CSE 280A
Assignment 2

Andrey Bzikadze

List 5 || (2)

(2) I've only done a) as b) is not required for Assignment 2.

The simulator code is on page 6.

The tree topology is generated in linear time (see function simulate_tree). The distribution of edges lens is ~~different~~ for exponent. increasing $N$. The simulation is done by inversing the cumulative distribution function.

Formally let $\ell_T = \ell_T^{(k)}$ be the length of the edge starting at some generation $T$ ($\Rightarrow N_T = e^{-\alpha T} N_0$) // for $k$ individuals.

$$P(\ell_T = 1) = \binom{k}{2} / N_T, \quad P(\ell_T = 2) = \frac{\binom{k}{2}}{N_T e^{-\alpha}} \left(1 - \frac{\binom{k}{2}}{N_T e^{-\alpha}}\right)$$

$$\Rightarrow P(\ell_T = \ell+1) = \frac{\binom{k}{2}}{N_T e^{-\alpha\ell}} \prod_{s=1}^{\ell} \left(1 - \frac{\binom{k}{2}}{N_T e^{-\alpha s}}\right).$$

The simulation idea comes from the fact that

$$\frac{P(\ell_T = \ell+1)}{P(\ell_T = \ell)} = e^{-\alpha} \left(1 - \frac{\binom{k}{2}}{e^{-\alpha k} N_T}\right)$$ - a very simple close form that allows one to go from $P(\ell_T = \ell)$ to $P(\ell_T = \ell+1)$ in short time.

The mutations are simulated by Poisson as was described on the lecture.

Simulator saves the tree in newick format and SNP matrix as np.savetxt result.

On page 7 I also saved SNP matrix and a tree for $n = 10$.