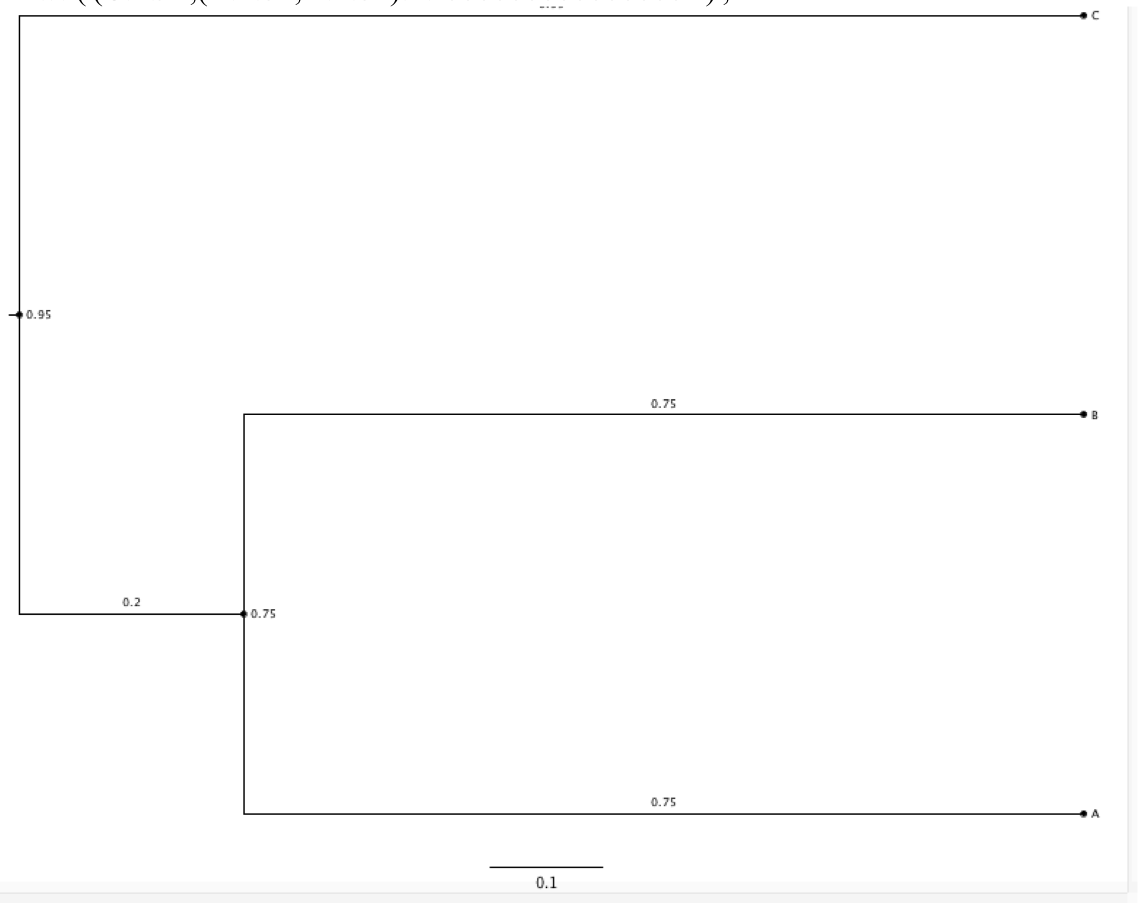
Matt Wong
MCB HW2

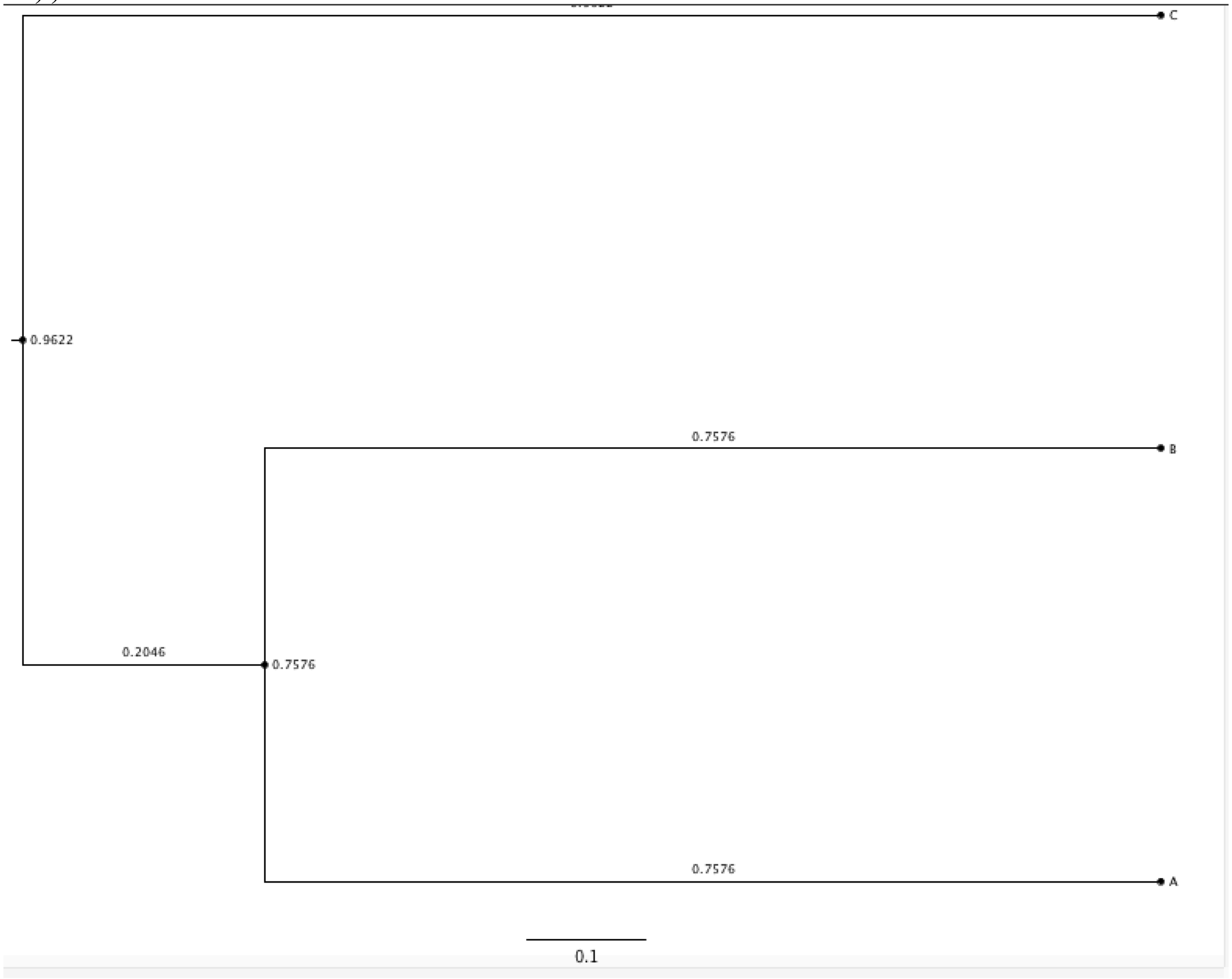
Question 2.1: *Question 2.1: In what way is the ultrametric tree built with the raw distance matrix incorrect? Why? How does the rate of change affect this, and why?*

The Ultrametric tree built with raw distance matrix is only accounting for the observed differences between the sequences, and therefore underestimating the true changes that have occurred. The rate of change will continue to increase the gap in this difference, because more unobserved changes will occur with an increased rate of change, therefore making the Ultrametric tree built with the raw distance matrix incorrect. As you can see in the figures below, the raw difference trees have distances that are significantly lower than that of the trees corrected with newlambda = (-(3/4))*math.log1p((-(4/3)*mymu)), or the corrected distance formula.

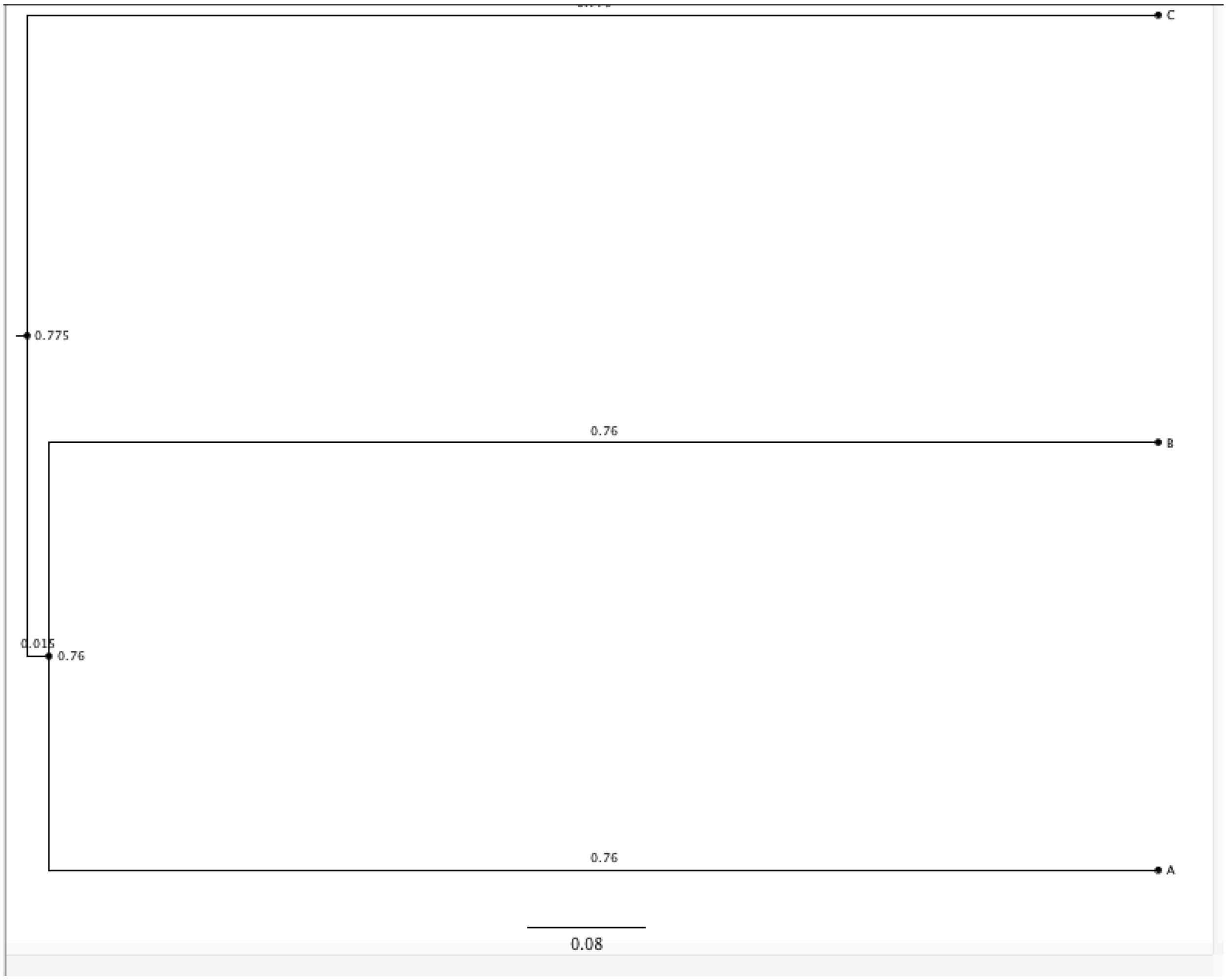
```
>>> simulateAndRecoverTrees(sampleTree1,.01,1000)
Raw: ('(C:0.95,(B:0.75,A:0.75):0.19999999999999996)',
```



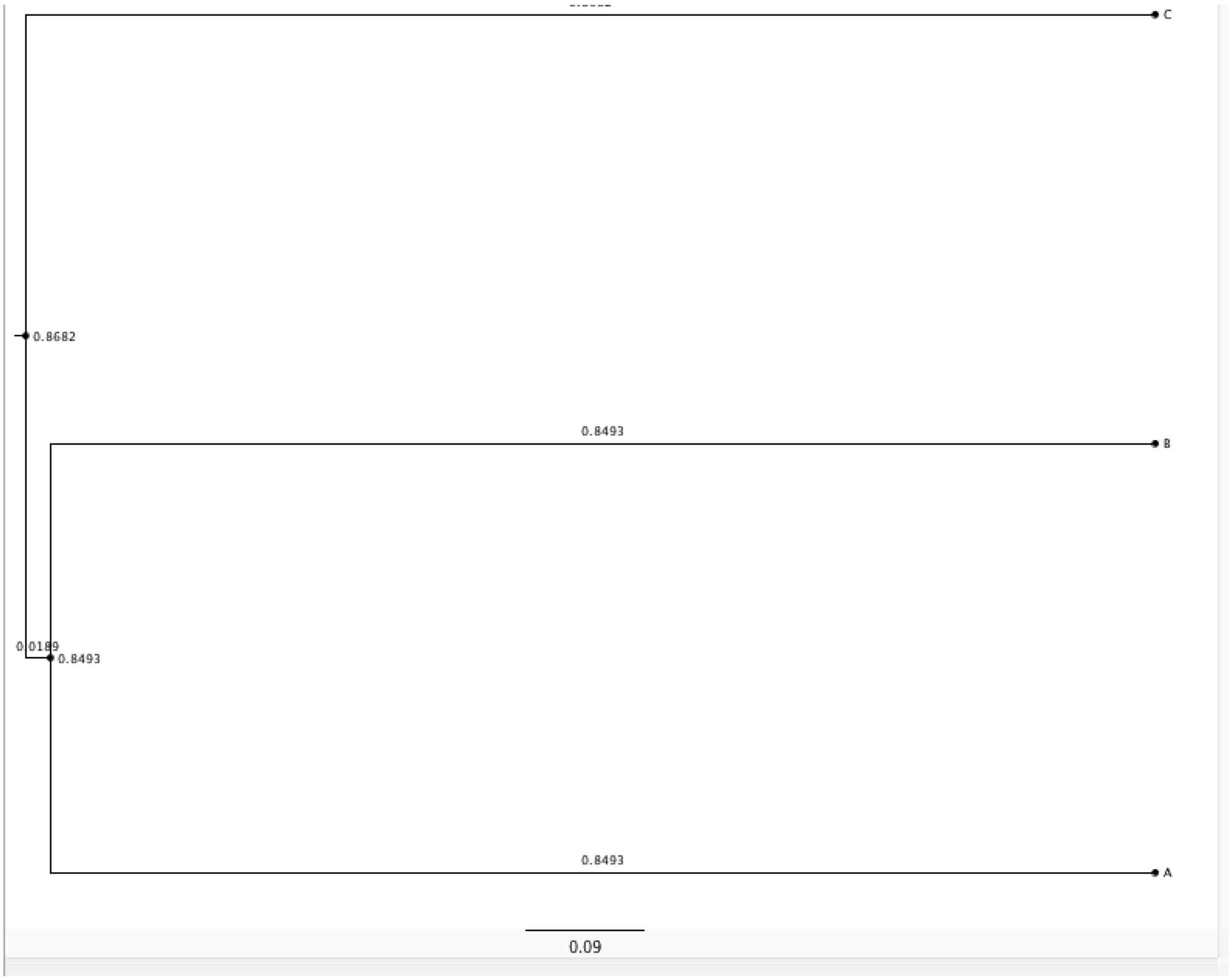
```
Corrected =
'(C:0.9622405042716609,(B:0.7576015244069793,A:0.7576015244069793):0.20463897986468
16)')
```



```
>>> simulateAndRecoverTrees(sampleTree1,.1,1000)
Raw =
('(C:0.7749999999999999,(B:0.7599999999999999,A:0.7599999999999999):0.0150000000000
00013)'
```



Corrected = '(C:0.8681692536927239,(B:0.8493091971739664,A:0.8493091971739664):0.01886005651875755)')



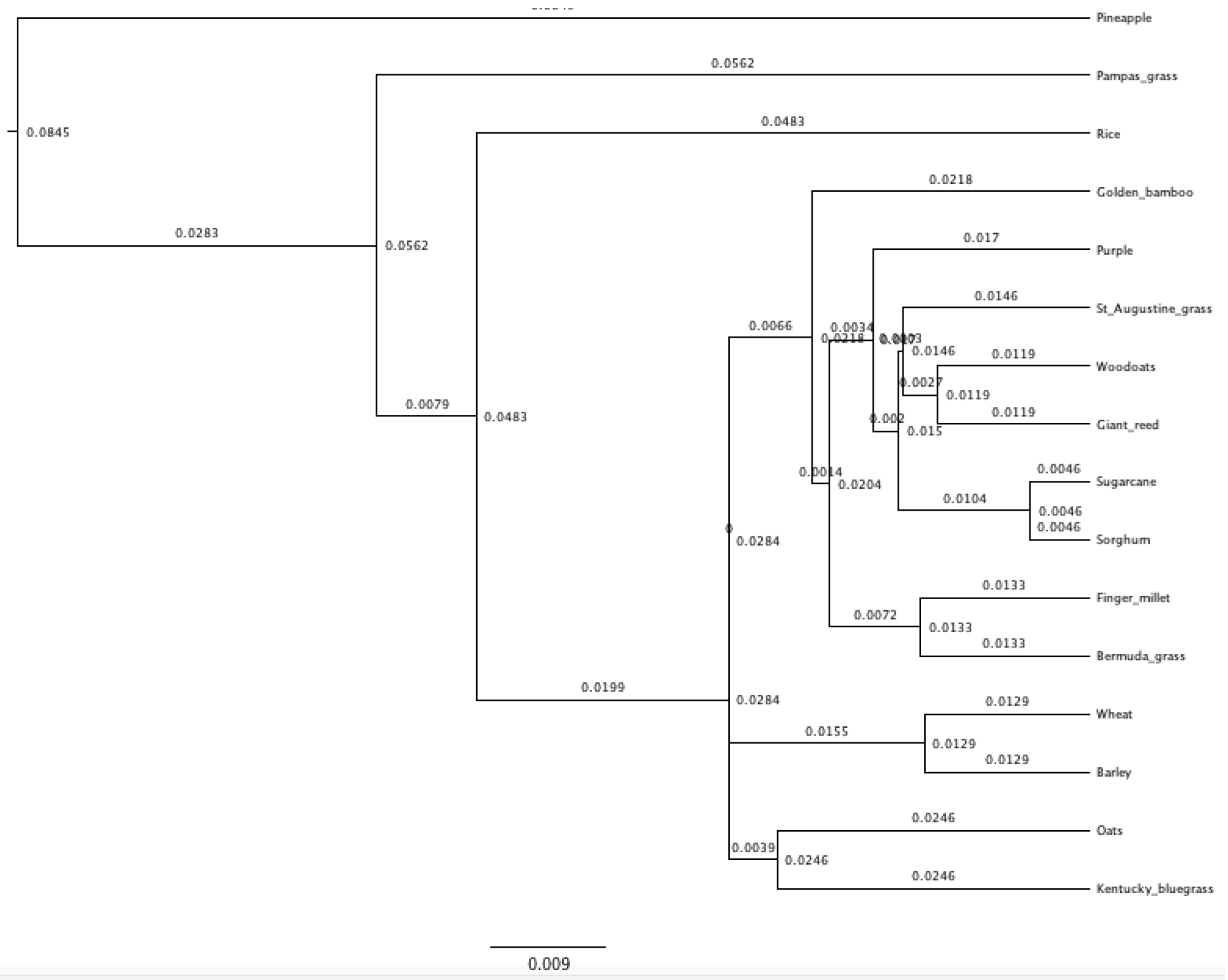
Question 2.2: In what way is the ultrametric tree different from the starting tree? Why? What effect does the rate of change and the Jukes Cantor correction have on this difference?

The Ultrametric tree will have branch lengths where the leaves all end at the same point in time, whereas in a normal starting tree that is not Ultrametric, the branches could have changed at different rates. The jukes cantor correction properly estimates the expected number of changes based on the observed changes, and will therefore help to minimize the difference between the genetic distances of the trees. The rate of change, however, will increase this change, as it will cause more room for the starting tree to acquire additional tree.

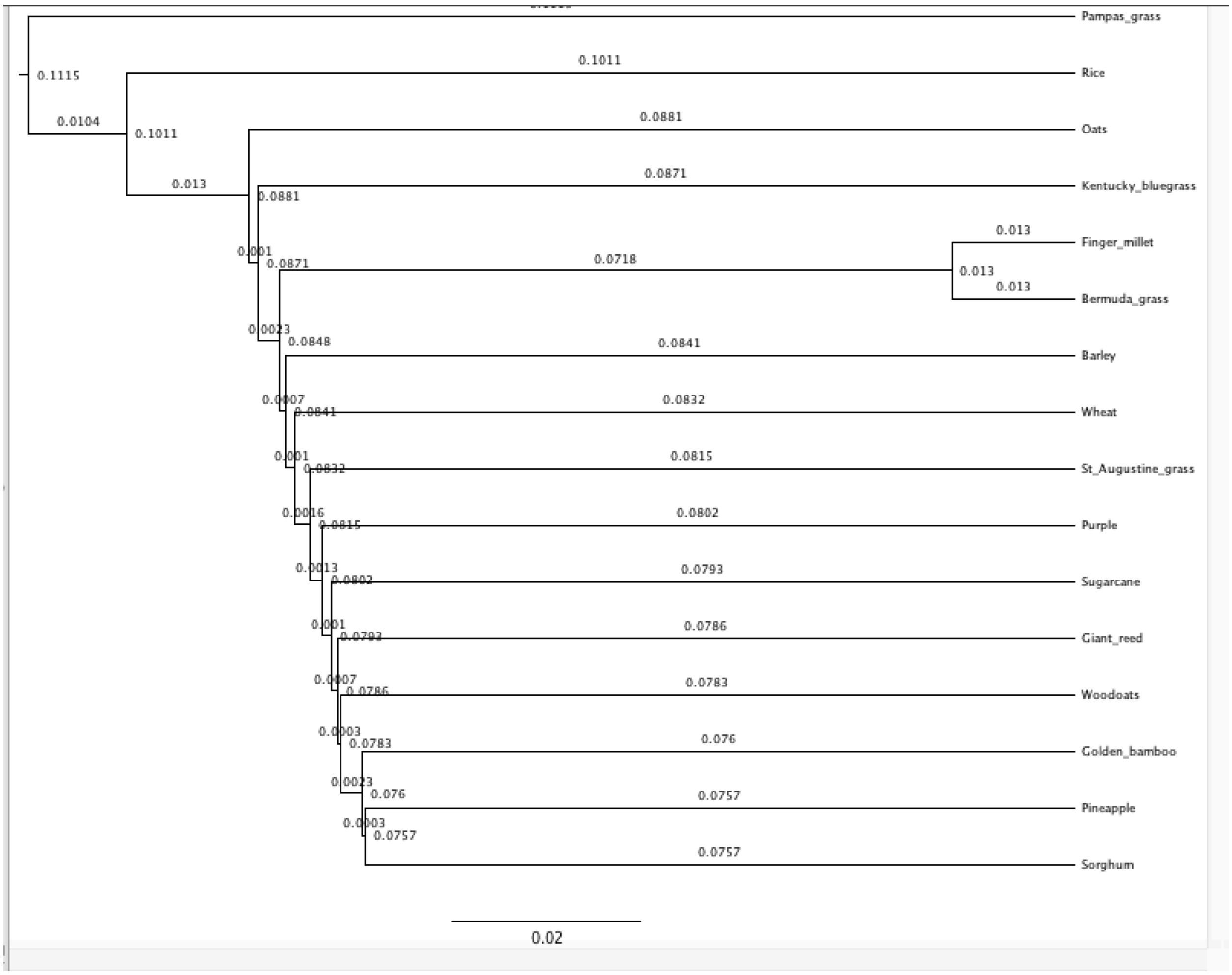
Question 3.1: How does your new tree differ from the old tree? Has the topology changed? If not, have the branch lengths changed? In what way? Explain why.

The new tree has a different topology, where the relationship of pairs of species is much more specified, and more nodes are included. Because of this, the respective branch lengths of the new tree are smaller, as the tree shows more points of species-species divergence. The original tree has many distances near the order of (0.07) illustrating a common point in time where all species pairs diverged, which seems much less accurate and unlikely.

Corrected Tree:



Uncorrected Tree:



Question 3.2: How does the topology of your ultrametric tree differ from the topology of the neighbor joining tree? In each tree, what is the smallest clade containing all C4 plants (purple three-awn, Bermuda grass, finger millet, St. Augustine grass, sorghum and sugarcane)? Explain how the two trees lead to different interpretations of the evolutionary history of C4.

The leaves in the tree created using the neighbor-joining algorithm are not all lined up in vertical unison like the Ultrametric trees. The smallest clade containing all C4 plants in the neighbor joining tree is Giant_reed, Purple_three_awn, Bermuda_grass, Finger_millet, Pampas_grass, Woodoats, St_Augustine_grass,Sorgum, and Sugarcane. The smallest clade containing all C4 plants in the corrected tree is Pampas grass, Rice, Golden bamboo, Purple, St.Augustine grass, Woodoats, Giant reed, sugarcane, sorghum finger millet, and Bermuda grass. The smallest clade containing all C4 plants in the uncorrected tree is all species. The corrected matrix shows a smaller clade for the history of C4, meaning the clade is still much conserved relative to the clade represented in the uncorrected tree. This tree overestimates the divergence of the clade, that sorghum is much less related to pampas grass than it should be, ultimately contributing to a larger clade. Additionally, the uncorrected tree claims that a large set of divergence events occurred at around the same time, leading to an immediate growth of the C4 clade, which is highly unlikely. In conclusion, the ultrametricly corrected tree that accounts for observable and non-observable changes provides a better represented phylogenetic tree.