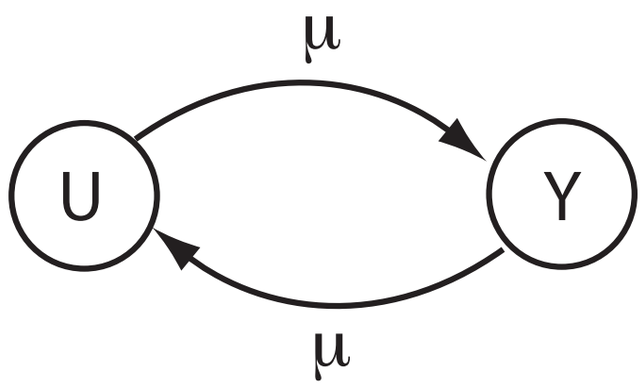
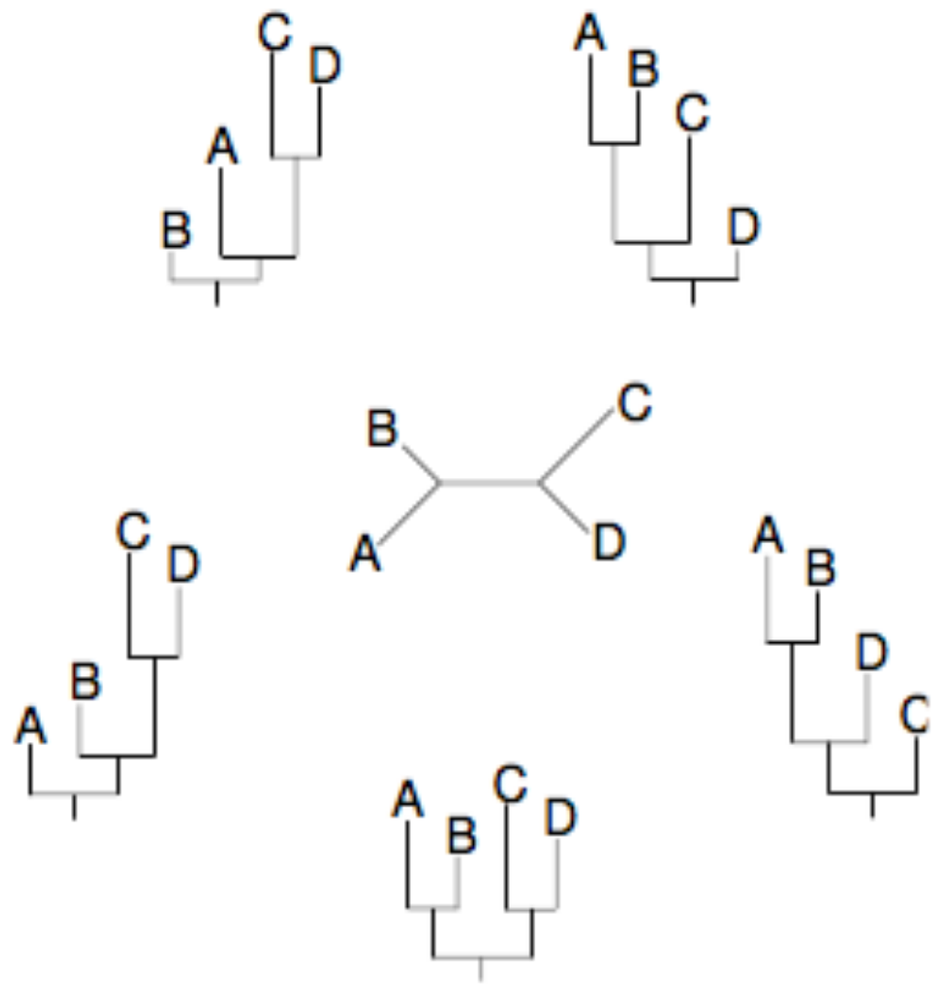
# Simulating sequence data on a phylogenetic tree

- **Tree relating samples to each other**
- **Branches with known length**
- **Mutation model**

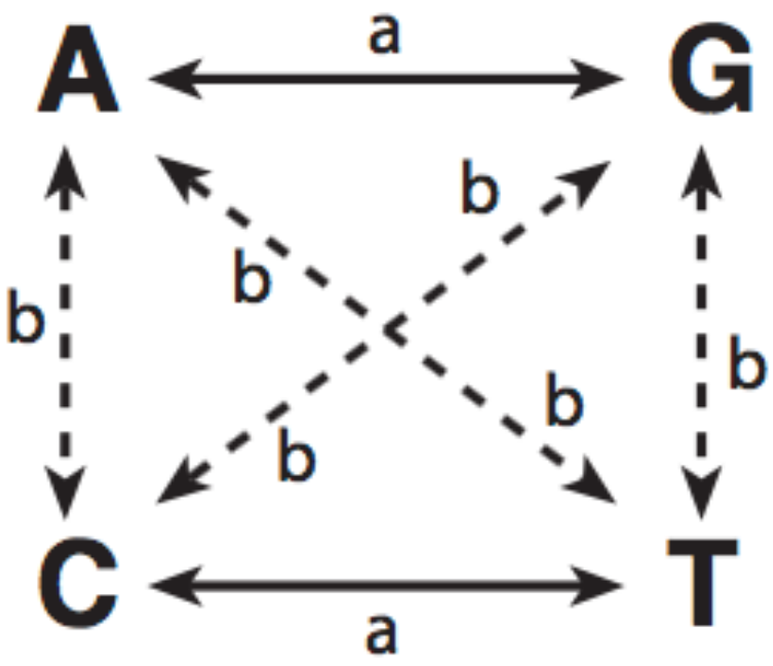


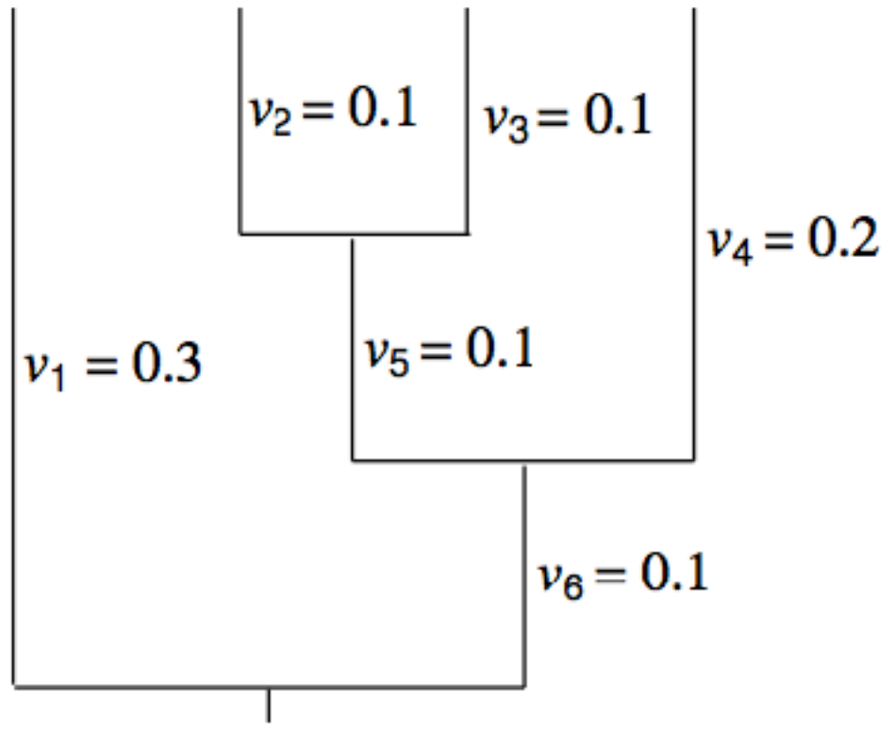



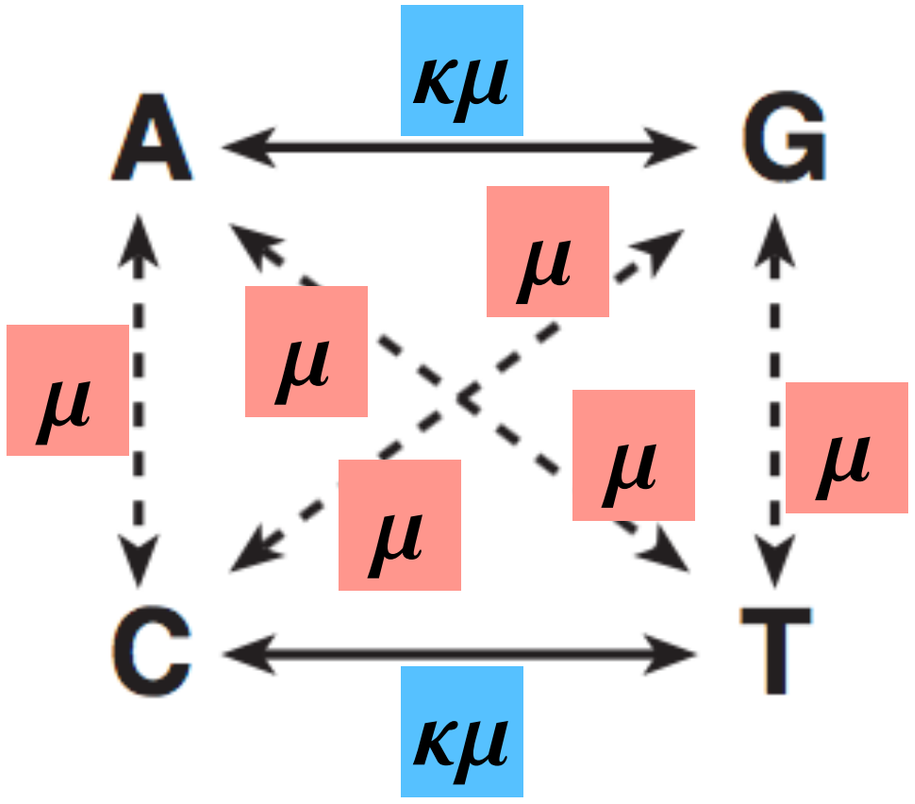

$$R = \begin{pmatrix} 1-\mu & \mu \\ \mu & 1-\mu \end{pmatrix}$$

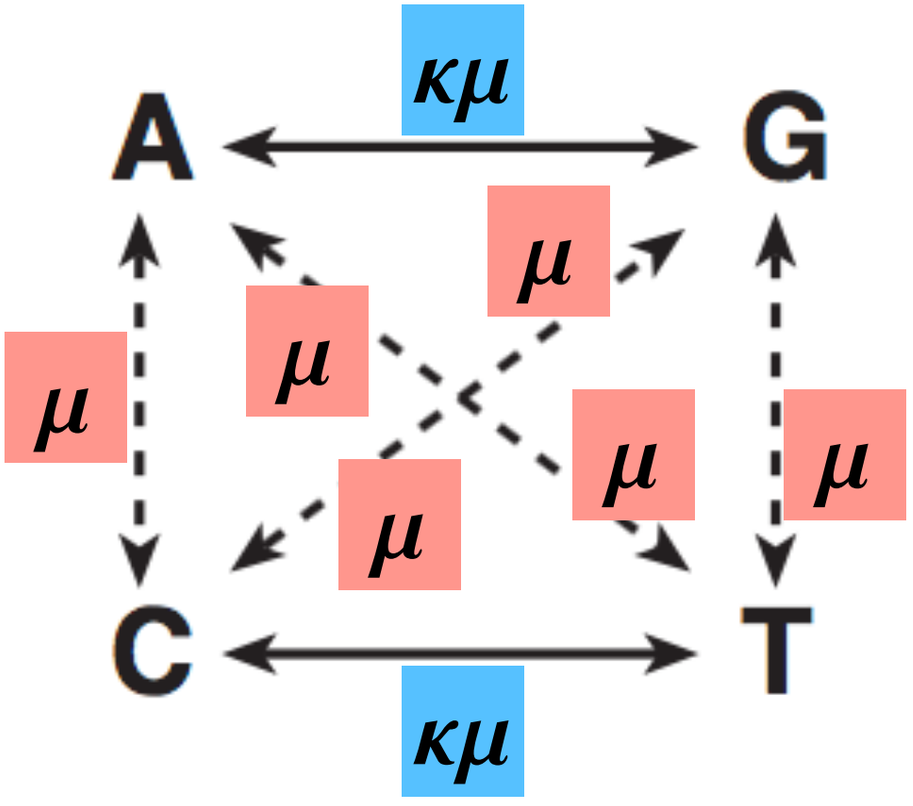
$$R_* = \begin{pmatrix} -\mu & \mu \\ \mu & -\mu \end{pmatrix}$$

$$R = I + R_*$$



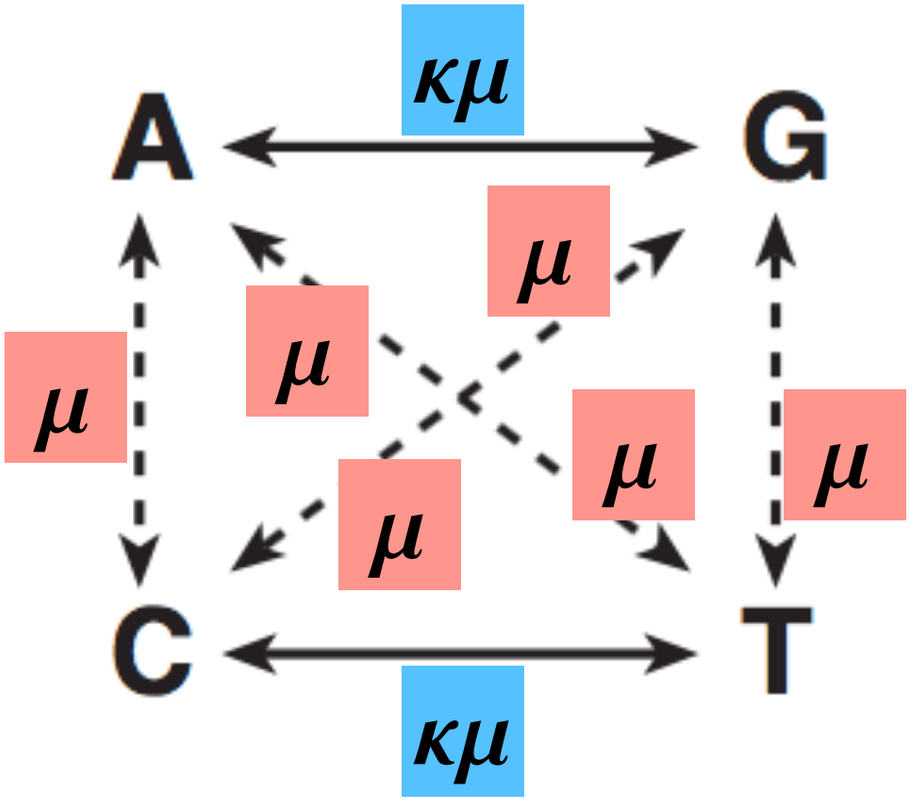




κμ
A
G
μ
μ
μ
μ
μ
μ
C
T
κμ




**Hasegawa-Kishino-Yano model**

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




**Hasegawa-Kishino-Yano model**

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

**We do not know $\mu$ and we substitute it with the expected number of nucleotide substitutions per site. For example we could assume that the mean rate of substitutions is 1.0**

$$\mu = \frac{1}{\sum_{i \in \{A,C,G,T\}} \pi_i q_{ii}}$$

- **Transition/Transversion**
- **Stationary distribution**

Let us consider a specific example of a rate matrix, with all of the parameters of the model taking specific values. For example, if we use the HKY85 model and fix the parameters to $\kappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$, we get the following matrix of instantaneous rates

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

Let us consider a specific example of a rate matrix, with all of the parameters of the model taking specific values. For example, if we use the HKY85 model and fix the parameters to $\kappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$, we get the following matrix of instantaneous rates

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

**Focusing on G**

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1.266 & 0.190 & -1.519 & 0.063 \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

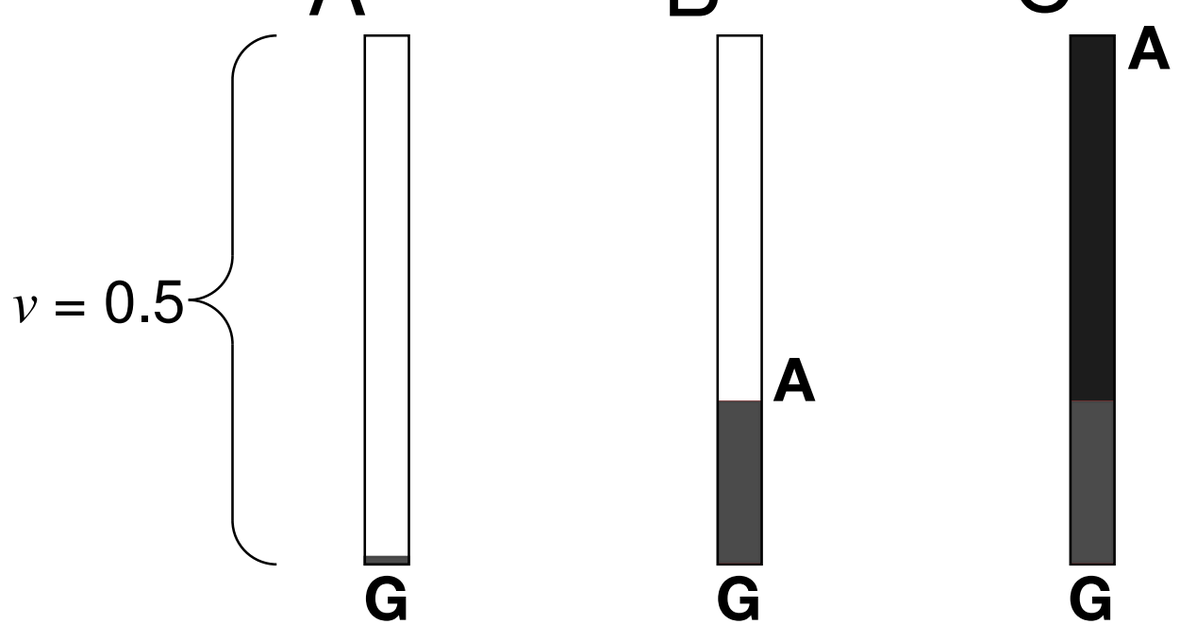



Figure 2: **Simulation under the HKY85 substitution process**. A single realization of the substitution process under the HKY85 model when $\kappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$. The length of the branch is $v = 0.5$ and the starting nucleotide is $G$ (light gray). A, The process starts in nucleotide G. B, The first change is 0.152 units up the branch. C, the change is from G to A (dark gray). The time at which the next change occurs exceeds the total branch length, so the process ends in state **A**.

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1.266 & 0.190 & -1.519 & 0.063 \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} \qquad \mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

**We draw a time using an exponential** $t = -\frac{1}{\lambda}\log_e(u)$

**We choose the next nucleotide**

$$G \to A : \frac{1.266}{1.519} = 0.833, \quad G \to C : \frac{0.190}{1.519} = 0.125, \quad G \to T : \frac{0.063}{1.519} = 0.042$$

**In the example we picked an A, now we continue with A: pick time, pick next nucleotide**

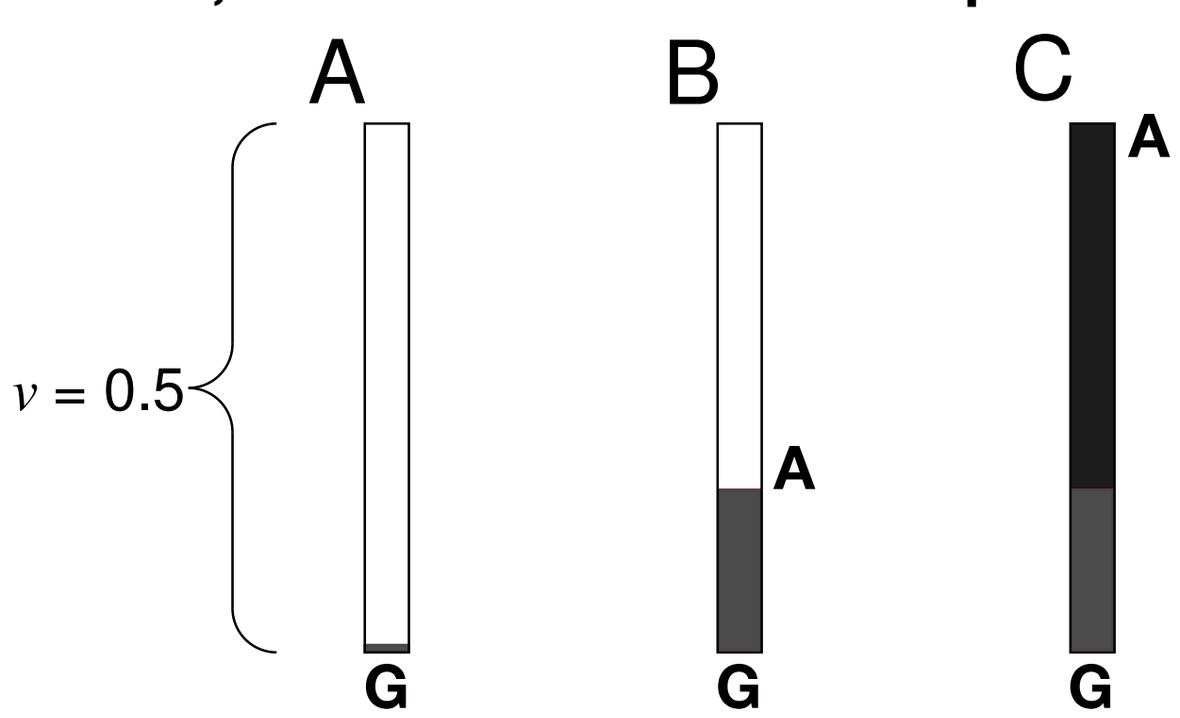



Figure 2: **Simulation under the HKY85 substitution process**. A single realization of the substitution process under the HKY85 model when $\kappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$. The length of the branch is $v = 0.5$ and the starting nucleotide is $G$ (light gray). A, The process starts in nucleotide G. B, The first change is 0.152 units up the branch. C, the change is from G to A (dark gray). The time at which the next change occurs exceeds the total branch length, so the process ends in state **A**.

The following table summarizes the results of 100 simulations, each of which started with the nucleotide $G$:

**Starting with G**

| Starting Nucleotide | Ending Nucleotide | Number of Replicates |
|---|---|---|
| G | A | 27 |
| G | C | 10 |
| G | G | 59 |
| G | T | 4 |

**Starting with A,C,G,T**

| | | Ending Nucleotide | | | |
|---|---|---|---|---|---|
| | | A | C | G | T |
| Starting Nucleotide | A | 0.67 | 0.13 | 0.20 | 0.00 |
| | C | 0.13 | 0.70 | 0.07 | 0.10 |
| | G | 0.27 | 0.10 | 0.59 | 0.04 |
| | T | 0.12 | 0.30 | 0.08 | 0.50 |

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

For the case we have been simulating, the exact transition probabilities (to four decimal places) are

$$\mathbf{P}(0.5) = \{p_{ij}(0.5)\} = \begin{pmatrix} 0.7079 & 0.0813 & 0.1835 & 0.0271 \\ 0.1085 & 0.7377 & 0.0542 & 0.0995 \\ 0.3670 & 0.0813 & 0.5244 & 0.0271 \\ 0.1085 & 0.2985 & 0.0542 & 0.5387 \end{pmatrix}$$

## Stationary Distribution

$$\mathbf{P}(0.00) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \quad \mathbf{P}(0.01) = \begin{pmatrix} 0.991 & 0.002 & 0.006 & 0.001 \\ 0.003 & 0.993 & 0.001 & 0.003 \\ 0.013 & 0.002 & 0.985 & 0.001 \\ 0.003 & 0.009 & 0.001 & 0.987 \end{pmatrix}$$

$$\mathbf{P}(0.10) = \begin{pmatrix} 0.919 & 0.018 & 0.056 & 0.006 \\ 0.024 & 0.934 & 0.012 & 0.029 \\ 0.113 & 0.018 & 0.863 & 0.006 \\ 0.025 & 0.086 & 0.012 & 0.877 \end{pmatrix} \quad \mathbf{P}(0.50) = \begin{pmatrix} 0.708 & 0.081 & 0.184 & 0.027 \\ 0.106 & 0.738 & 0.054 & 0.100 \\ 0.367 & 0.081 & 0.524 & 0.027 \\ 0.109 & 0.299 & 0.054 & 0.539 \end{pmatrix}$$

$$\mathbf{P}(1.00) = \begin{pmatrix} 0.580 & 0.141 & 0.232 & 0.047 \\ 0.188 & 0.587 & 0.094 & 0.131 \\ 0.464 & 0.141 & 0.348 & 0.047 \\ 0.188 & 0.394 & 0.094 & 0.324 \end{pmatrix} \quad \mathbf{P}(5.00) = \begin{pmatrix} 0.411 & 0.287 & 0.206 & 0.096 \\ 0.383 & 0.319 & 0.192 & 0.106 \\ 0.411 & 0.287 & 0.206 & 0.096 \\ 0.383 & 0.319 & 0.192 & 0.107 \end{pmatrix}$$

$$\mathbf{P}(10.0) = \begin{pmatrix} 0.401 & 0.299 & 0.200 & 0.099 \\ 0.399 & 0.301 & 0.199 & 0.100 \\ 0.401 & 0.299 & 0.200 & 0.099 \\ 0.399 & 0.301 & 0.199 & 0.100 \end{pmatrix} \quad \mathbf{P}(100) = \begin{pmatrix} 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \end{pmatrix}$$

## Stationary Distribution

$$\mathbf{P}(0.00) = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \quad \mathbf{P}(0.01) = \begin{pmatrix} 0.991 & 0.002 & 0.006 & 0.001 \\ 0.003 & 0.993 & 0.001 & 0.003 \\ 0.013 & 0.002 & 0.985 & 0.001 \\ 0.003 & 0.009 & 0.001 & 0.987 \end{pmatrix}$$

$$\mathbf{P}(0.10) = \begin{pmatrix} 0.919 & 0.018 & 0.056 & 0.006 \\ 0.024 & 0.934 & 0.012 & 0.029 \\ 0.113 & 0.018 & 0.863 & 0.006 \\ 0.025 & 0.086 & 0.012 & 0.877 \end{pmatrix} \quad \mathbf{P}(0.50) = \begin{pmatrix} 0.708 & 0.081 & 0.184 & 0.027 \\ 0.106 & 0.738 & 0.054 & 0.100 \\ 0.367 & 0.081 & 0.524 & 0.027 \\ 0.109 & 0.299 & 0.054 & 0.539 \end{pmatrix}$$

$$\mathbf{P}(1.00) = \begin{pmatrix} 0.580 & 0.141 & 0.232 & 0.047 \\ 0.188 & 0.587 & 0.094 & 0.131 \\ 0.464 & 0.141 & 0.348 & 0.047 \\ 0.188 & 0.394 & 0.094 & 0.324 \end{pmatrix} \quad \mathbf{P}(5.00) = \begin{pmatrix} 0.411 & 0.287 & 0.206 & 0.096 \\ 0.383 & 0.319 & 0.192 & 0.106 \\ 0.411 & 0.287 & 0.206 & 0.096 \\ 0.383 & 0.319 & 0.192 & 0.107 \end{pmatrix}$$

$$\mathbf{P}(10.0) = \begin{pmatrix} 0.401 & 0.299 & 0.200 & 0.099 \\ 0.399 & 0.301 & 0.199 & 0.100 \\ 0.401 & 0.299 & 0.200 & 0.099 \\ 0.399 & 0.301 & 0.199 & 0.100 \end{pmatrix} \quad \mathbf{P}(100) = \begin{pmatrix} 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \\ 0.400 & 0.300 & 0.200 & 0.100 \end{pmatrix}$$

We used HKY85 model with these parameters to $\varkappa = 5$, $\pi_A = 0.4$, $\pi_C = 0.3$, $\pi_G = 0.2$, and $\pi_T = 0.1$

# Simulating data

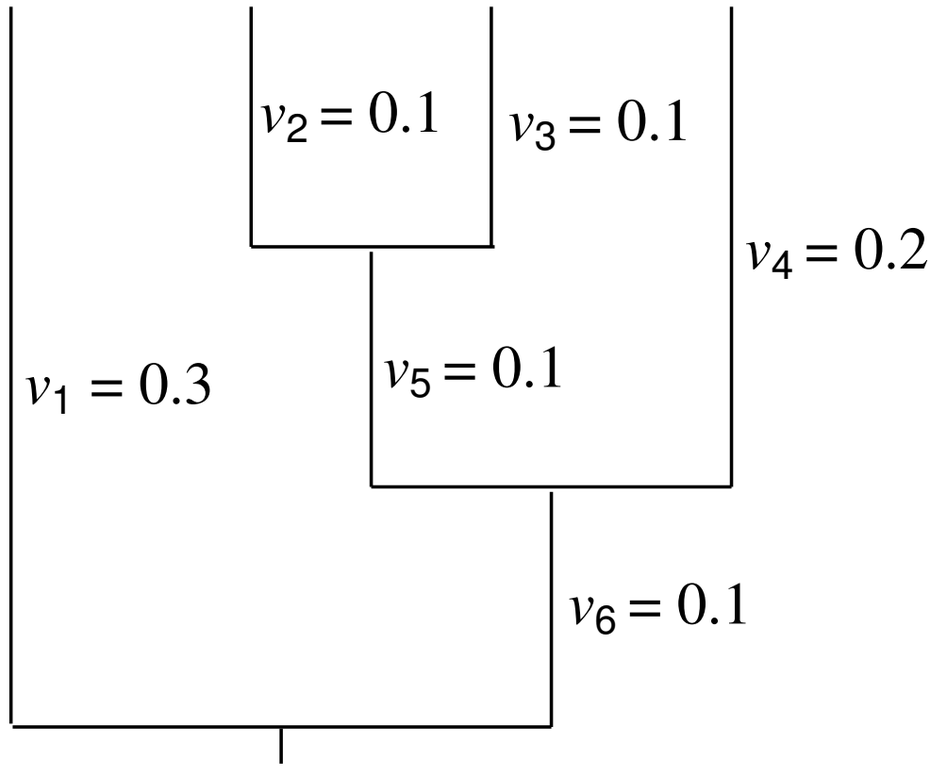



Figure 4: **The model tree for the simulations**. We simulate data on this tree. The branch lengths are denoted $v_i$.

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

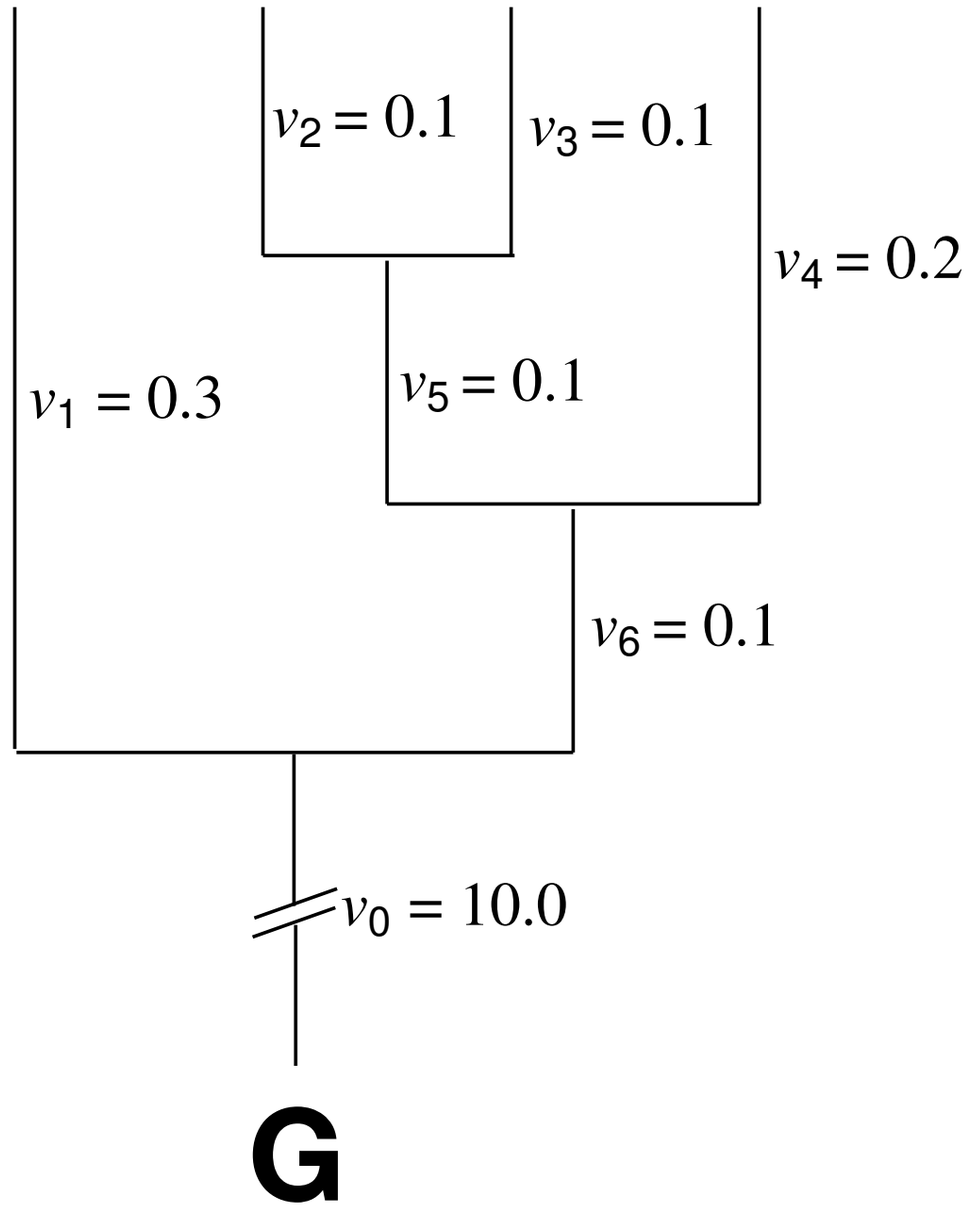
# Four methods to simulate data

1. we use our exponential random number generator, and add a tail to the tree and start with an A
2. No long tail, but pick root state from base frequencies
3. We have only three different branch length, we calculate the probability matrix and us

   that to generate the nucleotides at the nodes
4. We generate all site patterns and then pick among them (this will be discussed when know about likelihood.

## Exercise: generate at least two sites at the tips

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$




## Start with G

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1.266 & 0.190 & -1.519 & 0.063 \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

## We draw a time using an exponential

$$t = -\frac{1}{\lambda}\log_e(u)$$

## We choose the next nucleotide

$$G \to A : \frac{1.266}{1.519} = 0.833, \quad G \to C : \frac{0.190}{1.519} = 0.125, \quad G \to T : \frac{0.063}{1.519} = 0.042$$

The third simulation method does away with the need to generate exponential random variables. It takes advantage of our knowledge of the stationary distribution (as does the second method), but also takes advantage of our ability to calculate transition probabilities. There are only three different lengths of branches on our model tree (0.1, 0.2, and 0.3). The transition probabilities are

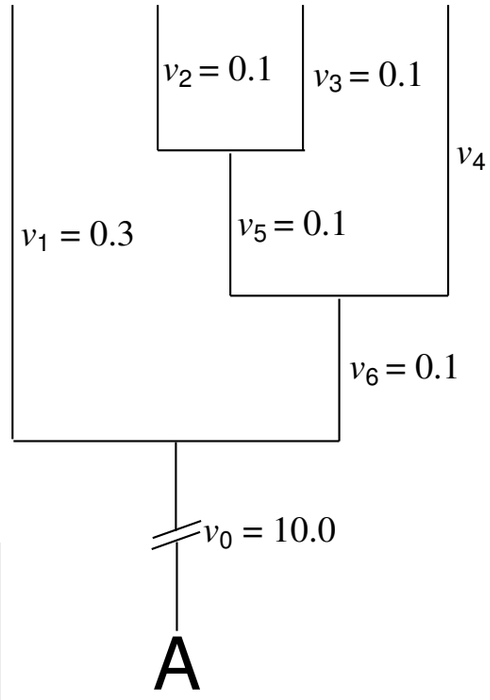
$$\mathbf{P}(0.10) = \begin{pmatrix} 0.919 & 0.018 & 0.056 & 0.006 \\ 0.024 & 0.934 & 0.012 & 0.029 \\ 0.113 & 0.018 & 0.863 & 0.006 \\ 0.025 & 0.086 & 0.012 & 0.877 \end{pmatrix} \quad \mathbf{P}(0.20) = \begin{pmatrix} 0.851 & 0.035 & 0.100 & 0.011 \\ 0.047 & 0.876 & 0.023 & 0.052 \\ 0.201 & 0.035 & 0.750 & 0.011 \\ 0.047 & 0.156 & 0.023 & 0.771 \end{pmatrix}$$

$$\mathbf{P}(0.30) = \begin{pmatrix} 0.795 & 0.051 & 0.135 & 0.017 \\ 0.069 & 0.824 & 0.034 & 0.071 \\ 0.270 & 0.051 & 0.659 & 0.017 \\ 0.069 & 0.214 & 0.034 & 0.681 \end{pmatrix}$$

Instead of drawing exponential random variables, and generating the process continuously across the entire tree, our simulation jumps from node to node on the tree. First, we generate the nucleotide at the root of the tree (the first split leading to all four taxa) by drawing from the stationary distribution. Then, we use the transition probabilities to simulate from one end to the other of each branch on the tree. We start from the root of the tree, and simulate up the tree to progressively higher branches until we have simulated a nucleotide at each tip of the tree.

# Algorithm to simulate data on a tree

Required: a tree with branch lengths and a mutation transition rate matrix




$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

1. $\tau = 0$; nuc $= \{\}$
2. Do forever
3. find $\lambda$ — if you are at **G** use the diagonal value row of $q_G$: 0.886
4. draw random number $r$ between 0 and 1 — $r_1 = 0.134$
5. calculate $t = -\log_e(r_1)\frac{1}{\lambda}$ — e.g. $t = -\log(0.134)\frac{1}{0.886} = 0.98521$
6. $\tau = \tau + t$ — $\tau = 0 + 0.98521$
7. if $\tau < v$ — $v = 10.0$

   return current nucleotide nuc
8. calculate change of nucleotide cumulative sum — $s = [0, \frac{q_{GA}}{q_{GG}}, \frac{q_{GA}}{q_{GG}} + \frac{q_{GC}}{q_{GG}}, 1.0]$

   $s = [0, \frac{1.266}{1.519}, \frac{1.266}{1.519} + \frac{0.190}{1.519}, 1.0]$
9. draw random number $r$ between 0 and 1 — $r_2 = 0.912$
10. pick interval in which $r_2$ lays — it is in the interval $(0.833, 0.958]$ and thus nuc=**C**
11. goto 2

Here is a second round of the example above: we are at **C** now

1. do forever
2. $\lambda = 0.696$ using row $q_C$
3. $r_1 = 0.449$
4. $t = -\log_e(0.449)\frac{1}{0.696} = 0.49964$
5. $\tau = 0.98521 + 0.49964 = 1.48485$
6. if $(\tau = 0.98521) < 10.0$

   return current nucleotide nuc
7. $s = [0, \frac{0.253}{0.696}, \frac{0.253}{0.696} + \frac{0.127}{0.696}, 1.0]$
8. $r_2 = 0.191$
9. pick interval in which $r_2$ lays it is in the interval $(0, 0.363]$ and thus nuc=**A**
10. goto 2

$$\begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

**GTR**

$$a, b, c, d, e, \pi_A, \pi_C, \pi_G$$
$$a + b + c + d + e + f = 1,$$
$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

**+all rates are different**

**TN**

$$a = c = d = f, b = e, \pi_A, \pi_C, \pi_G$$
$$T_i/T_v, Y/R \text{ ratio and three frequencies}$$
$$\text{since } \pi_A + \pi_C + \pi_G + \pi_T = 1$$

**+Purine/Pyrimidins**

**HKY F84**

$$a = f = c = d = 1, b = e, \pi_A, \pi_C, \pi_G$$
$$T_i/T_v \text{ ratio and three frequencies}$$
$$\text{since } \pi_A + \pi_C + \pi_G + \pi_T = 1$$

**+Transition/Transversion**

**F81**

$$a = b = c = d = e = f = 1, \pi_A, \pi_C, \pi_G$$
$$\text{and three frequencies since}$$
$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

**+uneqal base freq.**

**K2P**

$$a = f = c = d = 1, b = e,$$
$$\pi_A = \pi_C = \pi_G = \pi_T$$
$$T_i/T_b \text{ ratio}$$

**+uneqal base freq.**

**+Transition/Transversion**

**JC**

$$a = b = c = d = e = f = 1,$$
$$\pi_A = \pi_C = \pi_G = \pi_T$$