Fall 2022 Midterm

Daisy Fry-Brumit

2022-10-04

Question 1: What methods and models did I use to create the data and the reference alignment that I have provided you? (10 pts)

1A. Which of the 3 alignment methods available in the R msa() package did I use to make the provided reference alignment file? Be sure to explain your answer, and how you got there (don't just guess at one of the three). (5 pts)

To answer this question, I ran msa() using all 3 possible methods and output each result as a fasta file as seen in the code snippet below.

```
# initial sequence information
inputSeqs = readDNAStringSet('pumpkin-input.fa')

# start out by running msa() with all possible methods
cw_align = msa(inputSeqs, method = "ClustalW")
co_align = msa(inputSeqs, method = "ClustalOmega")
muscle_align = msa(inputSeqs, method = "Muscle")

# generate fasta files of the alignments
cw_align_align = msaConvert(cw_align, "bios2mds::align")
export.fasta(cw_align_align, outfile= 'cw_aln.fa')

co_align_align = msaConvert(co_align, "bios2mds::align")
export.fasta(co_align_align, outfile= 'co_aln.fa')

muscle_align_align = msaConvert(muscle_align, "bios2mds::align")
export.fasta(muscle_align_align, outfile= 'muscle_aln.fa')
```

After generating the .fa files, I used VerAlign online to compare my output files with *pumpkin-refaln.fa* as a reference, with results:

Clustal-W SP = 1.00 | CS = 0.99 | avg SPdist = 1.00

Clustal-O SP = 1.00 | CS = 1.00 | avg SPdist = 1.00

Muscle SP = 0.83 | CS = 0.60 | avg SPdist = 0.90

Although Clustal W generated a near-perfect match, Clustal-Omega generated an extact match to the rerence alignment. Thus, **the reference is a CLustal-Omega-generated alignment**.

1B. To create the input data set I provided you, I took real data and simulated additional mutations that would mimic the real evolutionary history of pumpkins. What mutation model did I use, and how did you come to that conclusion? Also, please provide a brief description of this model in terms of its parameters and assumptions. (5 pts)

To start I used modelTest() to return scores associated with each possible mutation model. To cover my bases, I did not subset the models used like we have previously in class.

```
# convert aligned sequences to phangorn friendly format
forPhang = msaConvert(co_align, type = "phangorn::phyDat")

# test all models on alignment
model_test = modelTest(forPhang)
```

Because I ran so many tests, I'm not going to print out the full table of results and eyeball the BIC values. I'll filter for the lowest BIC value instead:

```
model_test[which(model_test$BIC == min(model_test$BIC)), ]
```

```
##   Model df    logLik      AIC       AICw     AICc      AICcw      BIC
## 1    JC 55 -36188.77 72487.55 0.01152535 72491.81 0.01565718 72779.77
```

Based on the filter applied, it seems like the **Jukes Cantor model was used to generate mutations**. The JC model is a simple, one-parameter model that assumes

- base frequencies are equal
- rates of substitution among our bases are also equal (regardless of substitution type)

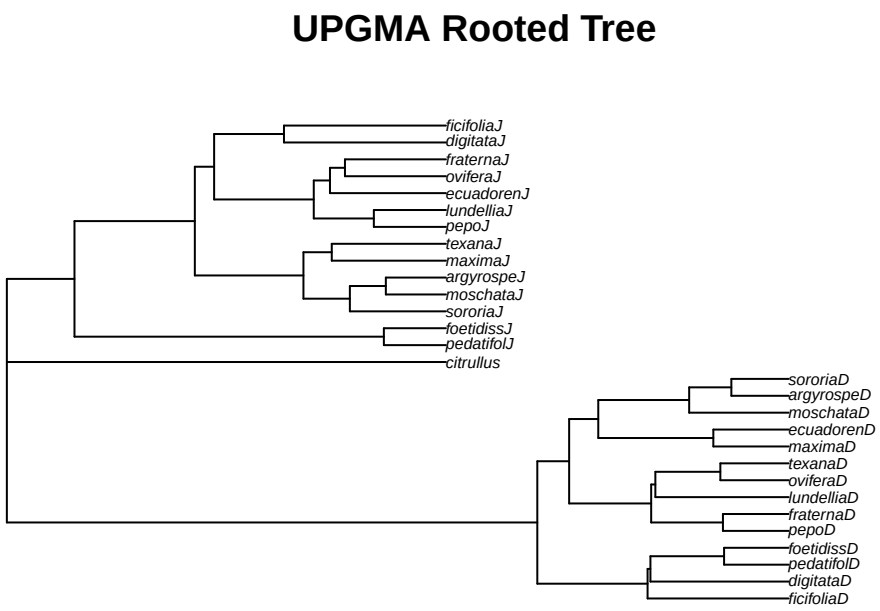
Question 2: Which set of orthologous genes best captures the evolutionary relationships shown in the reference tree that I provided? (10 pts)

To visualize the clades in the dataset and differentiate the orthologous groups, I have to build a tree. To do this, I started by using the JC69 model to get distance metrics, and then use those distance metrics to build and compare UPGMA and Neighbor-Joining tree methods.

```
# convert forPhang data into dist.dna useable format, then generate pairwise dist metrics
#names(forPhang) = gsub("D", "", names(forPhang))
#names(forPhang) = gsub("J", "", names(forPhang))
dna = as.DNAbin(forPhang)
D = dist.dna(dna, model='JC69')

# generate nj and upgma trees
upgma_tree = upgma(D)
nj_tree = nj(D)

# root using citrullus (watermelon) as outgroup
upgma_tree_root = root(phy=upgma_tree, outgroup="citrullus")
nj_tree_root = root(phy=nj_tree, outgroup="citrullus")

# input ref tree
refTree = ape::read.tree('pumpkin-refTree.nwk')
```

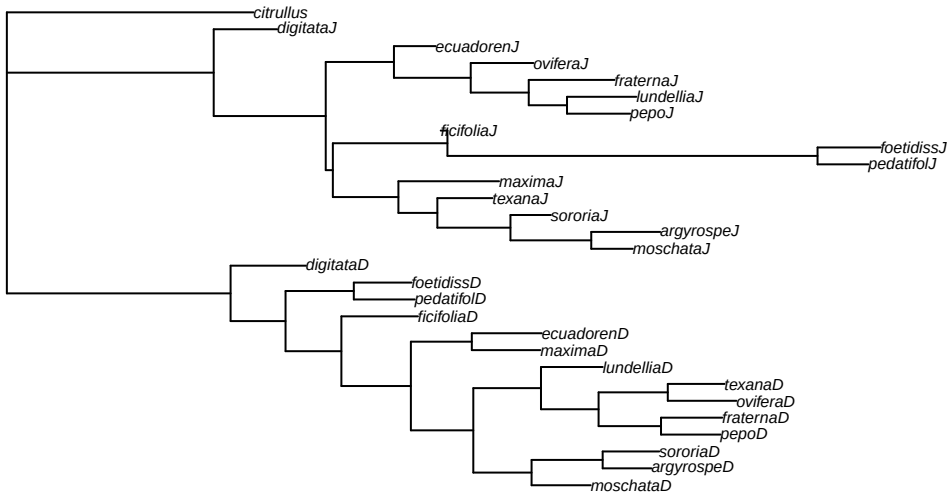
The resulting trees are as follows:

```
plot(upgma_tree_root, main="UPGMA Rooted Tree", cex=0.5)
```

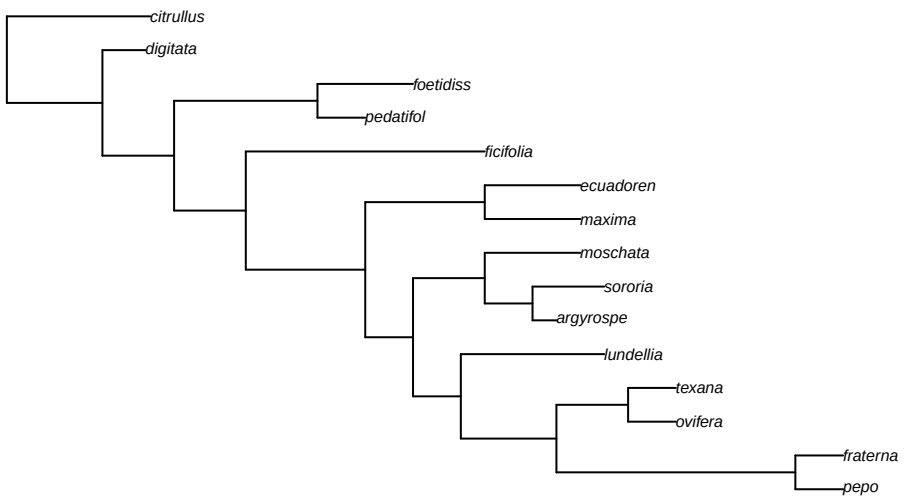



```
plot(nj_tree_root, main="NJ Rooted Tree", cex=0.5)
```

NJ Rooted Tree




```
plot(refTree, main="Reference Tree", cex = 0.5)
```



To me, it looks like the "D" orthologs generated by the Neighbor Joining method are the closest matches to the reference tree, but I want to use distance metrics to better evaluate the groups. To do this, I will compare the D and J groups from each tree building method to the reference tree (so 4 total comparisons using treedist). The results can be seen in the following table:

Table 1: D Group Comparison (NJ method)

	x
symmetric.difference	0.0000000
branch.score.difference	0.7377075
path.difference	0.0000000
quadratic.path.difference	5.1148061

Table 2: D Group Comparison (UPGMA method)

	x
symmetric.difference	8.0000000
branch.score.difference	0.8465216
path.difference	12.9614814
quadratic.path.difference	5.3590267

Table 3: J Group Comparison (NJ method)

	x
symmetric.difference	16.0000000
branch.score.difference	0.9430923
path.difference	20.4450483
quadratic.path.difference	4.7939323

Table 4: J Group Comparison (UPGMA method)

	x
symmetric.difference	20.0000000
branch.score.difference	0.9465929
path.difference	21.8632111
quadratic.path.difference	4.8352514

I can see from the trees and verify from the comparison tables that the "D" group generated by the Neighborhood Joining method has no unique clades when compared to the reference (symmetric.difference) and there is also no path difference. This leads me to believe that **the orthologous group containing "D" subunits on the chloroplast gene *psb* best captures the relationships provided by the reference.**

Note this part of the code took me a while to figure out (in terms of comparing each group separately to the reference) and so the code generating the displayed tables is longer and less efficient than I would care for. Thus, I'll upload the code for that separately so it doesn't ruin the flow of the current file.

Question 3: How many independent domestication events can you infer under the most likely transition cost model, and what is the most likely model? (10 pts)

Question 4: If you assume domestication is irreversible, then how many separate times did domestication occur in these species? (10 pts)

Question 5: Where was C. ficifolia (the fig-leaf gourd) most likely from? (10 pts)