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.

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

# start out by running msa() with all possible methods
cw_align = msa(inputSeqs, method = "ClustalW")
```

```
## use default substitution matrix
```

```r
co_align = msa(inputSeqs, method = "ClustalOmega")
```

```
## using Gonnet
```

```r
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 = | CS = | avg SPdist = **Muscle** SP = | CS = | avg SPdist =

Since Clustal-W has perfect SP and avg SPdist scores and a CS of 0.99, I feel safe in concluding that **the reference is a CLustal-W-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(cw_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 -36217.36 72544.72 0.01039447 72548.99 0.01440983 72836.99
```

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)

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)