Simulate data on a tree [first version]

Due data is 11:59pm Monday October 9, 2017.

Write a Python program and call it *simdata.py* that can read a tree in NEWICK format from a file (use your first assignment python code as a template for this exercise) and presents a set of datasets with simulated sequence data. Use the Jukes Cantor model to simulate data. the easiest way to do this is starting out at the root of the tree and evolve all sites on all branches and collect the site at the tip and then print phylip style dataset. Ideally the command line syntax for your python program is

```
python sim.py treefile phylipfile
```

Jukes-Cantor rate matrix

$$Q = \begin{pmatrix} -3/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & -3/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & -3/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & -3/4 \end{pmatrix} \tag{1}$$

The base frequencies are all $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$.

Algorithm

Tree traversal

1. start at the root
2. initialize the sequence: randomly draw from the base frequencies n times for n sites and add them to the sequence (you will need a function **randombase(π)** that returns A, C, G, or T)
3. preorder traverse the tree and evolve the sequence from one node to the other, stop when you reached the tip, you will also need to store the sequence in the interior nodes and the tips.

Evolve along a single branch

the function **evolve**, takes three arguments: start nucleotide, the Q matrix, and a branchlength v. the output is a nucleotide

1. $\tau = 0; \text{nuc} = \text{startnucleotide}$
2. Do forever
3. find λ if you are at **G** use the diagonal value row of q_G
4. draw random number r_1 between 0 and 1.
5. calculate $t = -\log_e(r_1)\frac{1}{\lambda}$
6. $\tau = \tau + t$
7. if $\tau > v$
   return current nucleotide nuc
8. calculate change of nucleotide cumulative sum for example: $s = [0, \frac{q_{GA}}{q_{GG}}, \frac{q_{GA}}{q_{GG}} + \frac{q_{GC}}{q_{GG}}, 1.0]$
9. draw random number r_2 between 0 and 1
10. pick interval in which r_2 lays, if we are at a G then first interval: A, second: C, third: T
11. goto "Do forever"

Print out the sequence

1. traverse the tree from the root to a tip
2. print the sequence at the tip in the format "name Sequence" (use 10 characters for the name)