Phylogenetic reconstruction

In this homework we will consider a simplified model for phylogenetic reconstruction. Broadly speaking, phylogenetics aims at reconstructing the evolutionary relationship among various species (both living and extinct). The main source of data for phylogenetics is provided by genome sequences of various species. Throughout this homework we’ll assume that the evolutionary relationships between various species are known (i.e. it is known which species descend from which, a.k.a. the ‘tree of life’). We will instead focus on the problem of reconstructing the genome of an ancestral (extinct) species on the basis of the sequences of living species.

In our model, the phylogenetic tree (the ‘tree of life’) is a binary tree, with a single ancestor and $\ell + 1$ generations (see Figure 1 for an example with $\ell = 3$). The generations will be denoted by $\{0, 1, \ldots, \ell\}$. Generation 0 has a single node (a common ancestor to all the species in the tree) that will be called ‘root’ and denoted by $\circ$. Generation $t \in \{1, \ldots, \ell\}$ has $2^t$ nodes. The nodes at the last generation (there is $2^\ell$ of them) will also be called ‘leaves’.

Each node correspond to a species. Each species (except leaves) has two species that are directed descendants. Each species, except the root, has one species that is its direct ancestor. A species is characterized by a DNA sequence, that is a sequence of length $n$, whereby each entry is one of the four letters (bases) $\{A, T, C, G\}$. The following are three examples of sequences with $n = 10$:

$$
\begin{align*}
s_1 &= (ATTTACAAGT), \\
s_2 &= (CATCTCTAGT), \\
s_3 &= (TAATACCAGT).
\end{align*}
$$

We assume the following model for mutations, with mutation parameter $p \in [0, 3/4]$:

1. The sequence at the root is uniformly random among all $4^n$ possible sequences.

2. If the species $i$ has sequence $s_i$, the sequences of each of the descendants of $i$ is generated independently as follows. Each base in $s_i$ mutates independently with probability $p$. It it mutates, one of the 3 remaining letters replaces the existing one with equal probability.
Figure 1: A phylogenetic tree with $\ell + 1 = 4$ generations.

2' An equivalent description of the mutation mechanism at the previous point is as follows. Each base is
tagged for ‘potential mutation’ independently with probability $4p/3$. Once it is tagged, its new value
is given by one of the 4 letters, each with probability $1/4$.

Leaves correspond to species that are alive at present time, and we now have their DNA sequence. Our
objective will be to infer the sequence of the root species.

(1) Show that the descriptions given at points 2 and 2’ above are indeed equivalent.

(2) Consider the case of sequences of length $n = 1$ (this is not interesting in practice, but is the basic building
block for what we will do). Choose one specific leaf node (e.g. the leftmost one). Compute the probability
that the sequence at that node is different from the one at the root (as a function of $p$ and $\ell$).

(3) Consider now the case of general $n$ and let $s_o$ be the root sequence. Let $s_i$ be the sequence at a leaf node
$i$, and denote by $X_i$ be the random variable that counts the number of positions in which $s_i$ differs form $s_o$.
Compute the expectation of $X_i$.

(4) Write a program that generates random samples from the above model (i.e. the sequences at each node
of the tree) for $\ell = 10$, $n = 100$. Generate one such a configuration for $p \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and plot in
each case the histogram of $X_i$, for $i$ varying over the set of leaves.

Compare these histograms with the formula for $E(X_i)$ obtained at the previous point.

(5) A possible approach for reconstructing the ancestral sequence $s_o$ is to consider the sequences on the
leaves and, for each base, take a majority vote. In other words, for the first base, we guess that it is given
by the letter that is most frequent among the first bases at the leaves. The same rule is repeated at each
position.

Let $\hat{s}_o^{\text{majority}}$ be the sequence estimated by this majority rule, and let $Z$ be a random variable that gives
the fraction of bases in which $s_o$ differs from $\hat{s}_o^{\text{majority}}$. Repeat the above simulation 10 times for each of
$p \in \{0.02, 0.04, \ldots, 0.48, 0.50\}$, evaluate $\hat{s}_o^{\text{majority}}$ and $Z$ for each new sample. Use these to estimate $E(Z)$ for
each of these values of $p$. Plot the results.