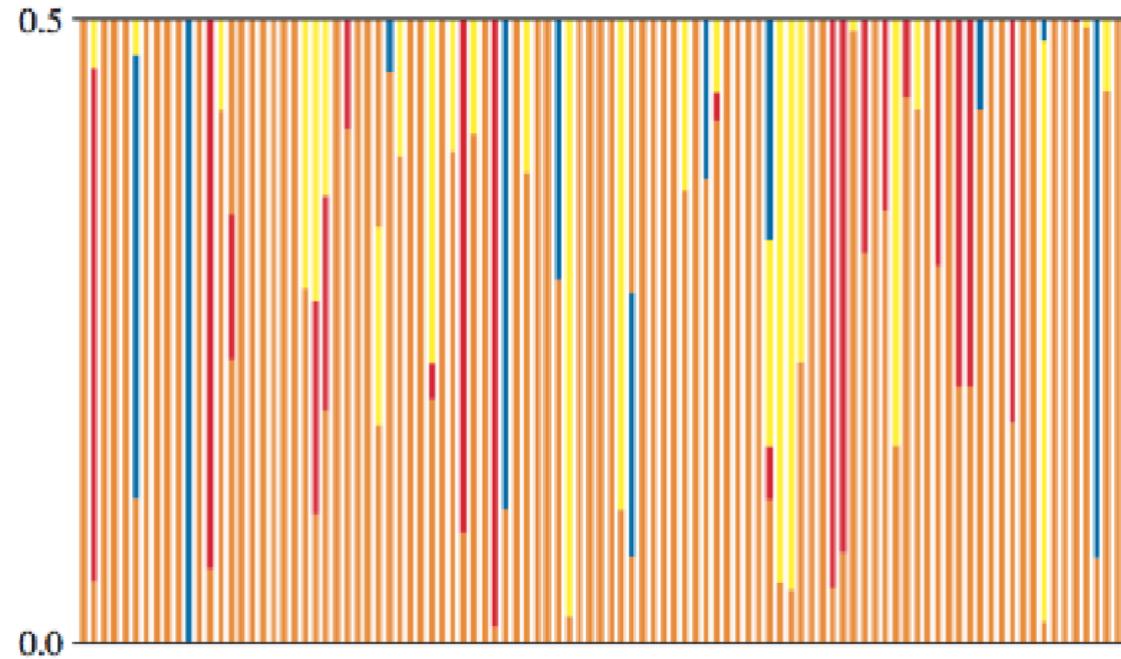
For simulations starting in G, the average number of changes was  $z = 0.352$

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The true branch length was  $\nu = 0.5$

We compute the average number of changes for each starting state,  $z$



For simulations starting in  $T$ , the average number of changes was  $z = 0.643$

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The weighted sum of the average (expected) number of changes for all four states when the average of the **Q** matrix is 1...

$i$	$\pi_i$
A	0.138
C	0.188
G	0.495
T	0.179

↑  
probability of  
each starting state

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The weighted sum of the average (expected) number of changes for all four states when the average of the  $\mathbf{Q}$  matrix is 1...

$i$	$\pi_i$	$z_i$
A	0.138	0.788
C	0.188	0.541
G	0.495	0.352
T	0.179	0.643

↑  
average number of  
changes away from each state

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The weighted sum of the average (expected) number of changes for all four states when the average of the  $\mathbf{Q}$  matrix is 1...

$i$	$\pi_i$	$z_i$	$\pi_i z_i$
A	0.138	0.788	0.109
C	0.188	0.541	0.102
G	0.495	0.352	0.174
T	0.179	0.643	0.155

↑  
product of  $\pi_i z_i$   
for each state

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The weighted sum of the average (expected) number of changes for all four states when the average of the  $\mathbf{Q}$  matrix is 1...

$i$	$\pi_i$	$z_i$	$\pi_i z_i$
A	0.138	0.788	0.109
C	0.188	0.541	0.102
G	0.495	0.352	0.174
T	0.179	0.643	0.155
		0.500	sum of terms for all states

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The weighted sum of the average (expected) number of changes for all four states when the average of the  $\mathbf{Q}$  matrix is 1...

$i$	$\pi_i$	$z_i$	$\pi_i z_i$
A	0.138	0.788	0.109
C	0.188	0.541	0.102
G	0.495	0.352	0.174
T	0.179	0.643	0.155
			0.500

i.e., the true branch length, in units expected  
(average) number of substitutions per site!

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The weighted sum of the average (expected) number of changes for all four states when the average of the  $\mathbf{Q}$  matrix is 1...

$i$	$\pi_i$	$z_i$	$\pi_i z_i$
A	0.138	0.788	0.109
C	0.188	0.541	0.102
G	0.495	0.352	0.174
T	0.179	0.643	0.155
			0.500

i.e., the true branch length, in units expected  
(average) number of substitutions per site!

Scaling the instantaneous-rate matrix such that the average substitution rate is 1 therefore allows us to interpret the branch lengths,  $\nu$ , in terms of the expected (average) number of substitutions per site.

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The weighted sum of the average (expected) number of changes for all four states when the average of the  $\mathbf{Q}$  matrix is 2...

$i$	$\pi_i$	$z_i$	$\pi_i z_i$
A	0.138	1.576	0.218
C	0.188	1.082	0.204
G	0.495	0.704	0.348
T	0.179	1.286	0.230
			1.000

i.e., the branch length, in units expected number of substitutions per site is 2 x the true value!

# Scaling the Instantaneous Rate Matrix

Can we correctly estimate the true branch length used in our simulation?

The weighted sum of the average (expected) number of changes for all four states when the average of the  $\mathbf{Q}$  matrix is 2...

$i$	$\pi_i$	$z_i$	$\pi_i z_i$
A	0.138	1.576	0.218
C	0.188	1.082	0.204
G	0.495	0.704	0.348
T	0.179	1.286	0.230
			1.000

i.e., the branch length, in units expected number of substitutions per site is 2 x the true value!

Scaling the instantaneous-rate matrix such that the average substitution rate is 2 would require that we divide the estimated branch lengths by 2 in order to interpret the branch lengths,  $\nu$ , in terms of the expected (average) number of substitutions per site.

# Outline

## → I. A brief review of Continuous-Time Markov models

### II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

### III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

### IV. Accommodating heterogeneity in the substitution process along the tree

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

→ What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# Stochastic Models of Nucleotide Substitution

General Time Reversible models of nucleotide substitution

This family of CTMC models makes several simplifying assumptions:

# Stochastic Models of Nucleotide Substitution

## General Time Reversible models of nucleotide substitution

This family of CTMC models makes several simplifying assumptions:

- the rate of the substitution process is constant across sites

# Stochastic Models of Nucleotide Substitution

## General Time Reversible models of nucleotide substitution

This family of CTMC models makes several simplifying assumptions:

- the rate of the substitution process is constant across sites
- the nature of the substitution process is constant across sites

# Stochastic Models of Nucleotide Substitution

## General Time Reversible models of nucleotide substitution

This family of CTMC models makes several simplifying assumptions:

- the rate of the substitution process is constant across sites
- the nature of the substitution process is constant across sites
- sites are independent

# Time Reversibility

Most continuous-time Markov models assume time reversibility

The probability of a sequence of states visited forward in time is equal to the series of states visited in the reverse order

$$\pi_j p_{ji}(t) = \pi_i p_{ij}(t)$$

# Time Reversibility

Most continuous-time Markov models assume time reversibility

The probability of a sequence of states visited forward in time is equal to the series of states visited in the reverse order

$$\pi_j p_{ji}(t) = \pi_i p_{ij}(t)$$

↑  
↑  
probability of  
starting state  $j$   
  
↓  
probability of ending in state  $i$   
given starting state  $j$  and  
running over duration  $t$

# Time Reversibility

Most continuous-time Markov models assume time reversibility

The probability of a sequence of states visited forward in time is equal to the series of states visited in the reverse order

$$\pi_j p_{ji}(t) = \pi_i p_{ij}(t)$$

↑  
↑  
probability of  
starting state  $i$   
  
probability of ending in state  $j$   
given starting state  $i$  and  
running over duration  $t$

# Time Reversibility

Most continuous-time Markov models assume time reversibility

Consider two rate matrices:

$$\mathbf{Q}_1 = \begin{pmatrix} \text{time irreversible} \\ -1.918 & 0.127 & 1.670 & 0.121 \\ 0.093 & -1.031 & 0.334 & 0.604 \\ 0.463 & 0.127 & -0.711 & 0.121 \\ 0.093 & 0.634 & 0.334 & -1.061 \end{pmatrix} \quad \mathbf{Q}_2 = \begin{pmatrix} \text{time reversible} \\ -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

# Time Reversibility

Most continuous-time Markov models assume time reversibility

Consider two rate matrices:

$$\mathbf{Q}_1 = \begin{pmatrix} \text{time irreversible} \\ -1.918 & 0.127 & 1.670 & 0.121 \\ 0.093 & -1.031 & 0.334 & 0.604 \\ 0.463 & 0.127 & -0.711 & 0.121 \\ 0.093 & 0.634 & 0.334 & -1.061 \end{pmatrix} \quad \mathbf{Q}_2 = \begin{pmatrix} \text{time reversible} \\ -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$

T  
|  
A

time irreversible:  $\mathbf{Q}_1$  0.0166716

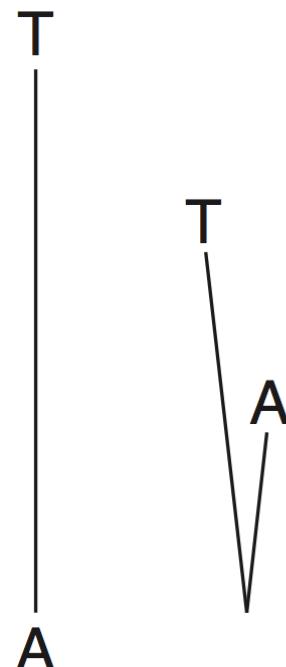
time reversible:  $\mathbf{Q}_2$  0.0045155

# Time Reversibility

Most continuous-time Markov models assume time reversibility

Consider two rate matrices:

$$\mathbf{Q}_1 = \begin{pmatrix} \text{time irreversible} \\ -1.918 & 0.127 & 1.670 & 0.121 \\ 0.093 & -1.031 & 0.334 & 0.604 \\ 0.463 & 0.127 & -0.711 & 0.121 \\ 0.093 & 0.634 & 0.334 & -1.061 \end{pmatrix} \quad \mathbf{Q}_2 = \begin{pmatrix} \text{time reversible} \\ -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$



time irreversible:  $\mathbf{Q}_1$     0.0166716    0.0172425

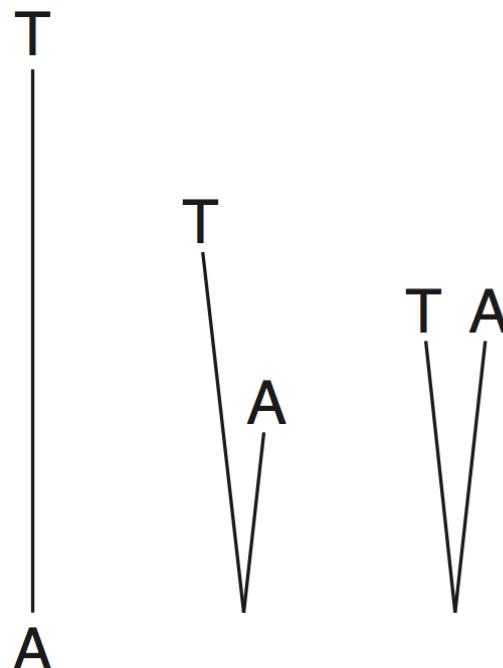
time reversible:  $\mathbf{Q}_2$     0.0045155    0.0045155

# Time Reversibility

Most continuous-time Markov models assume time reversibility

Consider two rate matrices:

$$\mathbf{Q}_1 = \begin{pmatrix} -1.918 & 0.127 & 1.670 & 0.121 \\ 0.093 & -1.031 & 0.334 & 0.604 \\ 0.463 & 0.127 & -0.711 & 0.121 \\ 0.093 & 0.634 & 0.334 & -1.061 \end{pmatrix} \quad \mathbf{Q}_2 = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$



time irreversible:  $\mathbf{Q}_1$  0.0166716 0.0172425 0.0175681

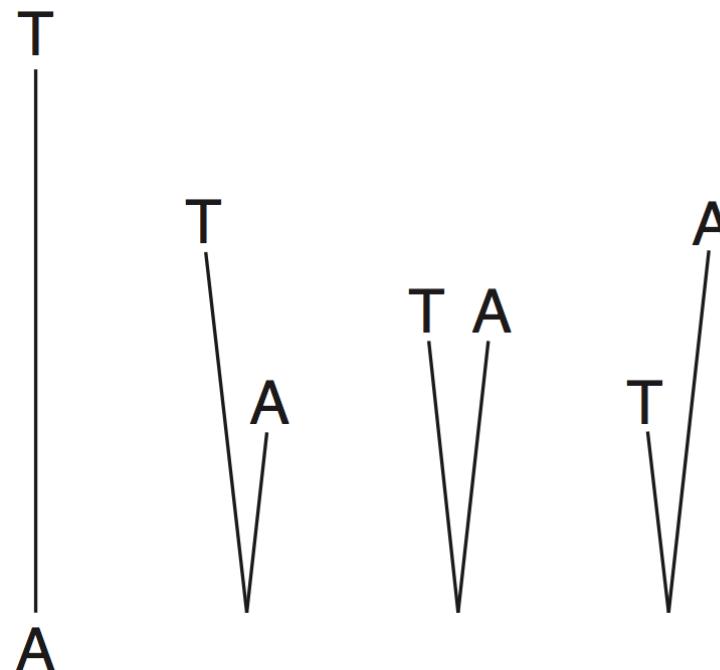
time reversible:  $\mathbf{Q}_2$  0.0045155 0.0045155 0.0045155

# Time Reversibility

Most continuous-time Markov models assume time reversibility

Consider two rate matrices:

$$\mathbf{Q}_1 = \begin{pmatrix} -1.918 & 0.127 & 1.670 & 0.121 \\ 0.093 & -1.031 & 0.334 & 0.604 \\ 0.463 & 0.127 & -0.711 & 0.121 \\ 0.093 & 0.634 & 0.334 & -1.061 \end{pmatrix} \quad \mathbf{Q}_2 = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$



time irreversible:  $\mathbf{Q}_1$  0.0166716 0.0172425 0.0175681 0.0179176

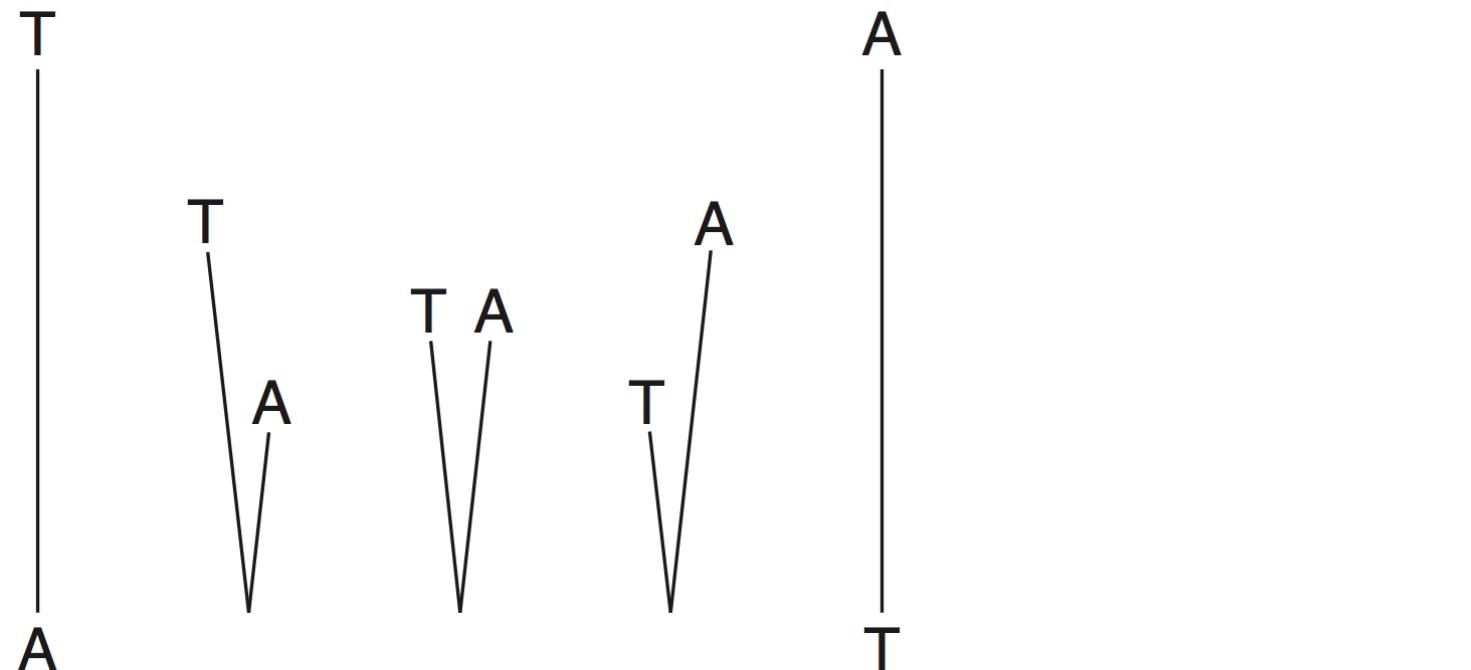
time reversible:  $\mathbf{Q}_2$  0.0045155 0.0045155 0.0045155 0.0045155

# Time Reversibility

Most continuous-time Markov models assume time reversibility

Consider two rate matrices:

$$\mathbf{Q}_1 = \begin{pmatrix} -1.918 & 0.127 & 1.670 & 0.121 \\ 0.093 & -1.031 & 0.334 & 0.604 \\ 0.463 & 0.127 & -0.711 & 0.121 \\ 0.093 & 0.634 & 0.334 & -1.061 \end{pmatrix} \quad \mathbf{Q}_2 = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$



time irreversible:  $\mathbf{Q}_1$     0.0166716    0.0172425    0.0175681    0.0179176    0.0186808

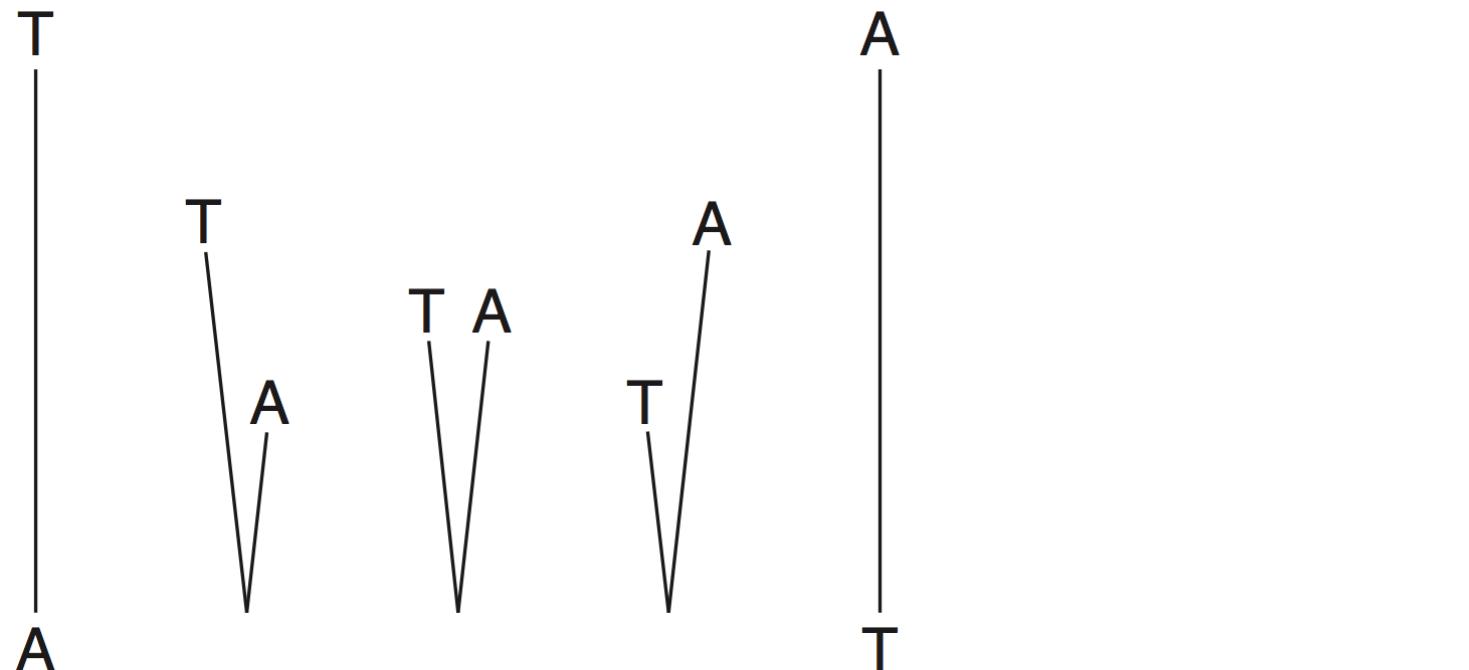
time reversible:  $\mathbf{Q}_2$     0.0045155    0.0045155    0.0045155    0.0045155    0.0045155

# Time Reversibility

Most continuous-time Markov models assume time reversibility

Consider two rate matrices:

$$\mathbf{Q}_1 = \begin{pmatrix} -1.918 & 0.127 & 1.670 & 0.121 \\ 0.093 & -1.031 & 0.334 & 0.604 \\ 0.463 & 0.127 & -0.711 & 0.121 \\ 0.093 & 0.634 & 0.334 & -1.061 \end{pmatrix} \quad \mathbf{Q}_2 = \begin{pmatrix} -1.916 & 0.541 & 0.787 & 0.588 \\ 0.148 & -1.069 & 0.415 & 0.506 \\ 0.286 & 0.170 & -0.591 & 0.135 \\ 0.525 & 0.236 & 0.594 & -1.355 \end{pmatrix}$$



time irreversible:  $\mathbf{Q}_1$  0.0166716 0.0172425 0.0175681 0.0179176 0.0186808

time reversible:  $\mathbf{Q}_2$  0.0045155 0.0045155 0.0045155 0.0045155 0.0045155

# Time Reversibility

Most continuous-time Markov models assume time reversibility

The probability of a sequence of states visited forward in time is equal to the series of states visited in the reverse order

$$\pi_j p_{ji}(t) = \pi_i p_{ij}(t)$$

Time-reversible models allow an unrooted tree to be rooted anywhere without changing the probability of the data

# Time Reversibility

Most continuous-time Markov models assume time reversibility

The probability of a sequence of states visited forward in time is equal to the series of states visited in the reverse order

$$\pi_j p_{ji}(t) = \pi_i p_{ij}(t)$$

Time-reversible models allow an unrooted tree to be rooted anywhere without changing the probability of the data

This allows us to perform inference under unrooted trees, which simplifies the space of trees that we have to evaluate, and the algorithms used to compute the likelihood

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

→ What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

 Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$ , where:

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$ , where:  $a = r_{AC}$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$ , where:  $a = r_{AC}$

$$\boxed{b = r_{AG}}$$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$ , where:  $a = r_{AC}$

$$b = r_{AG}$$

$$\boxed{c = r_{AT}}$$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$ , where:  $a = r_{AC}$

$$b = r_{AG}$$

$$c = r_{AT}$$

$$\boxed{d = r_{CG}}$$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$ , where:  $a = r_{AC}$

$$b = r_{AG}$$

$$c = r_{AT}$$

$$d = r_{CG}$$

$$\boxed{e = r_{CT}}$$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$ , where:  $a = r_{AC}$

$$b = r_{AG}$$

$$c = r_{AT}$$

$$d = r_{CG}$$

$$e = r_{CT}$$

$$\boxed{f = r_{GT}}$$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$

The **stationary frequencies** are in the columns of the matrix, reflecting the fact that the instantaneous rate of change to state  $j$  depends on its stationary frequency,  $\pi_j$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$

The **stationary frequencies** are in the columns of the matrix, reflecting the fact that the instantaneous rate of change to state  $j$  depends on its stationary frequency,  $\pi_j$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$

The **stationary frequencies** are in the columns of the matrix, reflecting the fact that the instantaneous rate of change to state  $j$  depends on its stationary frequency,  $\pi_j$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$

The **stationary frequencies** are in the columns of the matrix, reflecting the fact that the instantaneous rate of change to state  $j$  depends on its stationary frequency,  $\pi_j$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$

The **stationary frequencies** are in the columns of the matrix, reflecting the fact that the instantaneous rate of change to state  $j$  depends on its stationary frequency,  $\pi_j$

# Stochastic Models of Nucleotide Substitution

## Anatomy of General Time Reversible models

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

The **instantaneous rates** of change between states  $i$  and  $j$ ,  $q_{ij}$ , depend both on the relative substitution rates,  $r_{ij}$ , (represented here as  $a-f$ ), and also on the stationary frequencies of each state of the process,  $\pi_j$

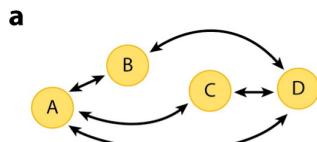
The **relative rates** allow for possible biases in the instantaneous rates of change between each pair of states,  $r_{ij}$

The **stationary frequencies** are in the columns of the matrix, reflecting the fact that the instantaneous rate of change to state  $i$  depends on its stationary frequency,  $\pi_j$

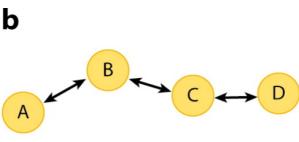
Recall that the average rate of substitution is scaled to be one:  $\mu = - \sum_i \pi_i q_{ii}$

# Stochastic Models of Nucleotide Substitution

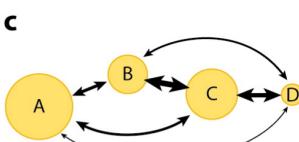
A phylogeographic intuition on relative rates and stationary frequencies



	[A]	[B]	[C]	[D]
[A]	—	$\mu$	$\mu$	$\mu$
[B]	$\mu$	—	$\mu$	$\mu$
[C]	$\mu$	$\mu$	—	$\mu$
[D]	$\mu$	$\mu$	$\mu$	—



	[A]	[B]	[C]	[D]
[A]	—	$\mu$	0	0
[B]	$\mu$	—	$\mu$	0
[C]	0	$\mu$	—	$\mu$
[D]	0	0	$\mu$	—

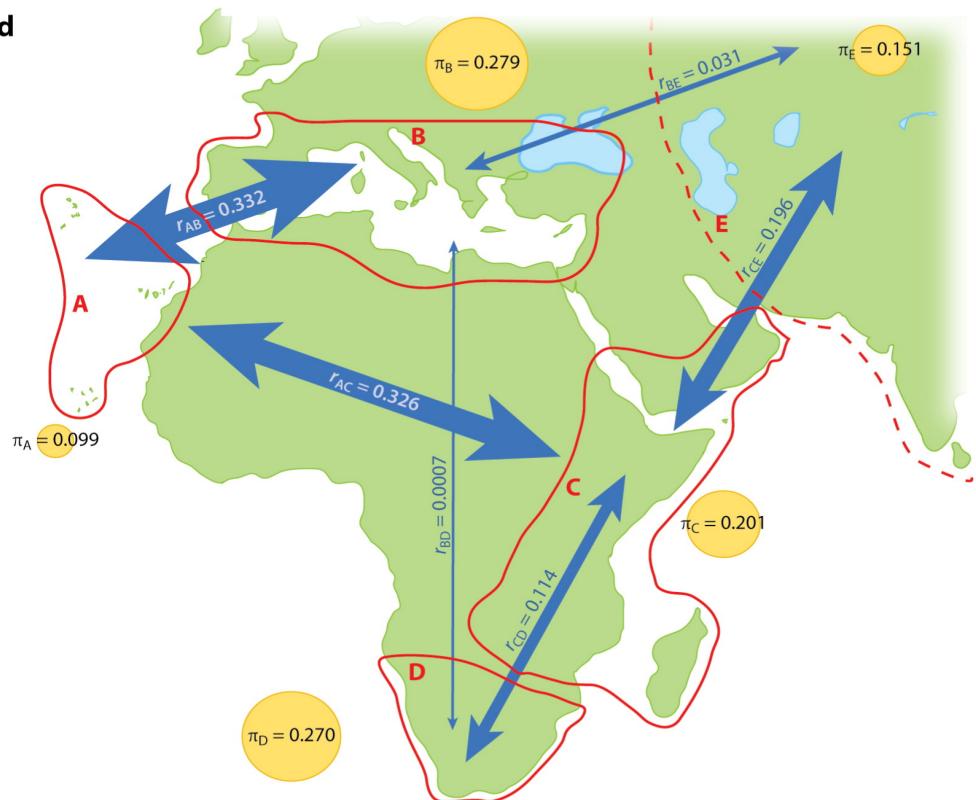


	[A]	[B]	[C]	[D]
[A]	—	$\pi_B r_{AB}$	$\pi_C r_{AC}$	$\pi_D r_{AD}$
[B]	$\pi_A r_{AB}$	—	$\pi_C r_{BC}$	$\pi_D r_{BD}$
[C]	$\pi_A r_{AC}$	$\pi_B r_{BC}$	—	$\pi_D r_{CD}$
[D]	$\pi_A r_{AD}$	$\pi_B r_{BD}$	$\pi_C r_{CD}$	—

Ronquist and Sanmartin used a GTR model to study migrations of plants between parts of the old world.

Exchangeabilities correspond to the intensities of biotic exchange (thickness of arrows)

Equilibrium frequencies correspond to the carrying capacities of the regions



# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969 (JC69) substitution model

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

Relative rates of change between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal, therefore:  $a = b = c = d = e = f$

# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969 (JC69) substitution model

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

Relative rates of change between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal, therefore:  $a = b = c = d = e = f$

# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969 (JC69) substitution model

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

Relative rates of change between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal, therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are assumed to be equal, therefore:  $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969 (JC69) substitution model

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

Relative rates of change between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal, therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are assumed to be equal, therefore:  $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969 (JC69) substitution model

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

Relative rates of change between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal, therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are assumed to be equal, therefore:  $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

A **free parameter** is a parameter that is free to vary (it is estimated from the data)

# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969 (JC69) substitution model

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

Relative rates of change between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal, therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are assumed to be equal, therefore:  $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

A **free parameter** is a parameter that is free to vary (it is estimated from the data)

How many free parameters are there in the JC69 substitution model?

# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969 (JC69) substitution model

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

Relative rates of change between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal, therefore:  $a = b = c = d = e = f$

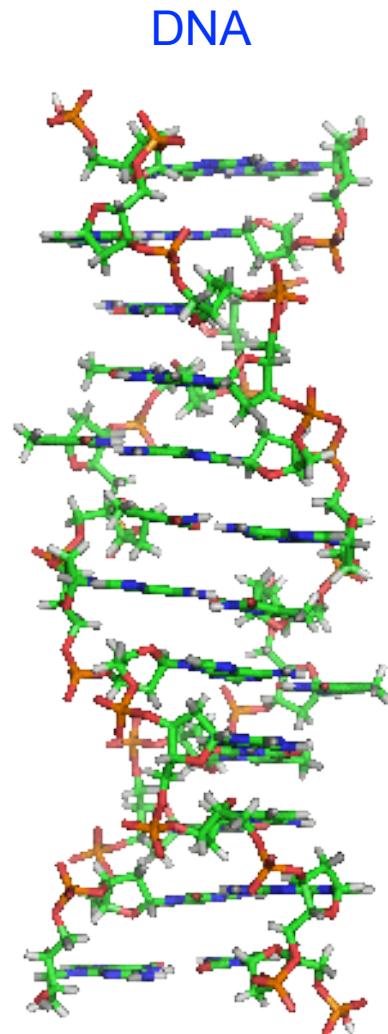
The stationary frequencies of each state are assumed to be equal, therefore:  $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

A **free parameter** is a parameter that is free to vary (it is estimated from the data)

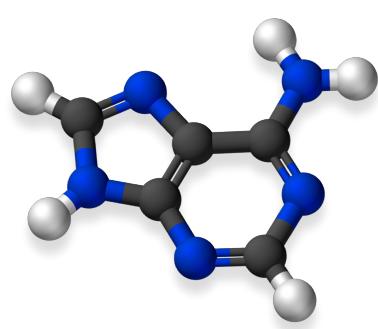
How many free parameters are there in the JC69 substitution model?

# Stochastic Models of Nucleotide Substitution

Biology motivates the extension of models

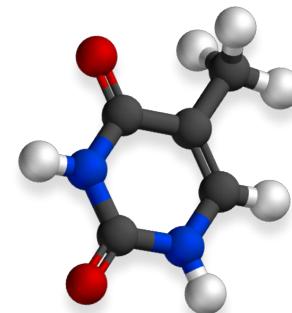


Purines



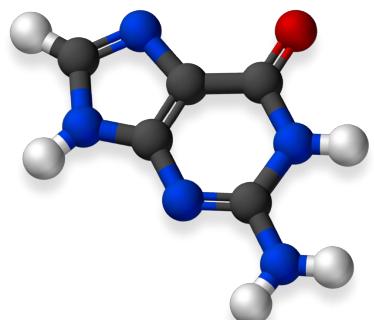
Adenine

Pyrimidines

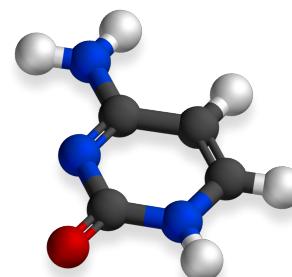


Thymine

Guanine

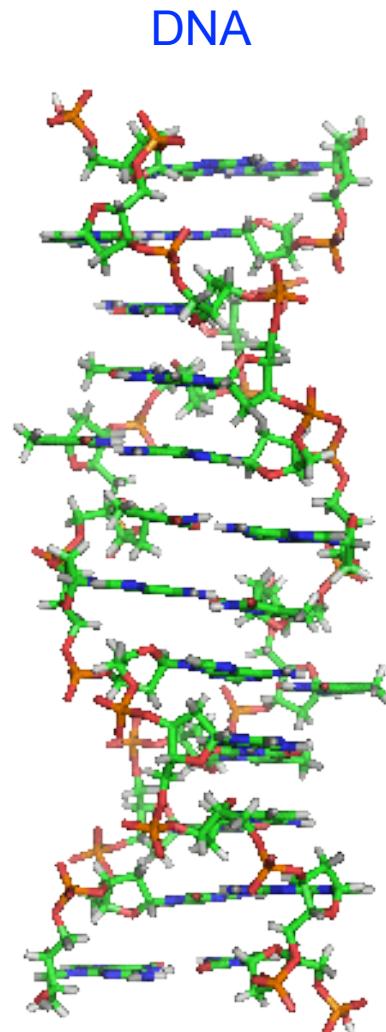


Cytosine

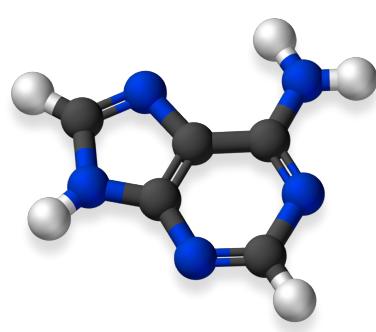


# Stochastic Models of Nucleotide Substitution

Biology motivates the extension of models

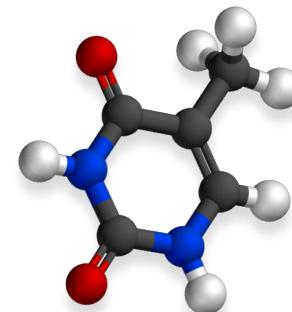


Purines



Adenine

Pyrimidines

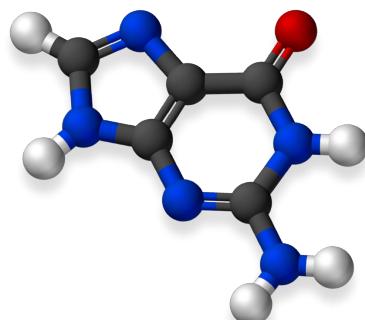


Thymine

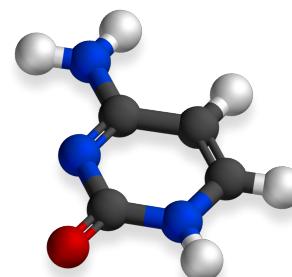
transition  
substitutions



Guanine

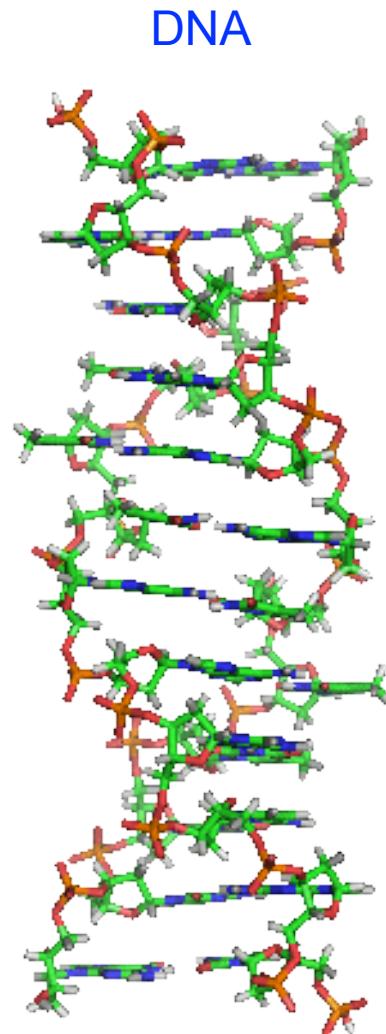


Cytosine

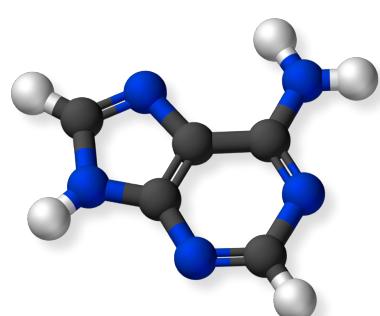


# Stochastic Models of Nucleotide Substitution

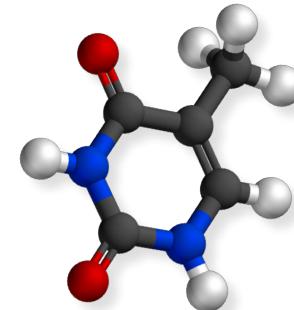
Biology motivates the extension of models



Purines



Pyrimidines

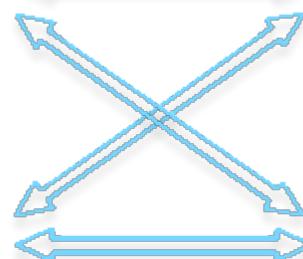


Adenine

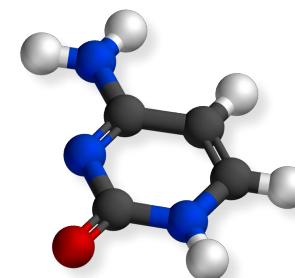
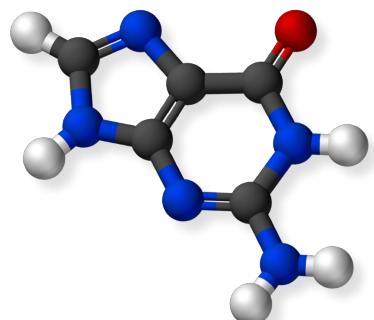


Thymine

Guanine



Cytosine

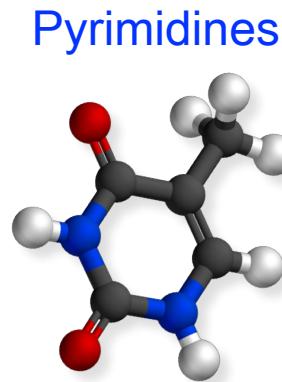
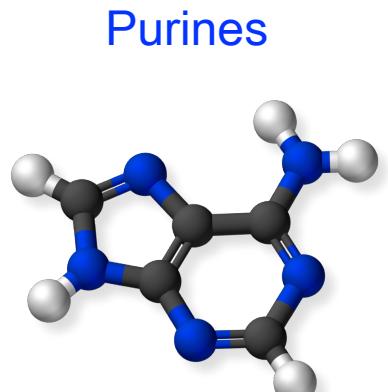
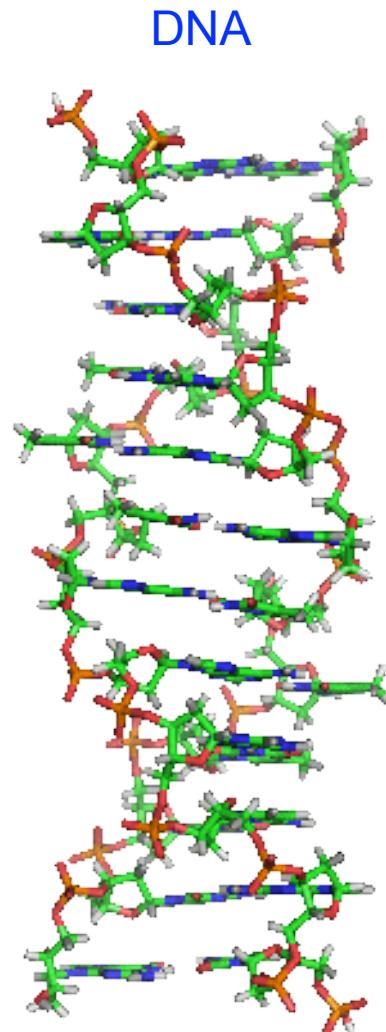


transversion  
substitutions

# Stochastic Models of Nucleotide Substitution

Biology motivates the extension of models

The molecular structure makes transitions more probable than transversions

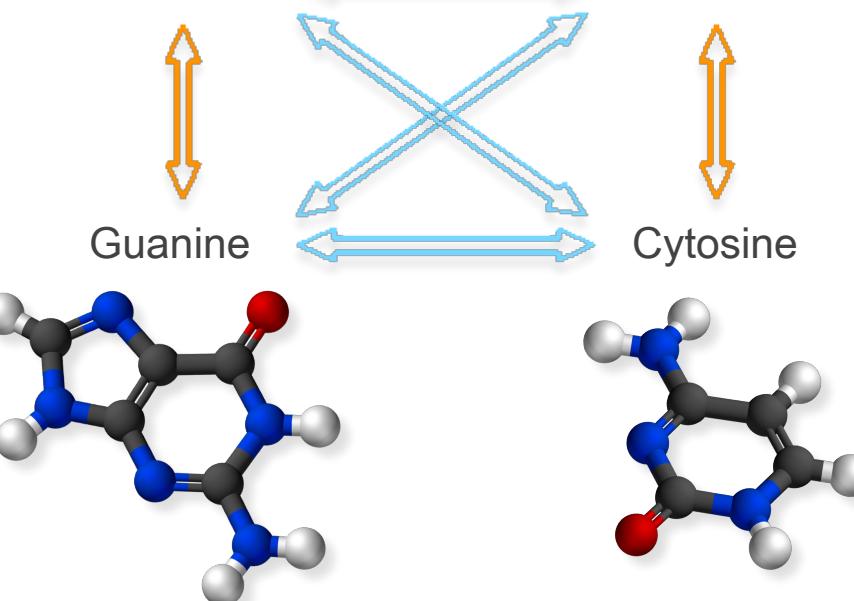


Adenine

Guanine

Thymine

Cytosine



transitions  
>  
transversions

# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \mu b 1/4 & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \mu e 1/4 \\ \mu b 1/4 & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \mu e 1/4 & \mu f 1/4 & - \end{pmatrix}$$

The stationary frequencies of all states are still assumed to be equal

# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \mu b 1/4 & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \mu e 1/4 \\ \mu b 1/4 & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \mu e 1/4 & \mu f 1/4 & - \end{pmatrix}$$

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

Purine      Pyrimidine  
A            T

G            C

# Stochastic Models of Nucleotide Substitution

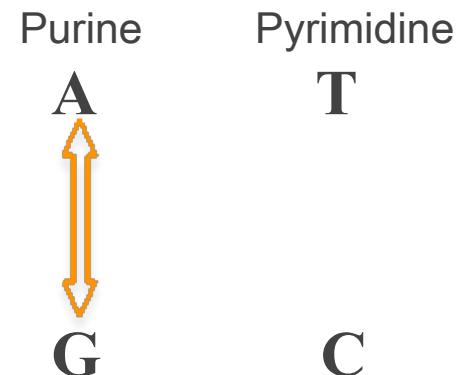
The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \boxed{\mu b 1/4} & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \mu e 1/4 \\ \boxed{\mu b 1/4} & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \mu e 1/4 & \mu f 1/4 & - \end{pmatrix}$$

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

$$b = r_{AG}$$



# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

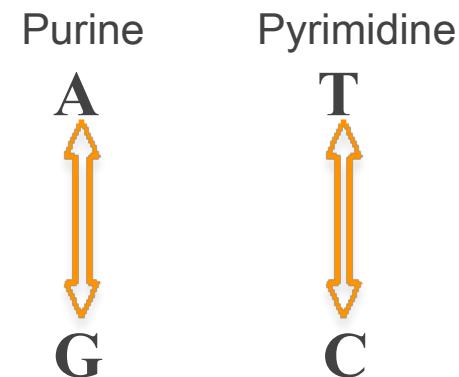
$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \boxed{\mu b 1/4} & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \boxed{\mu e 1/4} \\ \boxed{\mu b 1/4} & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \boxed{\mu e 1/4} & \mu f 1/4 & - \end{pmatrix}$$

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

$$b = r_{AG}$$

$$e = r_{CT}$$



# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \mu b 1/4 & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \mu e 1/4 \\ \mu b 1/4 & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \mu e 1/4 & \mu f 1/4 & - \end{pmatrix}$$

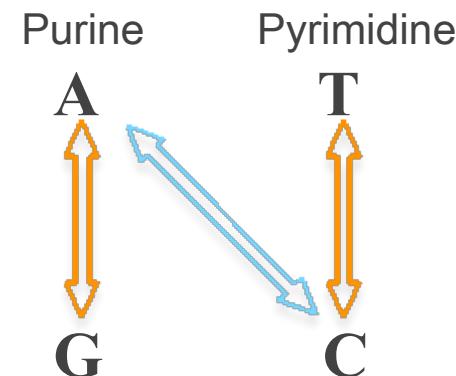
The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

$$b = r_{AG}$$

$$e = r_{CT}$$

$$a = r_{AC}$$



# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \mu b 1/4 & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \mu e 1/4 \\ \mu b 1/4 & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \mu e 1/4 & \mu f 1/4 & - \end{pmatrix}$$

The **stationary frequencies** of all states are still assumed to be equal

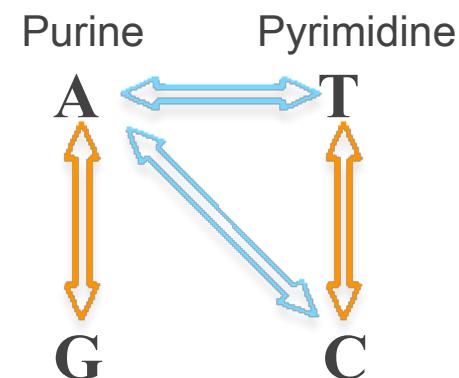
The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

$$b = r_{AG}$$

$$e = r_{CT}$$

$$a = r_{AC}$$

$$c = r_{AT}$$



# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \mu b 1/4 & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \mu e 1/4 \\ \mu b 1/4 & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \mu e 1/4 & \mu f 1/4 & - \end{pmatrix}$$

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

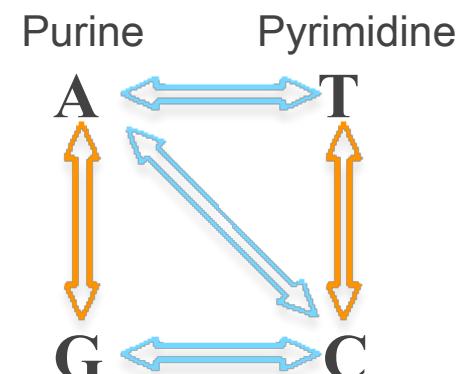
$$b = r_{AG}$$

$$e = r_{CT}$$

$$a = r_{AC}$$

$$c = r_{AT}$$

$$d = r_{CG}$$



# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \mu b 1/4 & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \mu e 1/4 \\ \mu b 1/4 & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \mu e 1/4 & \mu f 1/4 & - \end{pmatrix}$$

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

$$b = r_{AG}$$

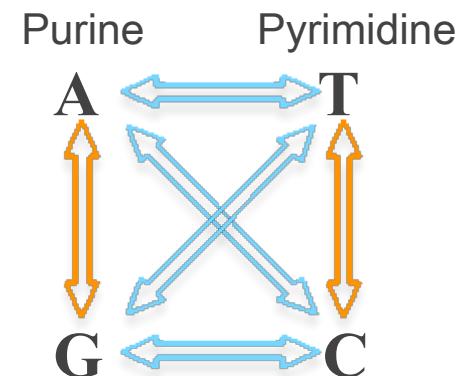
$$e = r_{CT}$$

$$a = r_{AC}$$

$$c = r_{AT}$$

$$d = r_{CG}$$

$$f = r_{GT}$$



# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \mu b 1/4 & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \mu e 1/4 \\ \mu b 1/4 & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \mu e 1/4 & \mu f 1/4 & - \end{pmatrix}$$

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

Under this model, all **transition** substitutions are assumed to have the same rate ( $b = e$ ) and all **transversion** substitutions are assumed to have the same rate ( $a = c = d = f$ )

# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \mu b 1/4 & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \mu e 1/4 \\ \mu b 1/4 & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \mu e 1/4 & \mu f 1/4 & - \end{pmatrix}$$

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

Under this model, all **transition** substitutions are assumed to have the same rate ( $b = e$ ) and all **transversion** substitutions are assumed to have the same rate ( $a = c = d = f$ )  
BUT we allow for a possible bias in the relative rate of transitions and transversions by introducing a new parameter,  $\kappa$ :

$$\kappa = \frac{r_{\text{transition}}}{r_{\text{transversion}}}$$

# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu a 1/4 & \boxed{\mu b 1/4} & \mu c 1/4 \\ \mu a 1/4 & - & \mu d 1/4 & \boxed{\mu e 1/4} \\ \boxed{\mu b 1/4} & \mu d 1/4 & - & \mu f 1/4 \\ \mu c 1/4 & \boxed{\mu e 1/4} & \mu f 1/4 & - \end{pmatrix}$$

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

Under this model, all **transition** substitutions are assumed to have the same rate ( $b = e$ ) and all **transversion** substitutions are assumed to have the same rate ( $a = c = d = f$ )  
BUT we allow for a possible bias in the relative rate of transitions and transversions by introducing a new parameter,  $\kappa$ :

$$\kappa = \frac{r_{\text{transition}}}{r_{\text{transversion}}}$$

# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

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

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

Under this model, all **transition** substitutions are assumed to have the same rate ( $b = e$ ) and all **transversion** substitutions are assumed to have the same rate ( $a = c = d = f$ )  
BUT we allow for a possible bias in the relative rate of transitions and transversions by introducing a new parameter,  $\kappa$ :

$$\kappa = \frac{r_{\text{transition}}}{r_{\text{transversion}}}$$

# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

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

The **stationary frequencies** of all states are still assumed to be equal

The **relative rates** accommodate possible bias in the instantaneous rates of transition and transversion substitutions

Under this model, all **transition** substitutions are assumed to have the same rate ( $b = e$ ) and all **transversion** substitutions are assumed to have the same rate ( $a = c = d = f$ )  
BUT we allow for a possible bias in the relative rate of transitions and transversions by introducing a new parameter,  $\kappa$ :

$$\kappa = \frac{r_{\text{transition}}}{r_{\text{transversion}}}$$

How many free parameters are there in the K80 substitution model?

# Stochastic Models of Nucleotide Substitution

Paparazzi time: Name that model!

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

# Stochastic Models of Nucleotide Substitution

The Kimura, 1980 (K80) substitution model

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

# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969 (JC69) substitution model

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

Relative rates of change between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal, therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are assumed to be equal, therefore:  $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969 (JC69) substitution model

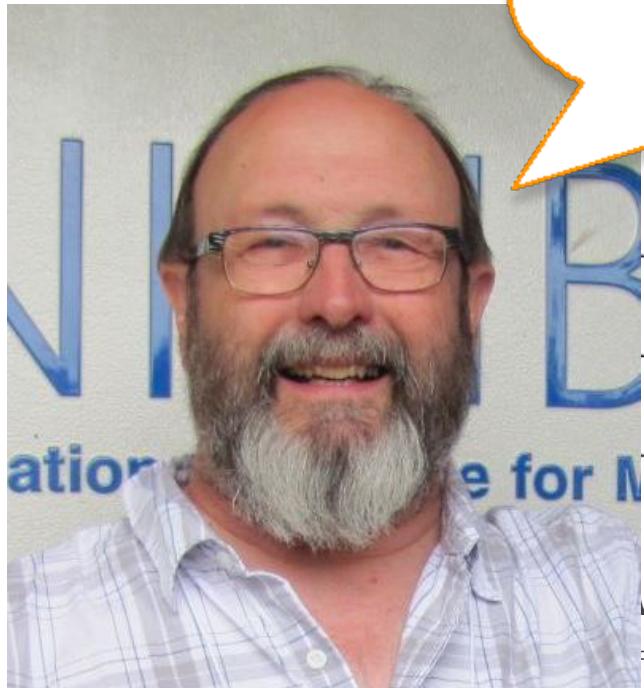
$$\mathbf{Q} = q_{ij} = \begin{pmatrix} - & \mu 1/4 & \mu 1/4 & \mu 1/4 \\ \mu 1/4 & - & \mu 1/4 & \mu 1/4 \\ \mu 1/4 & \mu 1/4 & - & \mu 1/4 \\ \mu 1/4 & \mu 1/4 & \mu 1/4 & - \end{pmatrix}$$

Relative rates of change between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal, therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are assumed to be equal, therefore:  $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

# Stochastic Models of Nucleotide Substitution

The Jukes and Cantor, 1969



We all know  
that's  
complete  
bullshit!

$$\begin{pmatrix} \mu 1/4 & \mu 1/4 & \mu 1/4 \\ \mu 1/4 & - & \mu 1/4 & \mu 1/4 \\ \mu 1/4 & \mu 1/4 & - & \mu 1/4 \\ \mu 1/4 & \mu 1/4 & \mu 1/4 & - \end{pmatrix}$$

Between each pair of states  $i$  and  $j$ ,  $r_{ij}$ , are assumed to be equal:  
 $r_{ij} = d = e = f$

The stationary frequencies of each state are assumed to be equal,  
therefore:  $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

# Stochastic Models of Nucleotide Substitution

Biology motivates the extension of models



$$A = 49,708 \div 151,015 = 32.9\%$$

$$G = 25,162 \div 151,015 = 16.7\%$$

$$A \neq C \neq G \neq T \neq 0.25$$

$$C = 41,073 \div 151,015 = 27.2\%$$

$$T = 34,972 \div 151,015 = 23.2\%$$

# Stochastic Models of Nucleotide Substitution

The Felsenstein, 1981 (F81) substitution model

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

Like the JC69 model, all relative rates are assumed to be equal,  
therefore:  $a = b = c = d = e = f$

# Stochastic Models of Nucleotide Substitution

The Felsenstein, 1981 (F81) substitution model

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

Like the JC69 model, all relative rates are assumed to be equal,  
therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are allowed to be unequal by adding  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Felsenstein, 1981 (F81) substitution model

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

Like the JC69 model, all relative rates are assumed to be equal,  
therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are allowed to be unequal by adding  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Felsenstein, 1981 (F81) substitution model

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

Like the JC69 model, all relative rates are assumed to be equal,  
therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are allowed to be unequal by adding  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Felsenstein, 1981 (F81) substitution model

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

Like the JC69 model, all relative rates are assumed to be equal,  
therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are allowed to be unequal by adding  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Felsenstein, 1981 (F81) substitution model

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

Like the JC69 model, all relative rates are assumed to be equal,  
therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are allowed to be unequal by adding  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Felsenstein, 1981 (F81) substitution model

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

Like the JC69 model, all relative rates are assumed to be equal,  
therefore:  $a = b = c = d = e = f$

The stationary frequencies of each state are allowed to be unequal by adding  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

How many free parameters are there in the F81 substitution model?

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

Like the K80 model, all **transitions** are assumed to have the same rate ( $b = e$ )  
and all **transversion** are assumed to have the same rate ( $a = c = d = f$ )  
and accommodates a possible bias in the transition/transversion rate ratio,  $\kappa$

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

Like the K80 model, all **transitions** are assumed to have the same rate ( $b = e$ )  
and all **transversion** are assumed to have the same rate ( $a = c = d = f$ )  
and accommodates a possible bias in the transition/transversion rate ratio,  $\kappa$

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

Like the K80 model, all **transitions** are assumed to have the same rate ( $b = e$ )  
and all **transversion** are assumed to have the same rate ( $a = c = d = f$ )  
and accommodates a possible bias in the transition/transversion rate ratio,  $\kappa$

Like the F81 model, stationary frequencies are free to vary by including  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

Like the K80 model, all **transitions** are assumed to have the same rate ( $b = e$ )  
and all **transversion** are assumed to have the same rate ( $a = c = d = f$ )  
and accommodates a possible bias in the transition/transversion rate ratio,  $\kappa$

Like the F81 model, stationary frequencies are free to vary by including  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

Like the K80 model, all **transitions** are assumed to have the same rate ( $b = e$ )  
and all **transversion** are assumed to have the same rate ( $a = c = d = f$ )  
and accommodates a possible bias in the transition/transversion rate ratio,  $\kappa$

Like the F81 model, stationary frequencies are free to vary by including  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

Like the K80 model, all **transitions** are assumed to have the same rate ( $b = e$ )  
and all **transversion** are assumed to have the same rate ( $a = c = d = f$ )  
and accommodates a possible bias in the transition/transversion rate ratio,  $\kappa$

Like the F81 model, stationary frequencies are free to vary by including  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

Like the K80 model, all **transitions** are assumed to have the same rate ( $b = e$ )  
and all **transversion** are assumed to have the same rate ( $a = c = d = f$ )  
and accommodates a possible bias in the transition/transversion rate ratio,  $\kappa$

Like the F81 model, stationary frequencies are free to vary by including  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

Like the K80 model, all **transitions** are assumed to have the same rate ( $b = e$ )  
and all **transversion** are assumed to have the same rate ( $a = c = d = f$ )  
and accommodates a possible bias in the transition/transversion rate ratio,  $\kappa$

Like the F81 model, stationary frequencies are free to vary by including  
a parameter for each state:

How many free parameters are there in the HKY85 substitution model?

# Stochastic Models of Nucleotide Substitution

The Hasegawa, Kishino, Yano, 1985 (HKY85) substitution model

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

Combines features of the K80 and F81 substitution models

Like the K80 model, all **transitions** are assumed to have the same rate ( $b = e$ )  
and all **transversion** are assumed to have the same rate ( $a = c = d = f$ )  
and accommodates a possible bias in the transition/transversion rate ratio,  $\kappa$

Like the F81 model, stationary frequencies are free to vary by including  
a parameter for each state:  $\pi_A, \pi_C, \pi_G, \pi_T$

How many free parameters are there in the HKY85 substitution model?

# Stochastic Models of Nucleotide Substitution

The General Time Reversible (GTR) substitution model

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

Allows all six relative substitution rates ( $a-f$ ) to vary

# Stochastic Models of Nucleotide Substitution

The General Time Reversible (GTR) substitution model

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

Allows all six relative substitution rates ( $a-f$ ) to vary

Like the F81 and HKY85 models, stationary frequencies are free to vary

# Stochastic Models of Nucleotide Substitution

The General Time Reversible (GTR) substitution model

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

Allows all six relative substitution rates ( $a-f$ ) to vary

Like the F81 and HKY85 models, stationary frequencies are free to vary

How many free parameters are there in the GTR substitution model?

# Stochastic Models of Nucleotide Substitution

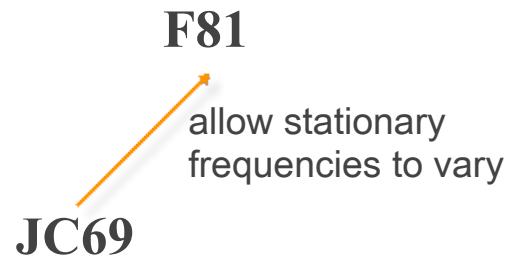
Members of the GTR substitution model family are hierarchically related

**JC69**

We can move from a simpler to a more general model by relaxing a constraint

# Stochastic Models of Nucleotide Substitution

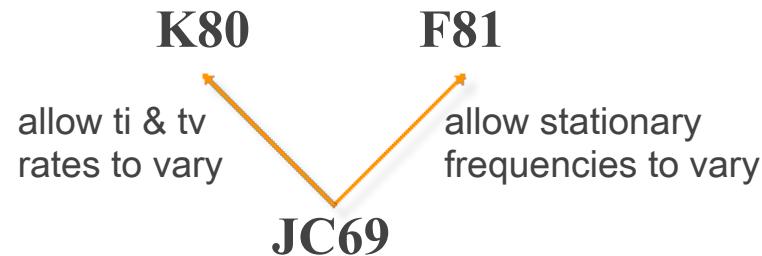
Members of the GTR substitution model family are hierarchically related



We can move from a simpler to a more general model by relaxing a constraint

# Stochastic Models of Nucleotide Substitution

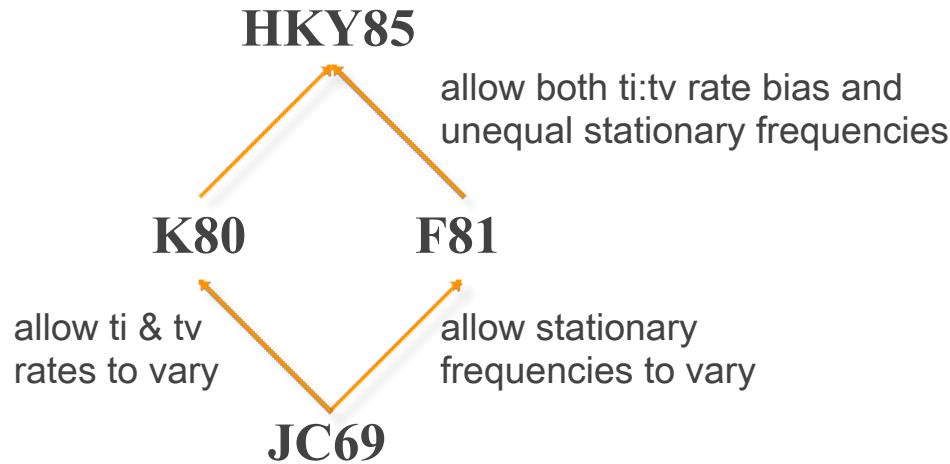
Members of the GTR substitution model family are hierarchically related



We can move from a simpler to a more general model by relaxing a constraint

# Stochastic Models of Nucleotide Substitution

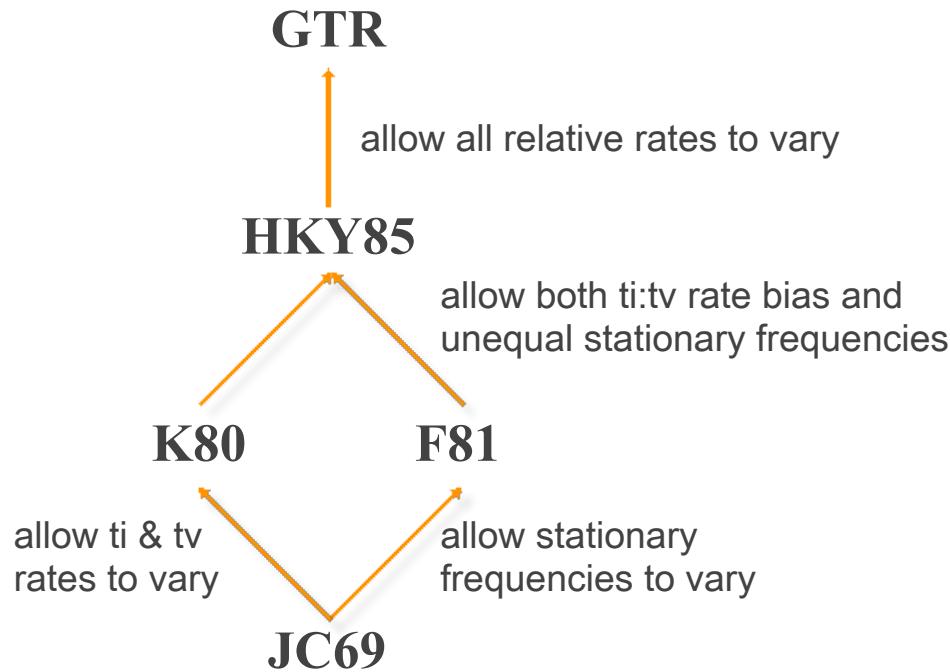
Members of the GTR substitution model family are hierarchically related



We can move from a simpler to a more general model by relaxing a constraint

# Stochastic Models of Nucleotide Substitution

Members of the GTR substitution model family are hierarchically related



We can move from a simpler to a more general model by relaxing a constraint

# Stochastic Models of Nucleotide Substitution

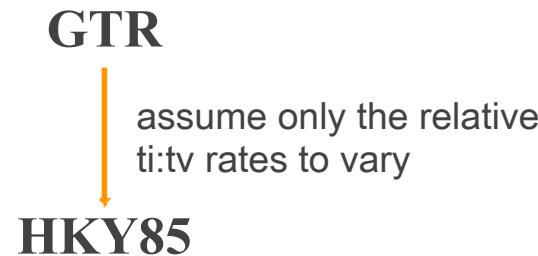
Members of the GTR substitution model family are hierarchically related

GTR

We can move from a more general model to a simpler model by imposing a constraint

# Stochastic Models of Nucleotide Substitution

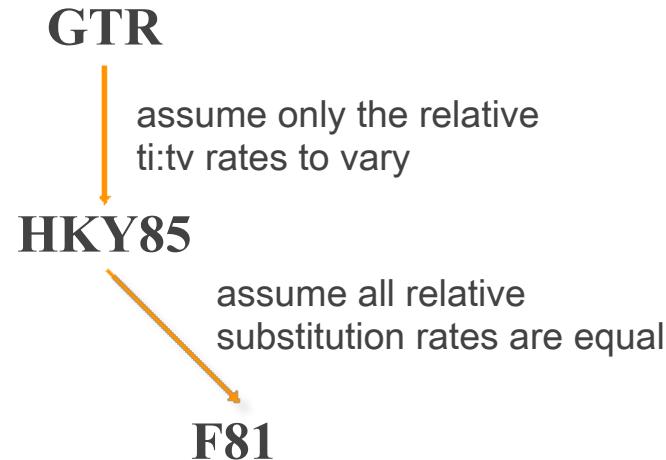
Members of the GTR substitution model family are hierarchically related



We can move from a more general model to a simpler model by imposing a constraint

# Stochastic Models of Nucleotide Substitution

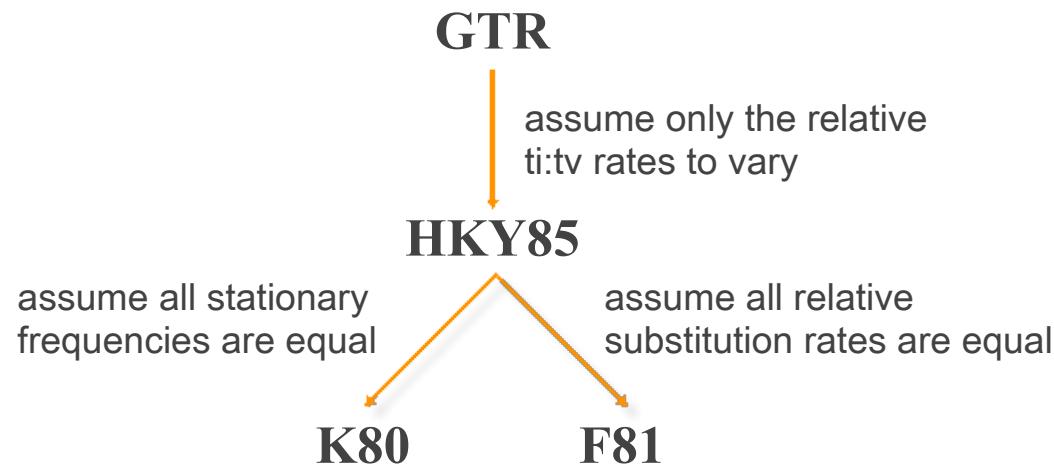
Members of the GTR substitution model family are hierarchically related



We can move from a more general model to a simpler model by imposing a constraint

# Stochastic Models of Nucleotide Substitution

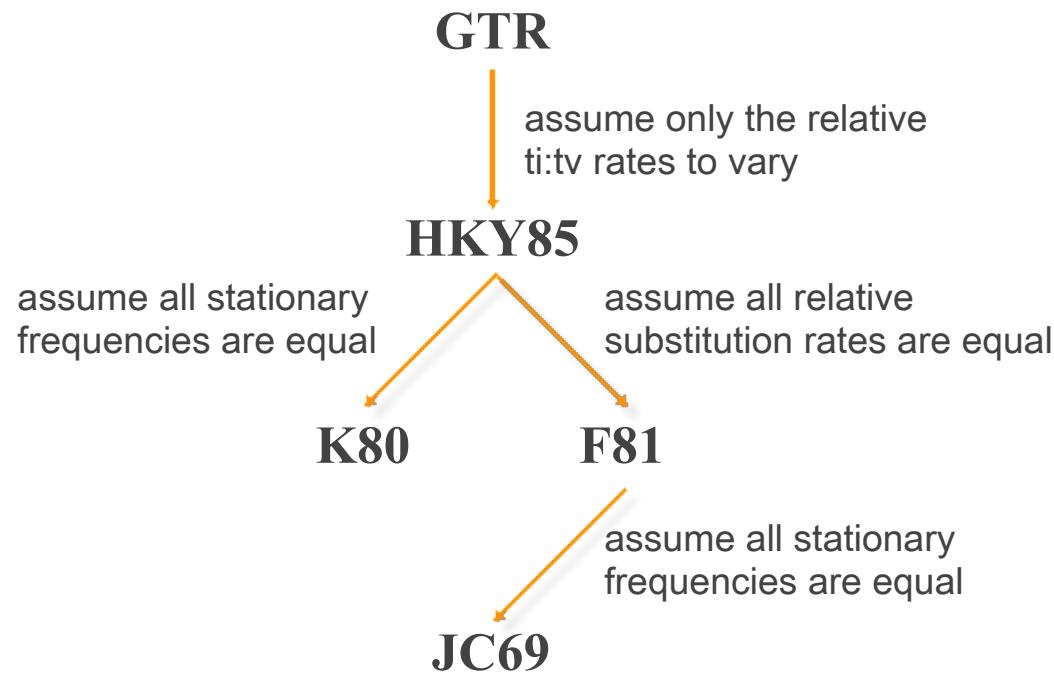
Members of the GTR substitution model family are hierarchically related



We can move from a more general model to a simpler model by imposing a constraint

# Stochastic Models of Nucleotide Substitution

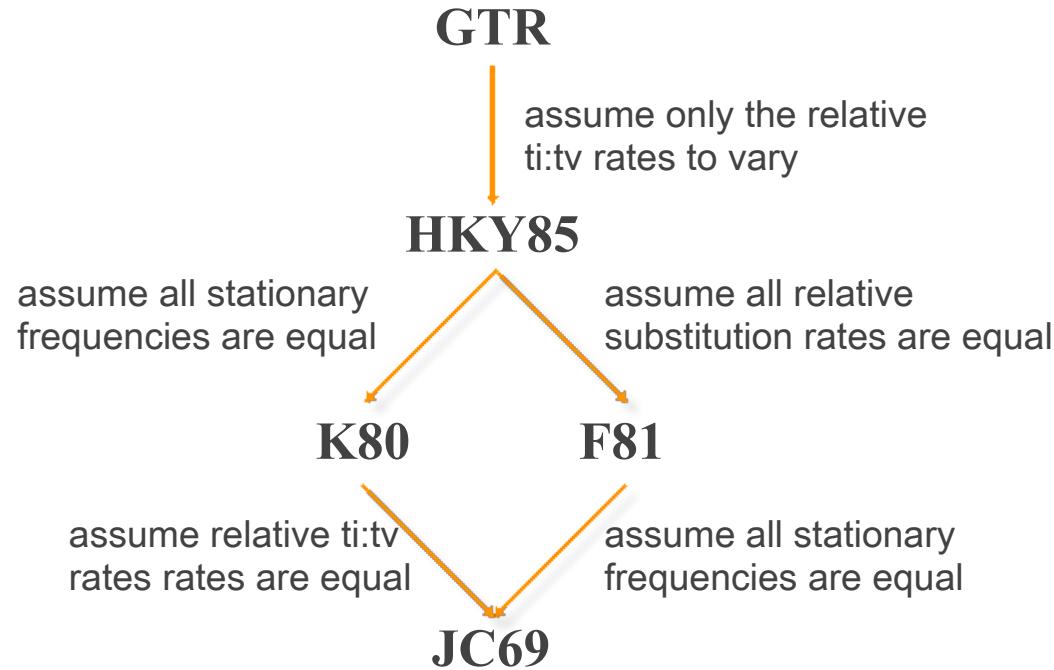
Members of the GTR substitution model family are hierarchically related



We can move from a more general model to a simpler model by imposing a constraint

# Stochastic Models of Nucleotide Substitution

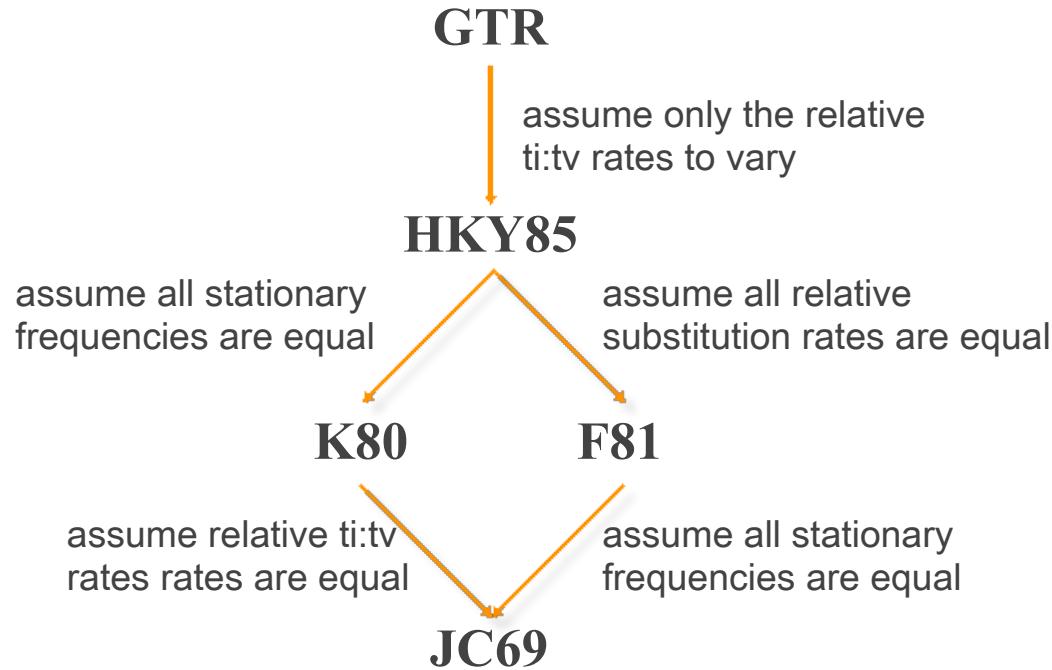
Members of the GTR substitution model family are hierarchically related



We can move from a more general model to a simpler model by imposing a constraint

# Stochastic Models of Nucleotide Substitution

Members of the GTR substitution model family are hierarchically related



We can move from a more general model to a simpler model by imposing a constraint  
The differences between models allow us to test hypotheses (to learn) about our data

# Stochastic Models of Nucleotide Substitution

How many members of the GTR model family are there?

The total number of unique models for  $k$  relative-rate parameters is described by Bell numbers:

$$\mathcal{B}(n) = \sum_{k=1}^n \mathcal{S}_2(n, k)$$

# Stochastic Models of Nucleotide Substitution

How many members of the GTR model family are there?

The total number of unique models for  $k$  relative-rate parameters is described by Bell numbers:

$$\mathcal{B}(n) = \sum_{k=1}^n \mathcal{S}_2(n, k)$$

which is the sum of the Sterling numbers of the second kind:

$$\mathcal{S}_2(n, k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^n$$

# Stochastic Models of Nucleotide Substitution

How many members of the GTR model family are there?

The total number of unique models for  $k$  relative-rate parameters is described by Bell numbers:

$$\mathcal{B}(n) = \sum_{k=1}^n \mathcal{S}_2(n, k)$$

which is the sum of the Sterling numbers of the second kind:

$$\mathcal{S}_2(n, k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^n$$

For  $K = 6$ :

$$\mathcal{B}(6) = 203$$

# Stochastic Models of Nucleotide Substitution

How many members of the GTR model family are there?

$K$	Models				
1	$M_1 = 111111$				
2	$M_2 = 122222$ $M_7 = 111112$ $M_{12} = 122221$ $M_{17} = 112211$ $M_{22} = 111122$ $M_{27} = 121122$ $M_{32} = 122211$	$M_3 = 121111$ $M_8 = 112222$ $M_{13} = 122111$ $M_{18} = 112121$ $M_{23} = 111222$ $M_{28} = 121212$	$M_4 = 112111$ $M_9 = 121222$ $M_{14} = 121211$ $M_{19} = 112112$ $M_{24} = 112122$ $M_{29} = 121221$	$M_5 = 111211$ $M_{10} = 122122$ $M_{15} = 121121$ $M_{20} = 111221$ $M_{25} = 112212$ $M_{30} = 122112$	$M_6 = 111121$ $M_{11} = 122212$ $M_{16} = 121112$ $M_{21} = 111212$ $M_{26} = 112221$ $M_{31} = 122121$
3	$M_{33} = 123333$ $M_{38} = 123111$ $M_{43} = 112131$ $M_{48} = 122333$ $M_{53} = 123232$ $M_{58} = 121333$ $M_{63} = 112322$ $M_{68} = 123221$ $M_{73} = 122321$ $M_{78} = 123311$ $M_{83} = 121133$ $M_{88} = 122131$ $M_{93} = 121213$ $M_{98} = 112133$ $M_{103} = 112132$ $M_{108} = 112233$ $M_{113} = 121332$ $M_{118} = 123132$	$M_{34} = 123222$ $M_{39} = 121311$ $M_{44} = 112113$ $M_{49} = 123233$ $M_{54} = 123223$ $M_{59} = 123133$ $M_{64} = 112232$ $M_{69} = 121322$ $M_{74} = 122132$ $M_{79} = 123131$ $M_{84} = 123211$ $M_{89} = 122113$ $M_{94} = 121132$ $M_{99} = 112321$ $M_{104} = 112123$ $M_{109} = 112323$ $M_{114} = 122133$ $M_{119} = 123213$	$M_{35} = 122322$ $M_{40} = 121131$ $M_{45} = 111231$ $M_{50} = 123323$ $M_{55} = 122332$ $M_{60} = 123313$ $M_{65} = 112223$ $M_{70} = 121232$ $M_{75} = 122123$ $M_{80} = 123113$ $M_{85} = 123121$ $M_{90} = 121321$ $M_{95} = 121123$ $M_{100} = 112312$ $M_{105} = 111233$ $M_{110} = 112332$ $M_{115} = 122313$ $M_{120} = 123231$	$M_{36} = 122232$ $M_{41} = 121113$ $M_{46} = 111213$ $M_{51} = 123332$ $M_{56} = 122323$ $M_{61} = 123331$ $M_{66} = 123122$ $M_{71} = 121223$ $M_{76} = 122231$ $M_{81} = 121331$ $M_{86} = 123112$ $M_{91} = 121312$ $M_{96} = 112331$ $M_{101} = 112231$ $M_{106} = 111232$ $M_{111} = 121233$ $M_{116} = 122331$ $M_{121} = 123312$	$M_{37} = 122223$ $M_{42} = 112311$ $M_{47} = 111123$ $M_{52} = 123322$ $M_{57} = 122233$ $M_{62} = 112333$ $M_{67} = 123212$ $M_{72} = 122312$ $M_{77} = 122213$ $M_{82} = 121313$ $M_{87} = 122311$ $M_{92} = 121231$ $M_{97} = 112313$ $M_{102} = 112213$ $M_{107} = 111223$ $M_{112} = 121323$ $M_{117} = 123123$ $M_{122} = 123321$
4	$M_{123} = 123444$ $M_{128} = 123242$ $M_{133} = 123411$ $M_{138} = 121134$ $M_{143} = 123344$ $M_{148} = 123442$ $M_{153} = 123432$ $M_{158} = 123144$ $M_{163} = 121334$ $M_{168} = 123341$ $M_{173} = 112342$ $M_{178} = 123142$ $M_{183} = 121324$	$M_{124} = 123433$ $M_{129} = 123224$ $M_{134} = 123141$ $M_{139} = 112341$ $M_{144} = 123434$ $M_{149} = 122344$ $M_{154} = 123243$ $M_{159} = 123414$ $M_{164} = 123413$ $M_{169} = 123314$ $M_{174} = 112324$ $M_{179} = 123124$ $M_{184} = 121234$	$M_{125} = 123434$ $M_{130} = 122342$ $M_{135} = 123114$ $M_{140} = 112314$ $M_{145} = 123443$ $M_{150} = 122343$ $M_{155} = 123234$ $M_{160} = 123441$ $M_{165} = 123431$ $M_{170} = 112344$ $M_{175} = 112234$ $M_{180} = 123241$ $M_{185} = 122341$	$M_{126} = 123334$ $M_{131} = 122324$ $M_{136} = 121341$ $M_{141} = 112134$ $M_{146} = 123244$ $M_{151} = 122334$ $M_{156} = 123342$ $M_{161} = 121344$ $M_{166} = 123143$ $M_{171} = 112343$ $M_{176} = 123412$ $M_{181} = 123214$ $M_{186} = 122314$	$M_{127} = 123422$ $M_{132} = 122234$ $M_{137} = 121314$ $M_{142} = 111234$ $M_{147} = 123424$ $M_{152} = 123423$ $M_{157} = 123324$ $M_{162} = 121343$ $M_{167} = 123134$ $M_{172} = 112334$ $M_{177} = 123421$ $M_{182} = 121342$ $M_{187} = 122134$
5	$M_{188} = 123455$ $M_{193} = 123345$ $M_{198} = 123451$	$M_{189} = 123454$ $M_{194} = 123452$ $M_{199} = 123415$	$M_{190} = 123445$ $M_{195} = 123425$ $M_{200} = 123145$	$M_{191} = 123453$ $M_{196} = 123245$ $M_{201} = 121345$	$M_{192} = 123435$ $M_{197} = 122345$ $M_{202} = 112345$
6	$M_{203} = 123456$				

# Stochastic Models of Nucleotide Substitution

Limitations of the GTR model family:

Homogeneity along the sequence and along the tree

- If the process were homogeneous along the sequence: all sites would be similar in their ACGT composition, and in the number of substitutions they have undergone
- If the process were homogeneous along the tree, the sequences would all be similar in their ACGT composition

# Stochastic Models of Nucleotide Substitution

Limitations of the GTR model family:

Homogeneity along the sequence and along the tree

- If the process were homogeneous along the sequence: all sites would be similar in their ACGT composition, and in the number of substitutions they have undergone
- If the process were homogeneous along the tree, the sequences would all be similar in their ACGT composition

Failure to account for heterogeneity along the tree and along the sequence can result in erroneous tree topologies, incorrect inferences of selection, etc.

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

 Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

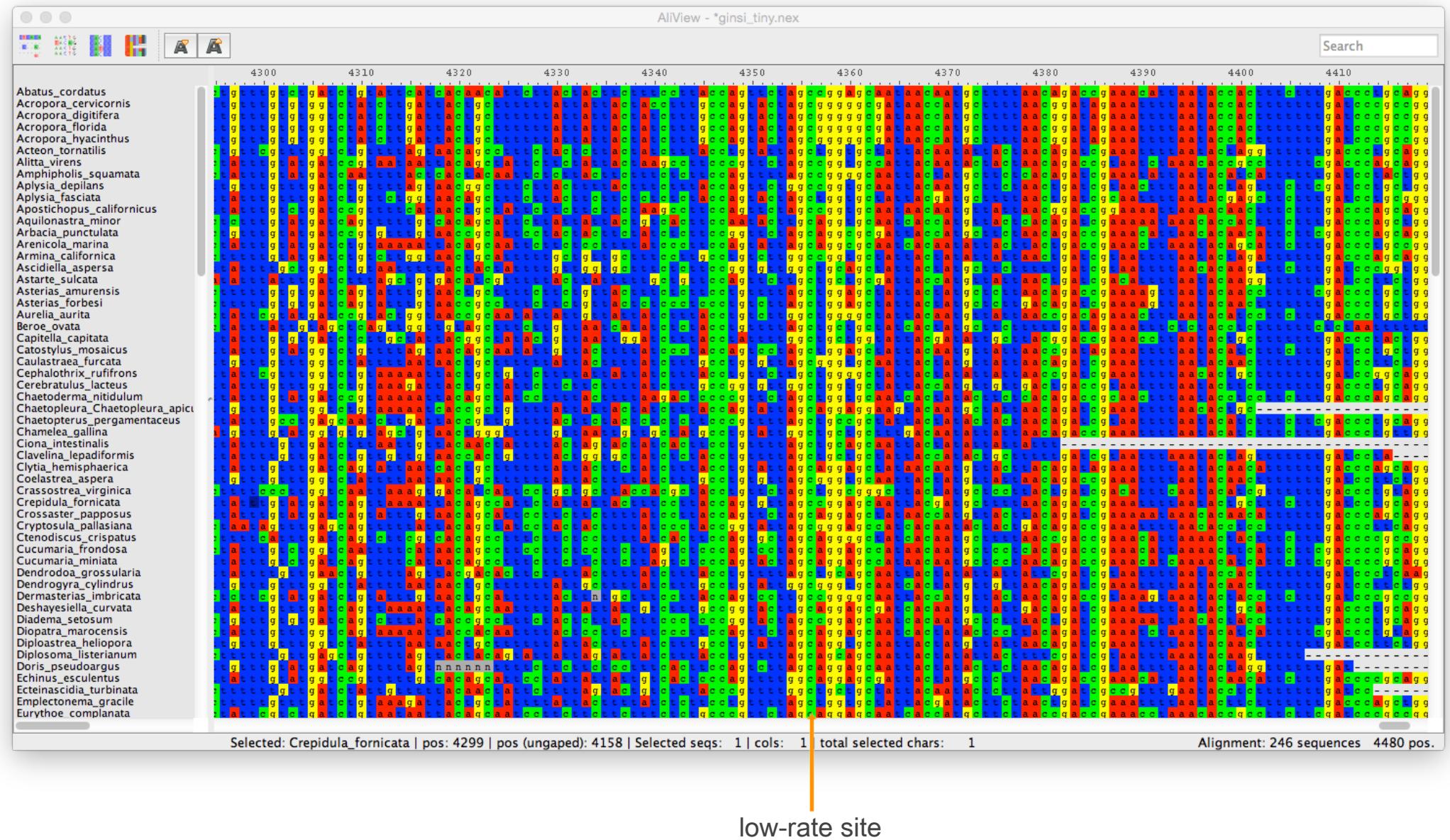
# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



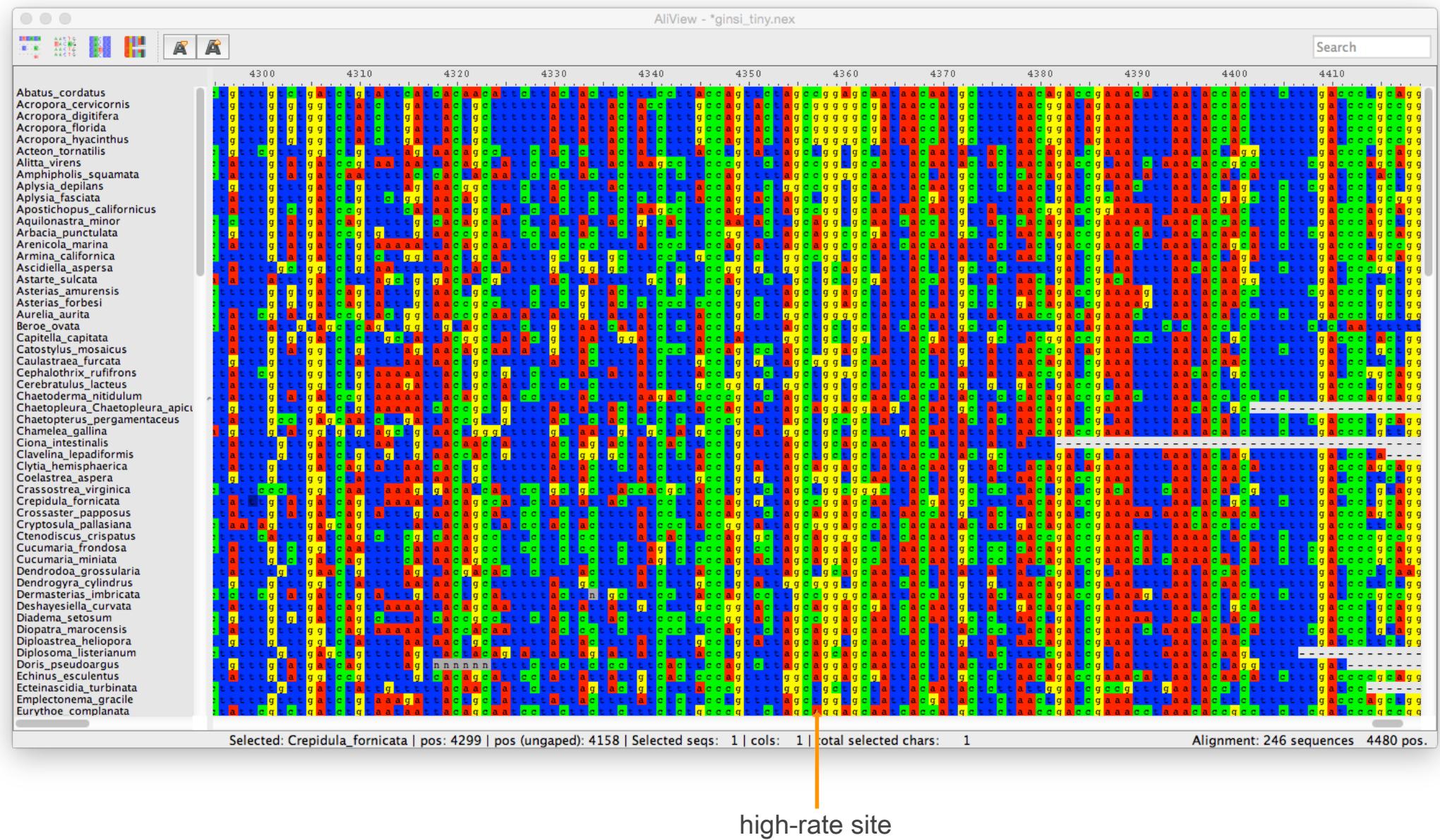
# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



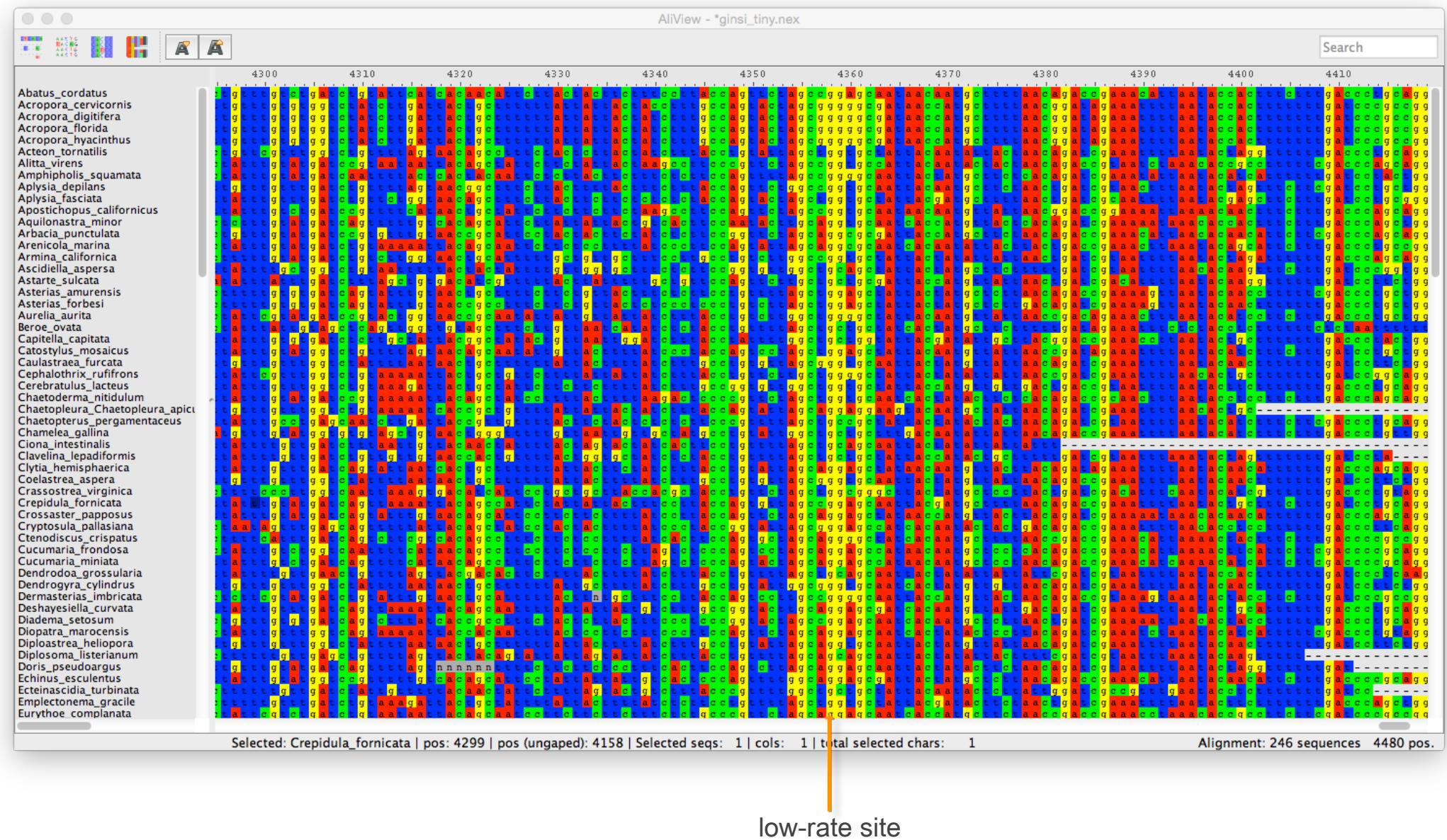
# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



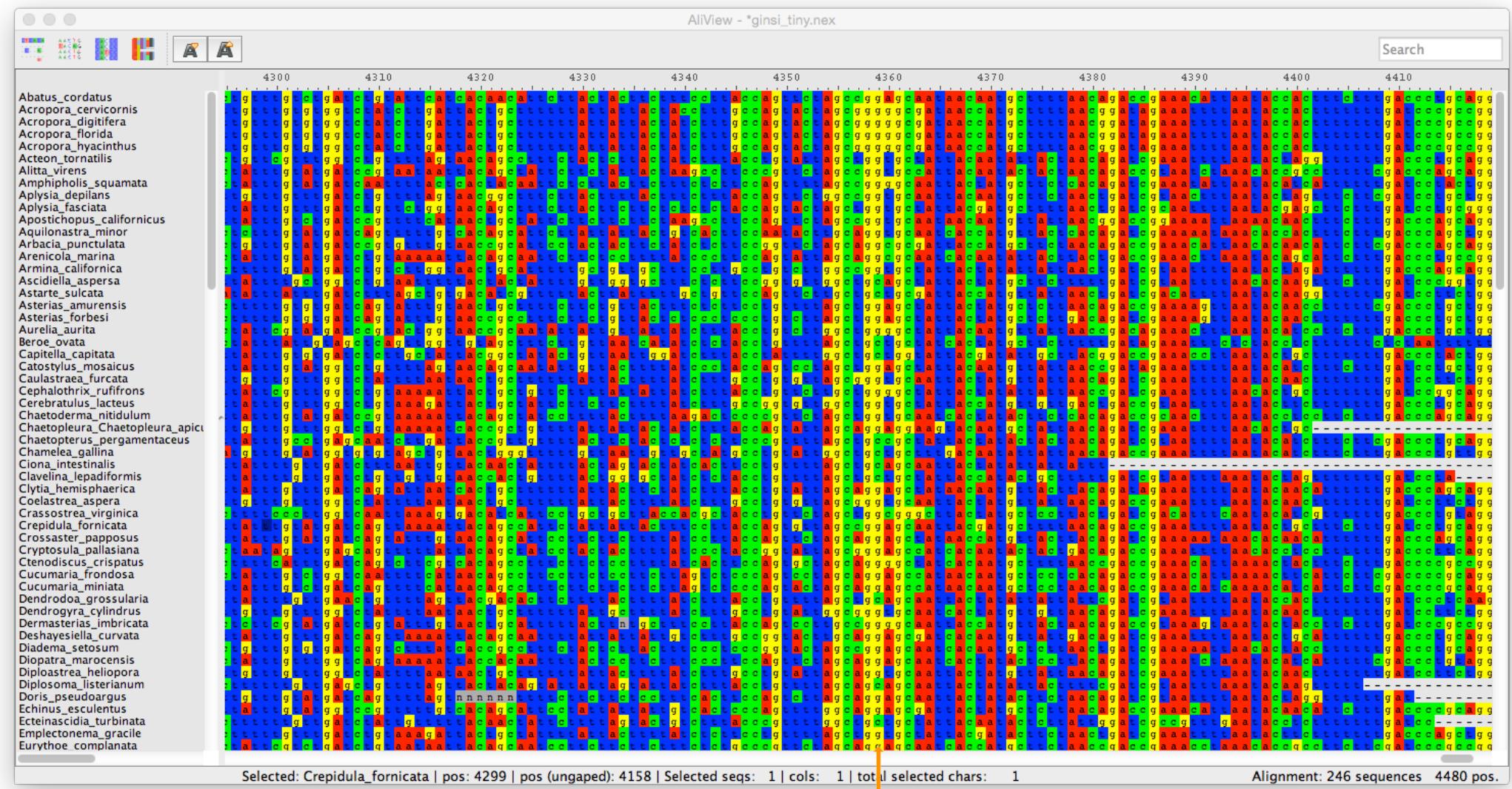
# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



### low-rate site

# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



high-rate site

# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



The substitution rates may differ between sites at different codon positions of protein-coding genes

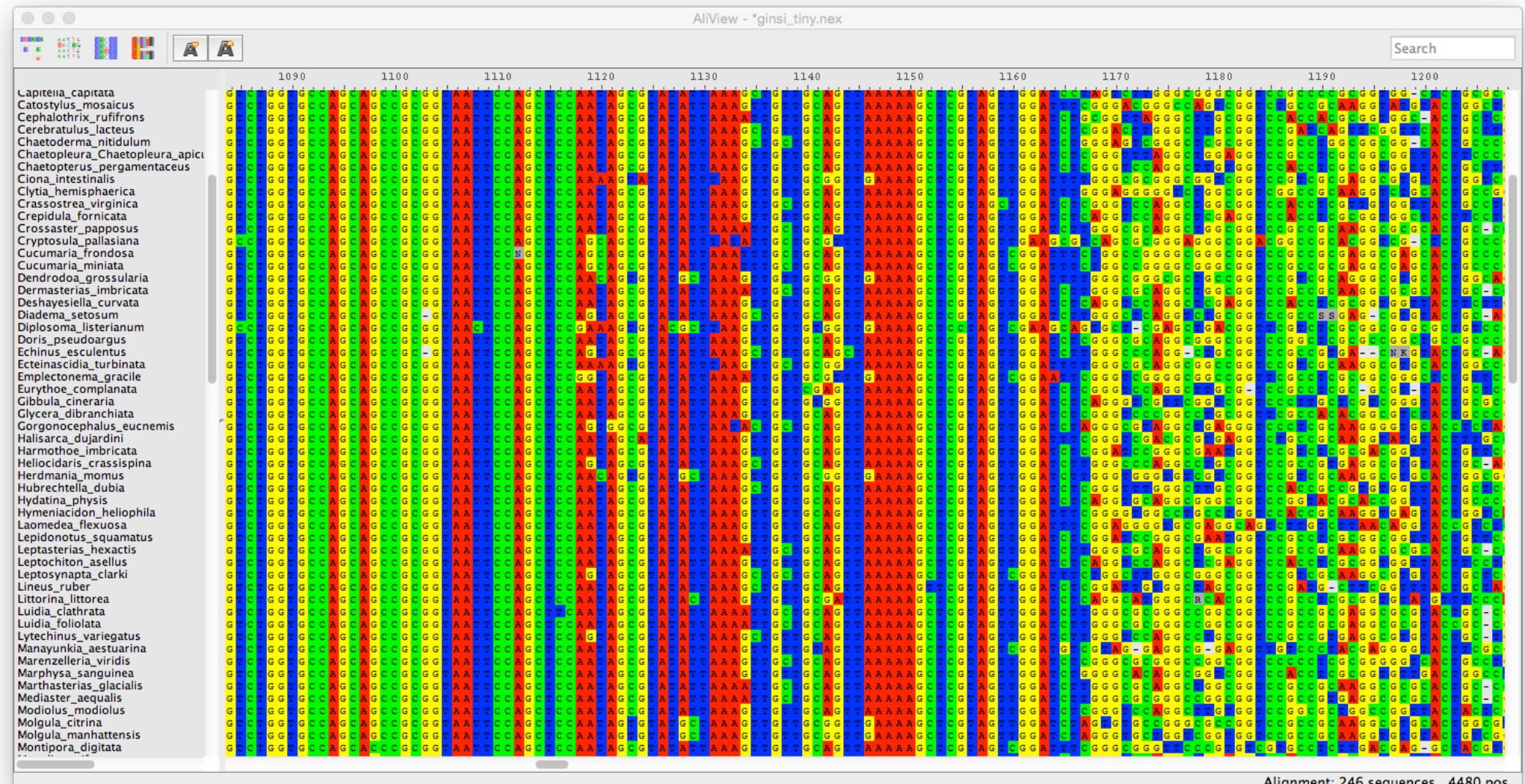
# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites

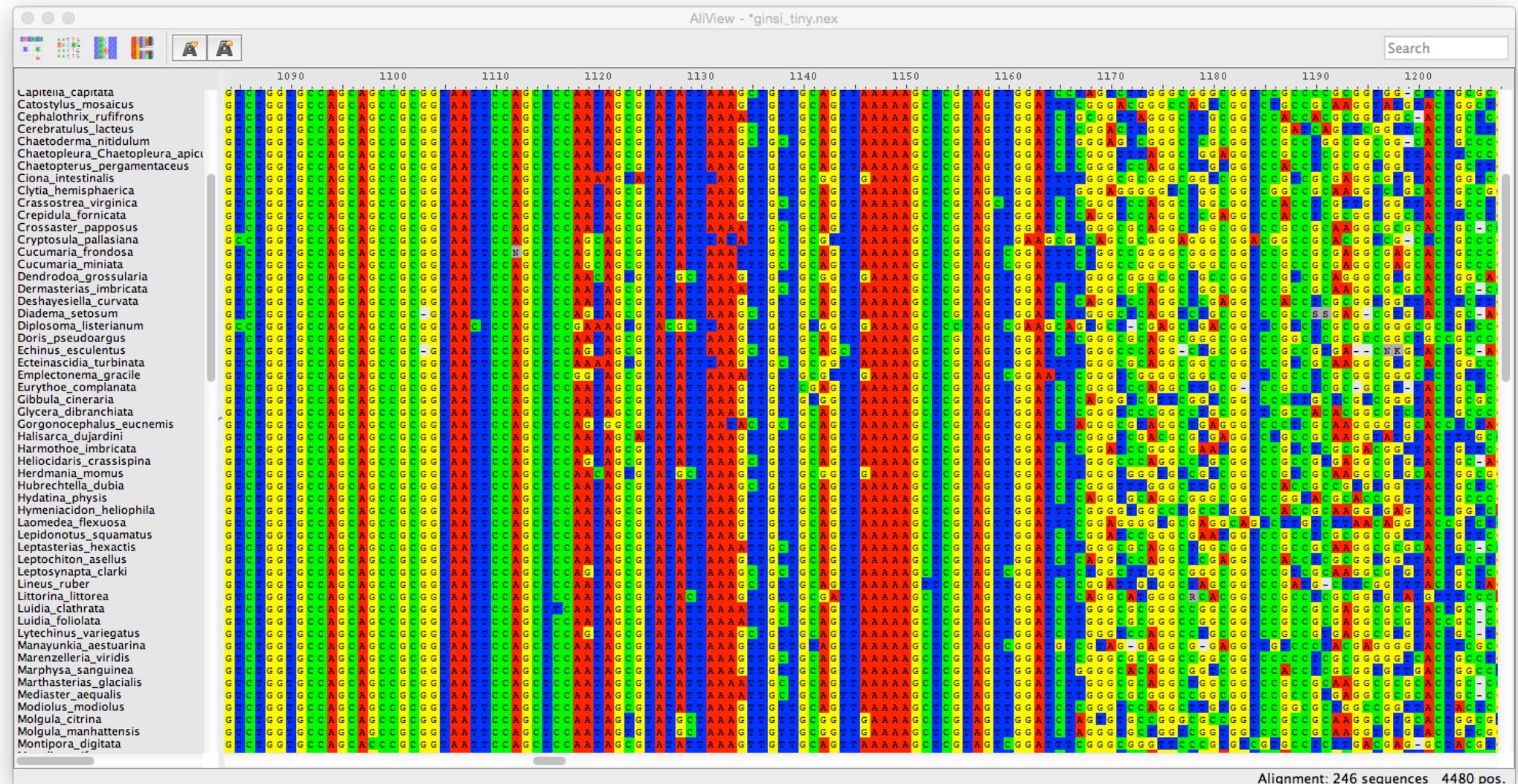


low-rate region

high-rate region

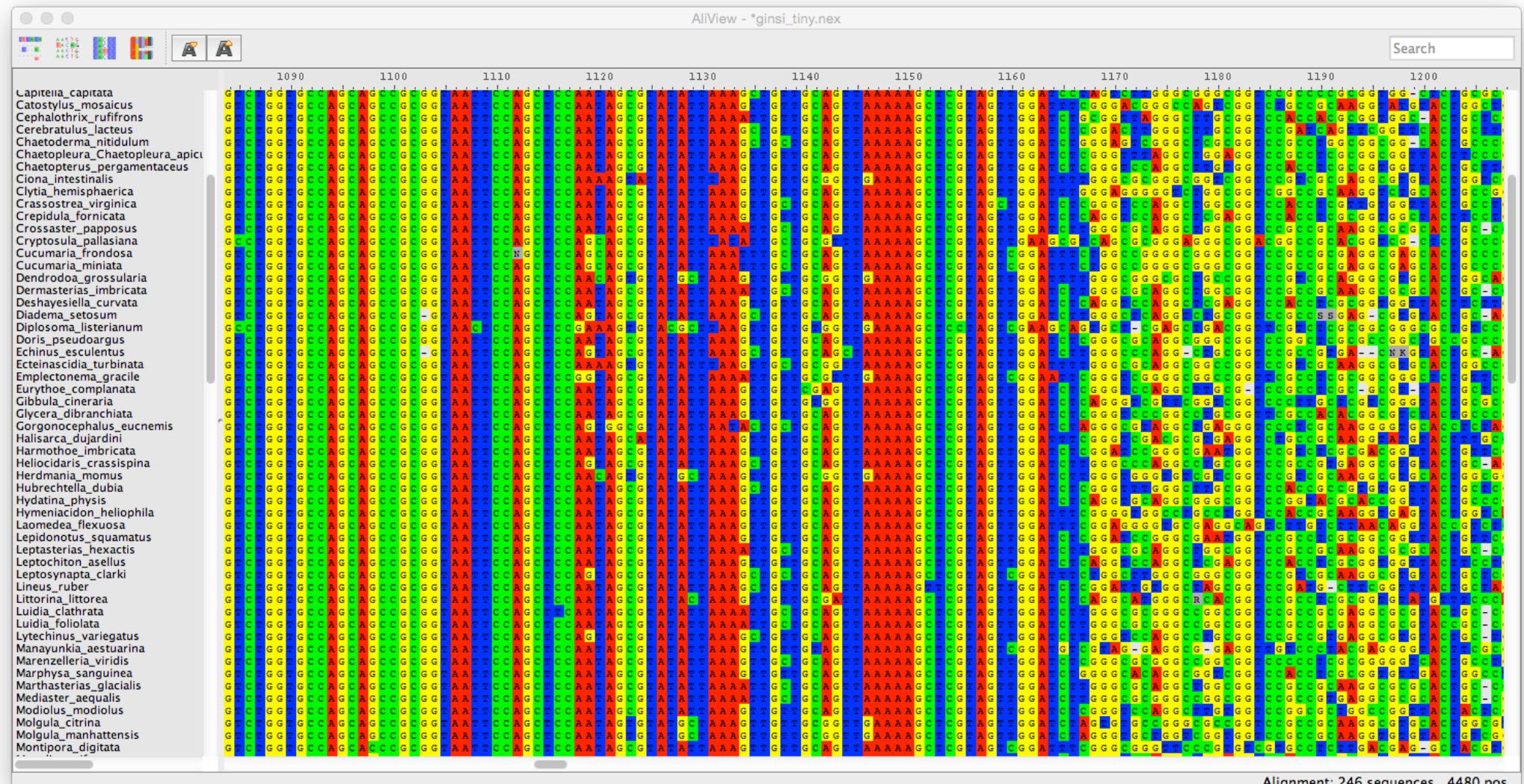
# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites



The substitution rates may differ between stem and loop regions of ribosomal gene regions

# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites

Variation in substitution rates across sites is a prevalent feature of empirical datasets

# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites

Variation in substitution rates across sites is a prevalent feature of empirical datasets

Substitution rates may vary across because of:

- systematic bias in mutation rates
- biased Gene conversion
- differences in functional constraints

# Accommodating Among Site Rate Variation

Standard models assume that the substitution rate is constant across sites

Variation in substitution rates across sites is a prevalent feature of empirical datasets

Substitution rates may vary across because of:

- systematic bias in mutation rates
- biased Gene conversion
- differences in functional constraints

Under simulation, failure to accommodate ASRV can cause biased estimates of:

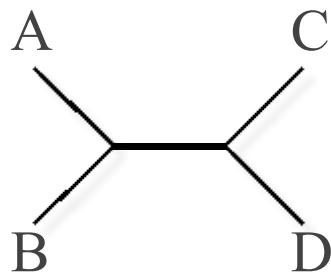
- tree topology
- branch lengths
- other substitution-model parameters

# Accommodating Among Site Rate Variation

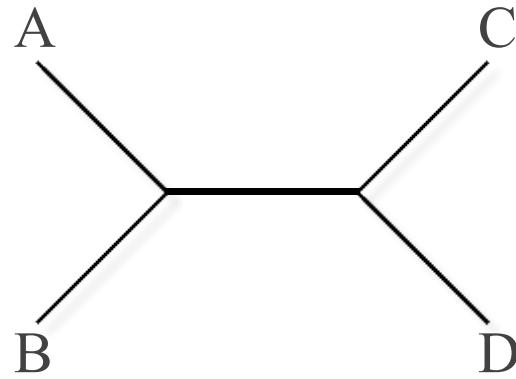
Biology motivates the extension of models

Substitution rates are reflected in the branch lengths of the tree

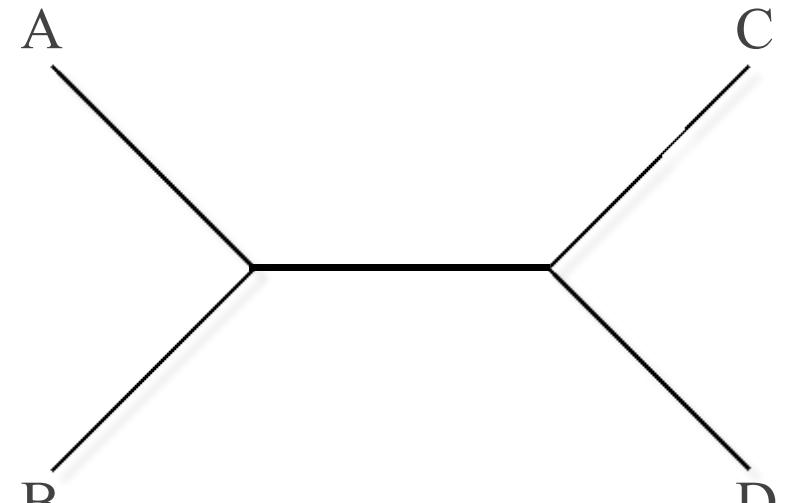
- the relative rate at a site proportionately stretches or compresses the branch lengths



low-rate site



intermediate-rate site



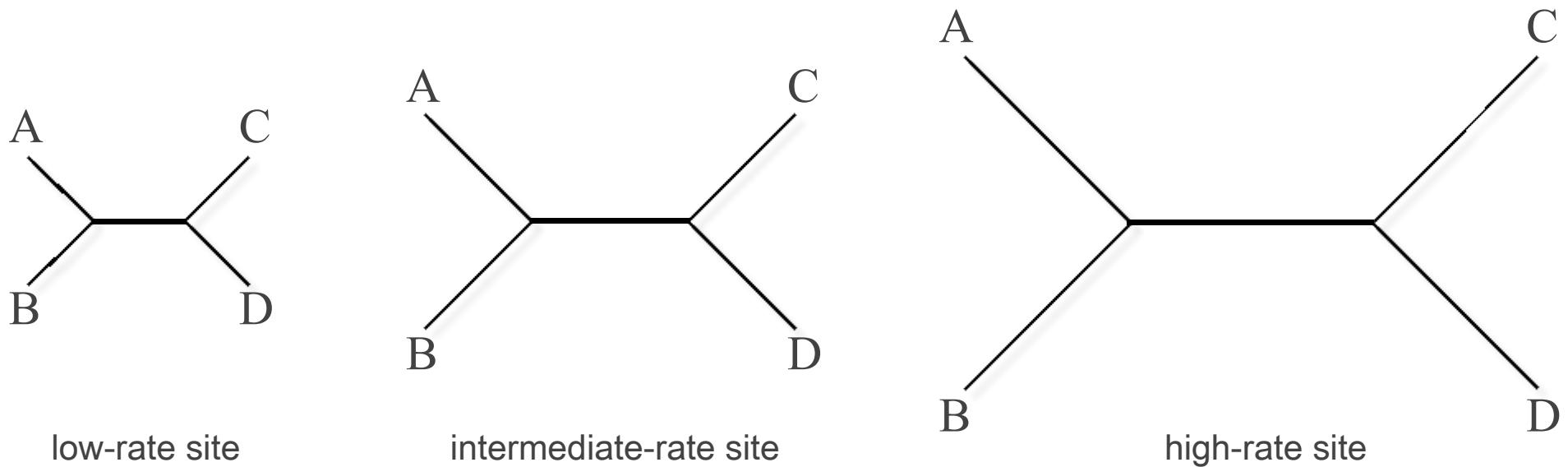
high-rate site

# Accommodating Among Site Rate Variation

Biology motivates the extension of models

Substitution rates are reflected in the branch lengths of the tree

- the relative rate at a site proportionately stretches or compresses the branch lengths



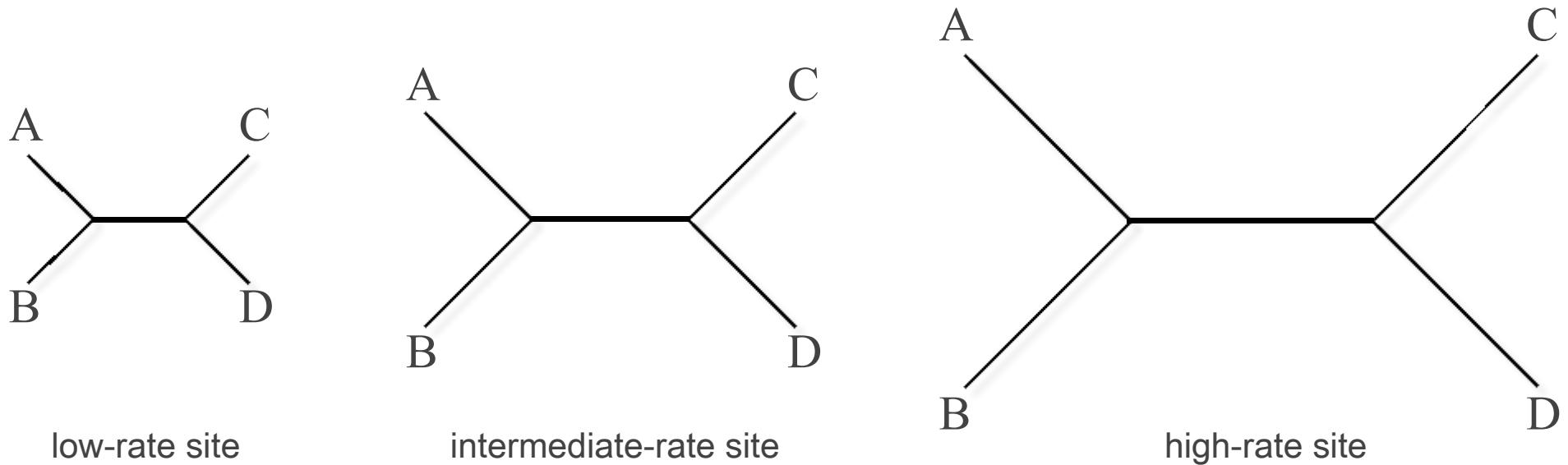
We can accommodate ASRV by incorporating relative substitution-rate multipliers

# Accommodating Among Site Rate Variation

Biology motivates the extension of models

Substitution rates are reflected in the branch lengths of the tree

- the relative rate at a site proportionately stretches or compresses the branch lengths



We can accommodate ASRV by incorporating relative substitution-rate multipliers

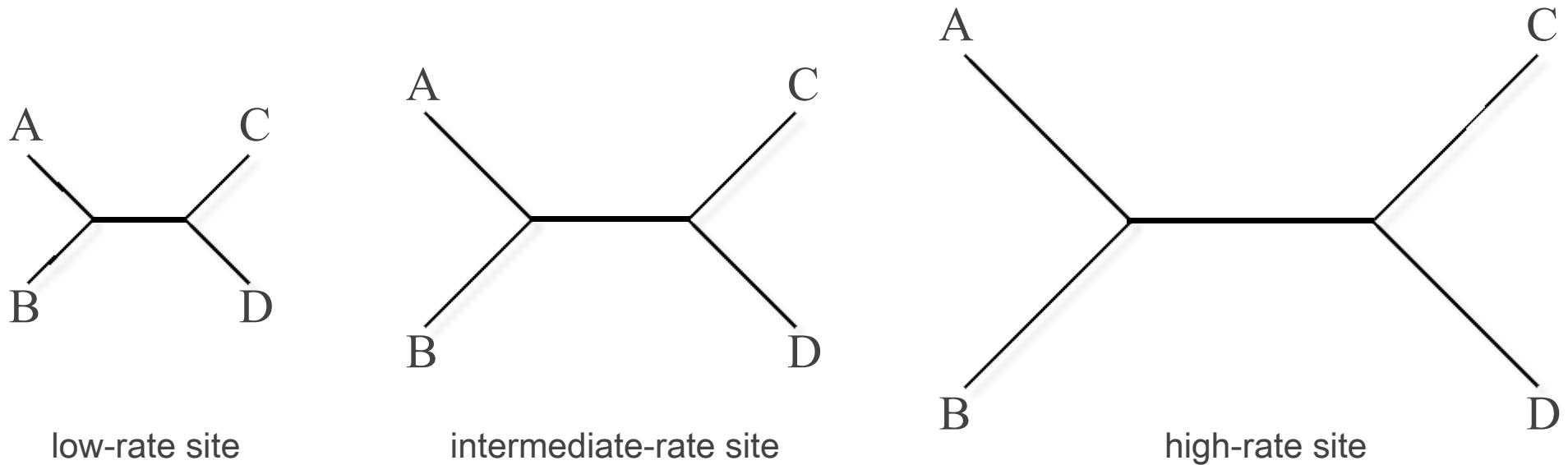
- the relative-rate multipliers are constrained to have a mean rate of one

# Accommodating Among Site Rate Variation

Biology motivates the extension of models

Substitution rates are reflected in the branch lengths of the tree

- the relative rate at a site proportionately stretches or compresses the branch lengths



We can accommodate ASRV by incorporating relative substitution-rate multipliers

- the relative-rate multipliers are constrained to have a mean rate of one
- the relative rate of a site can be assigned deterministically or treated as a random variable

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

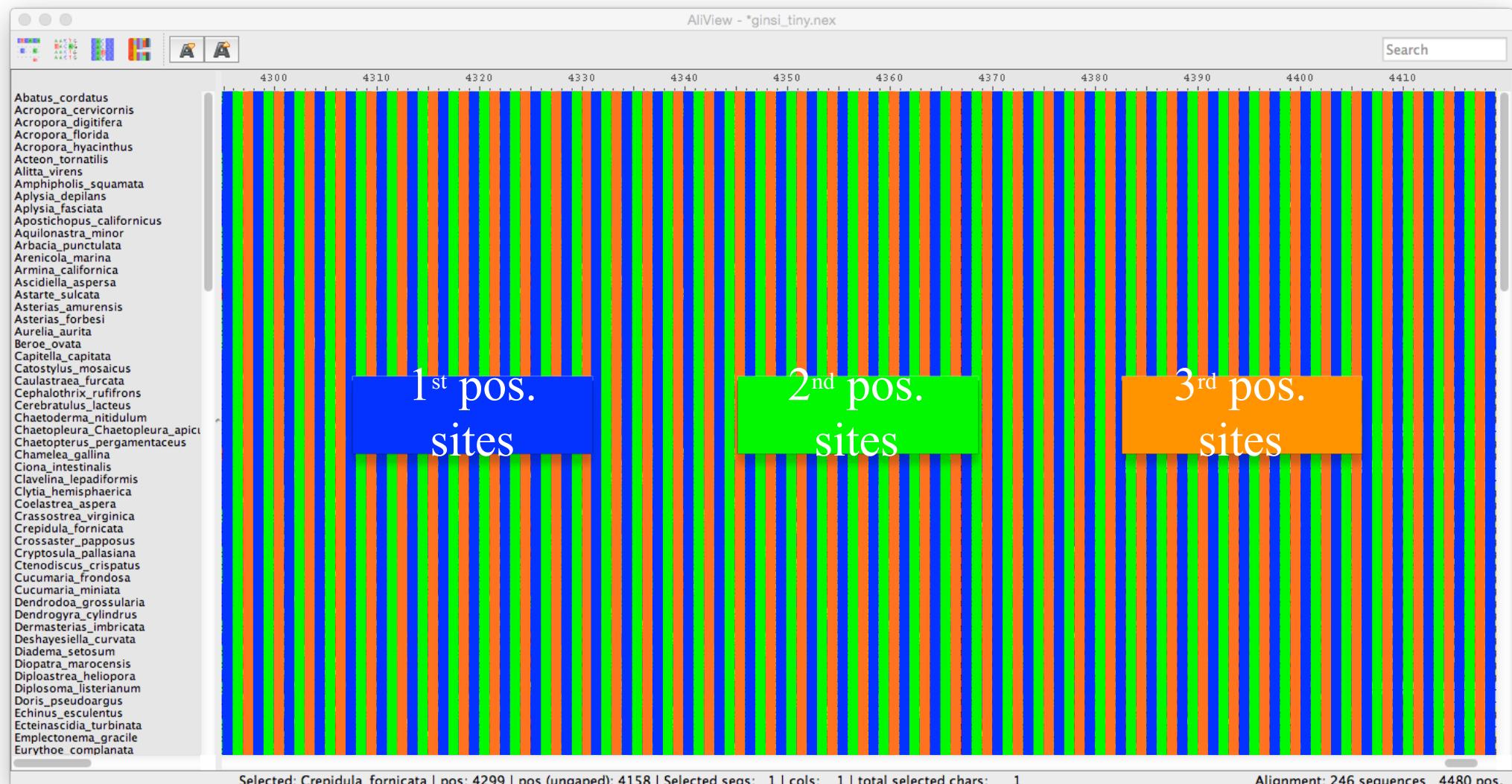
The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV



A site-specific model for protein-coding genes might have three rate categories, with sites in each codon position assigned to each of the three relative-rate categories

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

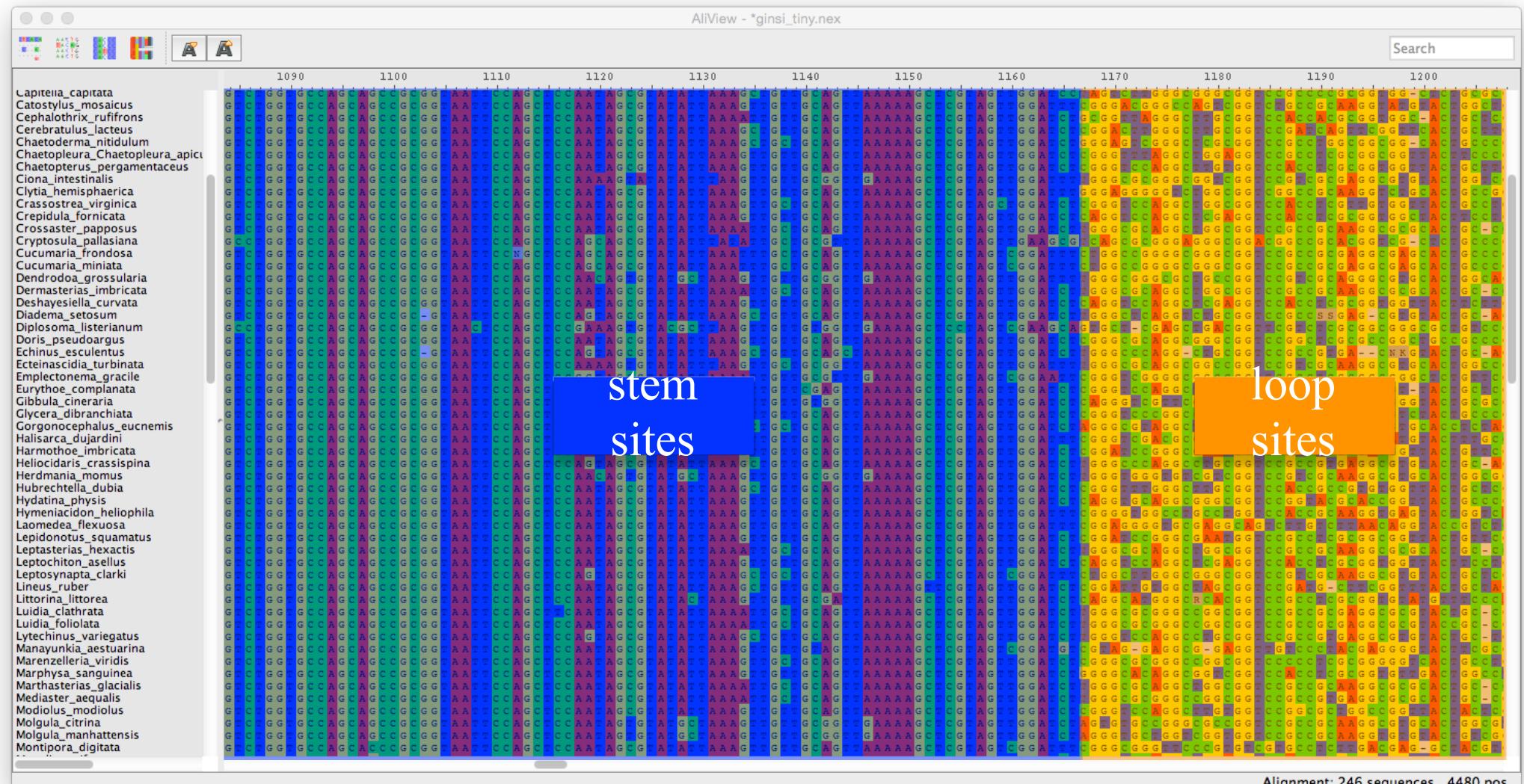
The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)
- two rate categories for ribosomal genes (one each for stem and loop regions)

# Accommodating Among Site Rate Variation

The Site-Specific (+SS) model of ASRV



A site-specific model for ribosomal genes might have two rate categories, with stem sites in one and loop sites in the other relative-rate category

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)
- two rate categories for ribosomal genes (one each for stem and loop regions)

All  $n_i$  sites within a given rate category are subject to the same rate multiplier,  $r_i$

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)
- two rate categories for ribosomal genes (one each for stem and loop regions)

All  $n_i$  sites within a given rate category are subject to the same rate multiplier,  $r_i$

For a model with  $k$  rate categories, we estimate  $(k - 1)$  relative-rate multipliers

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)
- two rate categories for ribosomal genes (one each for stem and loop regions)

All  $n_i$  sites within a given rate category are subject to the same rate multiplier,  $r_i$

For a model with  $k$  rate categories, we estimate  $(k - 1)$  relative-rate multipliers

The relative-rate multipliers are constrained to have a mean rate of one:

$$\frac{(r_1 \times n_1) + (r_2 \times n_2) + (r_3 \times n_3)}{n_1 + n_2 + n_3} = 1$$

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)
- two rate categories for ribosomal genes (one each for stem and loop regions)

All  $n_i$  sites within a given rate category are subject to the same rate multiplier,  $r_i$

For a model with  $k$  rate categories, we estimate  $(k - 1)$  relative-rate multipliers

The relative-rate multipliers are constrained to have a mean rate of one:

rate multiplier applied to sites assigned to category 1

$$\frac{(r_1 \times n_1) + (r_2 \times n_2) + (r_3 \times n_3)}{n_1 + n_2 + n_3} = 1$$

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)
- two rate categories for ribosomal genes (one each for stem and loop regions)

All  $n_i$  sites within a given rate category are subject to the same rate multiplier,  $r_i$

For a model with  $k$  rate categories, we estimate  $(k - 1)$  relative-rate multipliers

The relative-rate multipliers are constrained to have a mean rate of one:

rate multiplier applied to sites assigned to category 1

number of sites assigned to category 1

$$\frac{(r_1 \times n_1) + (r_2 \times n_2) + (r_3 \times n_3)}{n_1 + n_2 + n_3} = 1$$

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)
- two rate categories for ribosomal genes (one each for stem and loop regions)

All  $n_i$  sites within a given rate category are subject to the same rate multiplier,  $r_i$

For a model with  $k$  rate categories, we estimate  $(k - 1)$  relative-rate multipliers

The relative-rate multipliers are constrained to have a mean rate of one:

$$\frac{(r_1 \times n_1) + (r_2 \times n_2) + (r_3 \times n_3)}{n_1 + n_2 + n_3} = 1$$

↑  
rate multiplier applied to sites assigned to category 2  
↑  
number of sites assigned to category 2

# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)
- two rate categories for ribosomal genes (one each for stem and loop regions)

All  $n_i$  sites within a given rate category are subject to the same rate multiplier,  $r_i$

For a model with  $k$  rate categories, we estimate  $(k - 1)$  relative-rate multipliers

The relative-rate multipliers are constrained to have a mean rate of one:

$$\frac{(r_1 \times n_1) + (r_2 \times n_2) + (r_3 \times n_3)}{n_1 + n_2 + n_3} = 1$$

rate multiplier applied to sites assigned to category 3  
number of sites assigned to category 3



# Accommodating Among Site Rate Variation

## The Site-Specific (+SS) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+SS)

The number of rate categories is assumed to be known, e.g.:

- three rate categories for protein-coding genes (one for each codon position)
- two rate categories for ribosomal genes (one each for stem and loop regions)

All  $n_i$  sites within a given rate category are subject to the same rate multiplier,  $r_i$

For a model with  $k$  rate categories, we estimate  $(k - 1)$  relative-rate multipliers

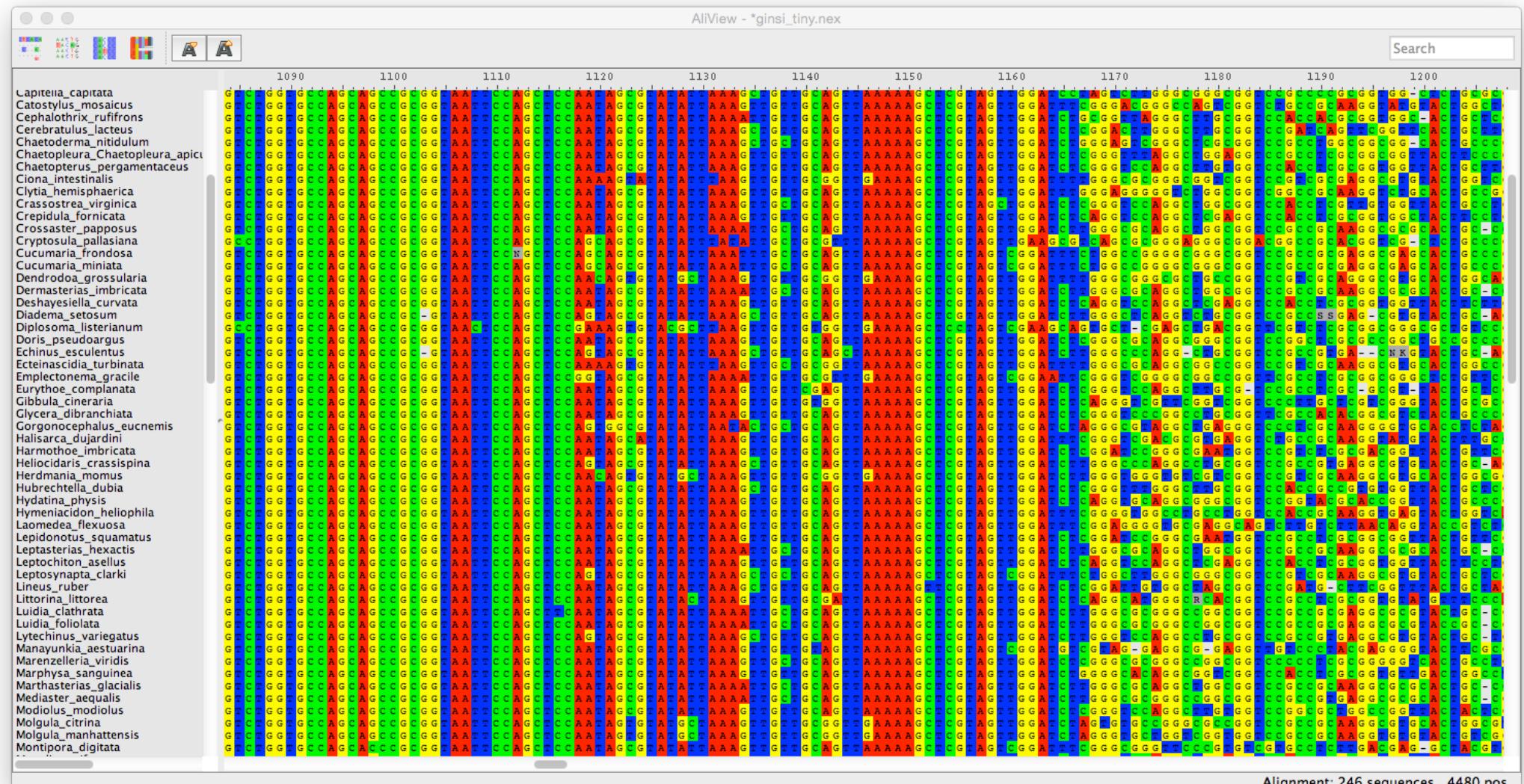
The relative-rate multipliers are constrained to have a mean rate of one:

$$\frac{(r_1 \times n_1) + (r_2 \times n_2) + (r_3 \times n_3)}{n_1 + n_2 + n_3} = 1$$

total number of sites

# Accommodating Among Site Rate Variation

The Proportion of Invariable Sites (+I) model of ASRV



The substitution process may be 'off' at some sites (but note that an invariant site  $\neq$  invariable site)

# Accommodating Among Site Rate Variation

## The Proportion of Invariable Sites (+I) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+I)

# Accommodating Among Site Rate Variation

## The Proportion of Invariable Sites (+I) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+I)

The rate of a site is assumed to be a random variable drawn from a discrete distribution:

- the discrete probability distribution has two states; the process is either 'off' or 'on'

# Accommodating Among Site Rate Variation

## The Proportion of Invariable Sites (+I) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+I)

The rate of a site is assumed to be a random variable drawn from a discrete distribution:

- the discrete probability distribution has two states; the process is either 'off' or 'on'
- the process is 'off' with probability  $p$

# Accommodating Among Site Rate Variation

## The Proportion of Invariable Sites (+I) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+I)

The rate of a site is assumed to be a random variable drawn from a discrete distribution:

- the discrete probability distribution has two states; the process is either 'off' or 'on'
- the process is 'off' with probability  $p$
- when the process is 'off', the relative-rate multiplier  $r_{off} = 0$

# Accommodating Among Site Rate Variation

## The Proportion of Invariable Sites (+I) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+I)

The rate of a site is assumed to be a random variable drawn from a discrete distribution:

- the discrete probability distribution has two states; the process is either 'off' or 'on'
- the process is 'off' with probability  $p$
- when the process is 'off', the relative-rate multiplier  $r_{off} = 0$
- the process is 'on' with probability  $(1 - p)$

# Accommodating Among Site Rate Variation

## The Proportion of Invariable Sites (+I) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+I)

The rate of a site is assumed to be a random variable drawn from a discrete distribution:

- the discrete probability distribution has two states; the process is either 'off' or 'on'
- the process is 'off' with probability  $p$
- when the process is 'off', the relative-rate multiplier  $r_{off} = 0$
- the process is 'on' with probability  $(1 - p)$
- when the process is 'on', the relative-rate multiplier  $r_{on} = 1/(1 - p)$

# Accommodating Among Site Rate Variation

## The Proportion of Invariable Sites (+I) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+I)

The rate of a site is assumed to be a random variable drawn from a discrete distribution:

- the discrete probability distribution has two states; the process is either 'off' or 'on'
- the process is 'off' with probability  $p$
- when the process is 'off', the relative-rate multiplier  $r_{\text{off}} = 0$
- the process is 'on' with probability  $(1 - p)$
- when the process is 'on', the relative-rate multiplier  $r_{\text{on}} = 1/(1 - p)$

The likelihood of site  $i$  is then integrated over the two possible rate multipliers:

$$L(x_i \mid \theta, p) = \underbrace{p \times L(x_i \mid \theta, r_{\text{off}})}_{\text{likelihood when process is 'off'}} + \underbrace{(1 - p)L(x_i \mid \theta, r_{\text{on}})}_{\text{likelihood when process is 'on'}}$$

Therefore, the pruning algorithm is iterated twice for each site:

# Accommodating Among Site Rate Variation

## The Proportion of Invariable Sites (+I) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+I)

The rate of a site is assumed to be a random variable drawn from a discrete distribution:

- the discrete probability distribution has two states; the process is either 'off' or 'on'
- the process is 'off' with probability  $p$
- when the process is 'off', the relative-rate multiplier  $r_{\text{off}} = 0$
- the process is 'on' with probability  $(1 - p)$
- when the process is 'on', the relative-rate multiplier  $r_{\text{on}} = 1/(1 - p)$

The likelihood of site  $i$  is then integrated over the two possible rate multipliers:

$$L(x_i \mid \theta, p) = \underbrace{p \times L(x_i \mid \theta, r_{\text{off}})}_{\text{likelihood when process is 'off'}} + \underbrace{(1 - p)L(x_i \mid \theta, r_{\text{on}})}_{\text{likelihood when process is 'on'}}$$

Therefore, the pruning algorithm is iterated twice for each site:

- once under the condition where the process is 'off', and the relative rate is  $r_{\text{off}} = 0$

# Accommodating Among Site Rate Variation

## The Proportion of Invariable Sites (+I) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+I)

The rate of a site is assumed to be a random variable drawn from a discrete distribution:

- the discrete probability distribution has two states; the process is either 'off' or 'on'
- the process is 'off' with probability  $p$
- when the process is 'off', the relative-rate multiplier  $r_{\text{off}} = 0$
- the process is 'on' with probability  $(1 - p)$
- when the process is 'on', the relative-rate multiplier  $r_{\text{on}} = 1/(1 - p)$

The likelihood of site  $i$  is then integrated over the two possible rate multipliers:

$$L(x_i \mid \theta, p) = \underbrace{p \times L(x_i \mid \theta, r_{\text{off}})}_{\text{likelihood when process is 'off'}} + \underbrace{(1 - p)L(x_i \mid \theta, r_{\text{on}})}_{\text{likelihood when process is 'on'}}$$

Therefore, the pruning algorithm is iterated twice for each site:

- once under the condition where the process is 'off', and the relative rate is  $r_{\text{off}} = 0$
- once under the condition where the process is 'on', and the relative rate is  $r_{\text{off}} = 1/(1 - p)$

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+G)

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+G)

The rate of a site is a random variable drawn from a discrete gamma distribution:

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+G)

The rate of a site is a random variable drawn from a discrete gamma distribution:

- the gamma distribution has two parameters: the shape parameter,  $\alpha$  and the scale parameter,  $\beta$

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+G)

The rate of a site is a random variable drawn from a discrete gamma distribution:

- the gamma distribution has two parameters: the shape parameter,  $\alpha$  and the scale parameter,  $\beta$
- the mean of the gamma distribution is  $\alpha/\beta$

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+G)

The rate of a site is a random variable drawn from a discrete gamma distribution:

- the gamma distribution has two parameters: the shape parameter,  $\alpha$  and the scale parameter,  $\beta$
- the mean of the gamma distribution is  $\alpha/\beta$
- the mean of the gamma distribution is constrained to have a mean of one:  $\alpha = \beta$

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The 'base' substitution model (a member of the GTR family) is extended to accommodate ASRV (e.g., HKY+G)

The rate of a site is a random variable drawn from a discrete gamma distribution:

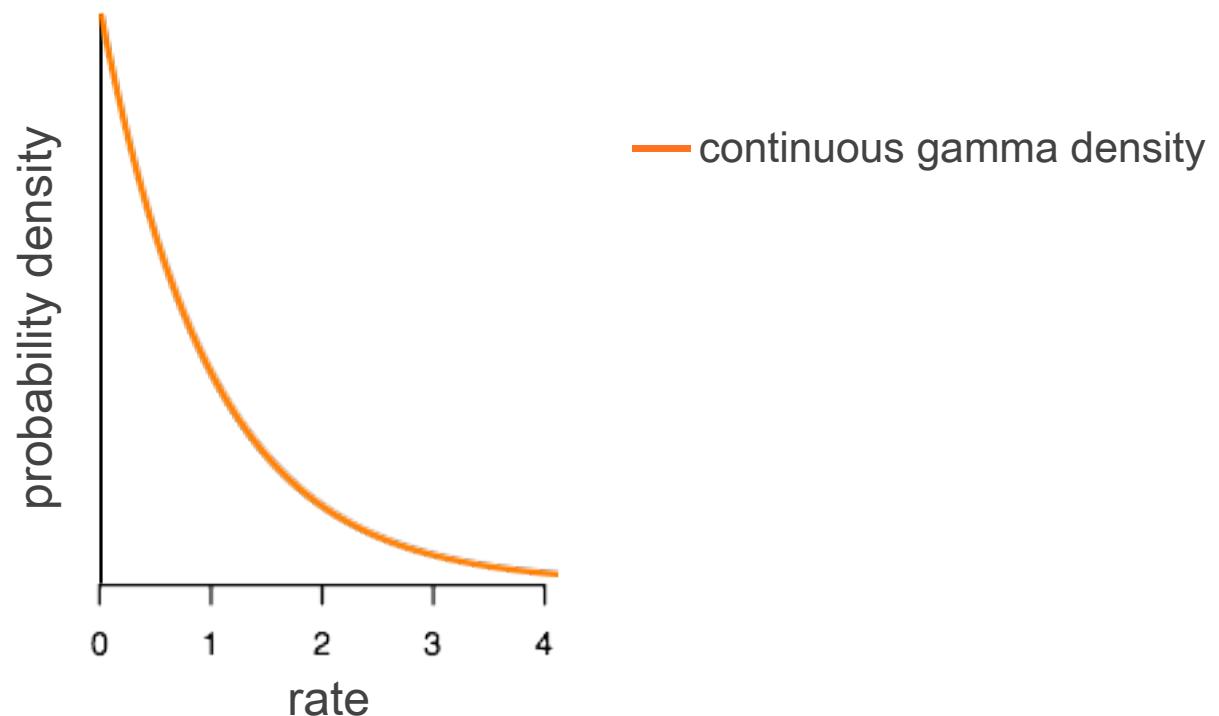
- the gamma distribution has two parameters: the shape parameter,  $\alpha$  and the scale parameter,  $\beta$
- the mean of the gamma distribution is  $\alpha/\beta$
- the mean of the gamma distribution is constrained to have a mean of one:  $\alpha = \beta$
- so, operationally, the gamma distribution involves the single shape parameter,  $\alpha$

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

- the gamma distribution is constrained to have a mean of one, such that  $\alpha = \beta$

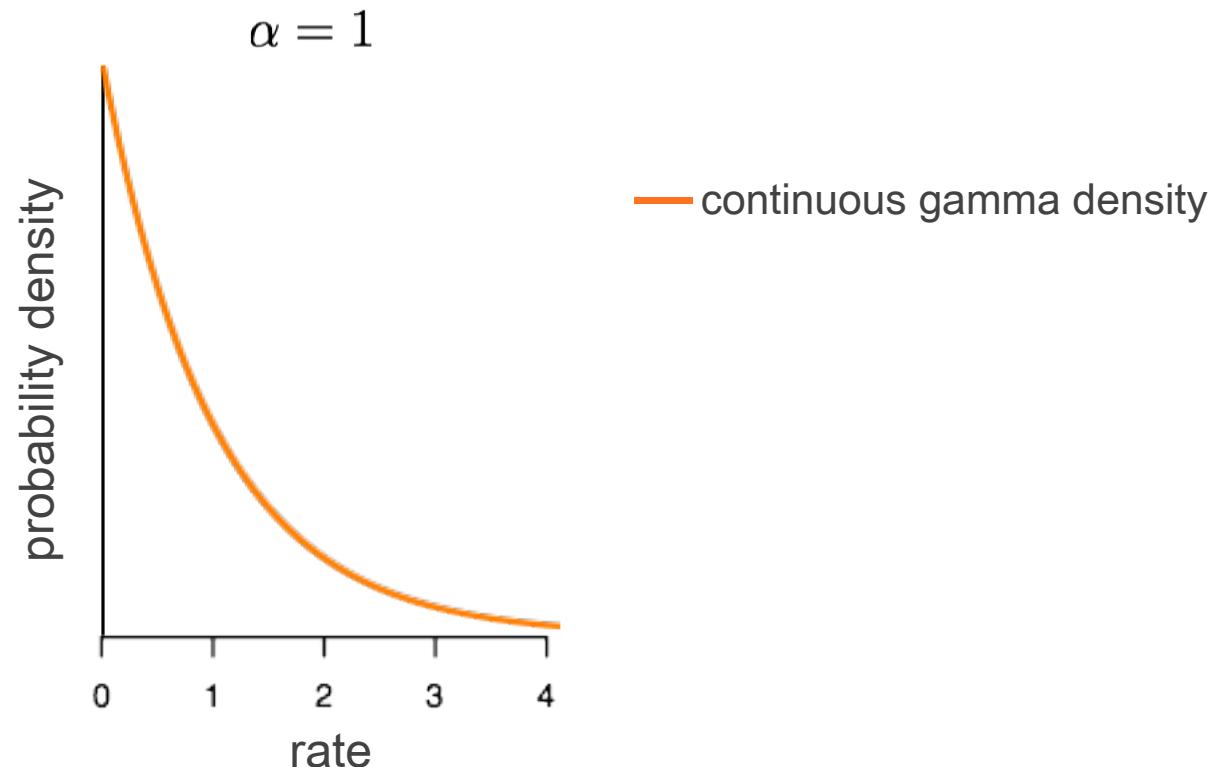


# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

- the gamma distribution is constrained to have a mean of one, such that  $\alpha = \beta$
- the shape of the distribution is defined by  $\alpha$

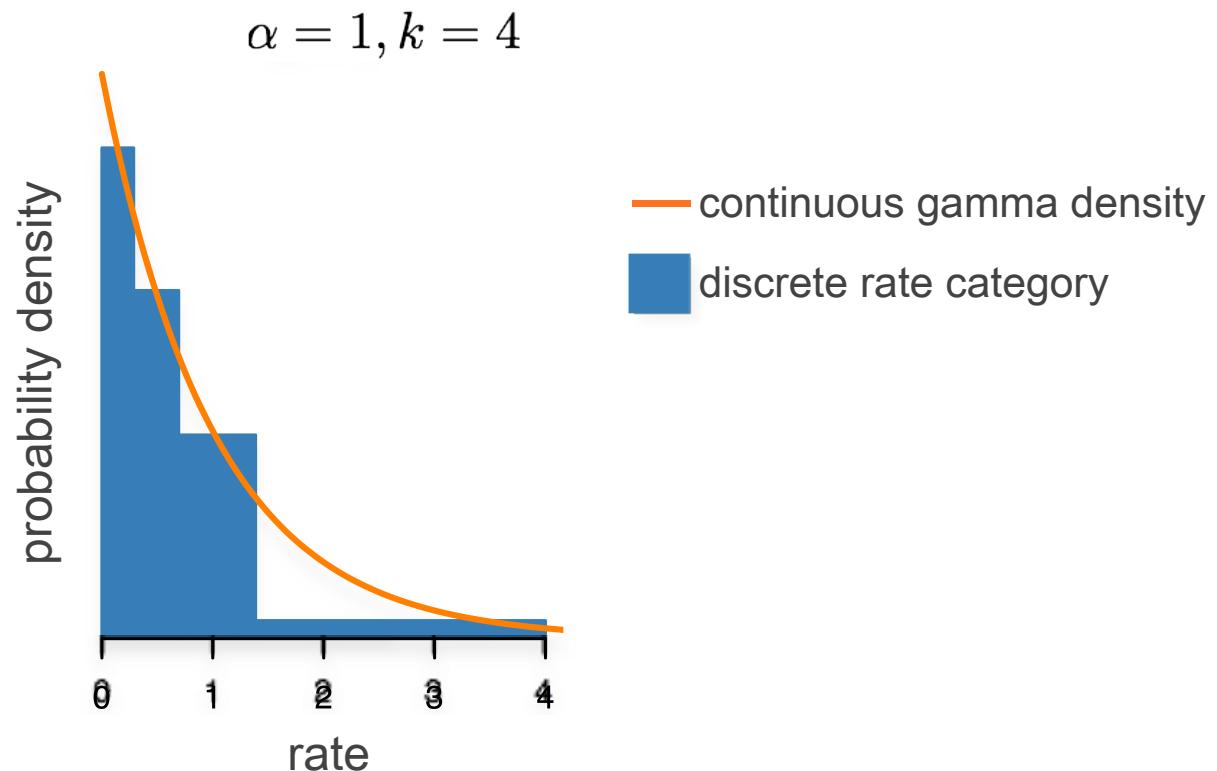


# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins



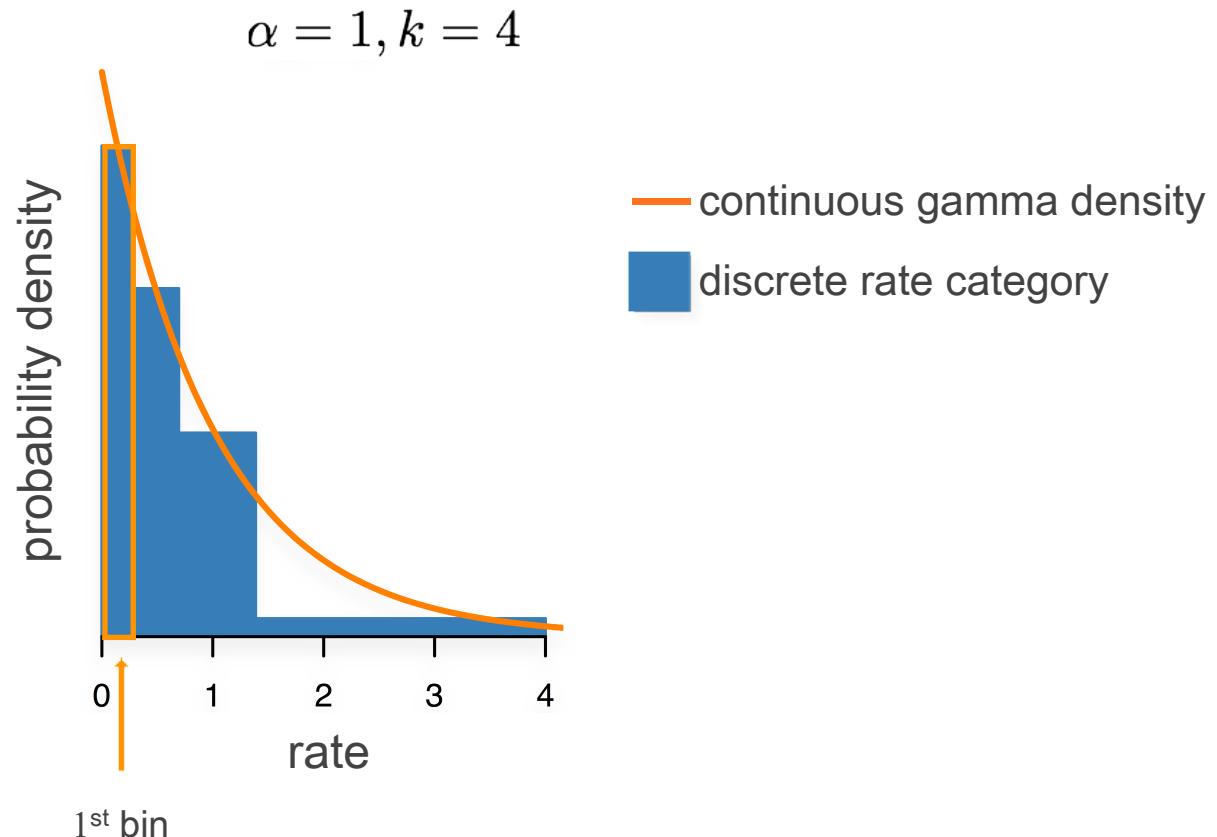
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

- bins are defined such that each encompasses an equal area of the gamma probability density



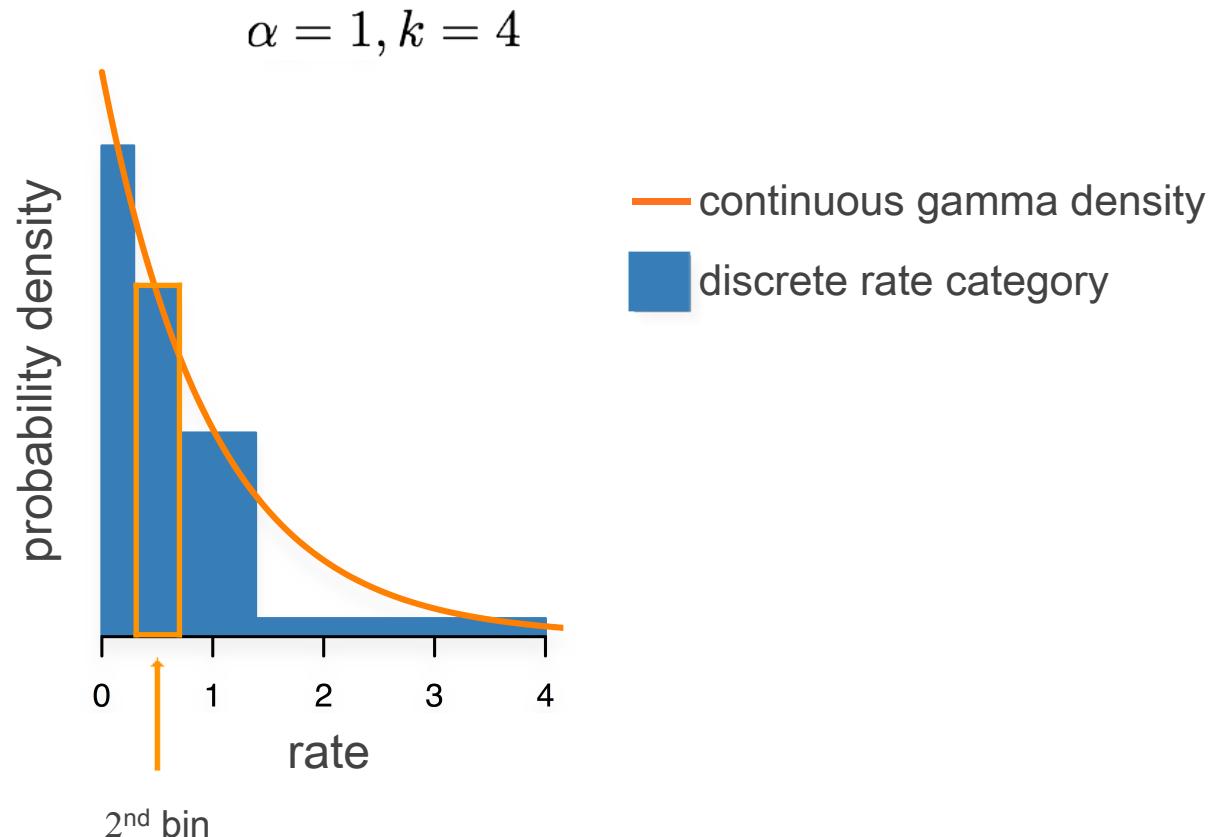
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

- bins are defined such that each encompasses an equal area of the gamma probability density



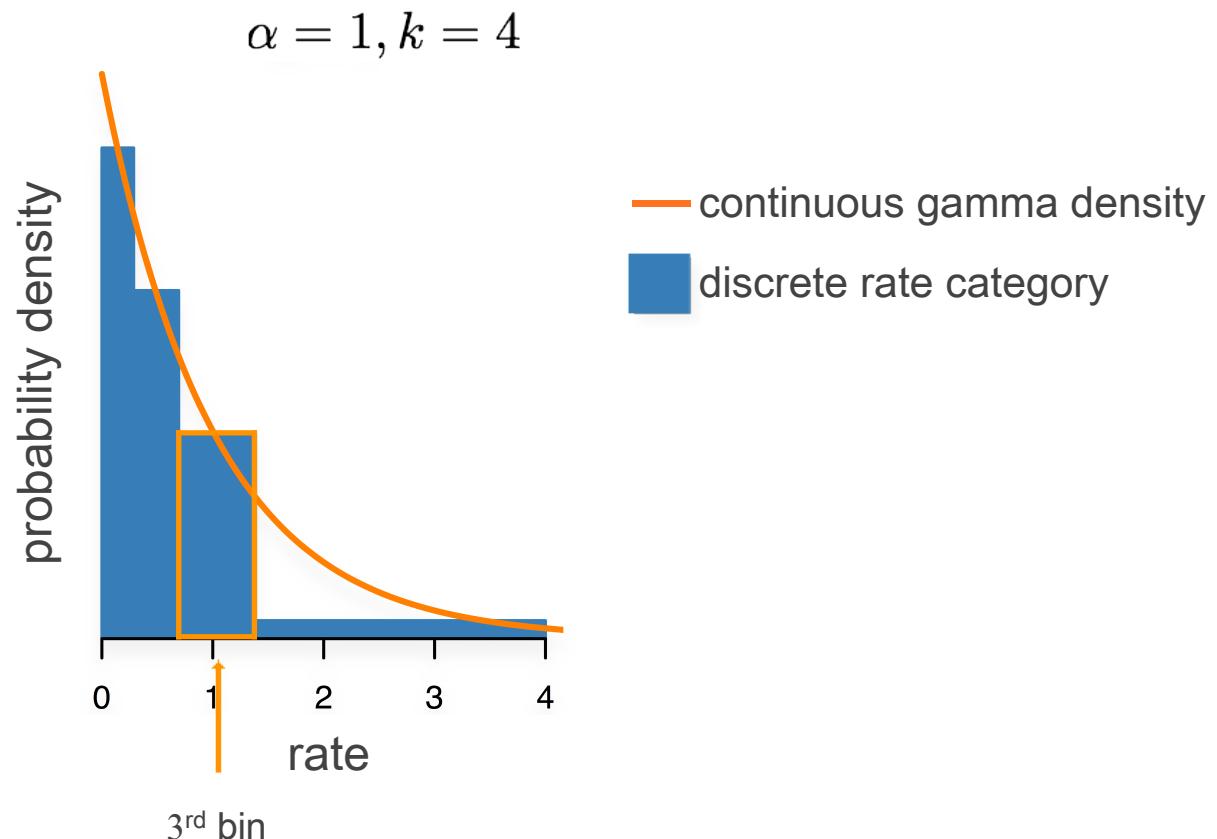
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

- bins are defined such that each encompasses an equal area of the gamma probability density



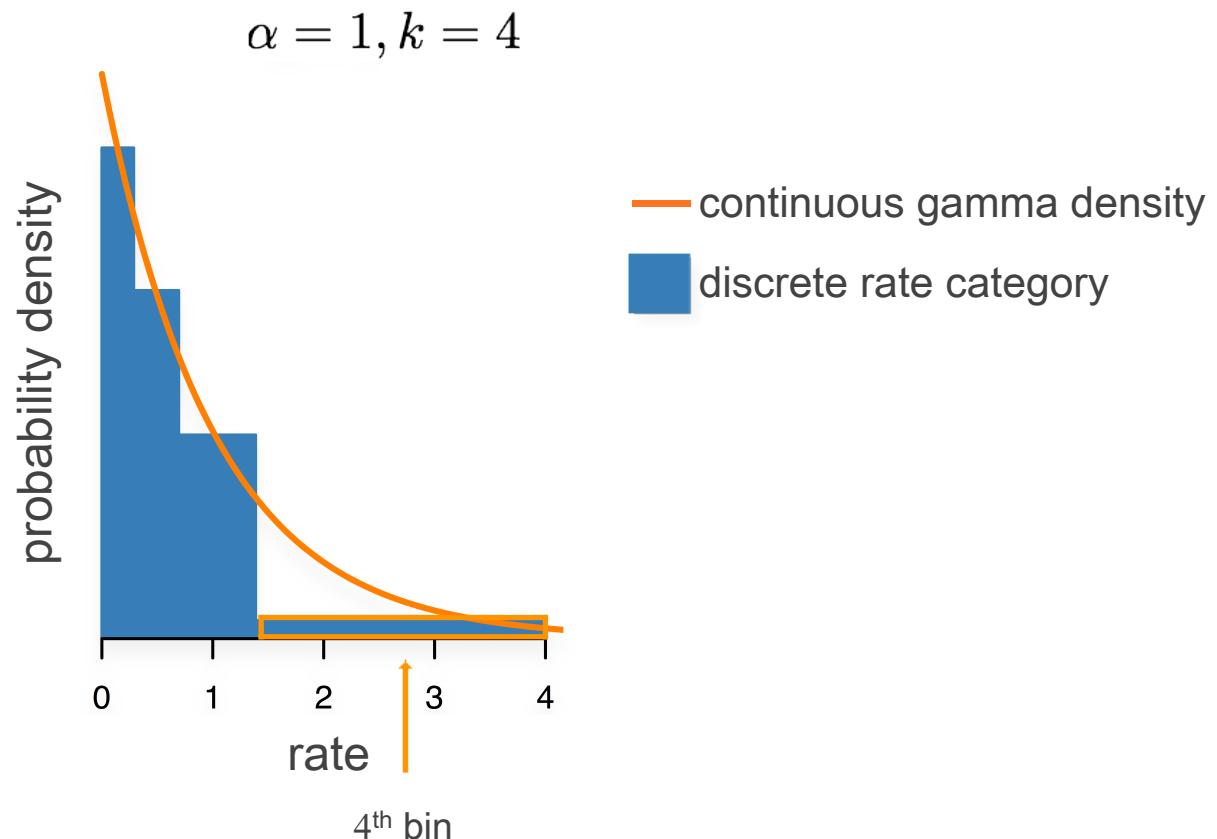
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

- bins are defined such that each encompasses an equal area of the gamma probability density



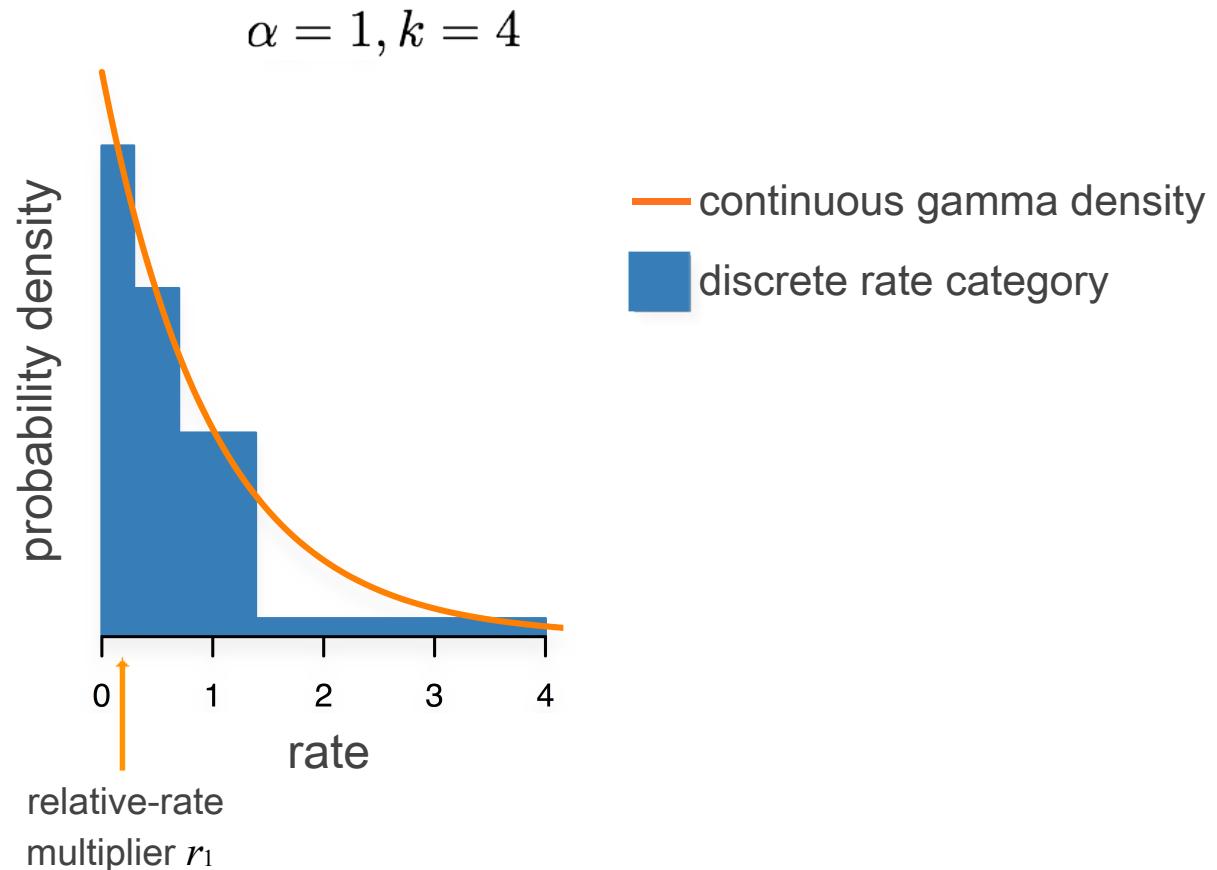
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

- bins are defined such that each encompasses an equal area of the gamma probability density
- there are  $k$  relative-rate multipliers; corresponding to the mean or median rate of each bin



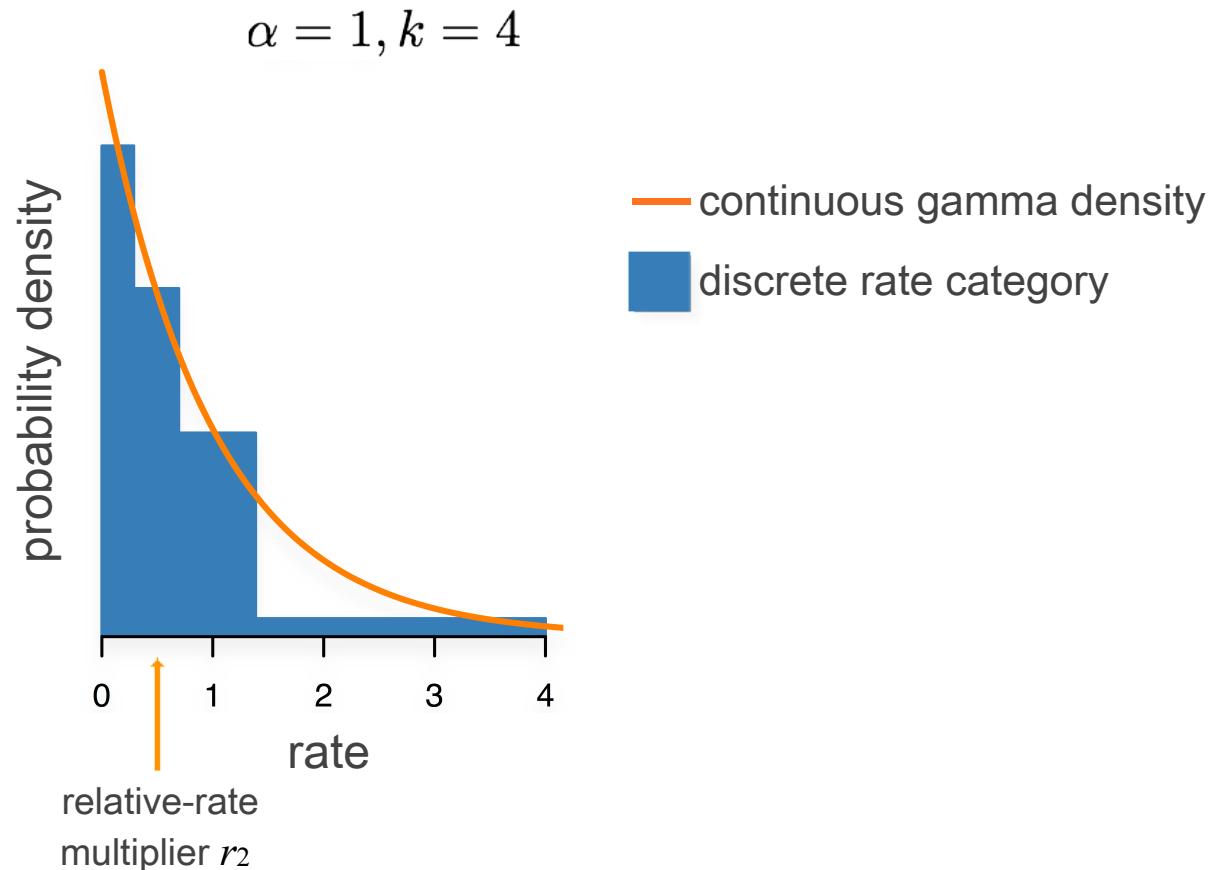
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

- bins are defined such that each encompasses an equal area of the gamma probability density
- there are  $k$  relative-rate multipliers; corresponding to the mean or median rate of each bin



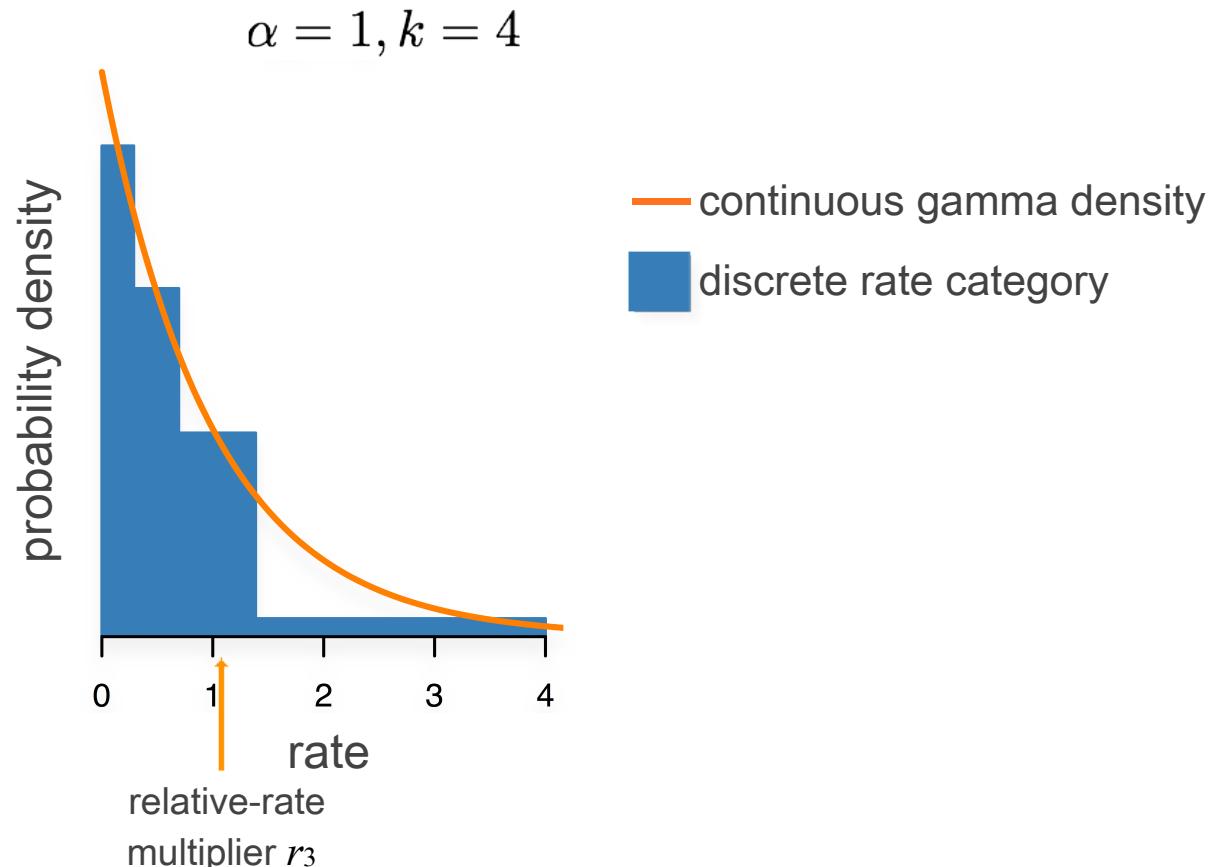
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

- bins are defined such that each encompasses an equal area of the gamma probability density
- there are  $k$  relative-rate multipliers; corresponding to the mean or median rate of each bin



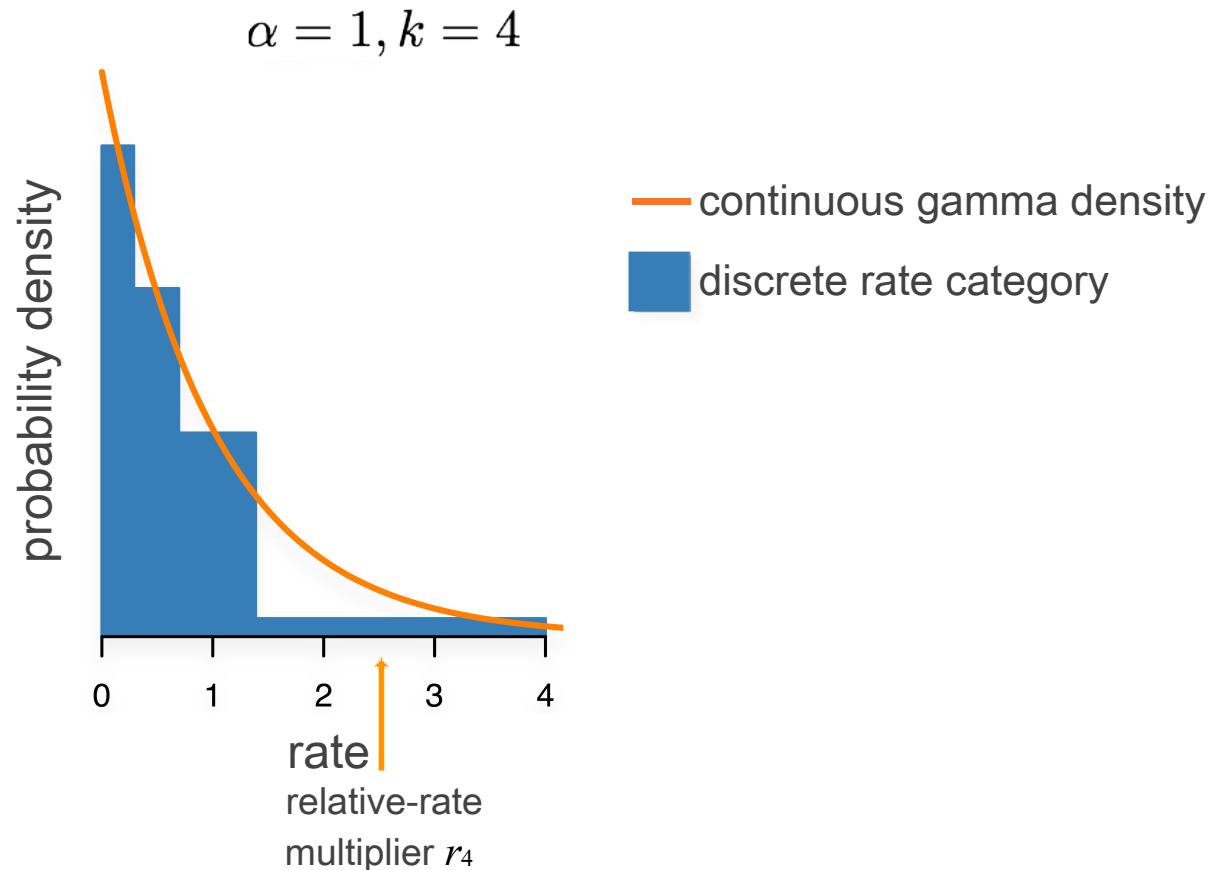
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

- bins are defined such that each encompasses an equal area of the gamma probability density
- there are  $k$  relative-rate multipliers; corresponding to the mean or median rate of each bin



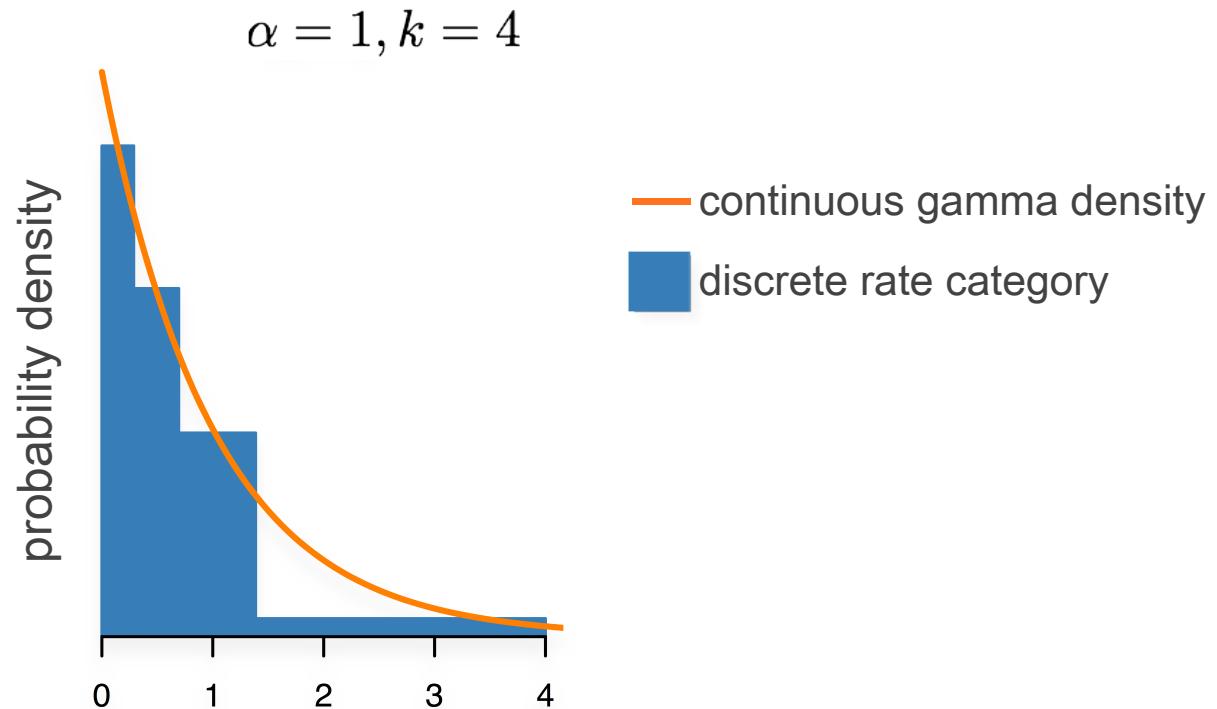
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$



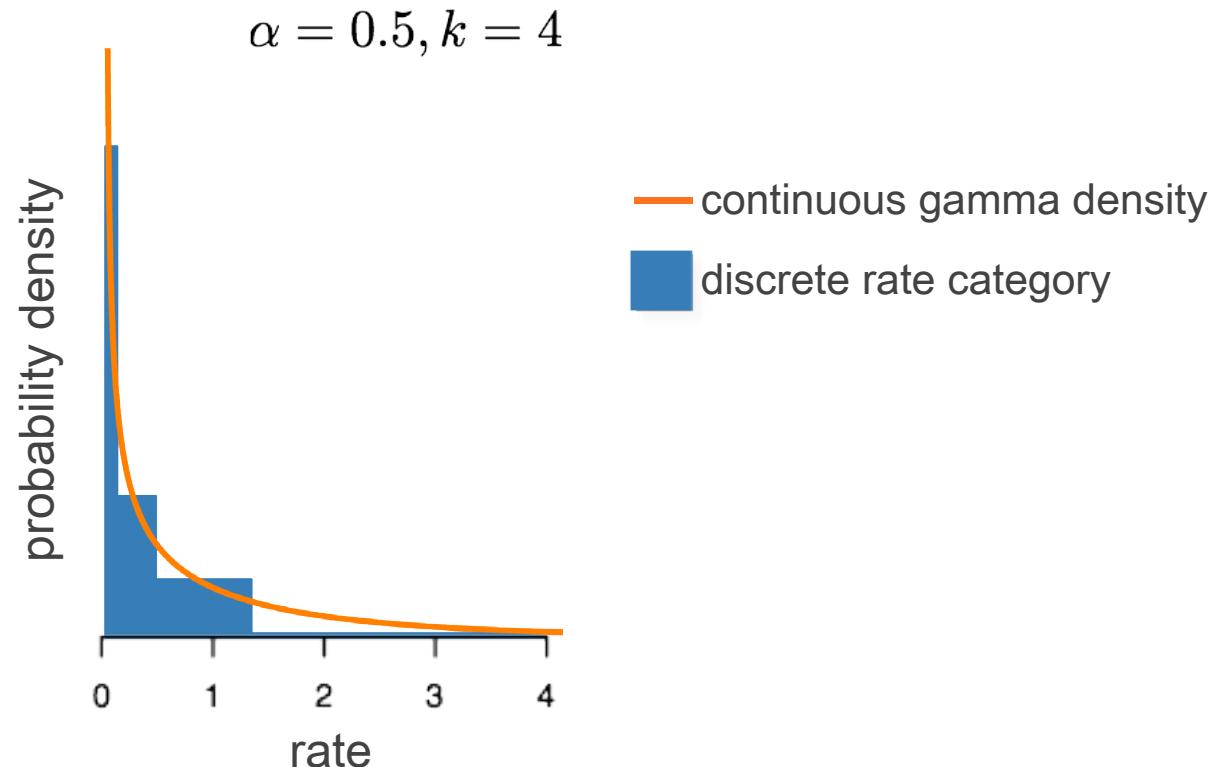
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$



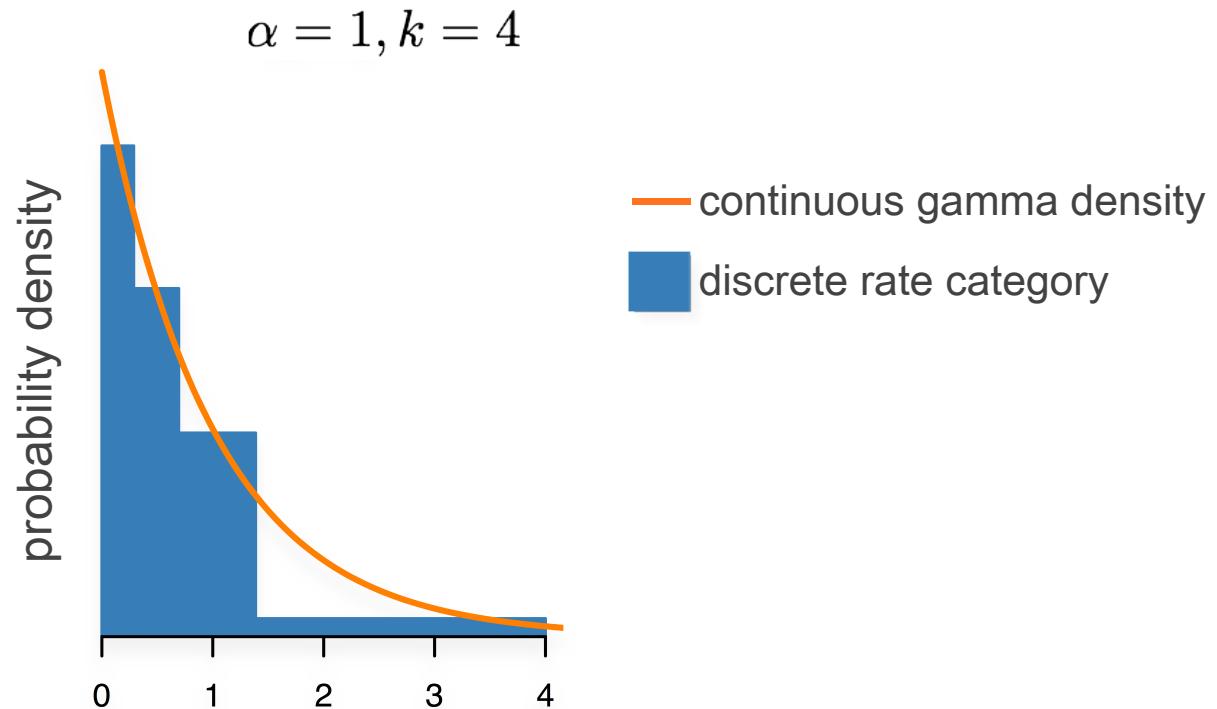
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$



# Accommodating Among Site Rate Variation

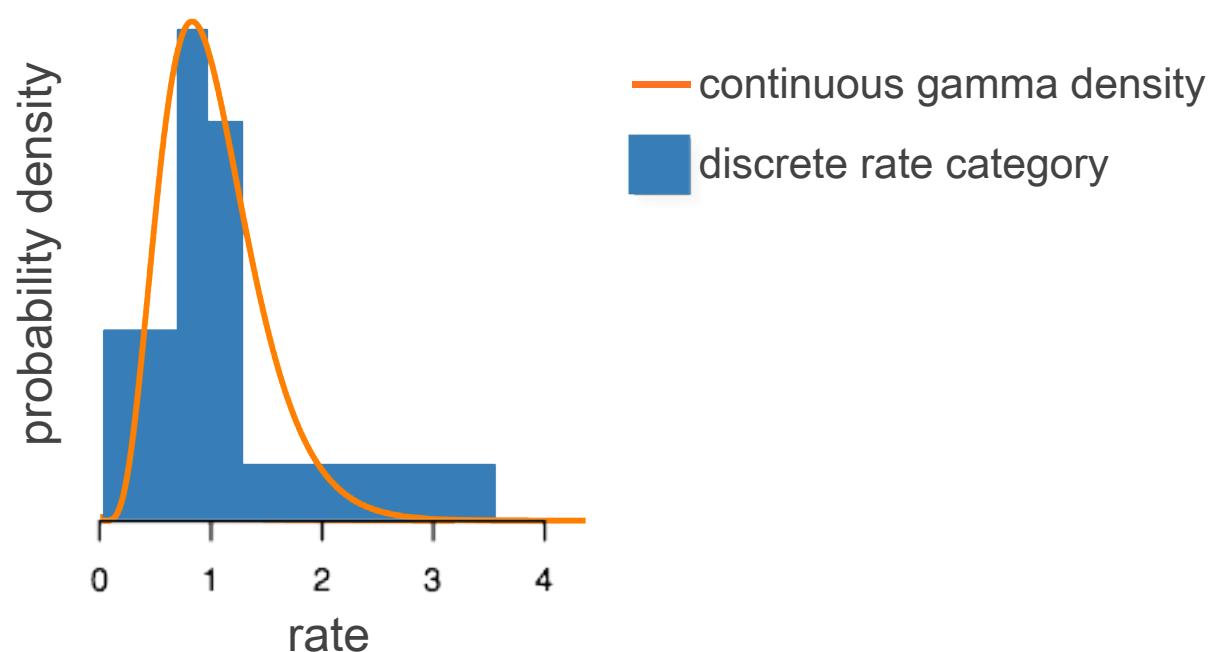
## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$

$$\alpha = 5, k = 4$$



# Accommodating Among Site Rate Variation

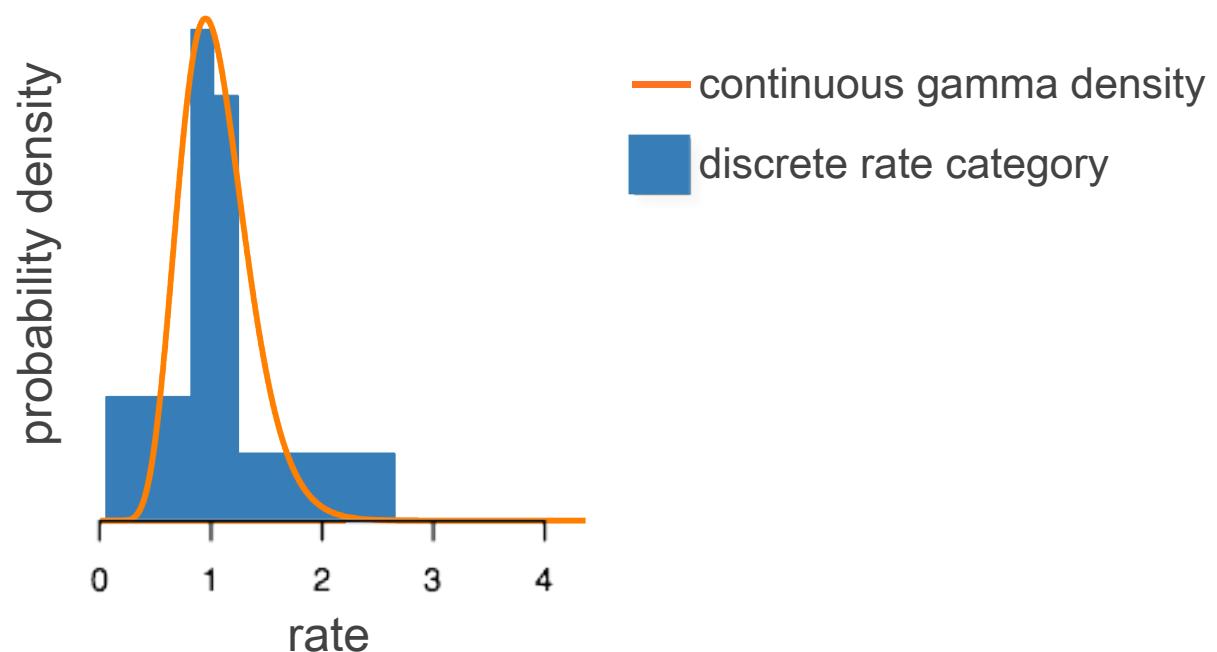
## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$

$$\alpha = 10, k = 4$$



# Accommodating Among Site Rate Variation

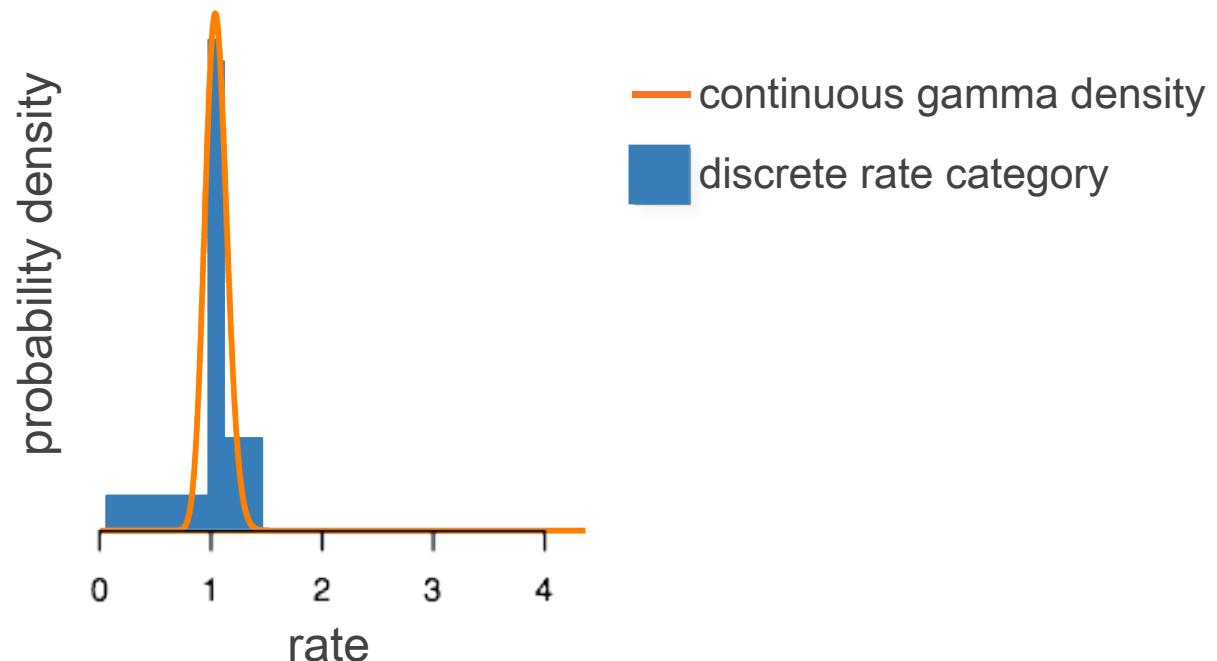
## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$

$$\alpha = 100, k = 4$$



# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$

The likelihood of site  $i$  is then integrated over the  $k$  discrete rate categories:

$$L(x_i \mid \theta, \alpha) = \sum_{j=1}^k f(x_i \mid \theta, r_j)$$

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$

The likelihood of site  $i$  is then integrated over the  $k$  discrete rate categories:

$$L(x_i \mid \theta, \alpha) = \sum_{j=1}^k f(x_i \mid \theta, r_j)$$

Therefore, the pruning algorithm is iterated  $k$  times for each site:

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$

The likelihood of site  $i$  is then integrated over the  $k$  discrete rate categories:

$$L(x_i \mid \theta, \alpha) = \sum_{j=1}^k f(x_i \mid \theta, r_j)$$

Therefore, the pruning algorithm is iterated  $k$  times for each site:

- where the site rate is multiplied by the relative rate,  $r_j$ , for each of the  $k$  bins

# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution provides a flexible means of accommodating ASRV

The gamma distribution is discretized into  $k$  discrete bins

The degree of ASRV is inversely related to the value of the shape parameter,  $\alpha$

The likelihood of site  $i$  is then integrated over the  $k$  discrete rate categories:

$$L(x_i \mid \theta, \alpha) = \sum_{j=1}^k f(x_i \mid \theta, r_j)$$

Therefore, the pruning algorithm is iterated  $k$  times for each site:

- where the site rate is multiplied by the relative rate,  $r_j$ , for each of the  $k$  bins
- the discrete gamma model therefore incurs a  $k$ -fold increase in computational burden

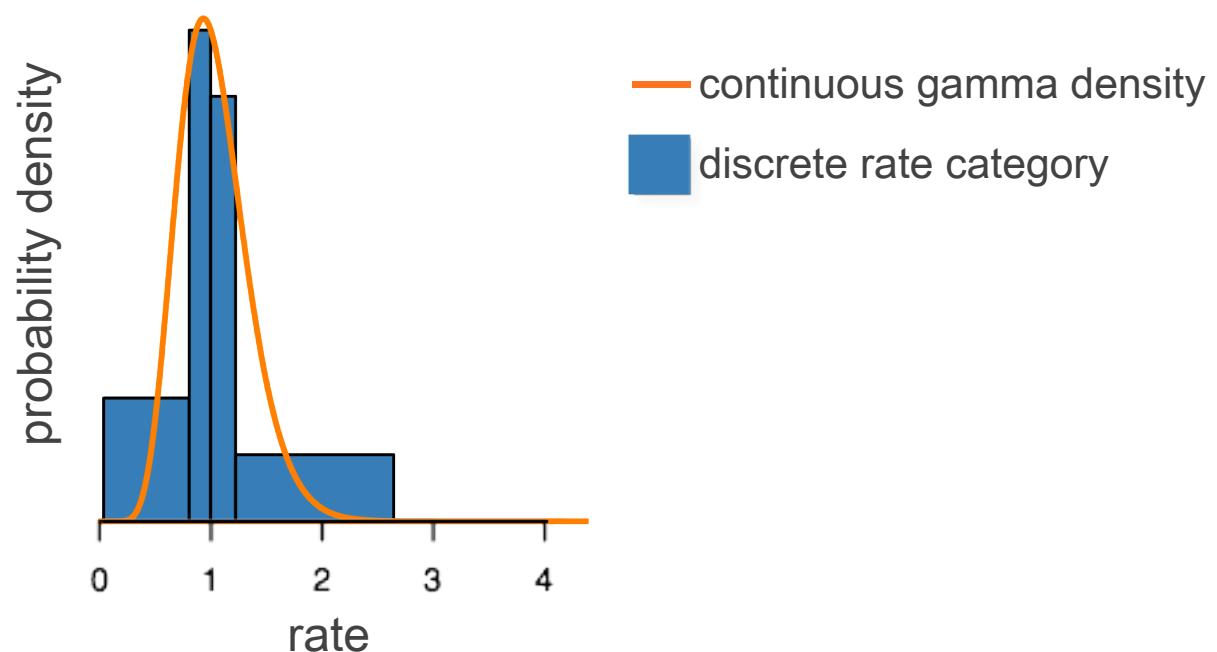
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution is discretized into  $k$  discrete bins

- there are  $k$  relative-rate multipliers; corresponding to the mean or median rate of each bin
- the approximation of the continuous gamma improves with increasing  $k$ , but so does the computational burden

$$\alpha = 10, k = 4$$



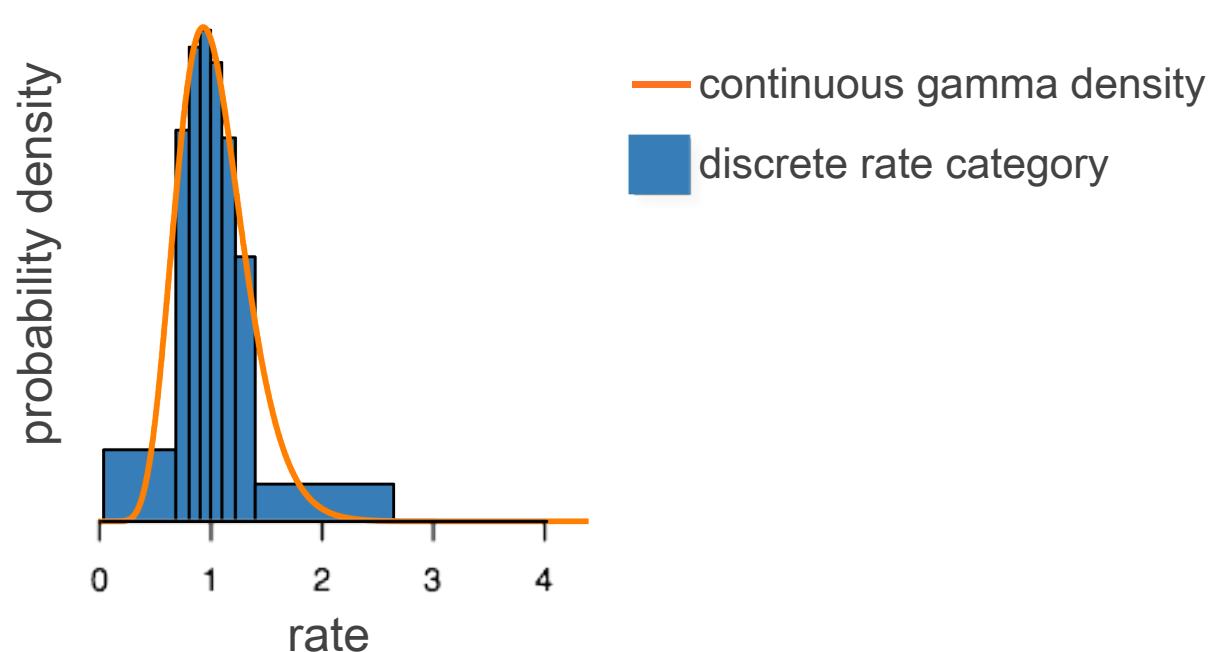
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution is discretized into  $k$  discrete bins

- there are  $k$  relative-rate multipliers; corresponding to the mean or median rate of each bin
- the approximation of the continuous gamma improves with increasing  $k$ , but so does the computational burden

$$\alpha = 10, k = 8$$



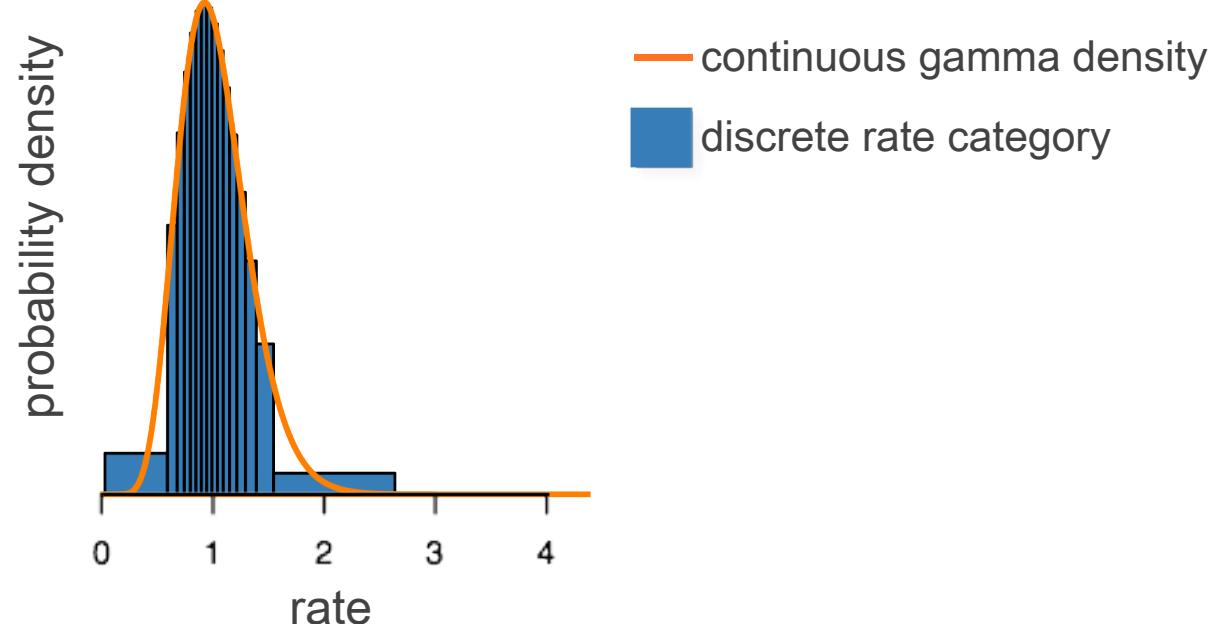
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution is discretized into  $k$  discrete bins

- there are  $k$  relative-rate multipliers; corresponding to the mean or median rate of each bin
- the approximation of the continuous gamma improves with increasing  $k$ , but so does the computational burden

$$\alpha = 10, k = 16$$



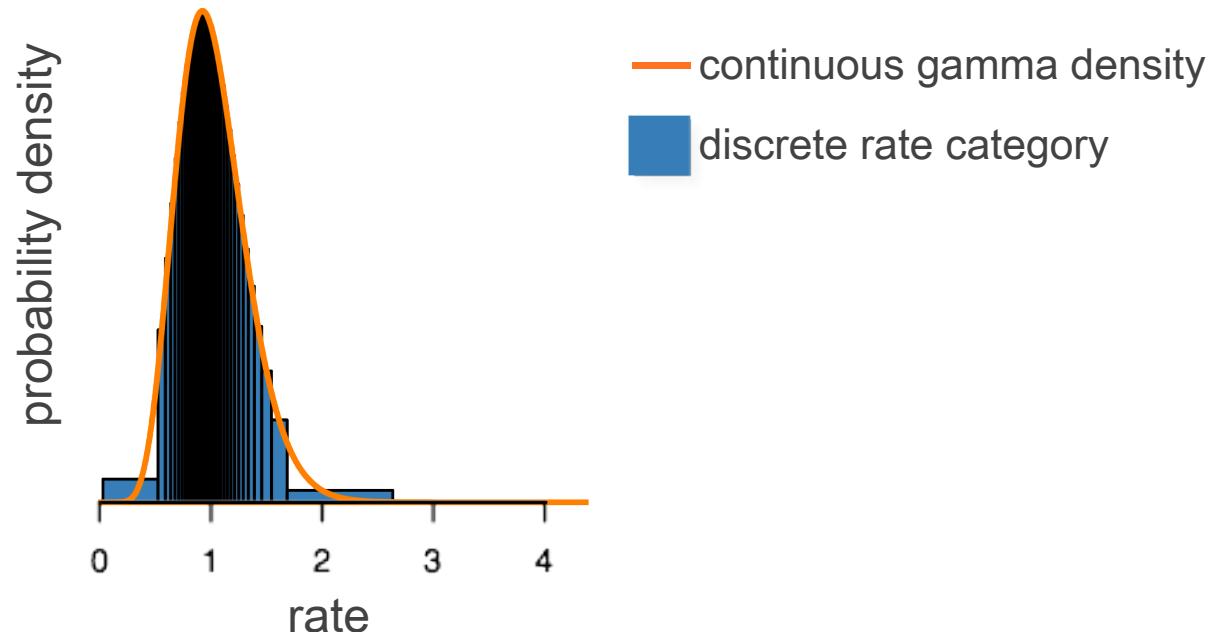
# Accommodating Among Site Rate Variation

## The Discrete Gamma Distribution (+G) model of ASRV

The gamma distribution is discretized into  $k$  discrete bins

- there are  $k$  relative-rate multipliers; corresponding to the mean or median rate of each bin
- the approximation of the continuous gamma improves with increasing  $k$ , but so does the computational burden

$$\alpha = 10, k = 32$$



# Accommodating Among Site Rate Variation

## Mixture models for accommodating ASRV

Site rates can be modeled as a mixture of the individual ASRV models (SS, I, and G):

# Accommodating Among Site Rate Variation

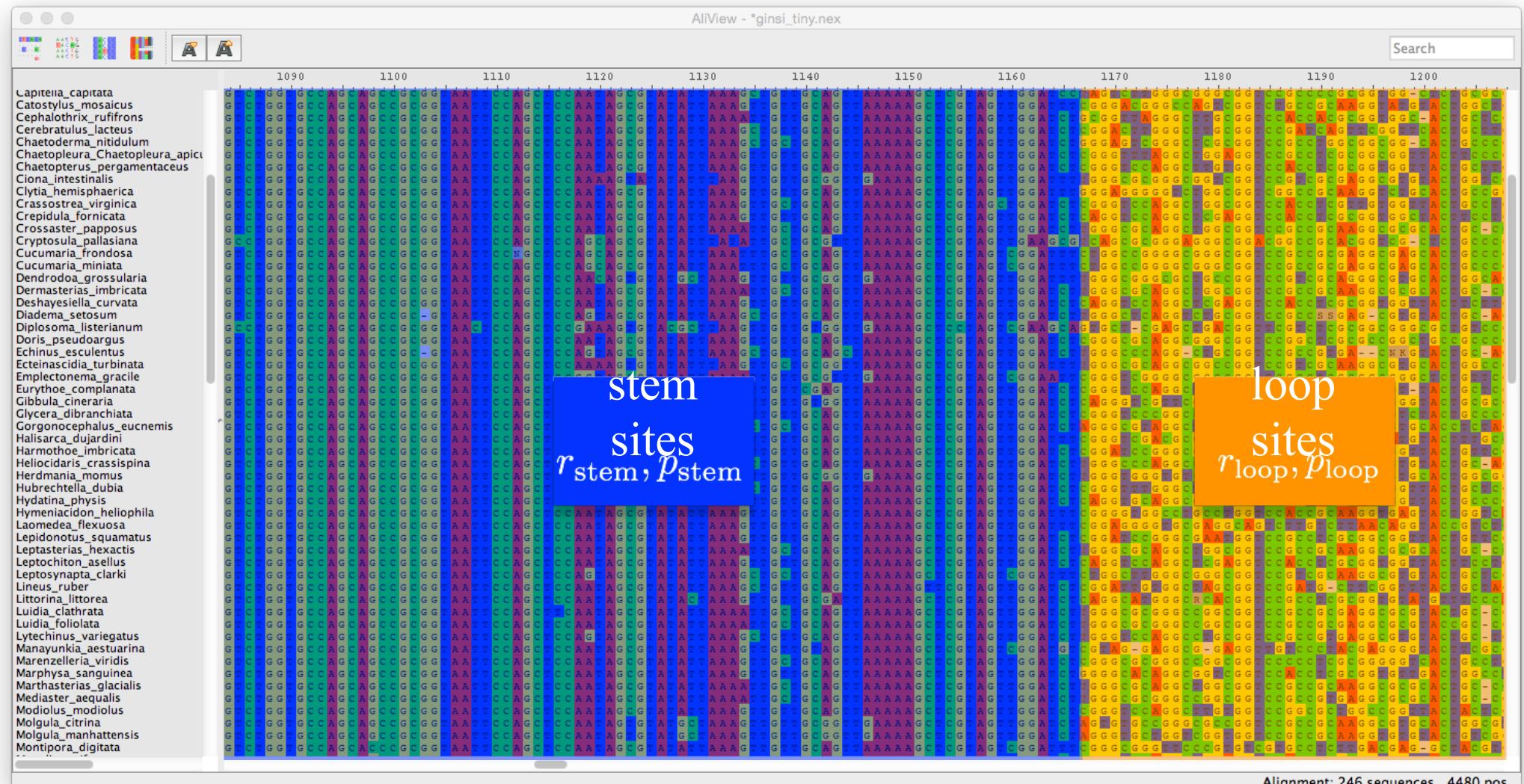
## Mixture models for accommodating ASRV

Site rates can be modeled as a mixture of the individual ASRV models (SS, I, and G):

- as a mixture of site-specific and proportion of invariable sites (+SS+I)

# Accommodating Among Site Rate Variation

## Mixture models for accommodating ASRV



Each pre-specified subset of sites has an overall relative-rate multiplier, and the rate of each site within each category is modeled as proportion of invariable sites model

# Accommodating Among Site Rate Variation

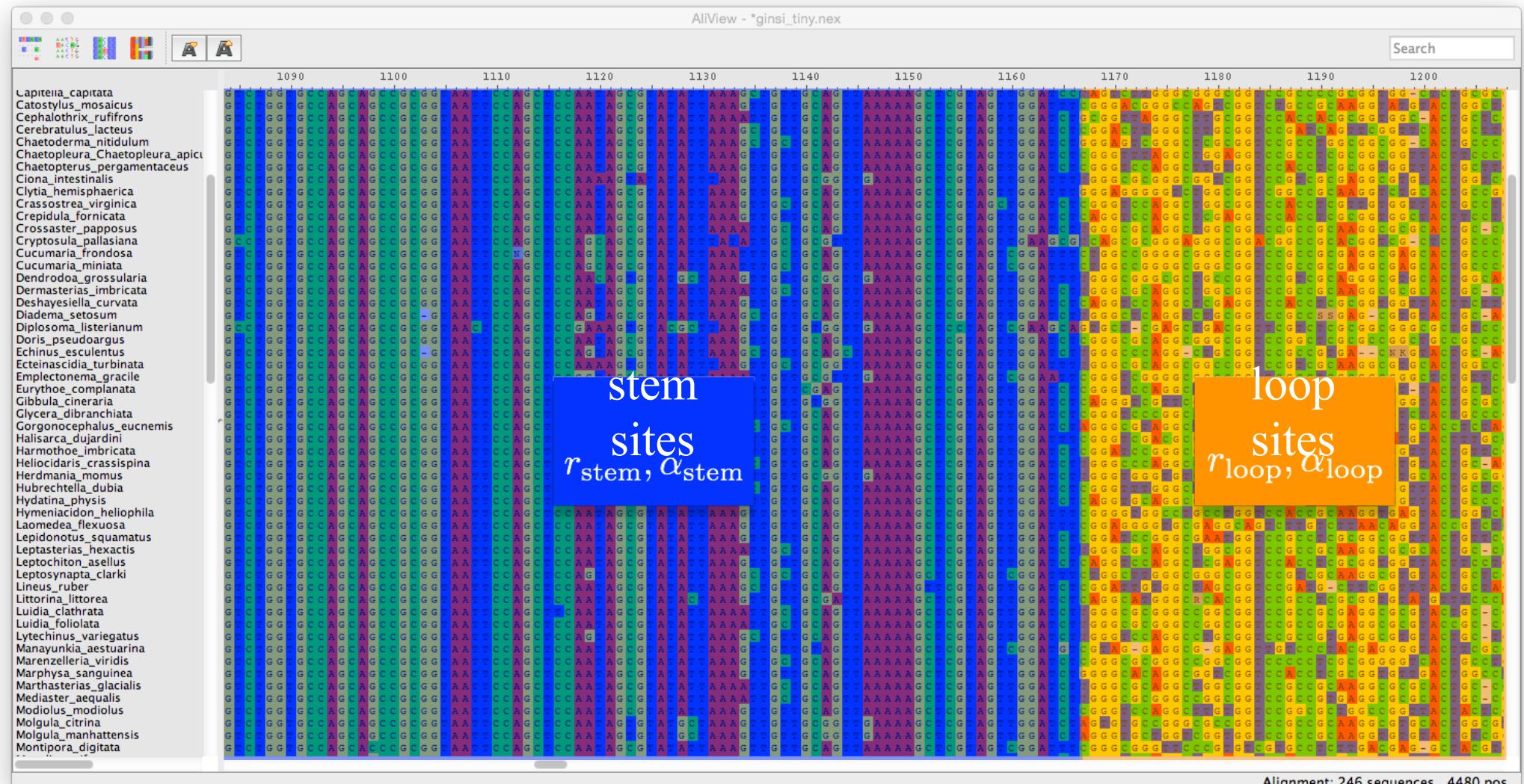
## Mixture models for accommodating ASRV

Site rates can be model as a mixture of the individual ASRV models (SS, I, and G):

- as a mixture of site-specific and proportion of invariable sites (+SS+I)
- as a mixture of site-specific and discrete-gamma distributed rates (+SS+G)

# Accommodating Among Site Rate Variation

## Mixture models for accommodating ASRV



Each pre-specified subset of sites has an overall relative-rate multiplier, and the rate of each site within each category is modeled as discrete-gamma distributed random variables

# Accommodating Among Site Rate Variation

## Mixture models for accommodating ASRV

Site rates can be model as a mixture of the individual ASRV models (SS, I, and G):

- as a mixture of site-specific and proportion of invariable sites (+SS+I)
- as a mixture of site-specific and discrete-gamma distributed rates (+SS+G)
- as a mixture of proportion of invariable sites and discrete-gamma distributed rates (+I+G)

# Accommodating Among Site Rate Variation

## Mixture models for accommodating ASRV



The rate of each site is modeled as a random variable drawn from a mixture of the discrete-gamma and proportion of invariable sites

# Accommodating Among Site Rate Variation

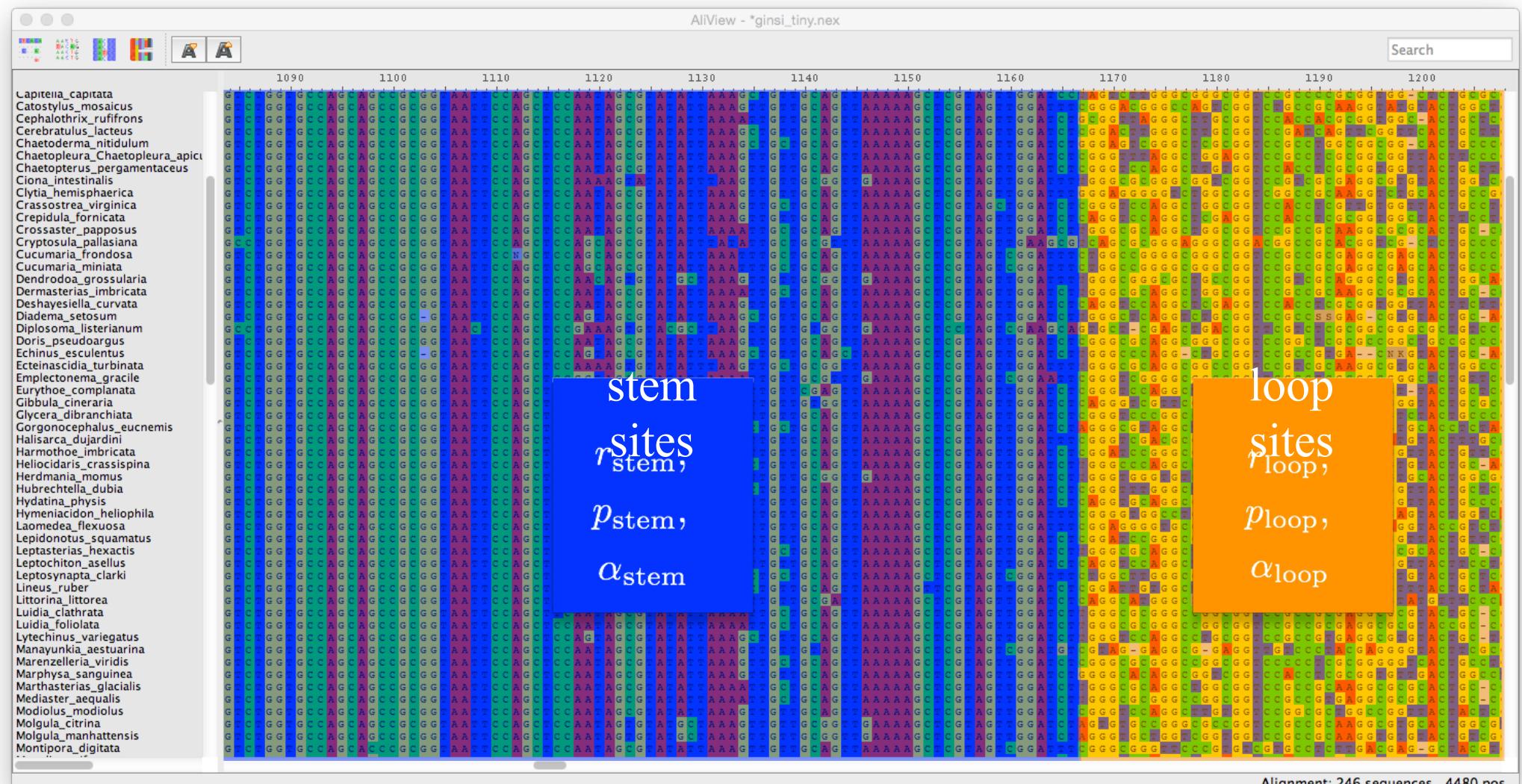
## Mixture models for accommodating ASRV

Site rates can be model as a mixture of the individual ASRV models (SS, I, and G):

- as a mixture of site-specific and proportion of invariable sites (+SS+I)
- as a mixture of site-specific and discrete-gamma distributed rates (+SS+G)
- as a mixture of proportion of invariable sites and discrete-gamma distributed rates (+I+G)
- as a mixture of all three models (+SS+I+G)

# Accommodating Among Site Rate Variation

## Mixture models for accommodating ASRV



Each pre-specified subset of sites has an overall relative-rate multiplier, and the rate of each site within each category is modeled as mixture of invariable and discrete-gamma distributed random variables

# Accommodating Among Site Rate Variation

## Mixture models for accommodating ASRV

Site rates can be model as a mixture of the individual ASRV models (SS, I, and G):

- as a mixture of site-specific and proportion of invariable sites (+SS+I)
- as a mixture of site-specific and discrete-gamma distributed rates (+SS+G)
- as a mixture of proportion of invariable sites and discrete-gamma distributed rates (+I+G)
- as a mixture of all three models (+SS+I+G)

Mixtures of I+G are potentially 'toxic'

- this mixture model is often 'non-identifiable', where there are an infinite number of parameter values for which the data are equally likely to be observed; e.g.,



# Accommodating Among Site Rate Variation

## Mixture models for accommodating ASRV

Site rates can be model as a mixture of the individual ASRV models (SS, I, and G):

- as a mixture of site-specific and proportion of invariable sites (+SS+I)
- as a mixture of site-specific and discrete-gamma distributed rates (+SS+G)
- as a mixture of proportion of invariable sites and discrete-gamma distributed rates (+I+G)
- as a mixture of all three models (+SS+I+G)

Mixtures of I+G are potentially 'toxic'

- this mixture model is often 'non-identifiable', where there are an infinite number of parameter values for which the data are equally likely to be observed; e.g.,
  - relatively large values of  $p$  and  $\alpha$



# Accommodating Among Site Rate Variation

## Mixture models for accommodating ASRV

Site rates can be model as a mixture of the individual ASRV models (SS, I, and G):

- as a mixture of site-specific and proportion of invariable sites (+SS+I)
- as a mixture of site-specific and discrete-gamma distributed rates (+SS+G)
- as a mixture of proportion of invariable sites and discrete-gamma distributed rates (+I+G)
- as a mixture of all three models (+SS+I+G)

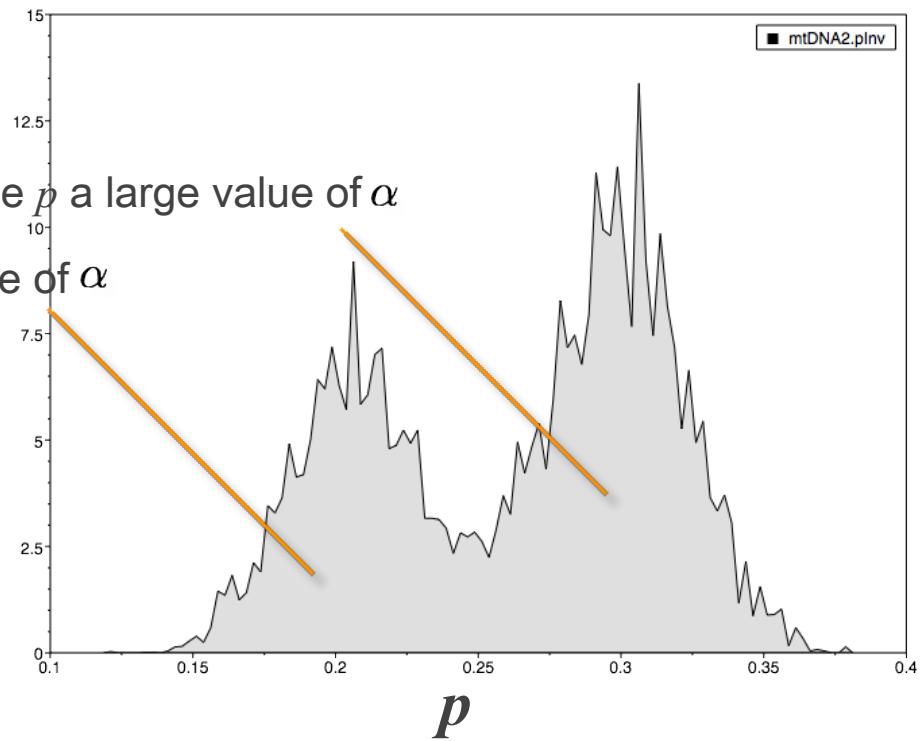
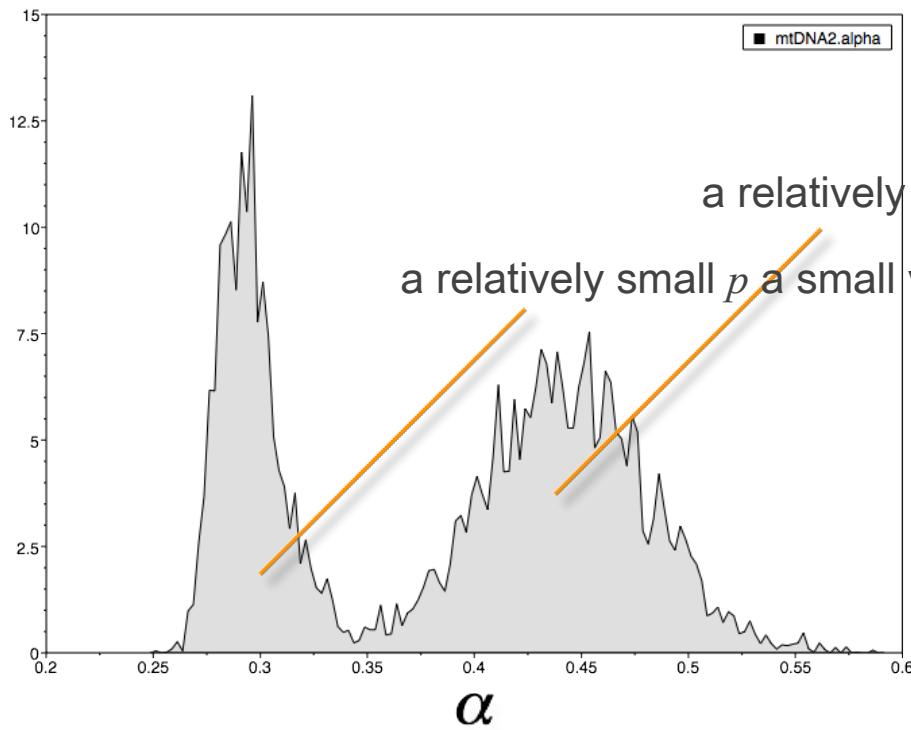
Mixtures of I+G are potentially 'toxic'

- this mixture model is often 'non-identifiable', where there are an infinite number of parameter values for which the data are equally likely to be observed; e.g.,
  - relatively large values of  $p$  and  $\alpha$
  - relatively small values of  $p$  and  $\alpha$



# Accommodating Among Site Rate Variation

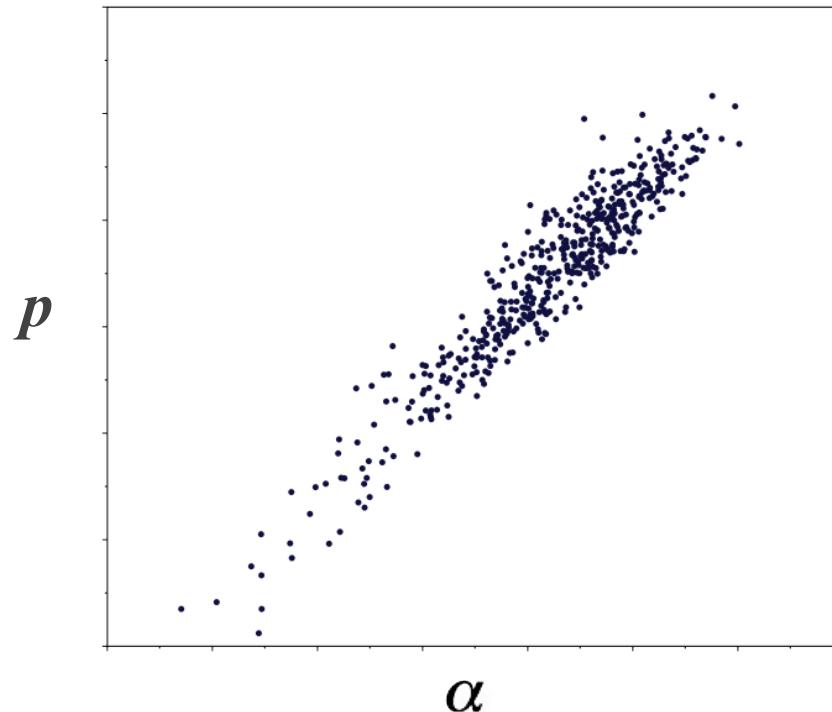
Mixture models of (I+G) for accommodating ASRV may be non-identifiable



These parameter interactions may cause pathologies that impact estimates of other model parameters

# Accommodating Among Site Rate Variation

Mixture models of (I+G) for accommodating ASRV may be non-identifiable



We can identify parameter interactions by plotting their joint posterior probability distribution (see Bayesian MCMC presentations!)

# Accommodating Among Site Rate Variation

## Why do we model ASRV?

ASRV can be due to a variety of processes (mutation, bGC, selection, epistatic interactions...)

ASRV models do not help us understand the data, they help us deal with its complexities to study another aspect of the data.

We model ASRV in this way because we do not use a mechanistic model of the processes creating ASRV

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

→ Among-site variation in substitution rates

Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

→ Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# Accommodating Among Site Process Heterogeneity

The substitution process may vary qualitatively across the alignment

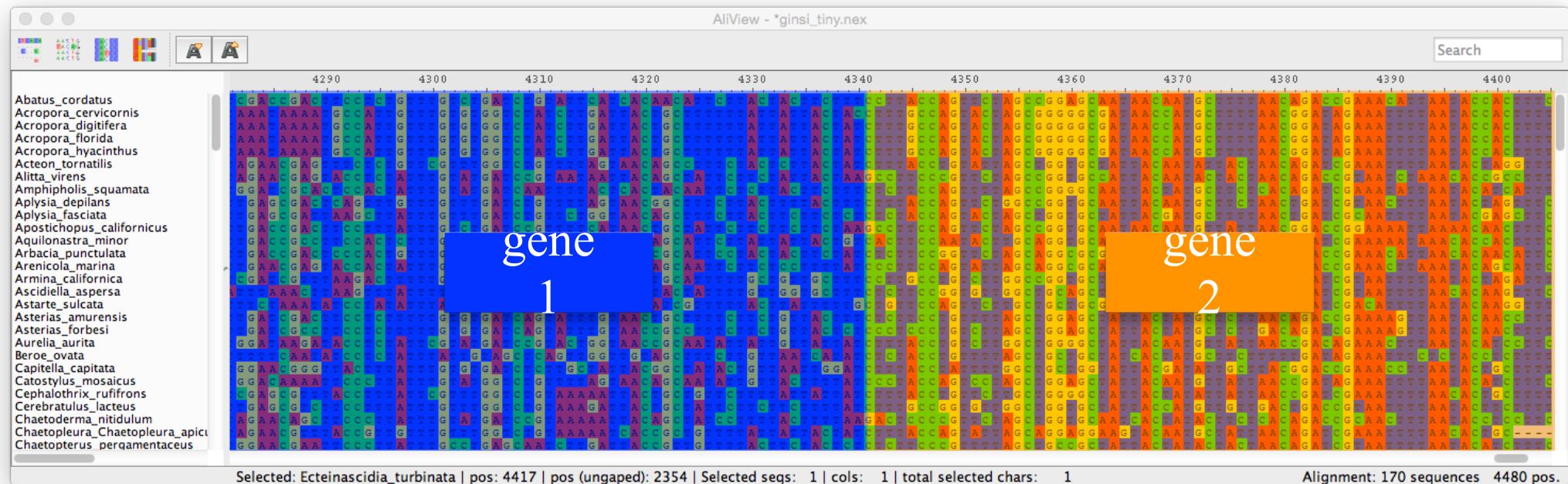


Alignments may be composed of (subsets of) sites from different:

- genes (e.g., cytb, COI)

# Accommodating Among Site Process Heterogeneity

The substitution process may vary qualitatively across the alignment



Alignments may be composed of (subsets of) sites from different:

- genes (e.g., cytb, COI)
- genomes (e.g., nuclear, mitochondrial, chloroplast)

# Accommodating Among Site Process Heterogeneity

The substitution process may vary qualitatively across the alignment



Alignments may be composed of (subsets of) sites from different:

- genes (e.g., cytb, COI)
- genomes (e.g., nuclear, mitochondrial, chloroplast)
- codon positions of protein-coding genes

# Accommodating Among Site Process Heterogeneity

The substitution process may vary qualitatively across the alignment

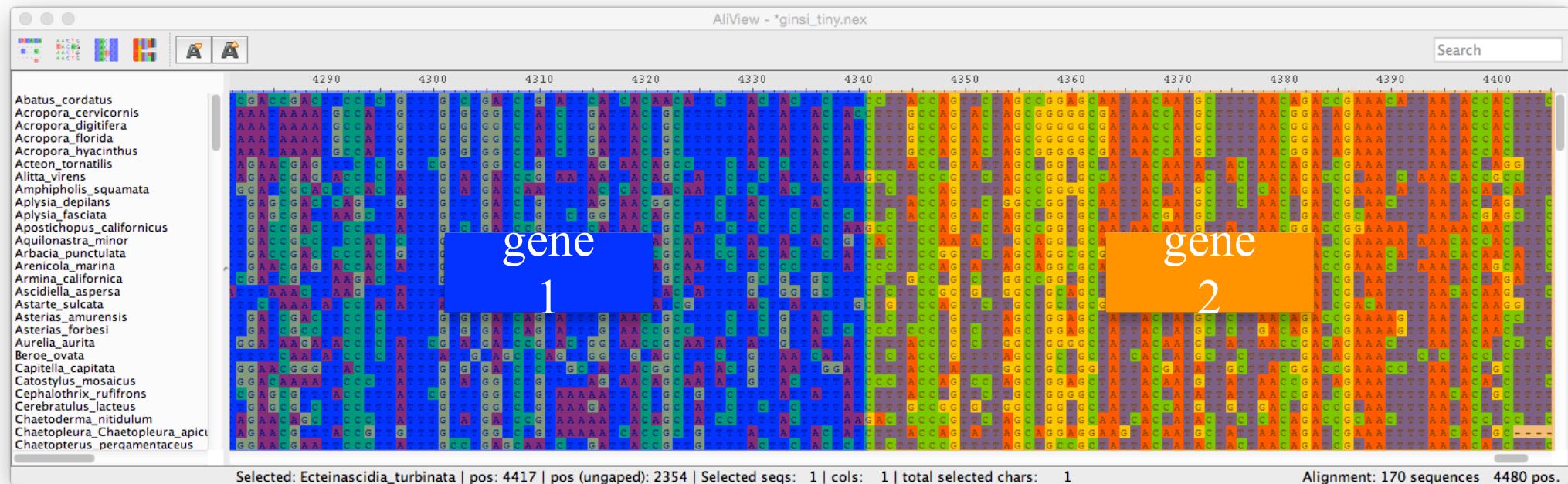


Alignments may be composed of (subsets of) sites from different:

- genes (e.g., cytb, COI)
- genomes (e.g., nuclear, mitochondrial, chloroplast)
- codon positions of protein-coding genes
- stem/loop regions of ribosomal genes

# Accommodating Among Site Process Heterogeneity

The substitution process may vary qualitatively across the alignment



The nature of the substitution process for these (subsets of) sites may differ in terms of:

- tree topology ('gene-tree conflict')

# Accommodating Among Site Process Heterogeneity

The substitution process may vary qualitatively across the alignment



The nature of the substitution process for these (subsets of) sites may differ in terms of:

- tree topology ('gene-tree conflict')
- branch-length proportions ('heterotachy')

# Accommodating Among Site Process Heterogeneity

The substitution process may vary qualitatively across the alignment



The nature of the substitution process for these (subsets of) sites may differ in terms of:

- tree topology ('gene-tree conflict')
- branch-length proportions ('heterotachy')
- stationary frequencies ( $\pi_i$ )

# Accommodating Among Site Process Heterogeneity

The substitution process may vary qualitatively across the alignment

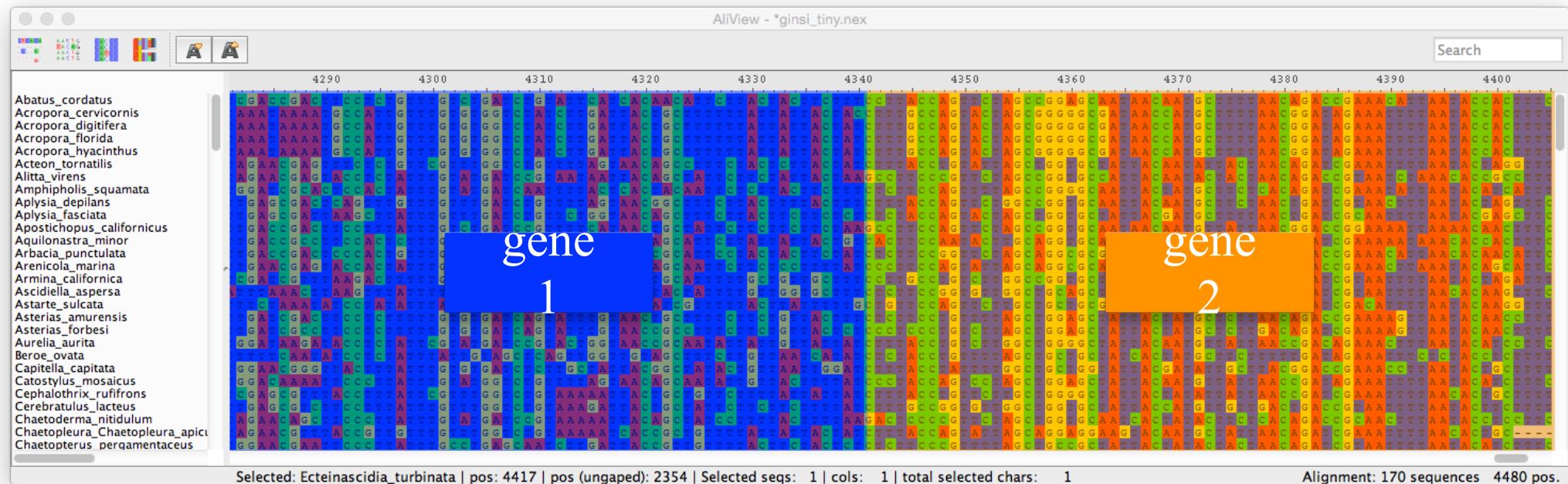


The nature of the substitution process for these (subsets of) sites may differ in terms of:

- tree topology ('gene-tree conflict')
- branch-length proportions ('heterotachy')
- stationary frequencies ( $\pi_i$ )
- relative rates ( $a-f$ )

# Accommodating Among Site Process Heterogeneity

The substitution process may vary qualitatively across the alignment



The nature of the substitution process for these (subsets of) sites may differ in terms of:

- tree topology ('gene-tree conflict')
- branch-length proportions ('heterotachy')
- stationary frequencies ( $\pi_i$ )
- relative rates ( $a-f$ )
- the nature/degree of ASRV ( $p, \alpha$ )

# Accommodating Among Site Process Heterogeneity

Biology motivates the extension of models

Variation in substitution process across sites is a prevalent feature of empirical datasets

# Accommodating Among Site Process Heterogeneity

Biology motivates the extension of models

Variation in substitution process across sites is a prevalent feature of empirical datasets

The tree topology may vary across the alignment ('gene-tree conflict') because of:

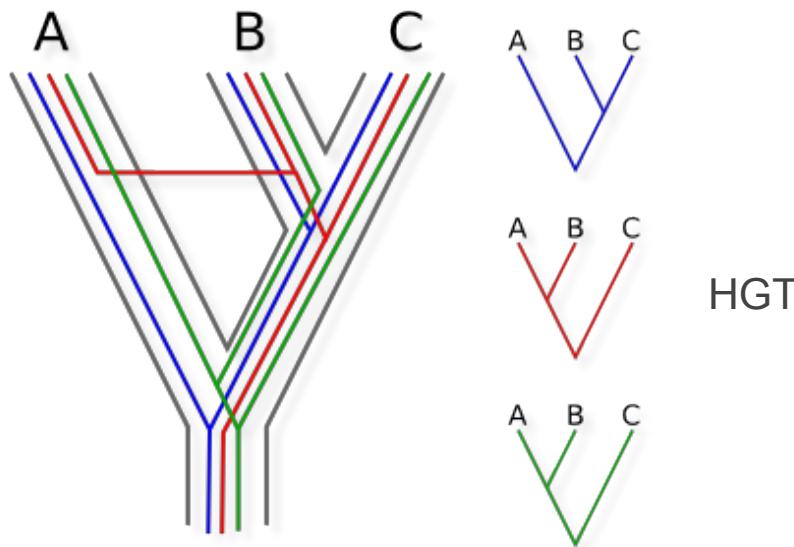
# Accommodating Among Site Process Heterogeneity

Biology motivates the extension of models

Variation in substitution process across sites is a prevalent feature of empirical datasets

The tree topology may vary across the alignment ('gene-tree conflict') because of:

- horizontal gene transfer/introgression



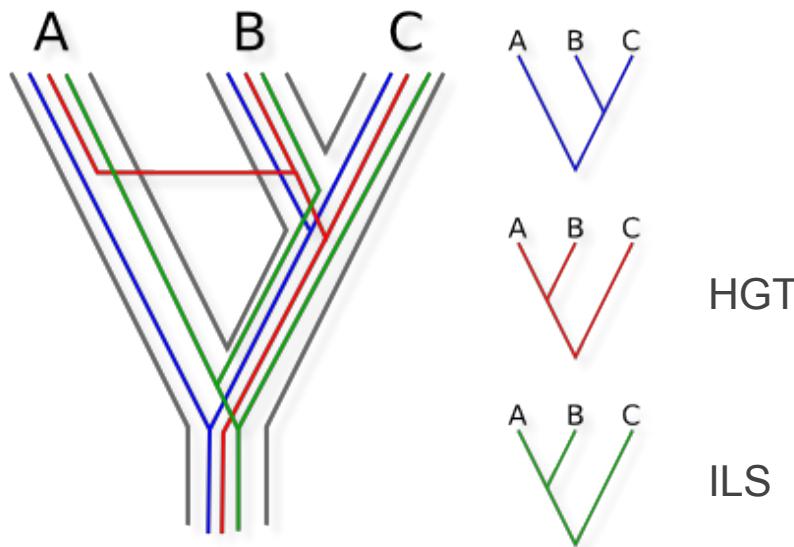
# Accommodating Among Site Process Heterogeneity

Biology motivates the extension of models

Variation in substitution process across sites is a prevalent feature of empirical datasets

The tree topology may vary across the alignment ('gene-tree conflict') because of:

- horizontal gene transfer/introgression
- deep coalescence (incomplete lineage sorting)



# Accommodating Among Site Process Heterogeneity

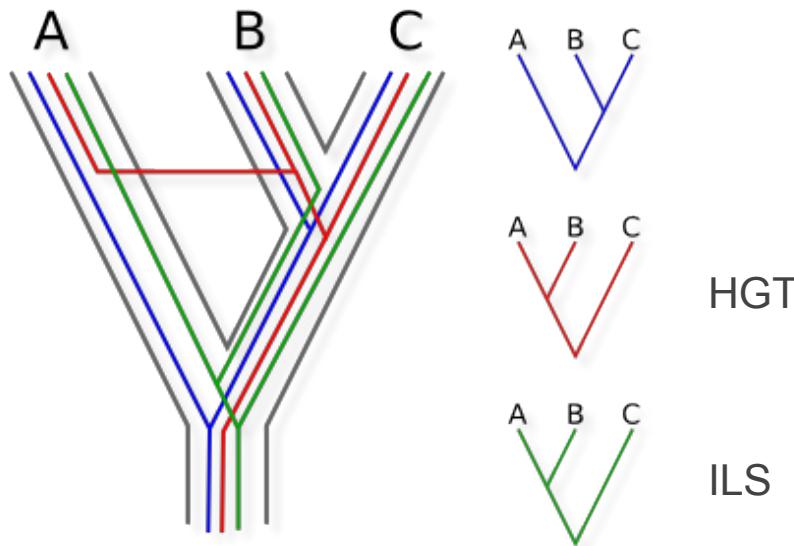
Biology motivates the extension of models

Variation in substitution process across sites is a prevalent feature of empirical datasets

The tree topology may vary across the alignment ('gene-tree conflict') because of:

- horizontal gene transfer/introgression
- deep coalescence (incomplete lineage sorting)

There are 'species-tree' methods that allow the tree topology to vary across sites



# Accommodating Among Site Process Heterogeneity

Biology motivates the extension of models

Even when all sites share the same tree topology, other aspects of the substitution process may vary across (subsets of) sites in the alignment

# Accommodating Among Site Process Heterogeneity

Biology motivates the extension of models

Even when all sites share the same tree topology, other aspects of the substitution process may vary across (subsets of) sites in the alignment

Under simulation, failure to accommodate substitution process heterogeneity can cause biased estimates of:

# Accommodating Among Site Process Heterogeneity

## Biology motivates the extension of models

Even when all sites share the same tree topology, other aspects of the substitution process may vary across (subsets of) sites in the alignment

Under simulation, failure to accommodate substitution process heterogeneity can cause biased estimates of:

- tree topology and nodal support

# Accommodating Among Site Process Heterogeneity

## Biology motivates the extension of models

Even when all sites share the same tree topology, other aspects of the substitution process may vary across (subsets of) sites in the alignment

Under simulation, failure to accommodate substitution process heterogeneity can cause biased estimates of:

- tree topology and nodal support
- branch lengths

# Accommodating Among Site Process Heterogeneity

## Biology motivates the extension of models

Even when all sites share the same tree topology, other aspects of the substitution process may vary across (subsets of) sites in the alignment

Under simulation, failure to accommodate substitution process heterogeneity can cause biased estimates of:

- tree topology and nodal support
- branch lengths
- divergence times

# Accommodating Among Site Process Heterogeneity

## Biology motivates the extension of models

Even when all sites share the same tree topology, other aspects of the substitution process may vary across (subsets of) sites in the alignment

Under simulation, failure to accommodate substitution process heterogeneity can cause biased estimates of:

- tree topology and nodal support
- branch lengths
- divergence times
- other substitution-model parameters

# Accommodating Among Site Process Heterogeneity

Mixed-model ('partitioned-data') approach for modeling ASPH

The basic idea is to assign (subsets of) sites to independent substitution models

# Accommodating Among Site Process Heterogeneity

## Mixed-model ('partitioned-data') approach for modeling ASPH

The basic idea is to assign (subsets of) sites to independent substitution models

The number of data subsets and the assignment of sites to subsets is assumed

- the number of partitions and assignment of subsets is informed by our biological knowledge

# Accommodating Among Site Process Heterogeneity

## Mixed-model ('partitioned-data') approach for modeling ASPH

The basic idea is to assign (subsets of) sites to independent substitution models

The number of data subsets and the assignment of sites to subsets is assumed

- the number of partitions and assignment of subsets is informed by our biological knowledge
- a substitution model is specified for each data subset

# Accommodating Among Site Process Heterogeneity

## Mixed-model ('partitioned-data') approach for modeling ASPH

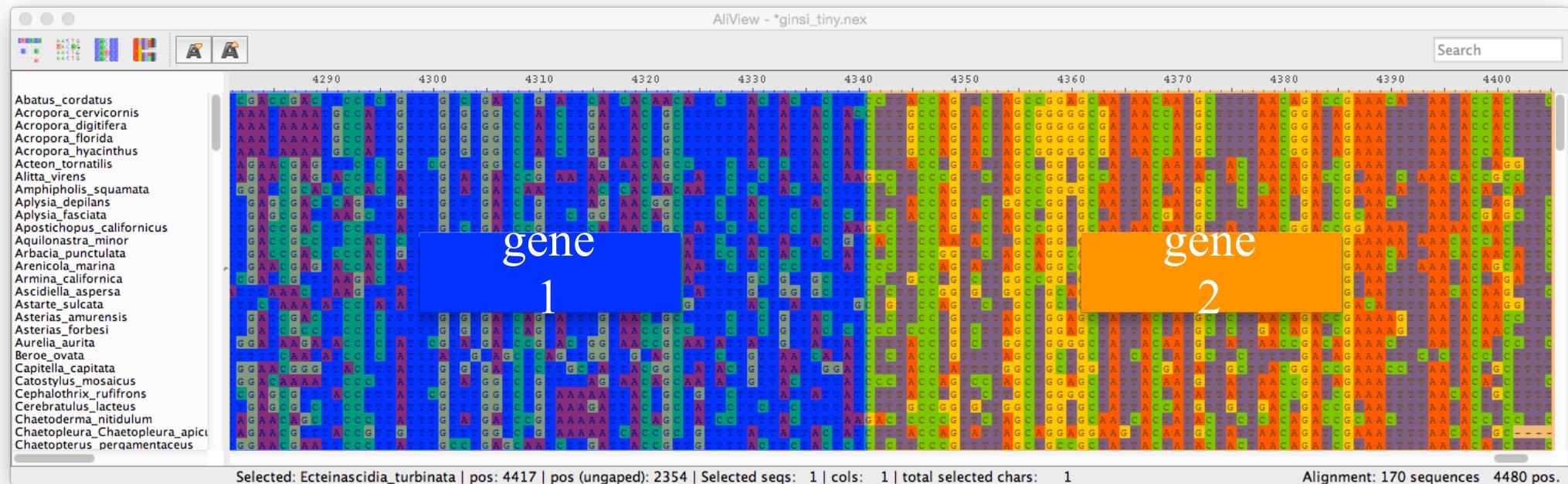
The basic idea is to assign (subsets of) sites to independent substitution models

The number of data subsets and the assignment of sites to subsets is assumed

- the number of partitions and assignment of subsets is informed by our biological knowledge
- a substitution model is specified for each data subset
- a substitution model is specified for each data subset

# Accommodating Among Site Process Heterogeneity

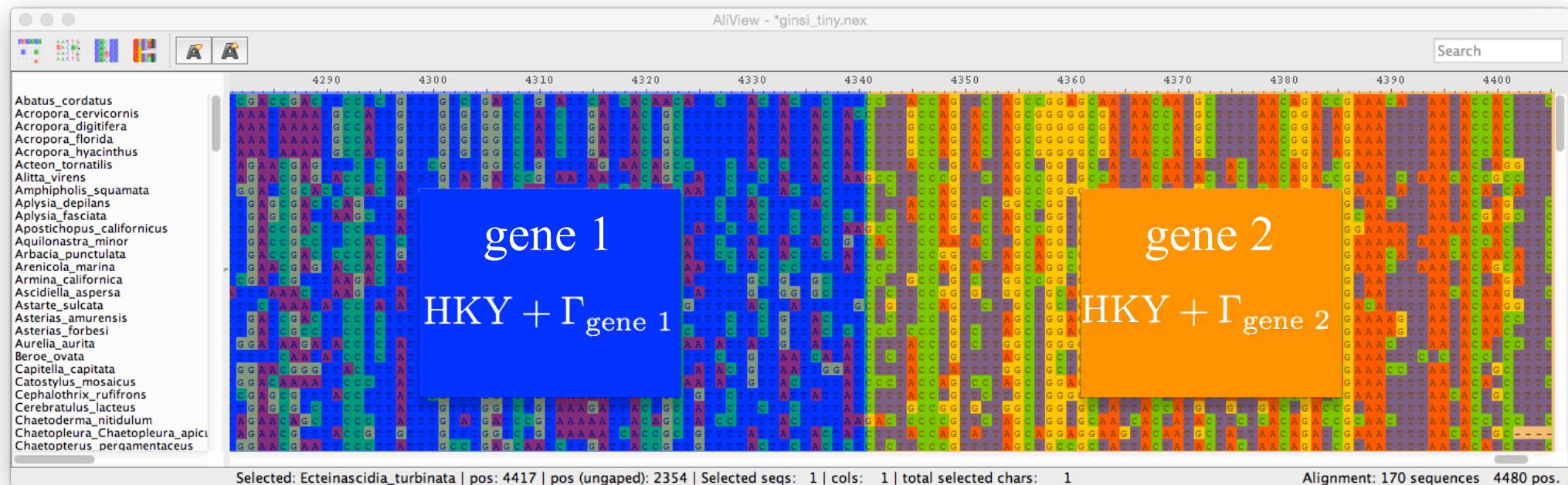
Mixed-model ('partitioned-data') approach for modeling ASPH



Define two or more data subsets that are likely to capture substitution process heterogeneity

# Accommodating Among Site Process Heterogeneity

Mixed-model ('partitioned-data') approach for modeling ASPH

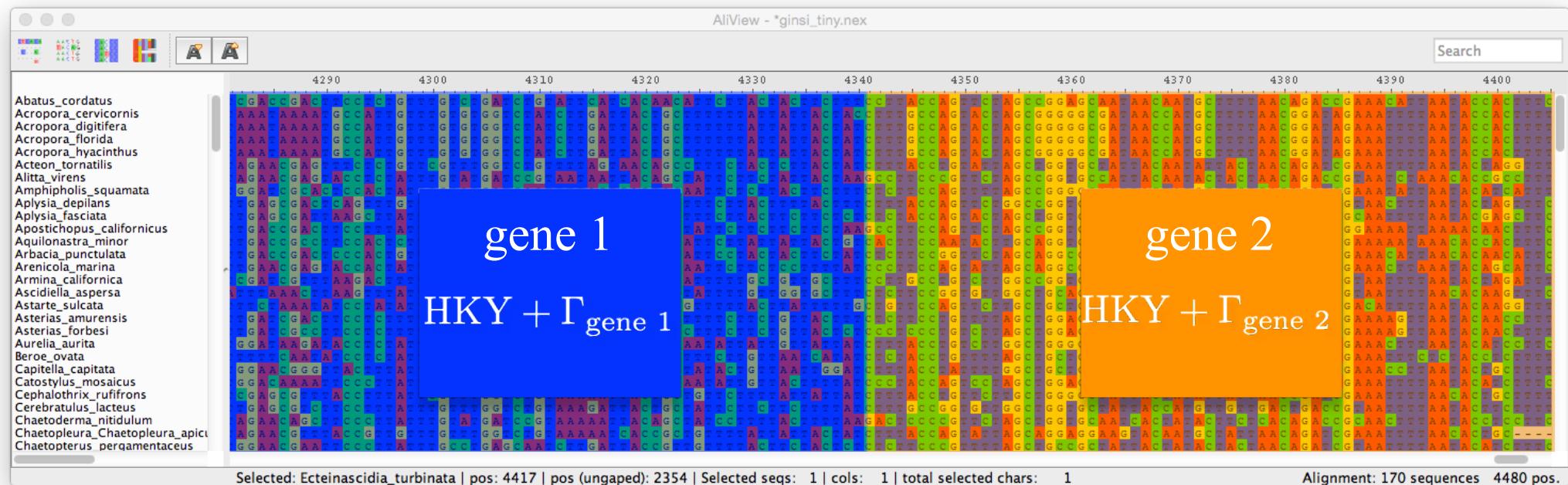


Define two or more data subsets that are likely to capture substitution process heterogeneity

Specify a substitution model for each pre-specified data subset

# Accommodating Among Site Process Heterogeneity

Mixed-model ('partitioned-data') approach for modeling ASPH



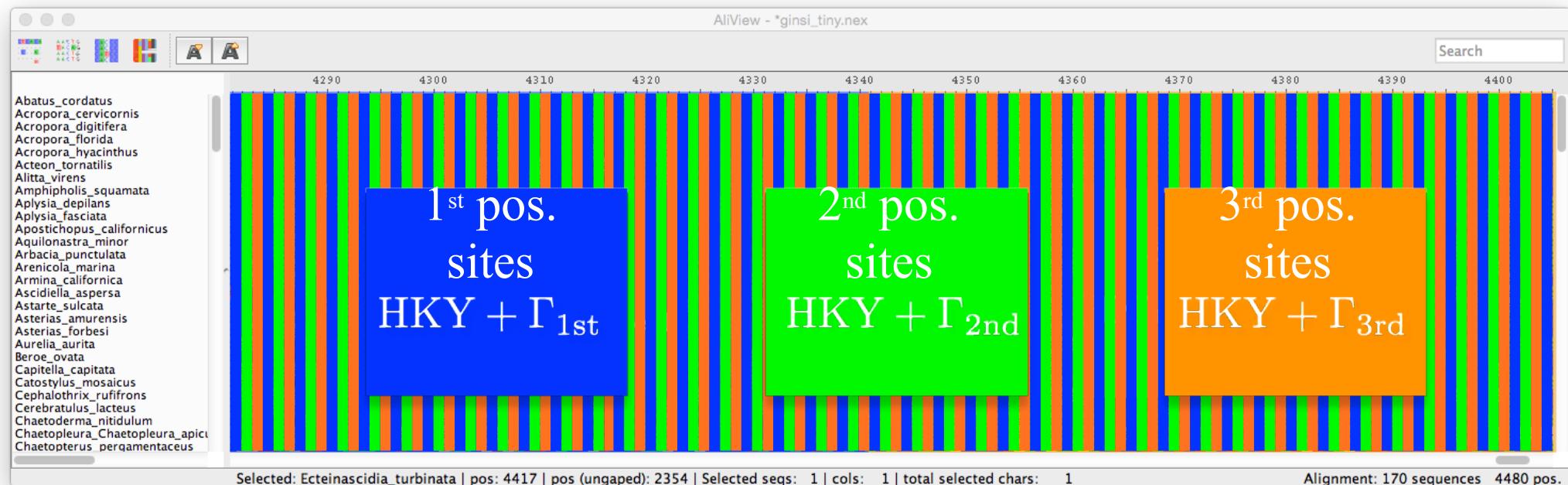
Define two or more data subsets that are likely to capture substitution process heterogeneity

Specify a substitution model for each pre-specified data subset

Estimate parameters of each model from sites in the corresponding data subset

# Accommodating Among Site Process Heterogeneity

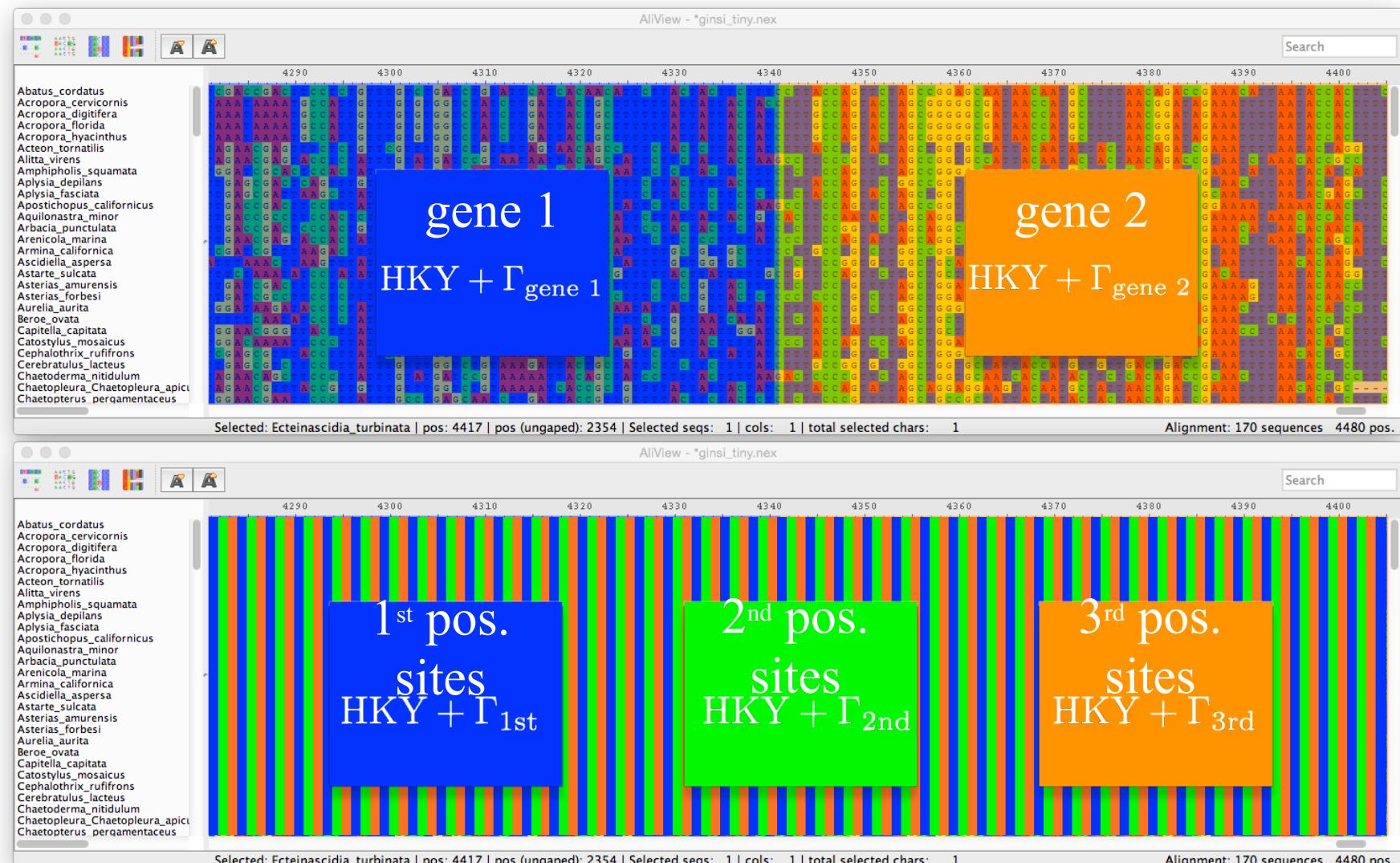
Mixed-model ('partitioned-data') approach for modeling ASPH



Alternative mixed-models are possible: for instance, the CAT model in Phylobayes (Lartillot et al.)

# Accommodating Among Site Process Heterogeneity

Mixed-model ('partitioned-data') approach for modeling ASPH



We can then compare the fit of the competing mixed-models to the data

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

→ Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# Outline

## I. A brief review of Continuous-Time Markov models

## II. Stochastic models of nucleotide substitution

What's the deal with time reversibility

Meet the GTR family

## III. Accommodating among-site heterogeneity in the substitution process

Among-site variation in substitution rates

Among-site variation in substitution process

## IV. Accommodating heterogeneity in the substitution process along the tree

# The Deinococcus-Thermus phylum and the effect of rRNA composition on phylogenetic tree construction.

Weisburg WG, Giovannoni SJ, Woese CR.

Syst Appl Microbiol. 1989;11:128-34.

sel=0	293	380
Aquifex	CTCCCCCTCGGGTAACGGGATGCGCACCGTAATGCATCCCGATGCCGATAGGCACCCCGCAGGC	CCGGCAGGTCTCGGAGAAGCAGC
Thermotoga	GGGCTGCCTGGTGTTCGGATGGTCACCGTGGCTCATCCGAATACCAGGTGGGGCGTAGAGAGACCGCGTGACGTATCGCGGGTGGCGT	
Thermus	GCGTTGCCTGGGGCTCCCATTGGAACCGTGGGACGTGGGATGCTCAGGAACGGTGGAGAGTGGTTCCGACAGGCACCGGGATCGAGC	
Deinococcus	GCGTTGATATTAACTGGTTTGAGACTCGAGTGA	CTGGATTGGATGTTACCTCTGAGATAACTCTGTGGGTACCAAGAAATCGAAG
Rickettsia	ATGCTGTTAGTAATTGGAAGGGCGTTCAATTCTTCAAATAACTAATAAGTGTATAGGATGATTCTATAATTATTAGAGGTGGGT	

"The (partial) sequence of *T. aquaticus* rRNA appears relatively close to those of other thermophilic eubacteria. [...] However, this closeness does not reflect a true evolutionary closeness; rather it is due to a "thermophilic convergence", the result of unusually high G+C composition in the rRNAs of thermophilic bacteria. Unless such compositional biases are taken into account, the branching order and root of phylogenetic trees can be incorrectly inferred."

# Widespread compositional heterogeneity

Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions.

Foster PG, Hickey DA.

J Mol Evol. 1999 Mar;48(3):284-90.

PMID: 10093217

Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae.

Tarrio R, Rodriguez-Trelles F, Ayala FJ.

Mol Biol Evol. 2001 Aug;18(8):1464-73.

PMID: 11470837 [Free Article](#)

Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria).

Herbeck JT, Degnan PH, Wernegreen JJ.

Mol Biol Evol. 2005 Mar;22(3):520-32. Epub 2004 Nov 3. Erratum in: Mol Biol Evol. 2005 Apr;22(4):1158.

PMID: 15525700 [Free Article](#)

The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods.

Foster PG, Cox CJ, Embley TM.

Philos Trans R Soc Lond B Biol Sci. 2009 Aug 12;364(1527):2197-207. doi: 10.1098/rstb.2009.0034.

PMID: 19571240 [Free PMC Article](#)

Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJ.

Heterogeneous models place the root of phylogeny.

Morgan CC, Foster PG, Webb AE, Pisani O'Connell MJ.

Mol Biol Evol. 2013 Sep;30(9):2145-56. doi: 10.1093/molbev/mst150. 2013 Jun 29.

PMID: 23813979 [Free PMC Article](#)

Evidence for GC-biased gene conversion lineage differences in avian base composition.

Weber CC, Boussau B, Romiguier J, Jarvis ED. Genome Biol. 2014;15(12):549.

PMID: 25496599 [Free PMC Article](#)

Conflicting phylogenies for early land plant groups: composition biases among synonymous codons.

Cox CJ, Li B, Foster PG, Embley TM, Ciferri A. Syst Biol. 2014 Mar;63(2):272-9. doi: 10.1093/sysbio/syt012. 6. No abstract available.

PMID: 24399481 [Free PMC Article](#)

Compositional biases among synonymous codons can conflict between gene and protein trees for some organisms.

Li B, Lopes JS, Foster PG, Embley TM, Ciferri A. Mol Biol Evol. 2014 Jul;31(7):1697-709. doi: 10.1093/molbev/msu132. 2014 May 2.

PMID: 24795089 [Free PMC Article](#)

Family-Level Sampling of Mitochondrial DNA Reveals Compositional Heterogeneity and Phylogenetic Conflict.

Timmermans MJ, Barton C, Haran J, Ahrens M, Ollikainen A, Dodsworth S, Foster PG, Boore AL. Genome Biol Evol. 2015 Dec 8;8(1):161-75. doi: 10.1093/gbe/evv250. 2015 Dec 8.

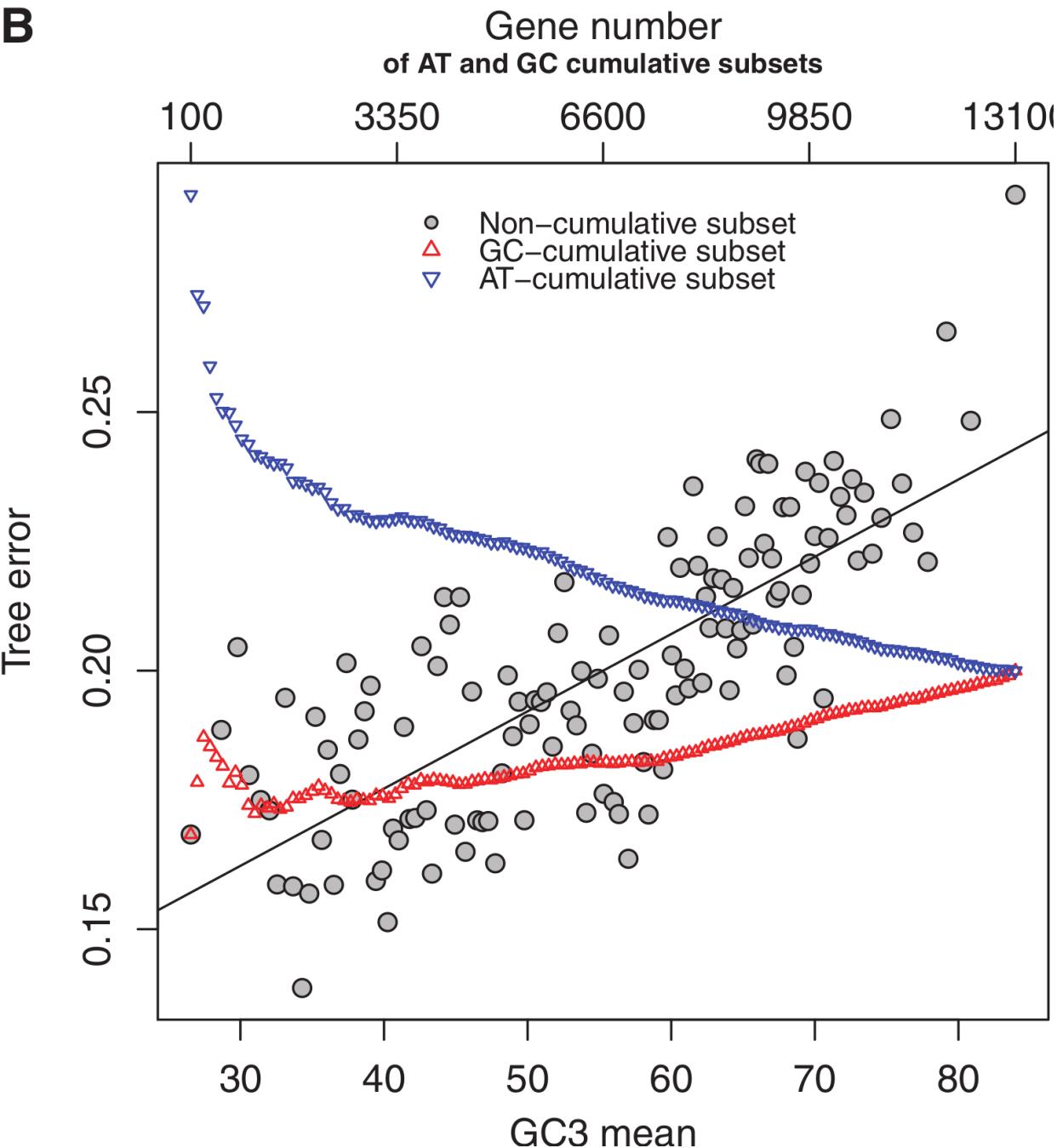
PMID: 26645679 [Free PMC Article](#)

The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated.

Jermini L, Ho SY, Ababneh F, Robinson J, O'Connell MJ. Mol Biol Evol. 2017 Jan 1;34(1):11-23. doi: 10.1093/molbev/msw301. 2016 Dec 1.

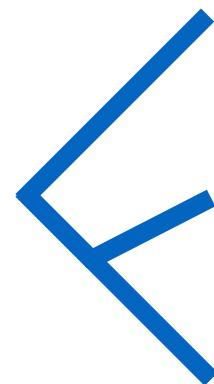
# The impact of compositional heterogeneity on mammalian phylogeny

B



- 131 sets of 100 genes ordered from low to high GC
- In GC rich gene families, GC content varies widely between genes
- GC rich genes have trees with more difference to the true tree

# Homogeneous models



Commonly used models inherit from GTR:

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

Where  $\Pi = (\pi_T, \pi_C, \pi_A, \pi_G)$  is the set of ***equilibrium*** frequencies

When a single matrix is used over a phylogeny, one assumes that sequences all tend towards

$$\Pi = (\pi_T, \pi_C, \pi_A, \pi_G)$$

*Does not fit many data sets*

# Software explicitly handling compositional heterogeneity

- Paml (Z. Yang)
- nhml (N. Galtier)
- nhPhyML (B. Boussau)
- bppML (J. Dutheil, L. Gueguen, B. Boussau)
- p4 (P. Foster)
- Hal-Has (V. Jayaswal)
- PhyloBayes ? (N. Lartillot)
- **NOT** PhyML, IQTree, RaXml, Garli, mrBayes, Beast(\*)...

# 2 major ways to model compositional heterogeneity

- The Barry-Hartigan model
- Mixtures of GTR matrices: transitions at nodes, or on branches

# Barry-Hartigan model

- ➊ Proposed in 1987
- ➋ Much work done by Lars Jermiin, John Robinson, Edward Susko, Barbara Holland, Jeremy Sumner, Gavin Huttley...
- ➌ fundamentally different from usual models: no branch lengths, no stationary frequencies

# Barry-Hartigan model vs GTR

Both models are reversible: no information about the position of the root.

## GTR

- Ubiquitous
- estimate 8 parameters of the Q matrix, then  $P = e^{Qt}$
- ancestral sequences:  
easy to reconstruct
- lack of closure

## Barry-Hartigan

- not implemented in any widely used software
- directly estimate the 12 parameters of the P matrix
- ancestral sequences: identifiability problem
- closure by matrix multiplication

# Mixtures of GTR matrices

Used for:

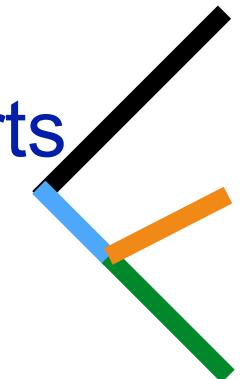
- model fitting (Jayaswal et al., 2011; Jayaswal et al., 2014; Dutheil et al., 2012)
- topology search

*Problem: lack of closure (Sumner et al. 2012)*

# Mixtures of GTR matrices

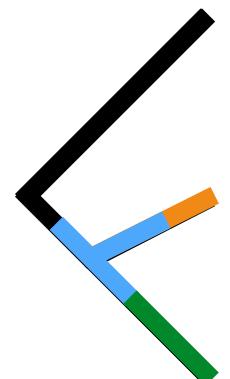
Changes at nodes:

- ➊  $n$  branches ->  $n$  matrices: Yang and Roberts 1995, Galtier and Gouy 1998, Boussau 2006, Heaps et al., 2014
- ➋  $n$  branches ->  $k < n$  matrices: Foster 2004



Changes along branches:

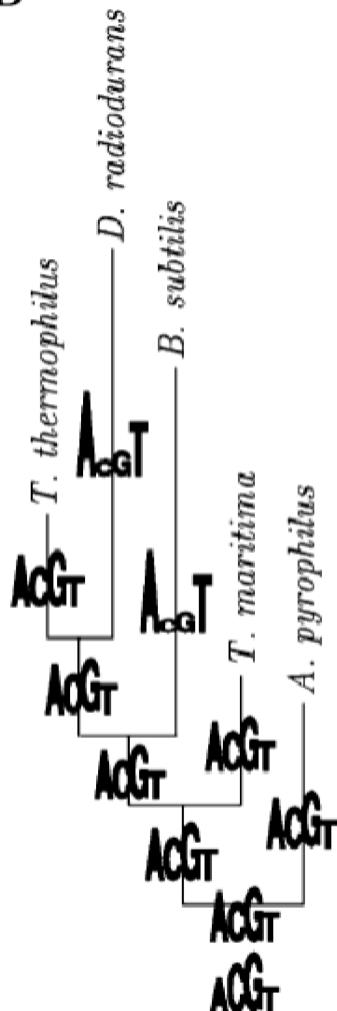
- ➌ Blanquart and Lartillot 2006, 2008



# Blanquart and Lartillot 2006

Bayesian MCMC sampling of histories of breakpoints between equilibrium frequencies along the phylogeny, and of the equilibrium frequencies themselves

B



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

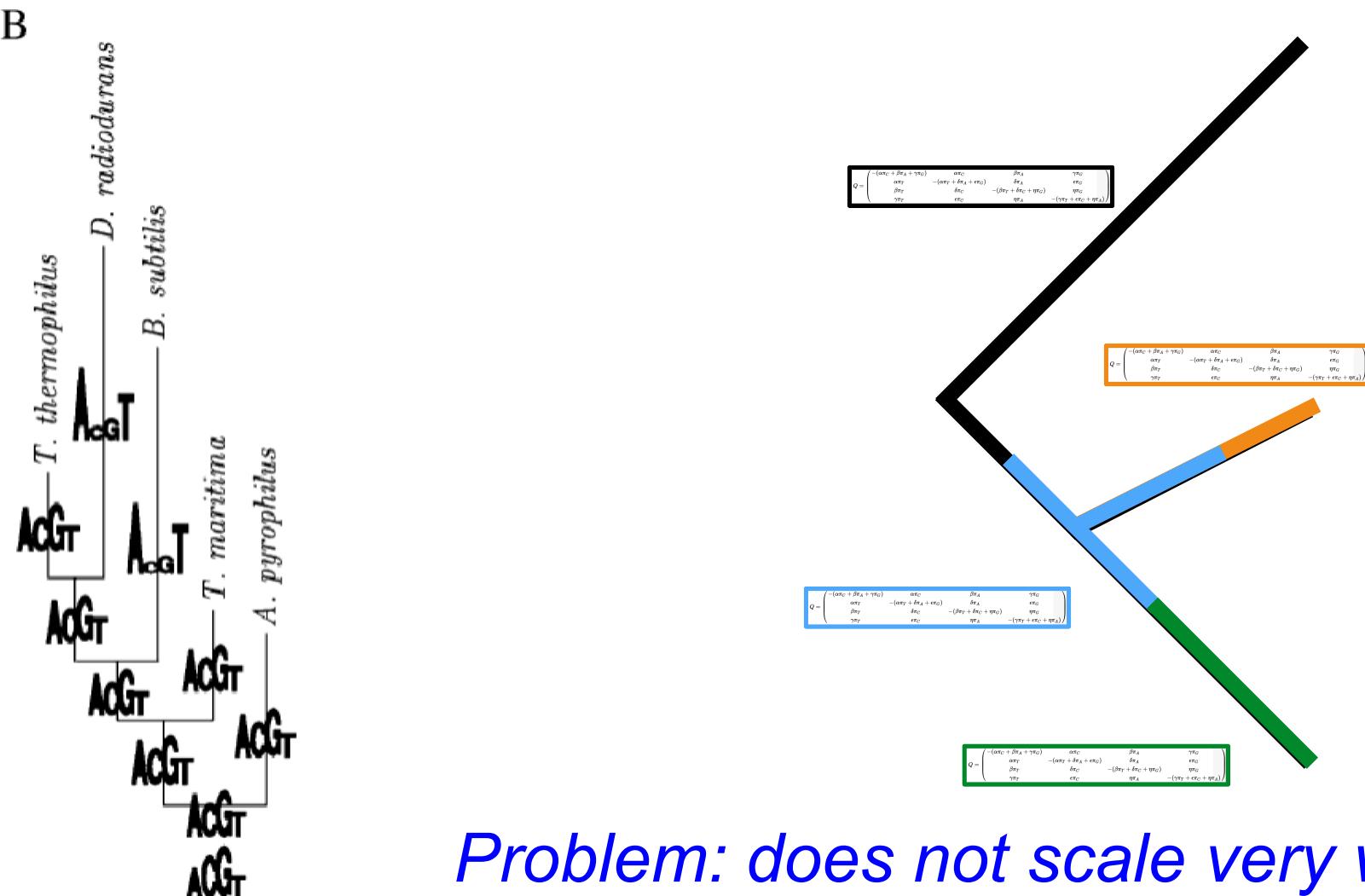
$$Q = \begin{pmatrix} -(\alpha\pi_C + \beta\pi_A + \gamma\pi_G) & \alpha\pi_C & \beta\pi_A & \gamma\pi_G \\ \alpha\pi_T & -(\alpha\pi_T + \delta\pi_A + \epsilon\pi_G) & \delta\pi_A & \epsilon\pi_G \\ \beta\pi_T & \delta\pi_T & -(\beta\pi_T + \delta\pi_A + \eta\pi_G) & \eta\pi_G \\ \gamma\pi_T & \epsilon\pi_T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

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

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

# Blanquart and Lartillot 2006

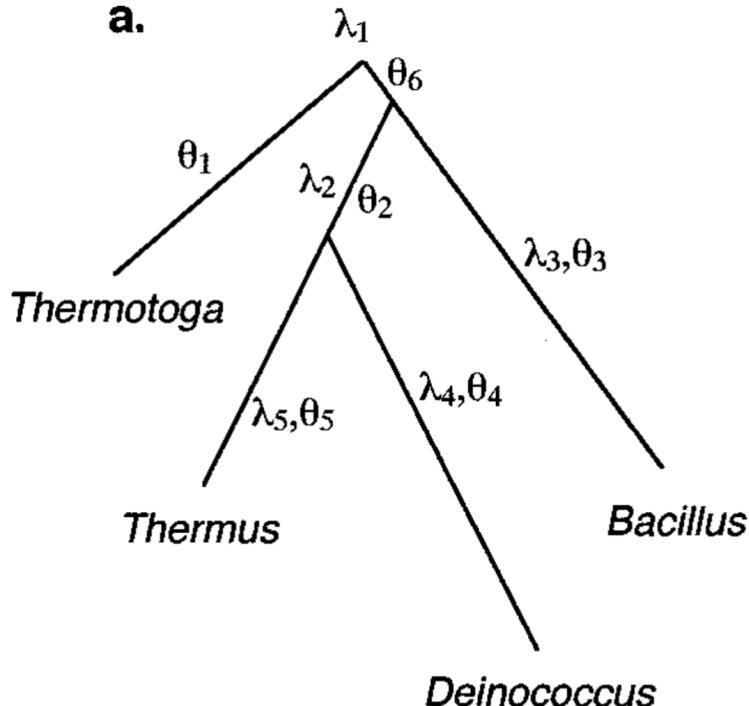
Bayesian MCMC sampling of histories of breakpoints  
between equilibrium frequencies along the phylogeny, and  
of the equilibrium frequencies themselves



# Galtier and Gouy 1998

## ML estimation of branch-wise GC equilibrium frequencies

a.



b.

	estimate	s.e.
$\kappa$	2.073	0.201
$\phi$	0.962	0.136
$\omega$	61.8%	1.5%
$\lambda_1$	0.097	0.012
$\lambda_2$	0.042	0.008
$\lambda_3$	0.111	0.012
$\lambda_4$	0.118	0.013
$\lambda_5$	0.067	0.010
$\theta_1$	81.4%	5.0%
$\theta_2$	69.7%	8.2%
$\theta_3$	28.6%	4.4%
$\theta_4$	28.3%	4.2%
$\theta_5$	71.4%	5.7%
$\theta_6$	100%	15.1%

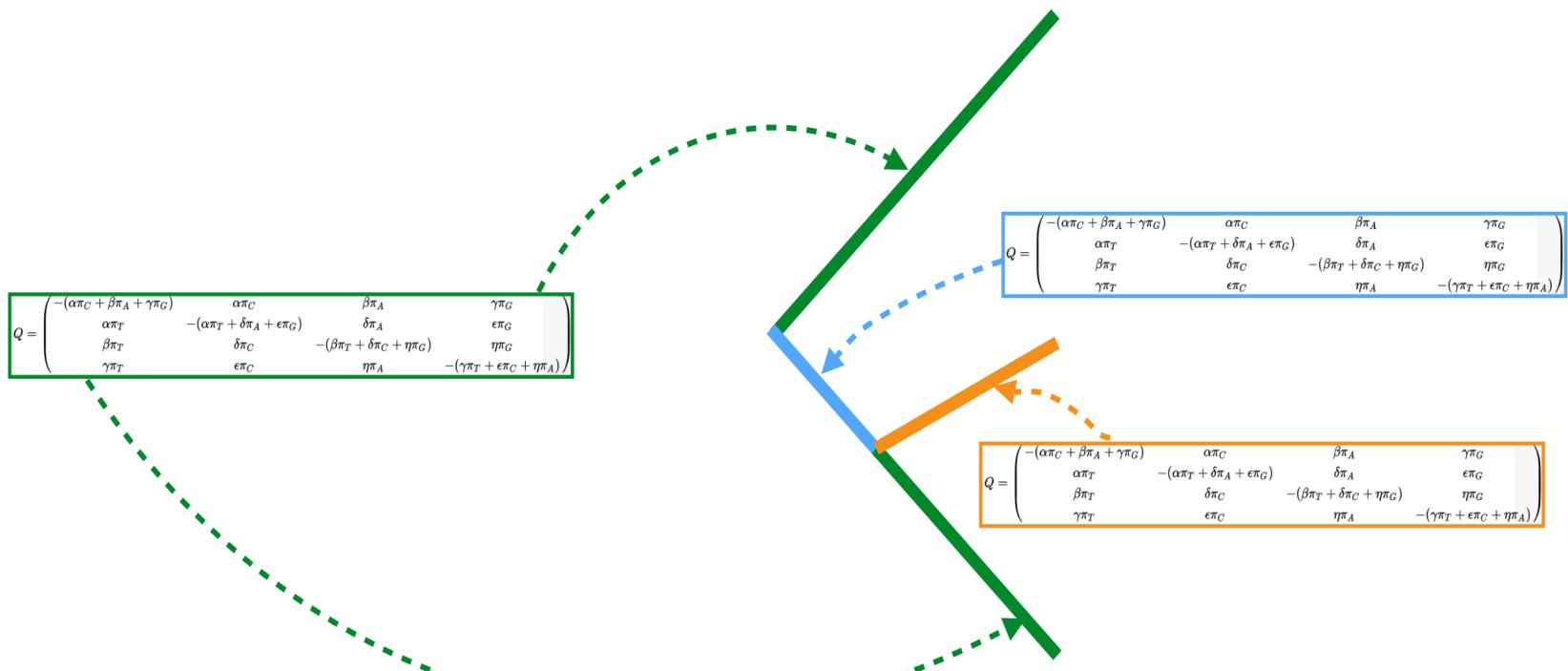
$$Q = \begin{pmatrix} -(\alpha\tau_T + \beta\tau_A + \gamma\tau_G) & \alpha\tau_C & \beta\tau_A & \gamma\tau_G \\ \alpha\tau_T & -(\alpha\tau_T + \delta\tau_A + \epsilon\tau_G) & \delta\tau_A & \epsilon\tau_G \\ \beta\tau_T & \delta\tau_C & -(\beta\tau_T + \delta\tau_C + \eta\tau_G) & \eta\tau_G \\ \gamma\tau_T & \epsilon\tau_C & \eta\tau_A & -(\gamma\tau_T + \epsilon\tau_C + \eta\tau_A) \end{pmatrix}$$

$$Q = \begin{pmatrix} -(\alpha\tau_T + \beta\tau_A + \gamma\tau_G) & \alpha\tau_C & \beta\tau_A & \gamma\tau_G \\ \alpha\tau_T & -(\alpha\tau_T + \delta\tau_A + \epsilon\tau_G) & \delta\tau_A & \epsilon\tau_G \\ \beta\tau_T & \delta\tau_C & -(\beta\tau_T + \delta\tau_C + \eta\tau_G) & \eta\tau_G \\ \gamma\tau_T & \epsilon\tau_C & \eta\tau_A & -(\gamma\tau_T + \epsilon\tau_C + \eta\tau_A) \end{pmatrix}$$

$$Q = \begin{pmatrix} -(\alpha\tau_T + \beta\tau_A + \gamma\tau_G) & \alpha\tau_C & \beta\tau_A & \gamma\tau_G \\ \alpha\tau_T & -(\alpha\tau_T + \delta\tau_A + \epsilon\tau_G) & \delta\tau_A & \epsilon\tau_G \\ \beta\tau_T & \delta\tau_C & -(\beta\tau_T + \delta\tau_C + \eta\tau_G) & \eta\tau_G \\ \gamma\tau_T & \epsilon\tau_C & \eta\tau_A & -(\gamma\tau_T + \epsilon\tau_C + \eta\tau_A) \end{pmatrix}$$

$$Q = \begin{pmatrix} -(\alpha\tau_T + \beta\tau_A + \gamma\tau_G) & \alpha\tau_C & \beta\tau_A & \gamma\tau_G \\ \alpha\tau_T & -(\alpha\tau_T + \delta\tau_A + \epsilon\tau_G) & \delta\tau_A & \epsilon\tau_G \\ \beta\tau_T & \delta\tau_C & -(\beta\tau_T + \delta\tau_C + \eta\tau_G) & \eta\tau_G \\ \gamma\tau_T & \epsilon\tau_C & \eta\tau_A & -(\gamma\tau_T + \epsilon\tau_C + \eta\tau_A) \end{pmatrix}$$

## Bayesian MCMC estimation of a limited set of equilibrium frequencies

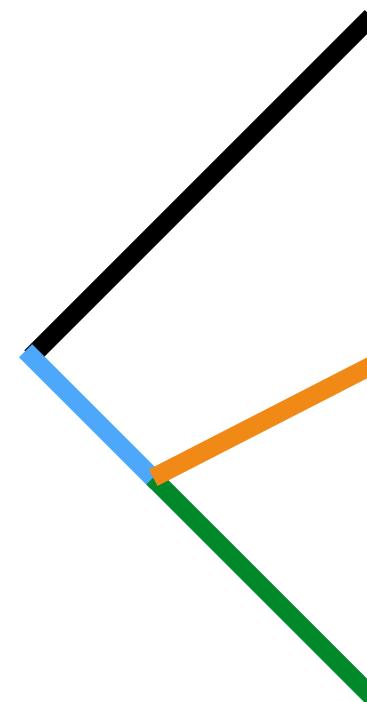


# Heaps 2014

Bayesian MCMC estimation

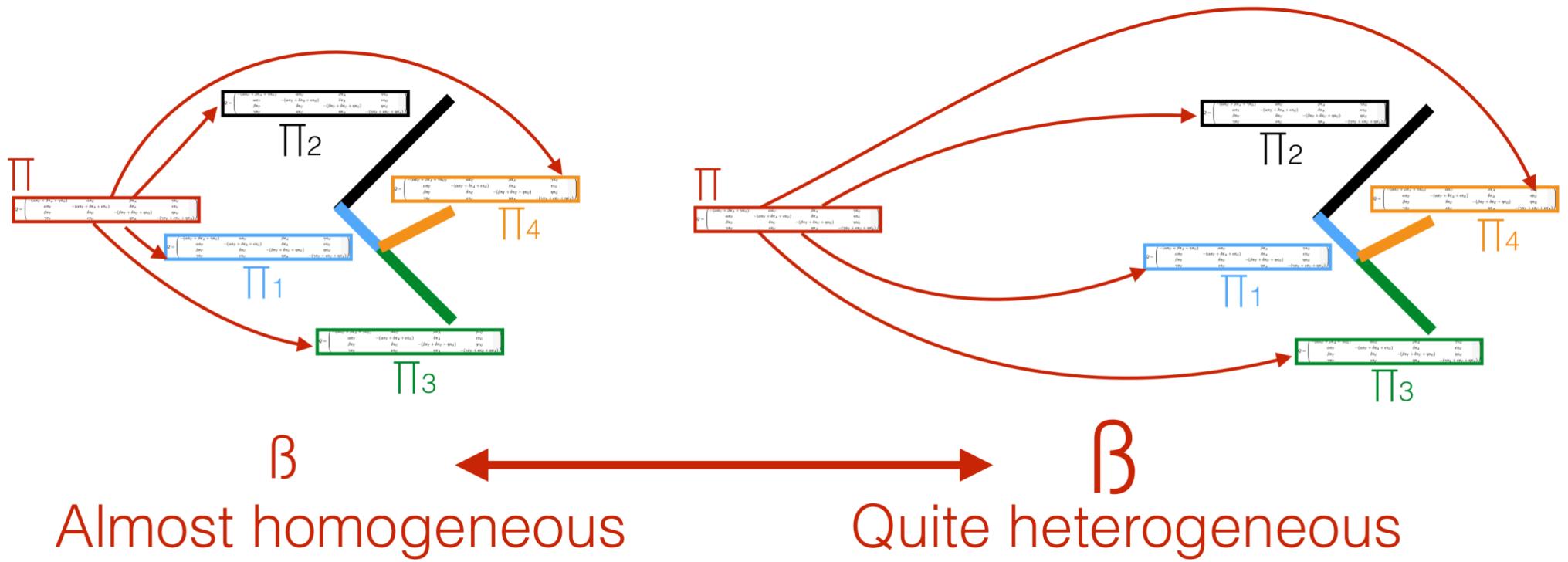
2 different models:

- ➊ a hierarchical model
- ➋ a correlated model



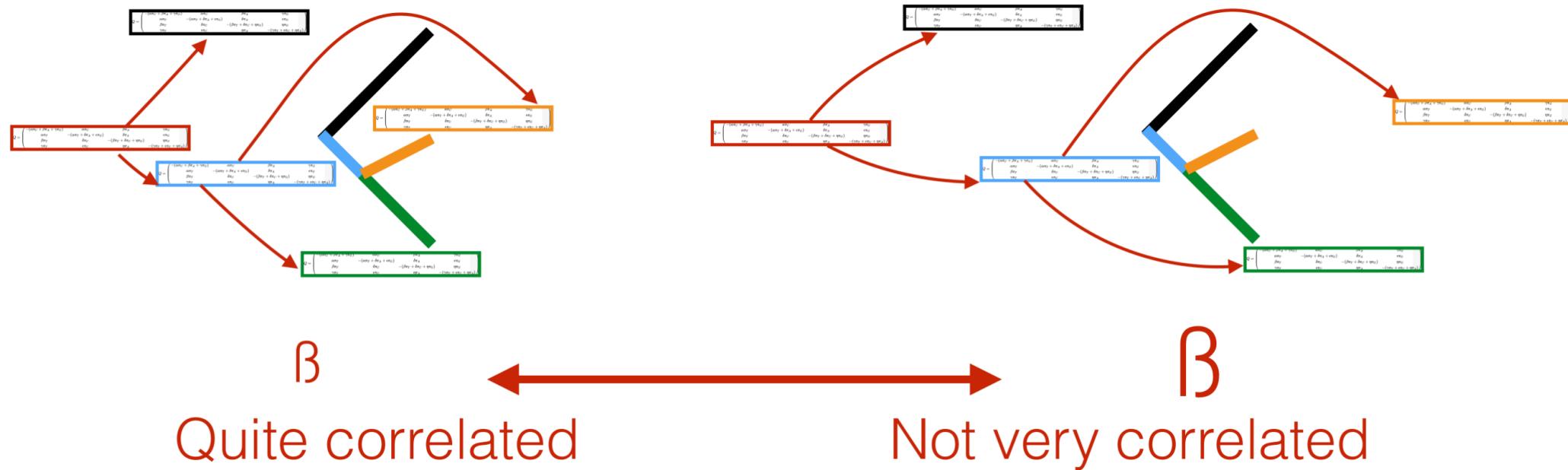
# Heaps 2014: Hierarchical model

- One central vector  $\Pi$
- Each branch  $i$  has its own vector  $\Pi_i$  in the neighbourhood of  $\Pi$



# Heaps 2014: Correlated model

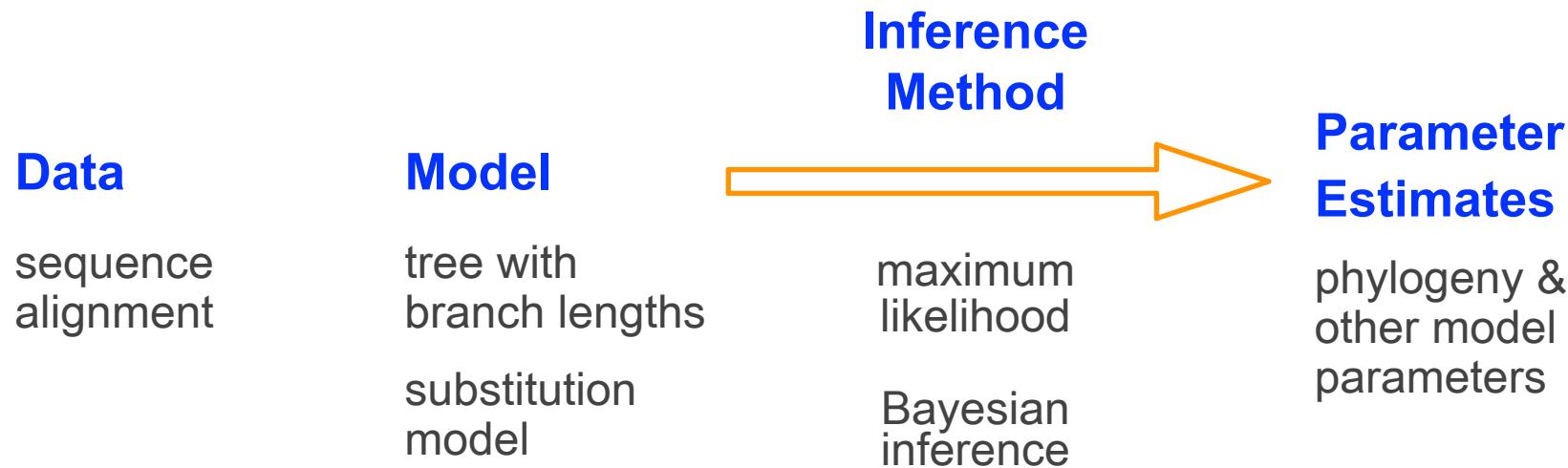
- One ancestral vector  $\Pi$
- Each branch  $i$  has its own vector  $\Pi_i$  in the neighbourhood of its parent vector  $\Pi_{pi}$



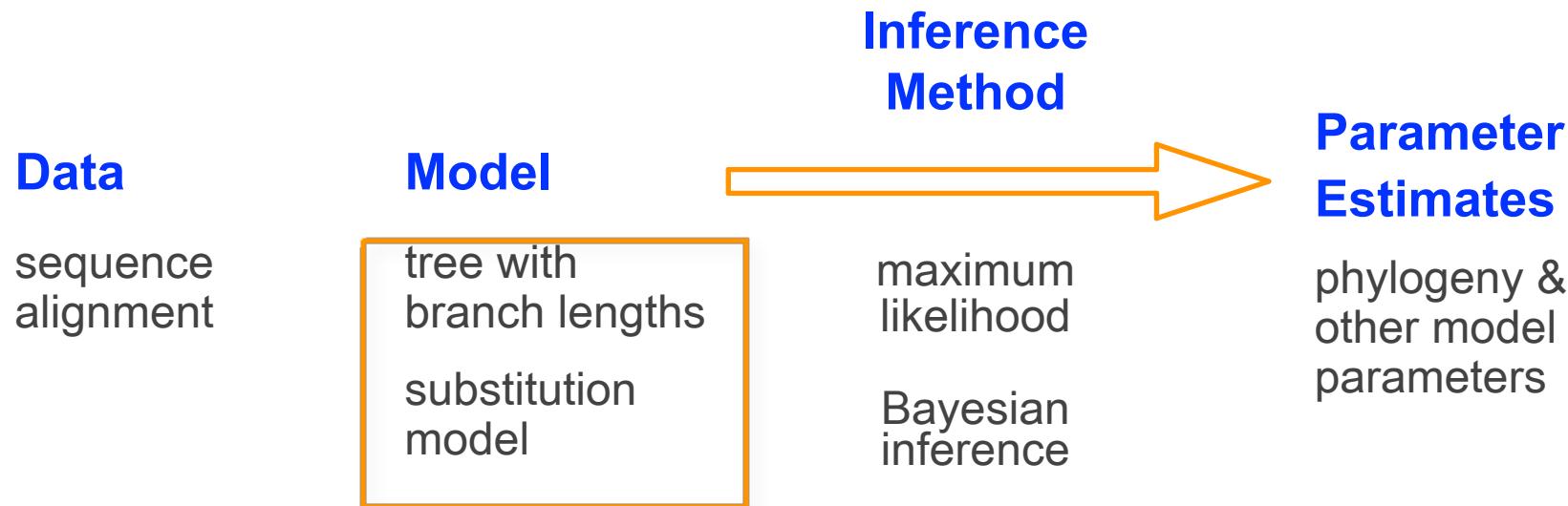
# Using models of compositional heterogeneity

- Not implemented in the most widely used software packages (except revBayes)
- Have not been used very often
- Have been used for ancestral sequence reconstruction
- Have rarely been combined with models of process heterogeneity across sites

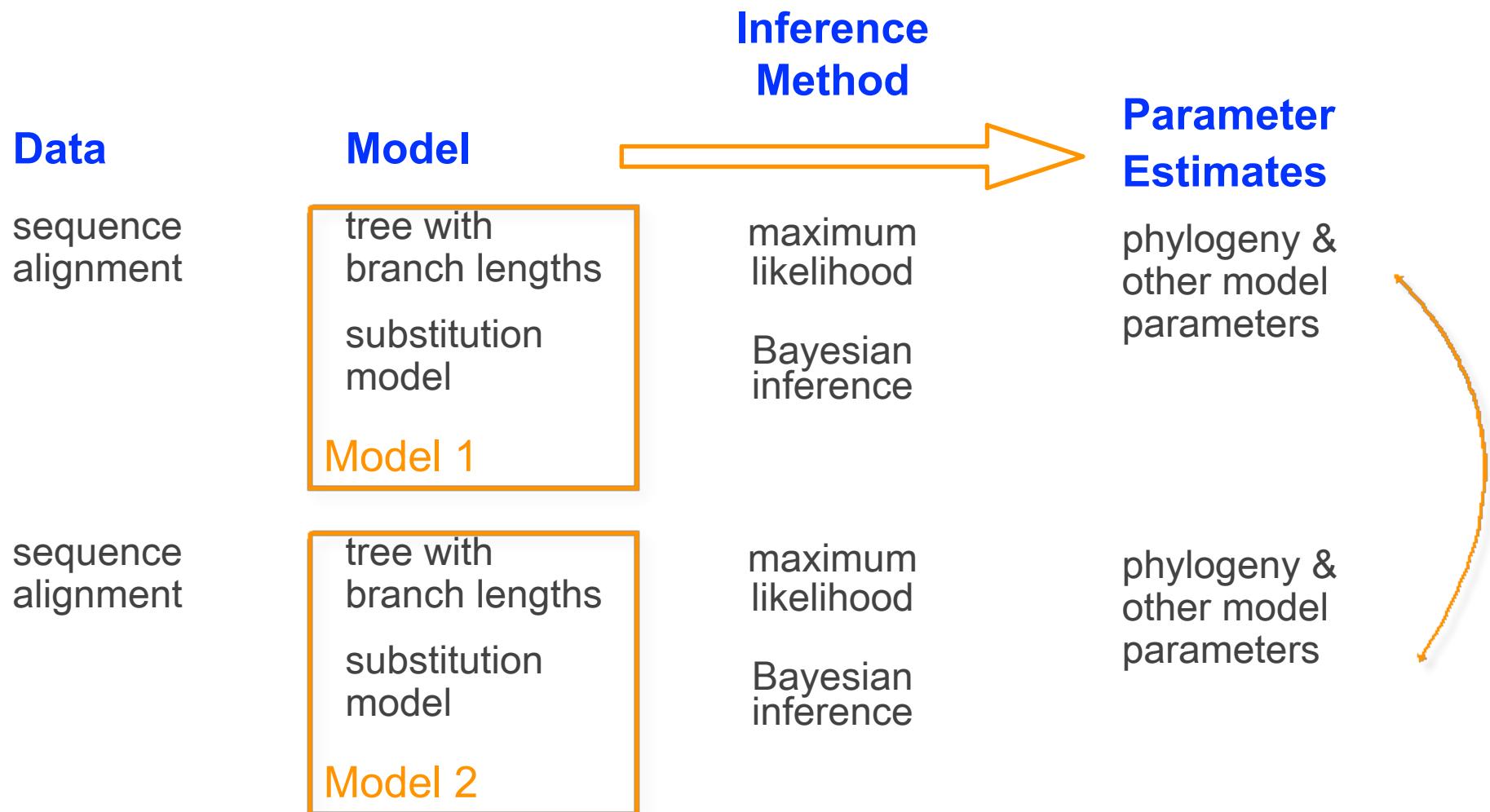
# Statistical Estimation of Phylogeny: An Outline



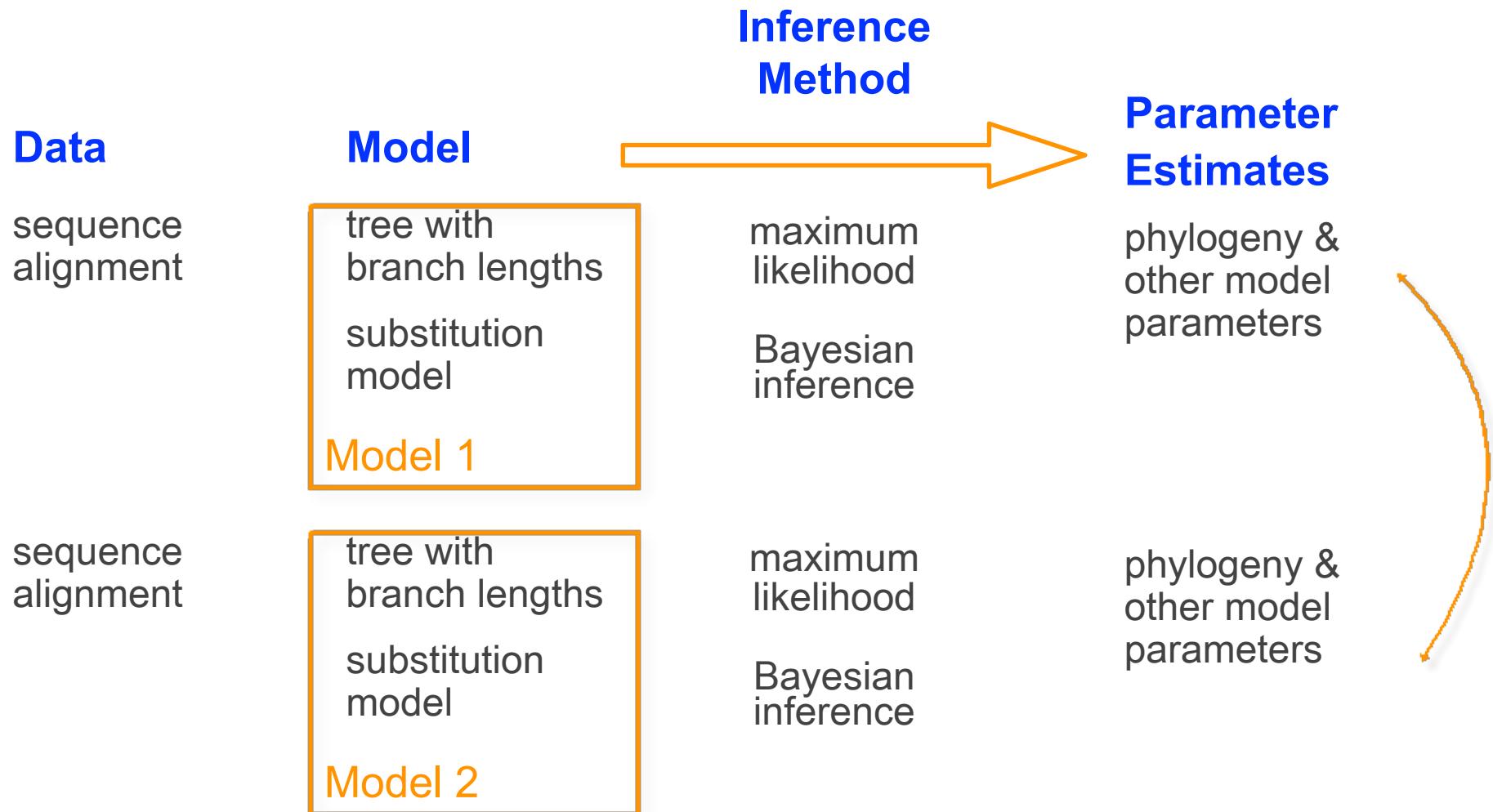
# Statistical Estimation of Phylogeny: An Outline



# Statistical Estimation of Phylogeny: An Outline

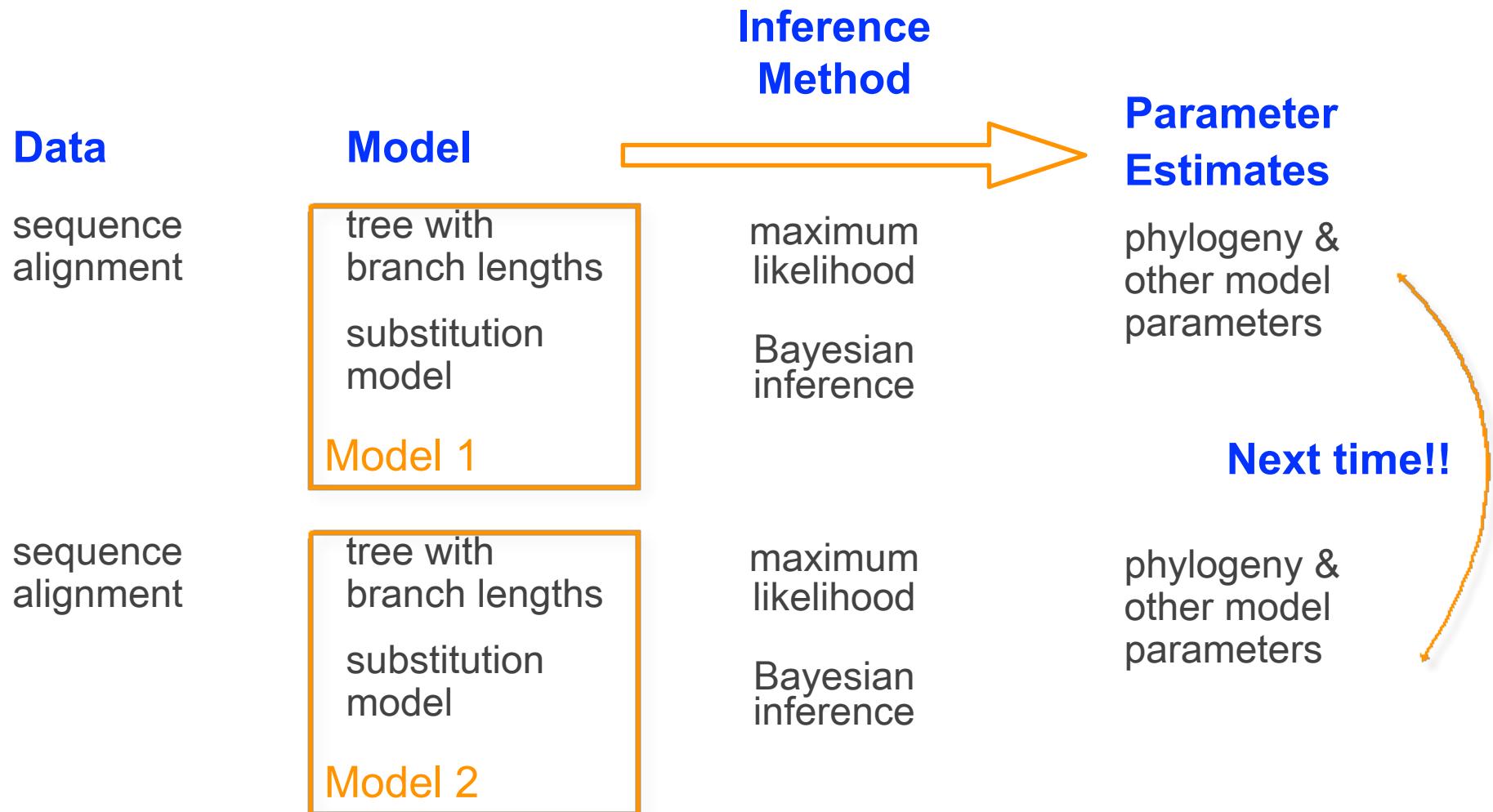


# Statistical Estimation of Phylogeny: An Outline



Model comparison is critical for obtaining unbiased parameter estimates

# Statistical Estimation of Phylogeny: An Outline



Model comparison is critical for obtaining unbiased parameter estimates

Model comparison is the means by which we test hypotheses about our data