**Assignment-3**

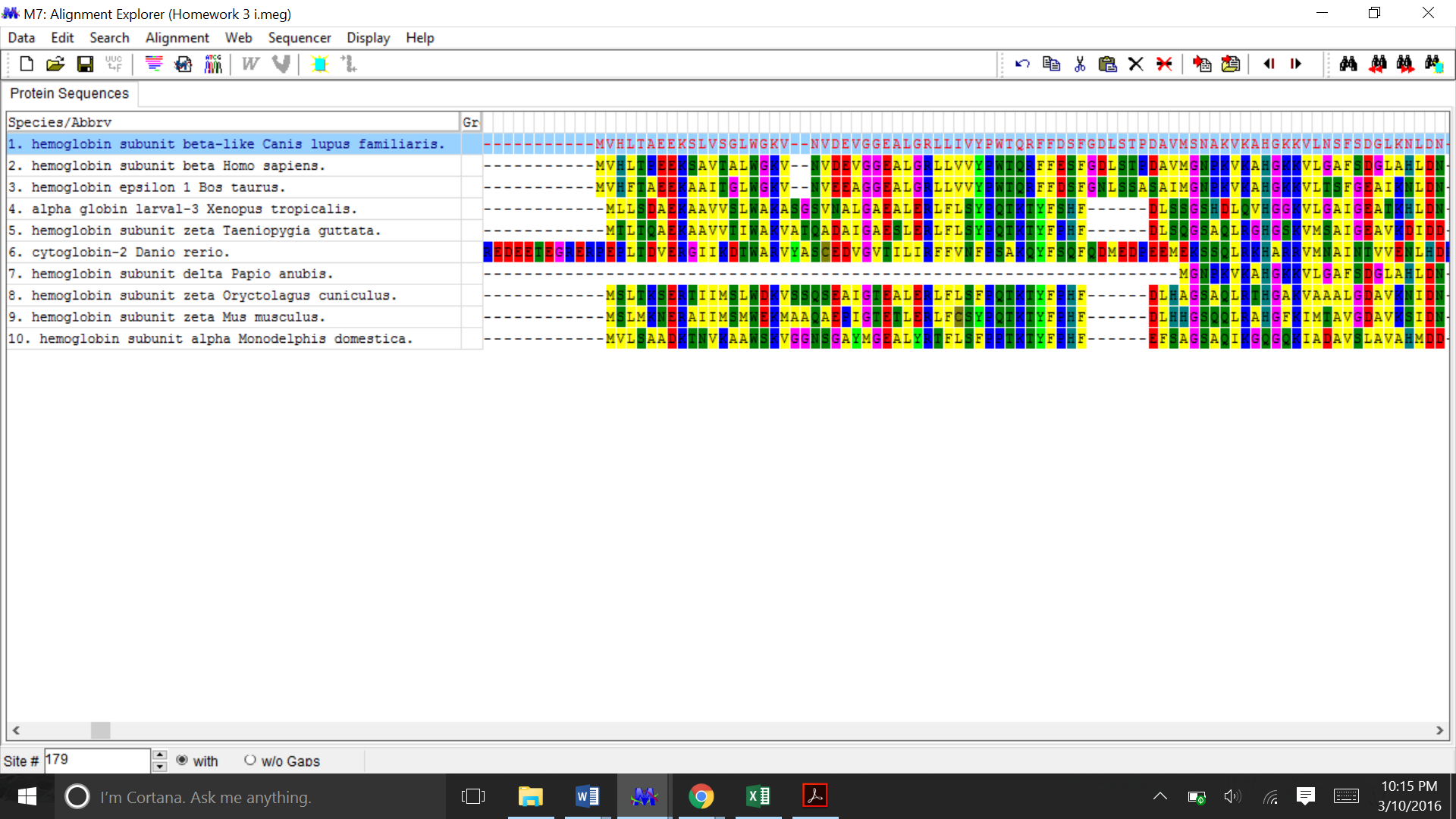
**Full marks = 100**

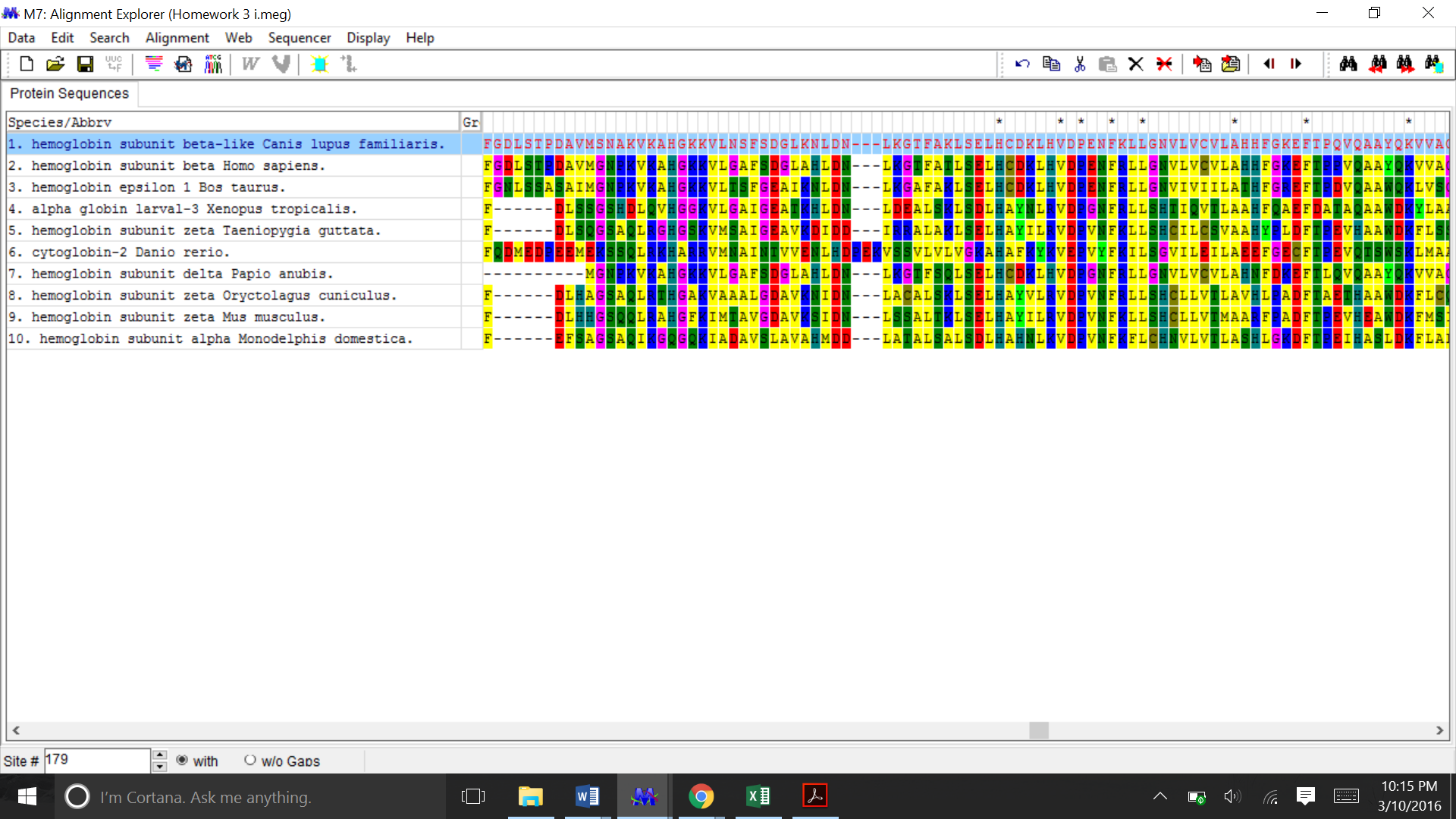
1. **Question 1.** Which sequence is your outgroup and why did you choose it? [ 5 points]

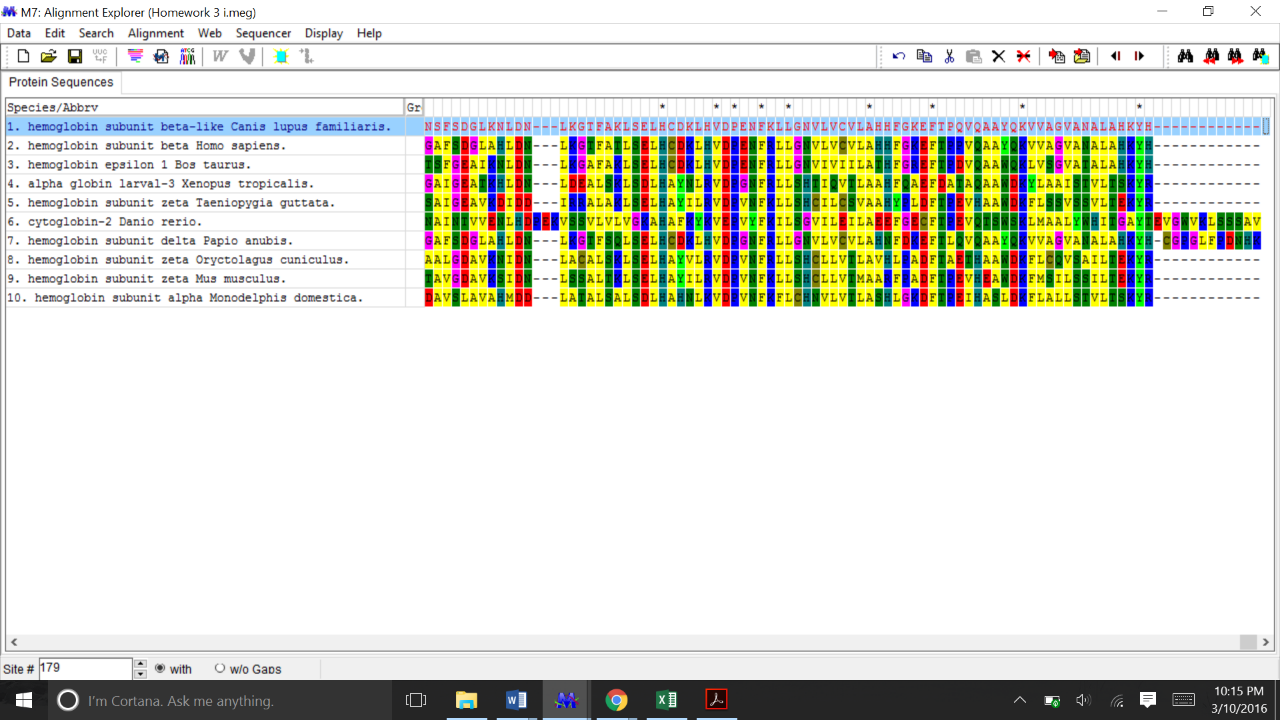
I have chosen cytoglobin-2 Danio rerio to be outgroup. I settled on the zebrafish due to the fact that its habitat is in freshwaters. I figured out that, its “globin-like” protein would be significantly different from the rest of the organisms on my list.

1. **Question 2.** Looking at the alignment, which amino acid regions are the most conserved across the 10 sequences? [10 points]

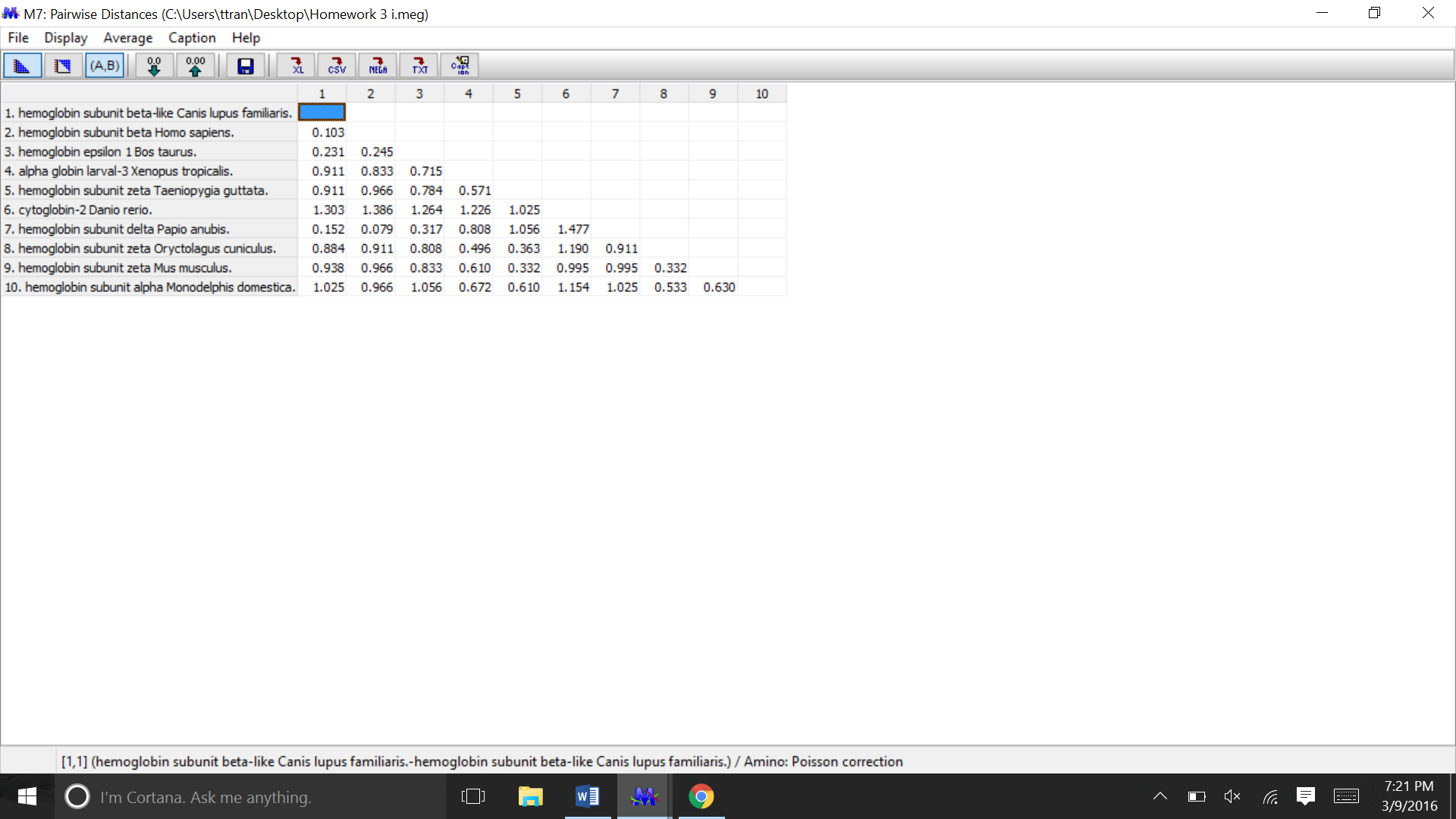
There seems to be a stretch of conserved hydrophobic amino acids between site 25-28, 49-51, 85-89, 130-136, 154- 162. There are also short sites that were near unanimously basic such as 24, 33, 48, and etc. These were some of the more obvious conserved regions between the 10 sequences.







Amino acid analysis

1. **Question 3.** Using the Pairwise Distances table, which two sequences are the most similar? Which two are the least similar? Include the table with your answer. [10 points]

Answer: According to the pairwise Distance table, *Hemoglobin subunit beta-like Canus lupus familiaris (domestic dog) and Hemoglobin subunit beta Homo sapiens (human) are the most similar (pairwise distance = 0.103) .* Applying the same logic with the table, *Hemoglobin subunit beta Homo sapiens (human) and Cytoglobin-2 Danio rerio (zebrafish) is suggested to be the least similar (pairwise distance = 1.303, 1.386, 1.264, 1.226, 1.025).*

1. **Question 4.** Are there sequences with drastically different amino acid compositions? Which ones? [7 points]

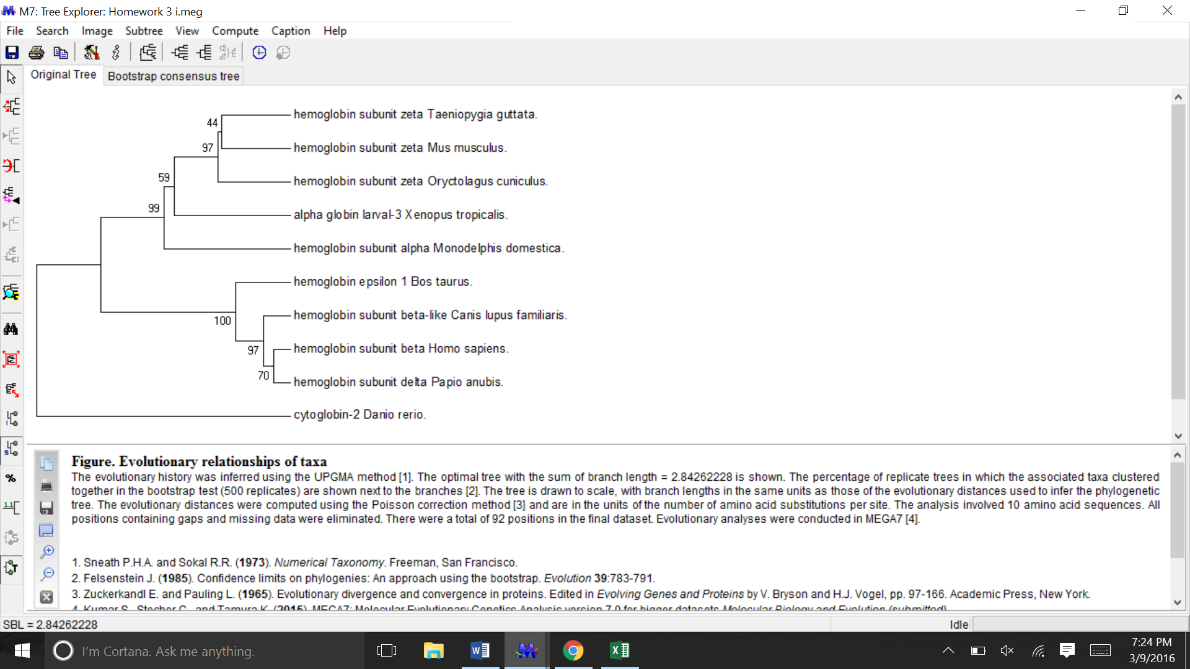
Answer: Analyzing the amino acid compositions, suggested that some amino acids have a lower frequency. The average Cys frequency was 1.256. However, some sequences range from 0.6 to 2.1. A more extreme example is the case of Isoleucine. Sequences 2 and 7 has zero frequency for isoleucine. But some sequences can range more of up to 4.08 to 2.11. Other amino acid that was consistently high across all sequences is lysine. All sequences have a high frequency for lysine.



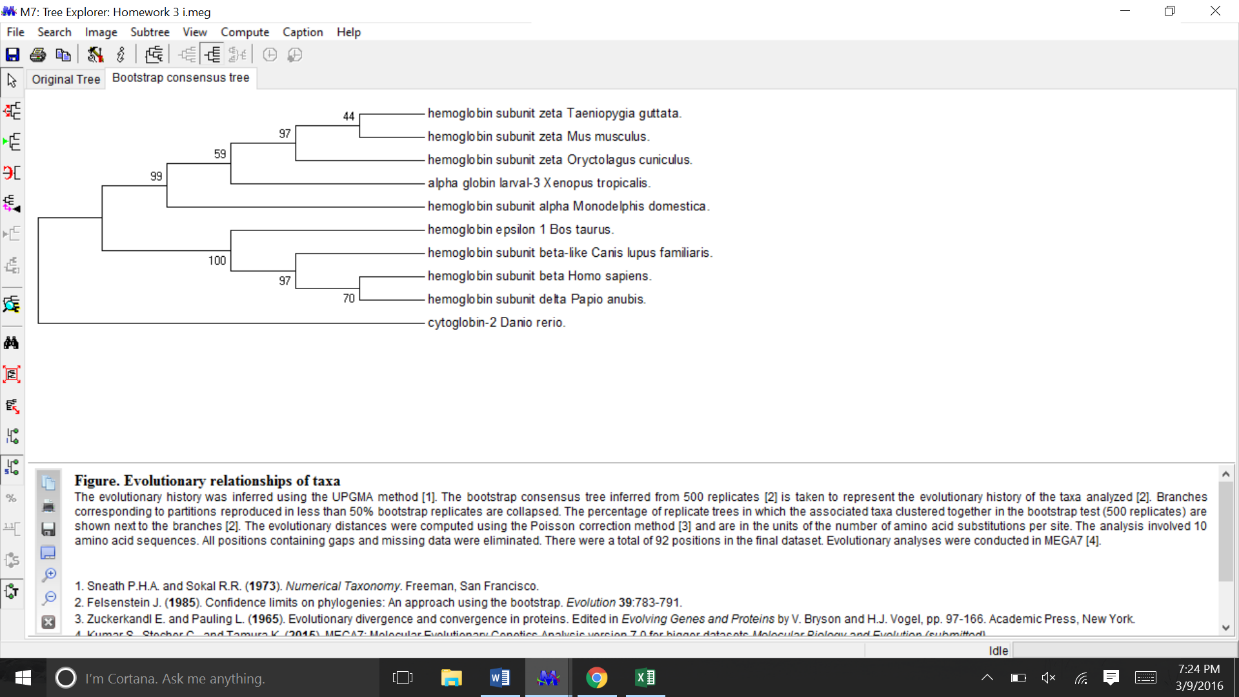
1. **Question 5.** How are these trees different? Include the original and bootstrap consensus trees for all four trees in your answer. [each trees carry 6 points, in total 30 points]

**Answer:**

* 1. **UPGMA** – Be sure to select “Bootstrap method” for the “Test for Phylogeny” option and enter 500 as the “No. of Bootstrap Replications”.

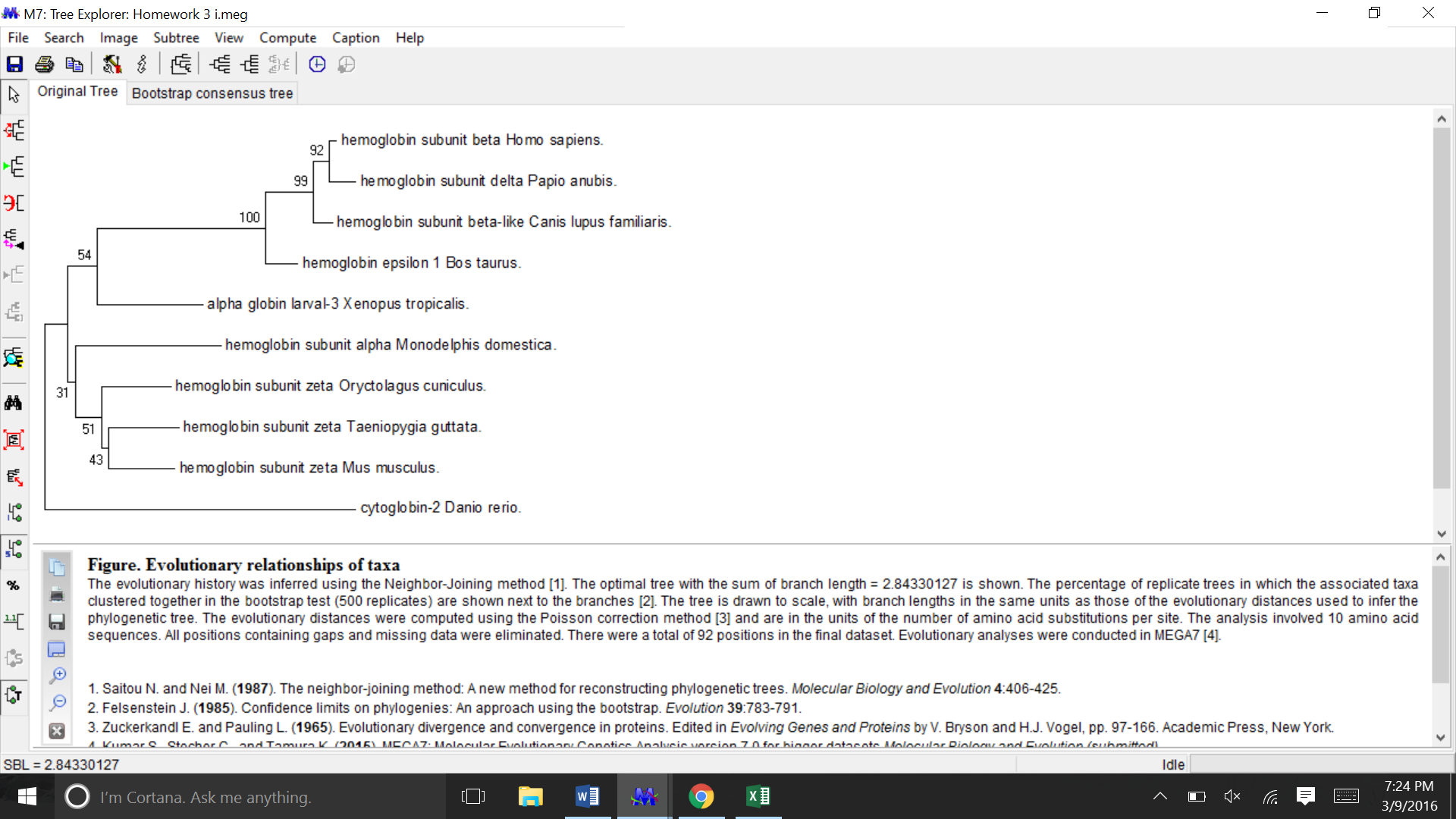


Original UPGMA tree

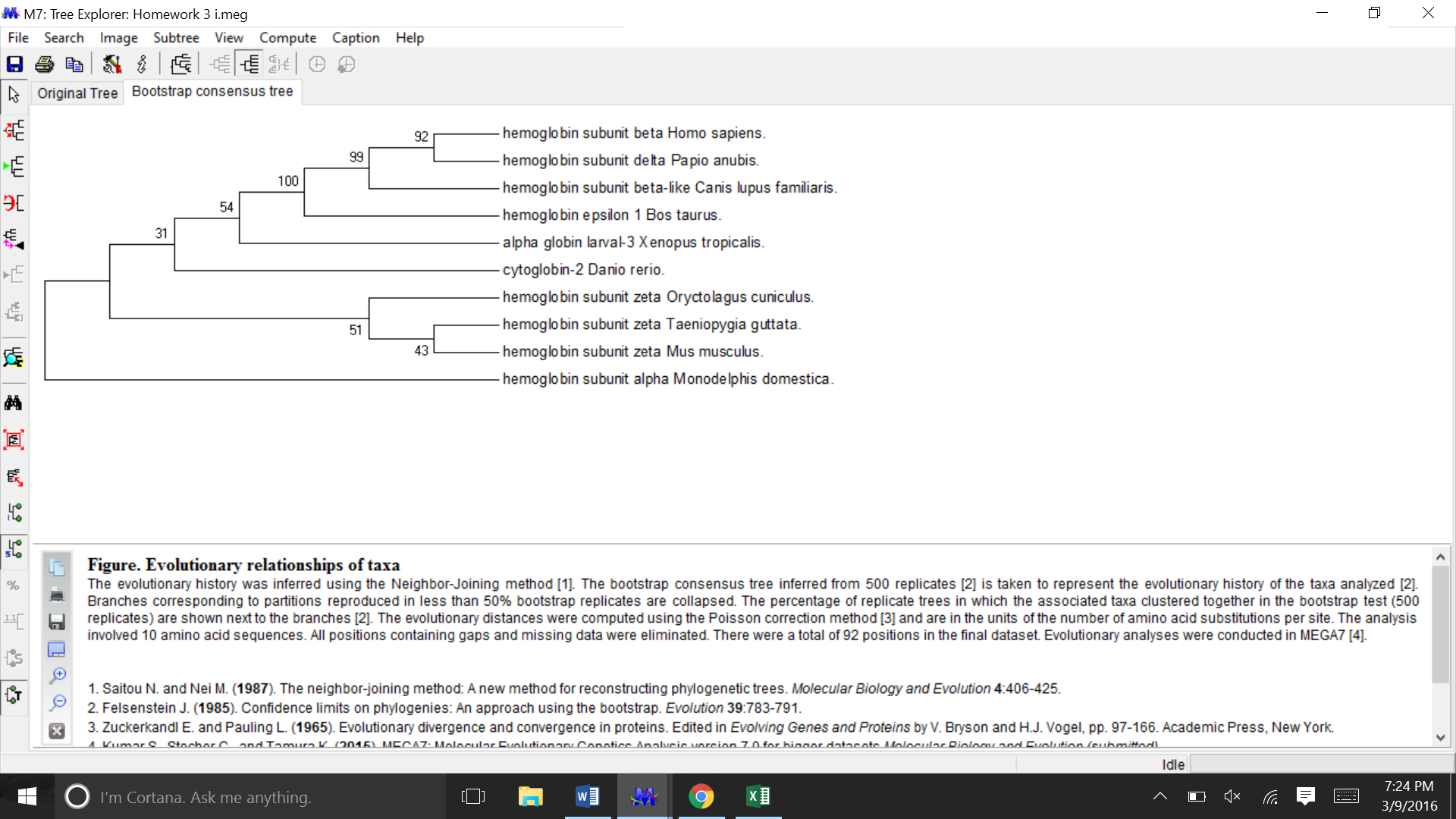


Boot strapped UPGMA consensus tree

* 1. **Neighbor-Joining** – Be sure to select “Bootstrap method” for the “Test for Phylogeny” option and enter 500 as the “No. of Bootstrap Replications”.

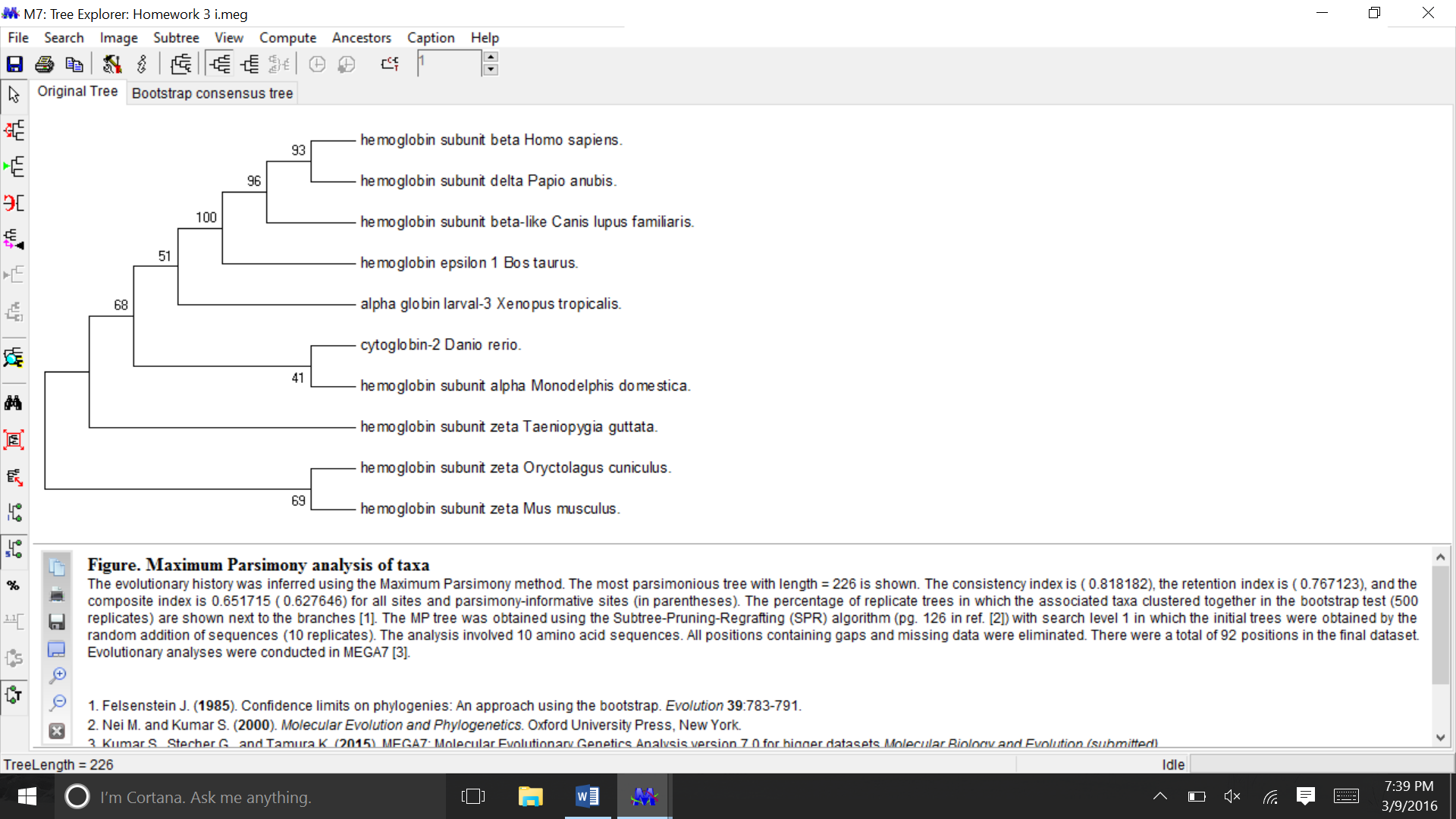


Original Neighbor-joining tree

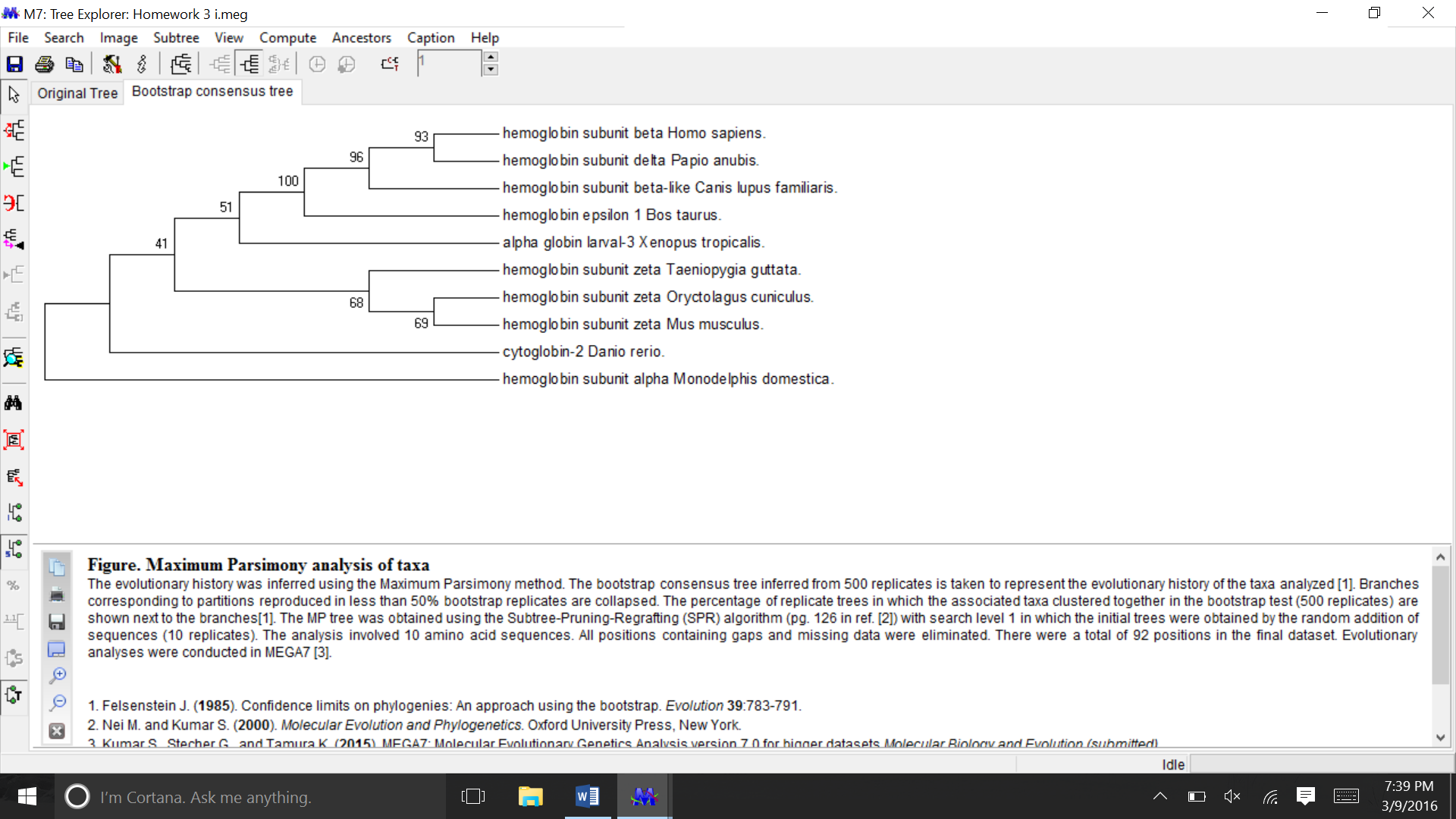


BootStrap consensus tree for Neighbor-joining.

* 1. **Maximum Parsimony** – Be sure to select “Bootstrap method” for the “Test for Phylogeny” option and enter 500 as the “No. of Bootstrap Replications”.

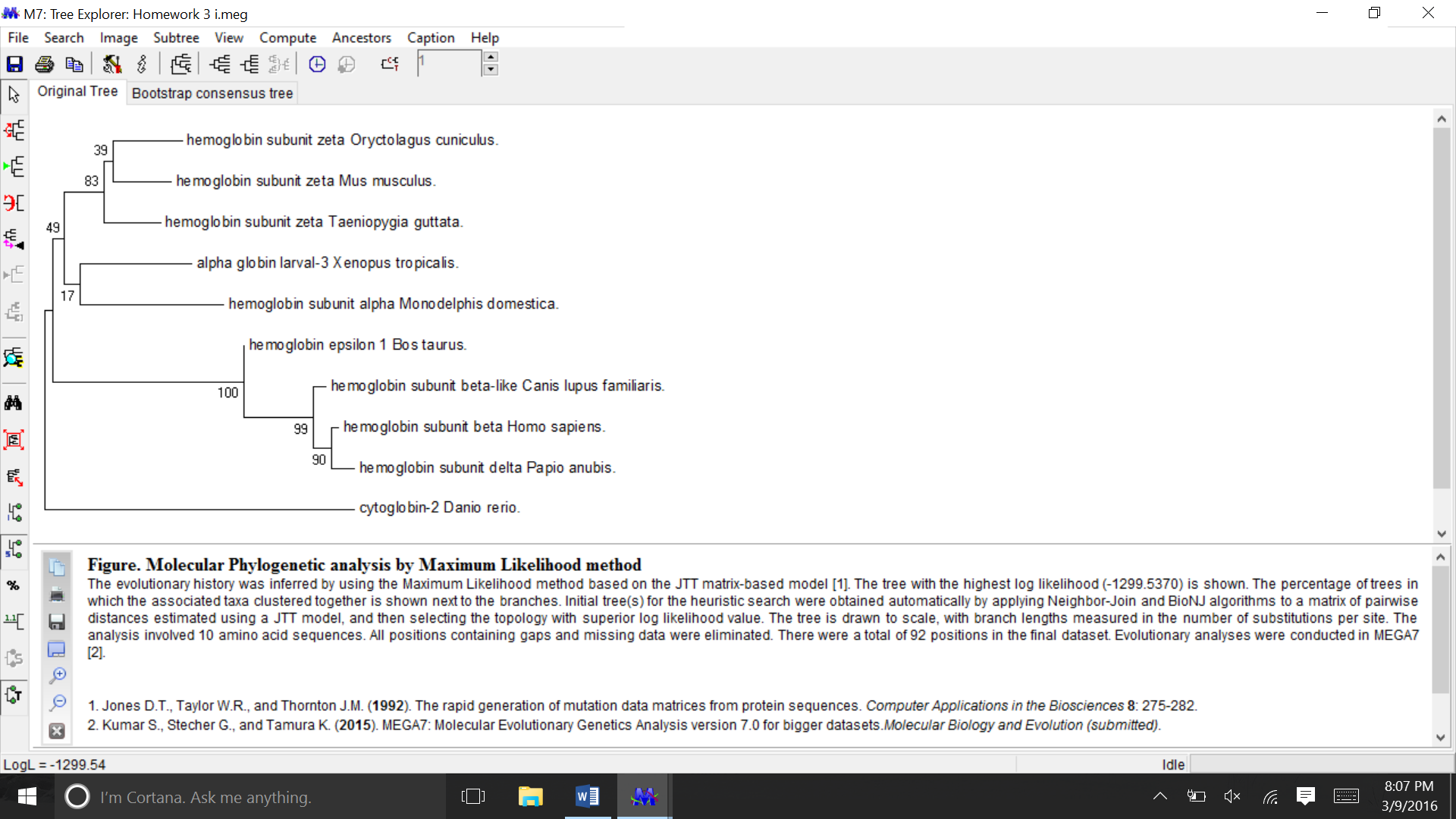


Original Maximum Parsimony tree

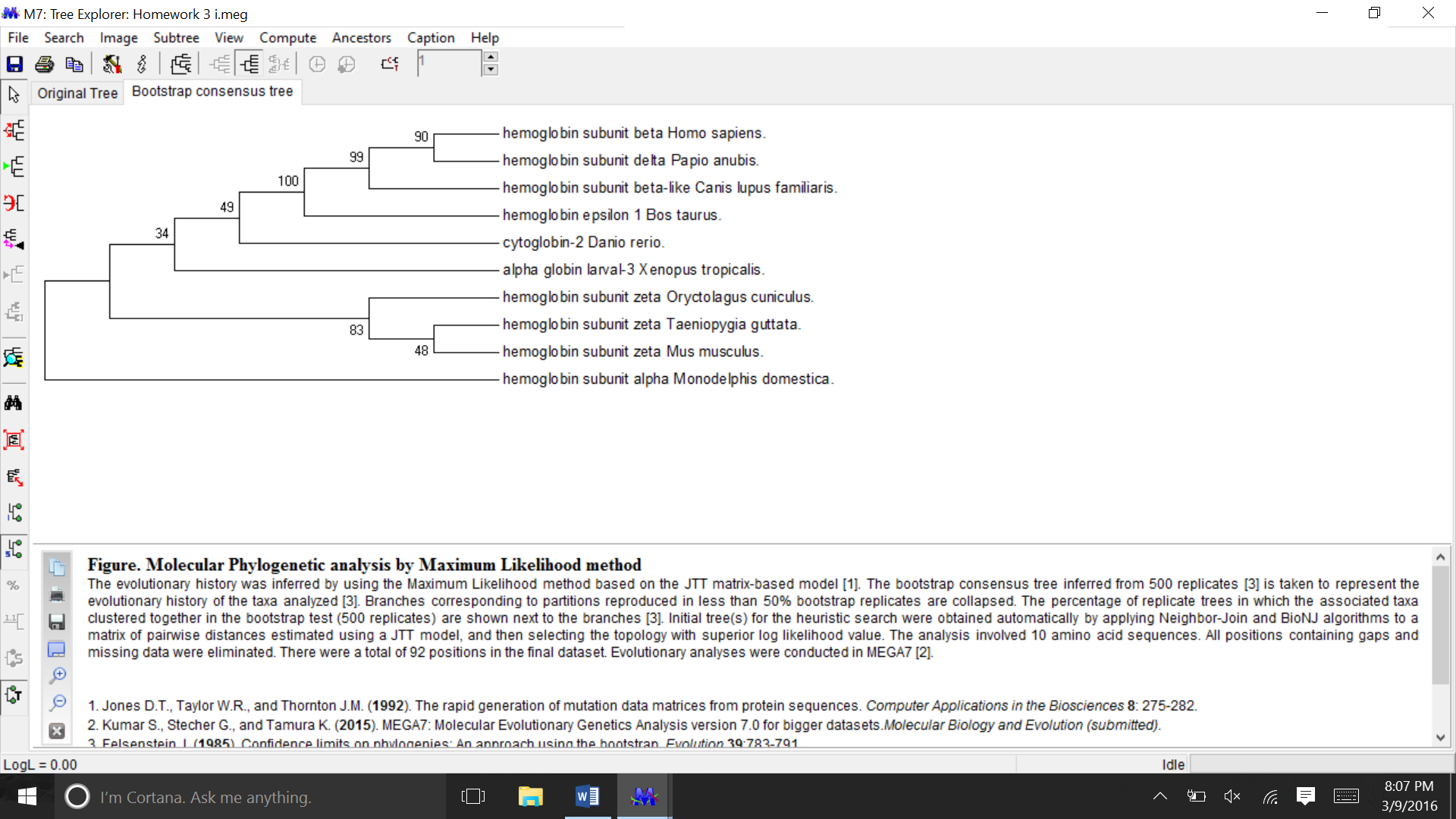


Bootstrap Consensus tree for Maximum Parsimony

* 1. **Maximum Likelihood** – Be sure to select “Bootstrap method” for the “Test for Phylogeny” option and enter 500 as the “No. of Bootstrap Replications”. (These methods may take a few minutes.)



Original Maximum Likelihood tree



Bootstrap consensus tree for Maximum Likelihood

Some of these trees are substantially different than the others. The UPGA trees seem to have reliable bootstrap values for some lineages (with the exception of Taenipygiu guttata and mus musculus). In addition, the outgroup which have been selected for this analysis, was the one that I predicted: cytoglobin-2 Dario rerio. However, in comparison with the neighbor-joining trees, some of the lineages were completely flipped. For example, the bootstrap consensus neighbor-joining trees, placed the outgroup as hemoglobin subunit alpha monodelphis domestica. This was the case with my maximum parsimony trees and maximum likelihood trees (boostrap consensus) as well. These methods (maximum parsimony and maximum likelihood trees) suggested that Homo sapiens and Papio Anubis are closely related (roughly 90 bootstrap value). But some algorithms such as the neighbor joining tree algorithm (bootstrap), were not as confident (roughly 70 bootstrap value).

1. **Question 6.**Are these trees rooted? Which ones are cladograms, additive, ultrametric, or non-ultrametric? [2+5 = 7 points]

**Answer:** These trees do not seem to be rooted at first glance, perhaps the presence of an outgroup can give the direction that evolution travel in an unrooted tree. For example, maximum likelihood trees have an outgroup of hemoglobin subunit alpha monodelphis domestica (bootstrap) and cytoglobin-2 Dario rerio (original).

* Cladogram are branch lengths that have no meaning. In the UPGA tree, both trees were cladograms. In the neighbor joining trees, only the bootstrap trees were non-cladograms. In the maximum parsimony trees, both seem to be cladograms (their branch lengths have no meaning or calibrated scale). Finally, in the maximum likelihood trees, the original seem to be an additive tree (branch lengths show evolutionary divergence) while the bootstrap tree was a cladogram.
* Ultrametric trees assume a “constant” mutation rate reflected in the evolutionary tree branches.
* The only non-ultrametric trees observed from this analysis, were the neighbor joining (original) and maximum likelihood tree (original). All the other trees were ultrametric trees.

1. **Question 7.** Which branches in the trees do you NOT trust because of the bootstrap number? [5 points]

**Answer:** Due to low bootstrap values, I do not trust the taeniopygia guttata and mus musculus in my UPGMA trees (~40 bootstrap). There seems to be a trend that species that Mus musculus was paired up with, the bootstrap value tend to be very low.

1. **Question 8.** Is your chosen outgroup always the outgroup in each generated tree? [5 points]

**Answer:** In the beginning, I chose cytoglobin-2 danio rerio as my predicated outgroup. However, this was only consistent with the UPGMA algorithm. In my neighbor joining trees, only the original tree predicated cytoglobin-2 danio rerio as my predicted outgroup.

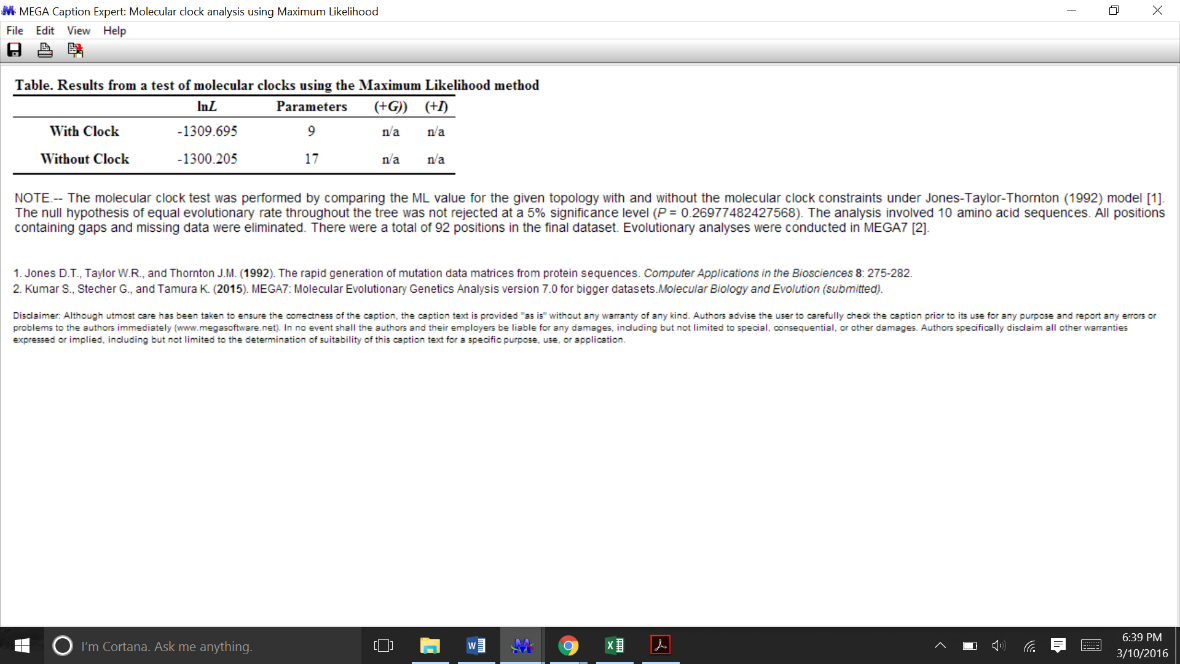
1. **Question 9.** Which tree do you think is the best and why? [ 2+5 = 7 points]

**Answer:** Estimating from the bootstrap scores, the maximum likelihood consensus tree seems to have higher bootstrap scores across its branches. However, this is hard to evaluate. Certain algorithms seem to place higher bootstrap values at different nodes. I liked UPGMA bootstrap values earlier in the nodes (suggesting better ancestry and descendants). In addition, some high bootstrap values tend to be biased towards certain branches. For example, my boostrap consensus neighbor joining tree had a separate branch with really low boostrap values, while its other branches have really high bootstrap values.

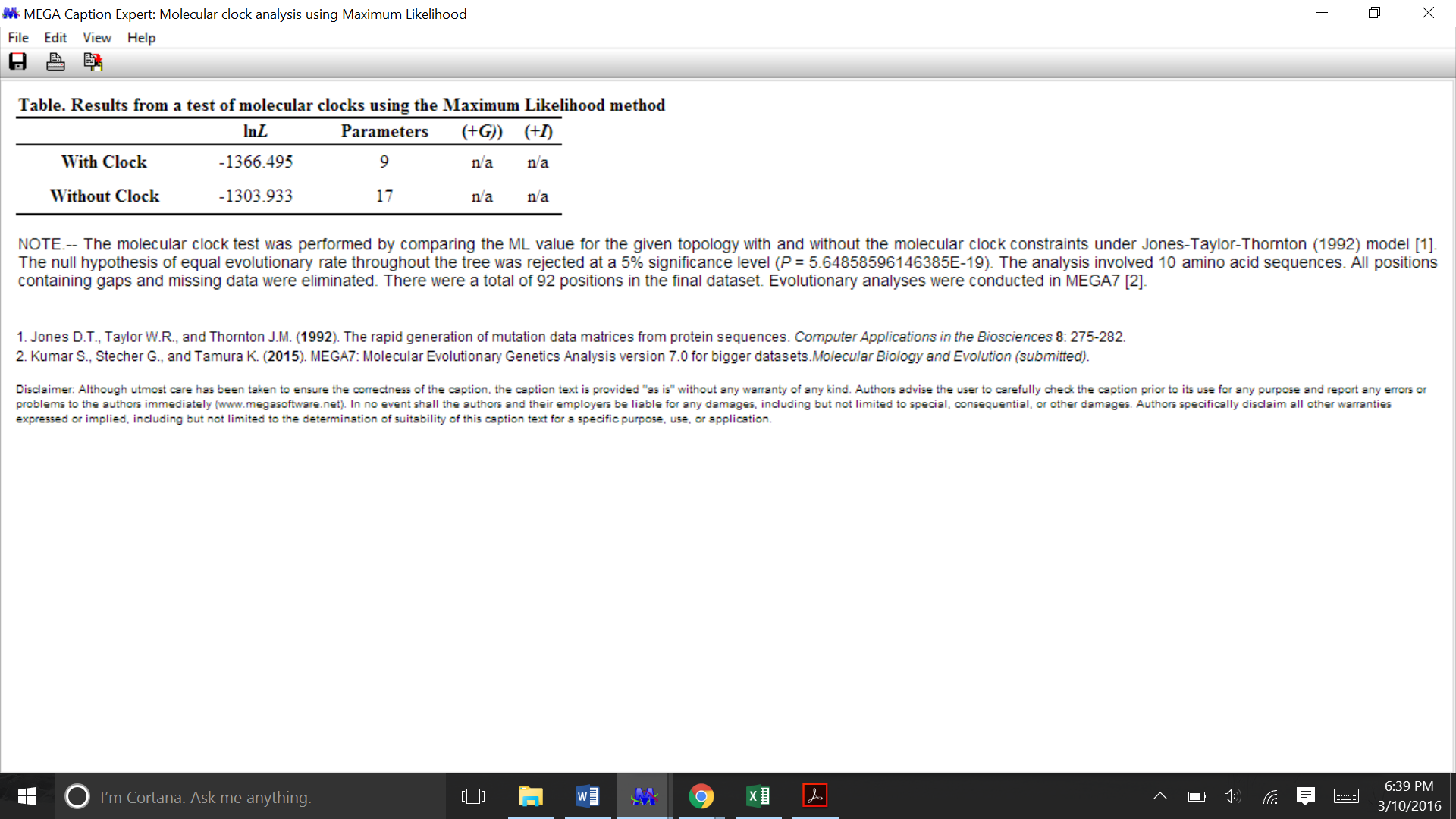
Overall though, it’s worth noting that most phylogeny experts prefer neighbor-joining as their distance-based algorithms. Maximum parsimony trees try to build trees that are character based; therefore, the trees will often try to have the shortest branch lengths possible. Finally, the maximum likelihood algorithm determines tree topology and branch lengths that bests reflect the observed data set; because the algorithm utilizes statistical methods more modern than UPGMA, neighbor joining, and maximum parsimony, I would expect it to create the best consensus tree.

1. **Question 10:** Based on the p-value, do we reject or not reject the molecular clock hypothesis? [10 points]

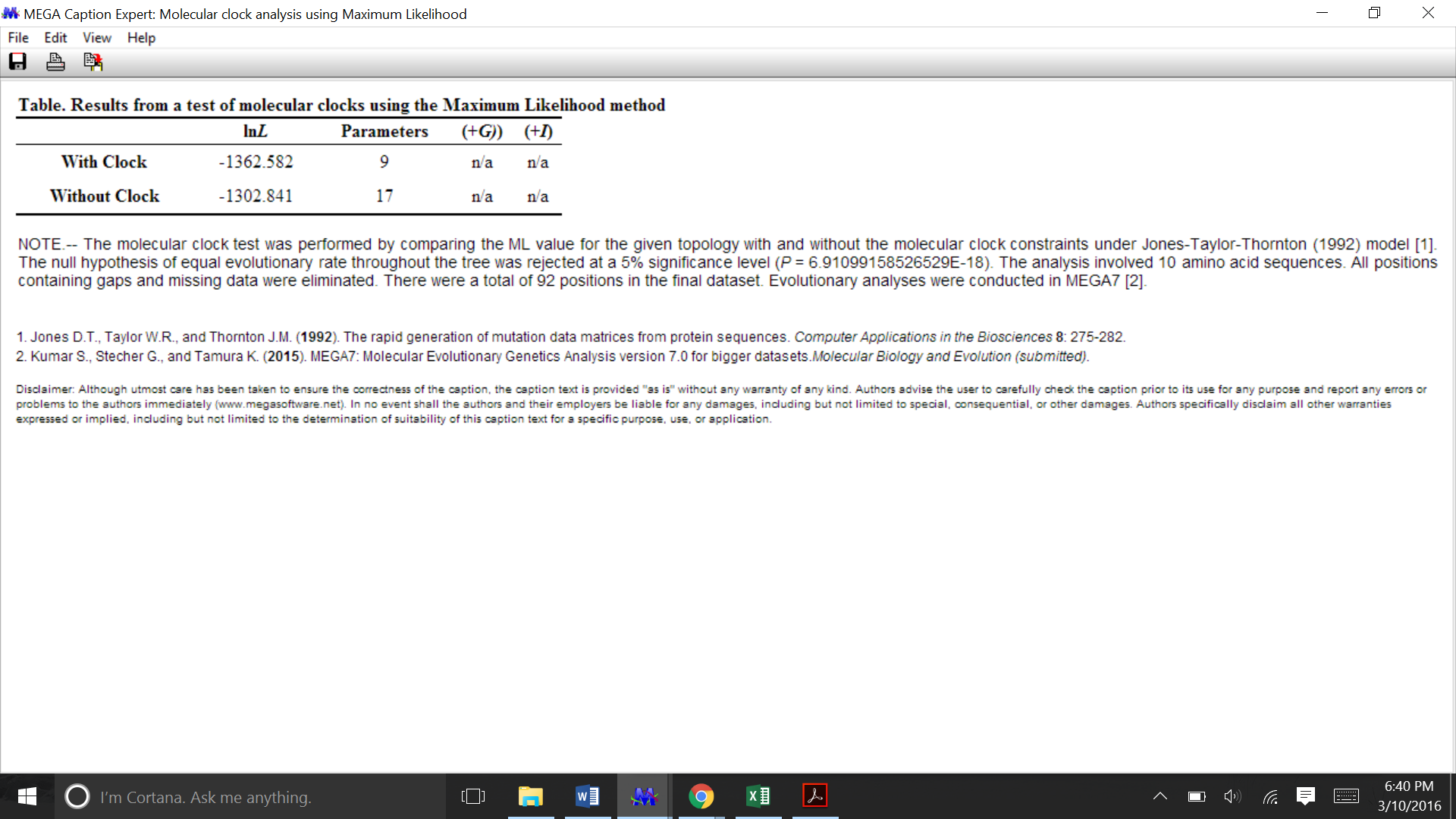
Answer:



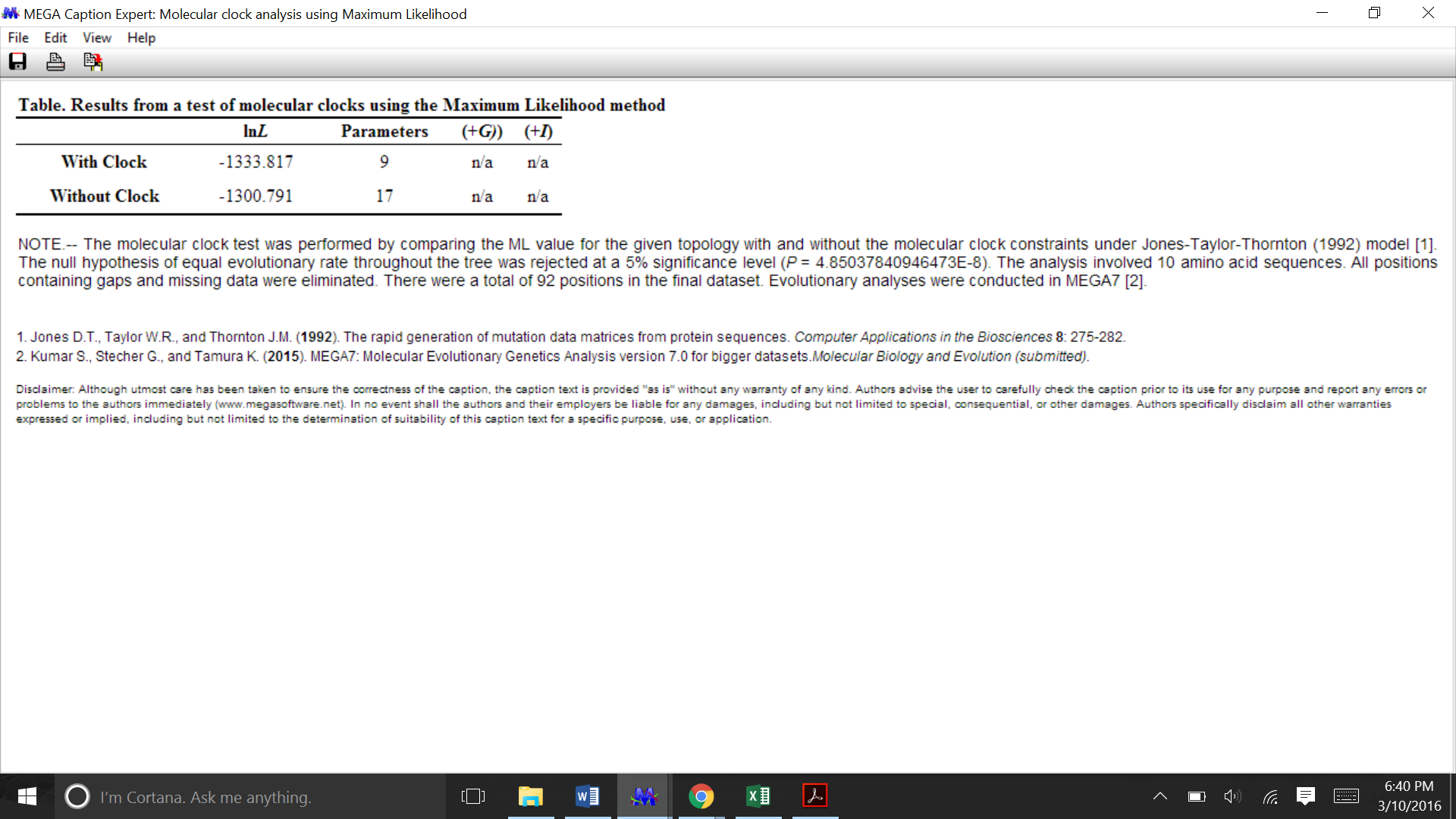
UPGMA : null hypothesis states equal evolutionary rate throughout the tree (the concept of the molecular clock hypothesis). Based on this tree (5% significance level), the p-value was 0.26978. Therefore the molecular clock hypothesis (null hypothesis) was not rejected.



Neighbor joining: null hypothesis states equal evolutionary rate throughout the tree (the concept of the molecular clock hypothesis). Based on this tree (5% significance level), the p-value was 5.648E-19. Therefore, the molecular clock hypothesis (null hypothesis) was rejected.



Maximum Parsimony: null hypothesis states equal evolutionary rate throughout the tree (the concept of the molecular clock hypothesis). Based on this tree (5% significance level), the p-value was 6.9E-18. Therefore, the molecular clock hypothesis (null hypothesis) was rejected.



Maximum likelihood: null hypothesis states equal evolutionary rate throughout the tree (the concept of the molecular clock hypothesis). Based on this tree (5% significance level), the p-value was 4.85E-8. Therefore, the molecular clock hypothesis (null hypothesis) was rejected.