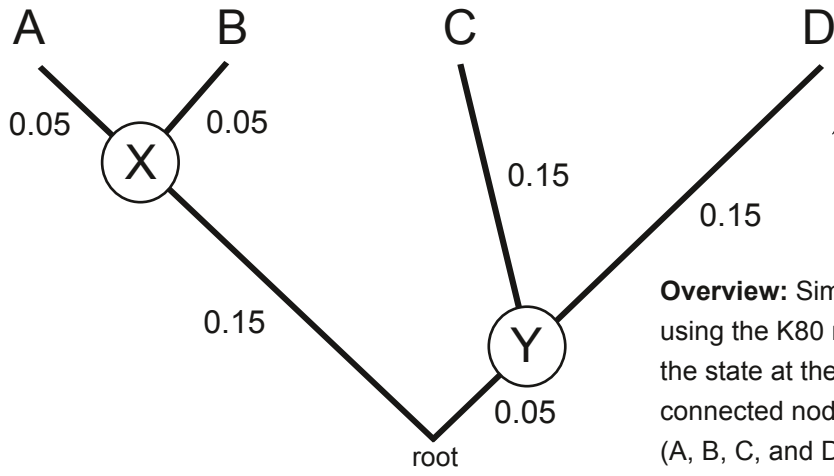
Homework 4: Simulating data on a tree



$$\kappa = 5$$

$$\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$$

Overview: Simulate data for one DNA site on the tree shown using the K80 model with the kappa value 5. Start by choosing the state at the root, then proceed to choose the state at each connected node until you have chosen a state for all four taxa (A, B, C, and D) as well as the ancestral nodes X and Y.

Start by using R to generate some pseudorandom numbers. These will be used to choose the state at each node. **Start by choosing a seed (totally your choice!) and recording it in the box provided on the next page.** Use the command **set.seed(12345)** to set the seed (replace 12345 with a number you chose). Use the command **runif(7)** to output 7 random variates from a uniform(0,1) distribution. Please **use each number in the order they were generated**. Don't skip any numbers because I will use your seed to check your answers and it complicates things if you don't use them in the order generated.

For each node, you will fill out a table like the one shown below, listing the four possible DNA states (A, C, G, T), the probability of seeing each state, and the cumulative probability from left to right.

Use your first uniform random variate to decide the starting state at the root. For example, you would choose A if your random number u is less than 0.25. You would choose C if 0.25 < u < 0.50, choose G if 0.50 < u < 0.75, and choose T if u > 0.75.

	A	C	G	T
probability	0.25	0.25	0.25	0.25
cumulative	0.25	0.50	0.75	1.00

For example, if the first random uniform number is u = 0.67981, then you would choose G as the root state. This is analogous to spinning the "Picker Wheel" (https://pickerwheel.com/) that I showed in class. The picker wheel does not allow anyone to repeat our choices, and would be incredibly tedious if you had to choose more than a handful of states, so using a sequence of pseudorandom numbers is better.

Proceed to use each subsequent uniform random variate to **choose the state at a node** connected to one you've already determined. You will need to **fill out each table** on the next page using the edge length appropriate for the state being chosen. The transition probabilities for the K80 model are provided below.

K80 Transition Probabilities

$\frac{1}{4} - \frac{1}{4}e^{-\frac{4v}{\kappa+2}}$	$\frac{1}{4} + \frac{1}{4}e^{-\frac{4v}{\kappa+2}} - \frac{1}{2}e^{-\frac{2v(\kappa+1)}{\kappa+2}}$	$\frac{1}{4} + \frac{1}{4}e^{-\frac{4v}{\kappa+2}} + \frac{1}{2}e^{-\frac{2v(\kappa+1)}{\kappa+2}}$
transversion-type substitution	transition-type substitution	no change across edge

Use **5 decimal places** when recording values that are not whole numbers.

Seed:

Root state:

1st uniform variate

X state:

2nd uniform variate

edge length = 0.15	A	C	G	T
probability				
cumulative				

Y state:

3rd uniform variate

edge length = 0.05	A	C	G	T
probability				
cumulative				

A state:

4th uniform variate

edge length = 0.05	A	C	G	T
probability				
cumulative				

B state:

5th uniform variate

edge length = 0.05	A	C	G	T
probability				
cumulative				

C state:

6th uniform variate

edge length = 0.15	A	C	G	T
probability				
cumulative				

D state:

7th uniform variate

edge length = 0.15	A	C	G	T
probability				
cumulative				