Temporal Modeling and Gene Flow Analysis in Jumping Spiders

Samantha Taylor

01 December 2021

Overview

In this assignment, you will perform a molecular clock analysis in MrBayes to look at the history of divergence among a group of spider species. This analysis will involve multiple genes and partitions, and will rely on the node dating method for calibrating the molecular clock. Once you have obtained your tree with predicted divergence dates, you will be able to look at whether or not the diversification rates in this genus have remained constant over time.

For the second part of the exercise, you will focus on a small subset of closely-related individuals to test hypotheses about ongoing gene flow between some species.

About the Data Set

The sequence data that you will be working with comes from a collection of species belonging to the genus Habronattus, more commonly known as the jumping spiders. These spiders are of interest to evolutionary biologists because of their diverse color patterns, distinct courtship displays, and unusual sex chromosome evolution. Some people even keep and sell jumping spiders as pets, because of their interesting behaviors and "cute" faces.

The species in this genus have also diverged relatively recently (in the last 5 million years), and there is likely ongoing hybridization in certain clades that has made fully resolving the phylogeny historically challenging.

The dataset for the molecular clock analysis consists of 15 different Habronattus species, plus 1 outgroup species: Pellenes canadensis. There are 63 sequences for each species, which come from a set of noncoding loci.

Perform a Molecular Clock Analysis in MrBayes

IMPORTANT!!! Due to the number of partitions, the increased complexity of the clock analysis, and the fact that we need at least 10 million iterations to achieve convergence, this run takes roughly 8 hours to complete. Please plan accordingly!

Get a Concatenated Alignment and Partition File

Use R to create a concatenated alignment of all 63 loci, and save the final alignment in nexus format to run in MrBayes. Generate the partition information for MrBayes, and upload both files to the cluster. You can use MUSCLE as your alignment method.

```
options(stringsAsFactors=FALSE)

my.files = list.files(path="./spider-data", pattern="*.fasta", full.names=TRUE)
seqs = readDNAStringSet(my.files[1])
seqs = seqs[order(names(seqs))]
x = msa(seqs, method="Muscle", order="input")
outfile = gsub("fasta", "phy", my.files[1])
write.phylip(x, outfile)

my.lengths = rep(0, length(my.files))
my.lengths[1] = ncol(x)

for (i in 2:length(my.files)) {
    new = readDNAStringSet(my.files[i])
    new = new[order(names(new))]
    y = msa(new, method="Muscle", order="input")
    out = gsub("fasta", "phy", my.files[i])
    write.phylip(y,out)
    my.lengths[i] = ncol(y)
    temp = paste0(x,y)
    names(temp) = names(new)
    x = DNAStringSet(temp)
}
concat.aln = DNAMultipleAlignment(x)
write.nexus.data(as.DNAbin(concat.aln), file="SpidersConcat.nex")

### Create the partition file for MrBayes
my.starts = c(1)
for (i in 1:(length(my.lengths)-1)) {
     my.starts[i+1] = my.lengths[i] + my.starts[i]
}
my.ends = my.starts + (my.lengths-1)
my.loci = gsub("\\.fasta", "", basename(file.path("./spider-data/",my.files)))
info1 = paste(my.starts, my.ends, sep="-")
info2 = paste(my.loci, info1, sep="=")
info3 = paste("charset", info2, sep=" ")
info3 = paste(info3, ";", sep="")
loc.list = paste(my.loci, collapse=",")
list.w.length = paste(c(length(my.loci), loc.list), collapse=":")
info4 = paste(c("partition", "byLocus", "=", list.w.length), collapse=" ")
info4 = paste(info4, ";", sep="")
write.table(c(info3,info4), file="Spiders-MrB-part.txt", quote=FALSE, row.names=FALSE, col.names=FALSE)
```

Go ahead and paste the partition information into a MrBayes configuration file (i.e., the mb_input file we've used in past examples). Make sure you also edit your configuration file to read the correct input nexus alignment and use the partitioning scheme that you've set up for the 63 loci. Since these are non-coding loci, you do not need to worry about setting up different partitions for 1st, 2nd, and 3rd codons, etc.

Set your Outgroup as Pellenes.

Define the Model Parameters and Settings for the Partitions

Edit the model settings such that you will use the GTR + I model on all of your partitions. Ensure that all of your substitution model parameters are unlinked for all of your loci. Set your rate prior to be variable. Use [sed -i '/execute .*/ r Spiders-MrB-part.txt' mb_input].

Question 1:

Copy and paste the lines of code you used to specify your substitution model, the linking/unlinking of partitions, and the rate prior. Please do NOT copy and paste all of your partition information (i.e. all of the charset lines)! (1 point)

```
lset app=(all) nst=6 rates=propinv;
unlink revmat=(all) pinvar=(all) statefreq=(all) shape=(all);
prset applyto=(all) ratepr=variable;
Outgroup Pellenes;
mcmc ngen=10000000000 samplefreq=100 printfreq=1000;
sumt relburnin=yes burninfrac=0.25;
sumt relburnin=yes burninfrac=0.25;
```

Define the Node Dating Constraints

In order to calibrate our clock, we need to set up constraints based on some known divergence times of certain clades. Here is what we already know about this group:

1. The split between the outgroup, Pellenes, and the entire Habronattus clade is estimated to be about 5 million years old (±1.2 my). The most recent it could possibly be is 2.5 million years ago.

2. H. conjunctus, H. signatus, and H. hirsutus diverged from all of the other Habronattus species right after the split from Pellenes. This is estimated to have happened 4.8 million years ago (±1.4 my), and could not have happened any more recently than 2.1 million years ago.
3. H. signatus and H.hirsutus eventually split from H. conjunctus around 3.25 million years ago (±1.3 my). This time is relatively uncertain, and could have even been as recent as 500,000 years ago.
4. After the conjunctus, signatus, hirsutus clade diverged from the remaining Habronattus species, H. zapotecanus was the next to split off. This happened around 2.7 million years ago (± 0.5 my), and was at least 1.7 million years ago.
5. Finally, we also know that H. jucundus and H. borealis formed their own clade about 1.5 million years ago (± 0.25 my). The minimum age of this split is 1 million years.

Based on the information that I've given you, edit your MrBayes configuration file to define the constraint nodes, calibrate the ones we have divergence time information for, and adjust the prior settings to use the constraints and calibration times. You will want to look closely at the practice exercise to remember all of the settings you will need.

A few important notes:

1. In the practice exercise, we used ranges of numbers (e.g. 2-10) to define which taxa were grouped together at each constraint node. However, if you want to define a list of individual taxa rather than a range, you just separate each list element with a space. As an example, I'll give you the code for the first constraint, which contains the entire Habronattus clade (i.e. this has everyone except Pellenes, which is taxa #13): 'constraint Habronattus = 1-12 14-16;'

2. You can name each node constraint whatever you want, but you cannot use the exact same name as one of the species; each defined label in MrBayes needs to be unique.
3. Use a truncatednormal() distribution for ALL of your node age calibrations.
4. Do NOT add any prior settings for samplestrat or sampleprob (we will leave this at their default settings for this analysis).
5. Use lognorm(-5.5, 0.5) for the clockratepr.
6. You can use the same brlenspr and clockvarpr settings as we used in the practice exercise.

Question 2:

Copy and paste the lines of code you used to set up all of the node dating constraints, calibrations, and priors. (5 points)

```
constraint one = 1-12 14-16;
constraint two = conjunctus hirsutus signatus;
constraint three = hirsutus signatus;
constraint four = zapoteca;
constraint five = jucundus borealis;
calibrate one = truncatednormal(2.5,5,1.2);
calibrate two = truncatednormal(2.1,4.8,1.4);
calibrate three = truncatednormal(0.5,3.25,1.3);
calibrate four = truncatednormal(1.7,2.7,0.5);
calibrate five = truncatednormal(1,1.5,0.25);
prset nodeagepr = calibrated;
prset topologypr = constraint(one,two,three,four,five);
prset brlenspr = clock:uniform;
prset clockratepr = lognorm(-5.5, 0.5);
prset clockvarpr = igr;
```

Define the MCMC settings and Start the Run

Finish editing your configuration file by adding your mcmc settings and the sump and sumt commands. Use 10 million iterations for the chain, with a sampling frequency every 100 runs and a print frequency every 1,000. Make sure to not include the conformat=simple option in your sumt command.

Before submitting your job, you will likely also need to edit your slurm script to increase the time limit and the number of processors. I used 8 cores for my analysis (make sure to update the -np option in the mrbayes command line as well if you want to use more processors). With 8 threads and 64GB RAM, my job took a little over 8 hours to complete.

Plot the Results

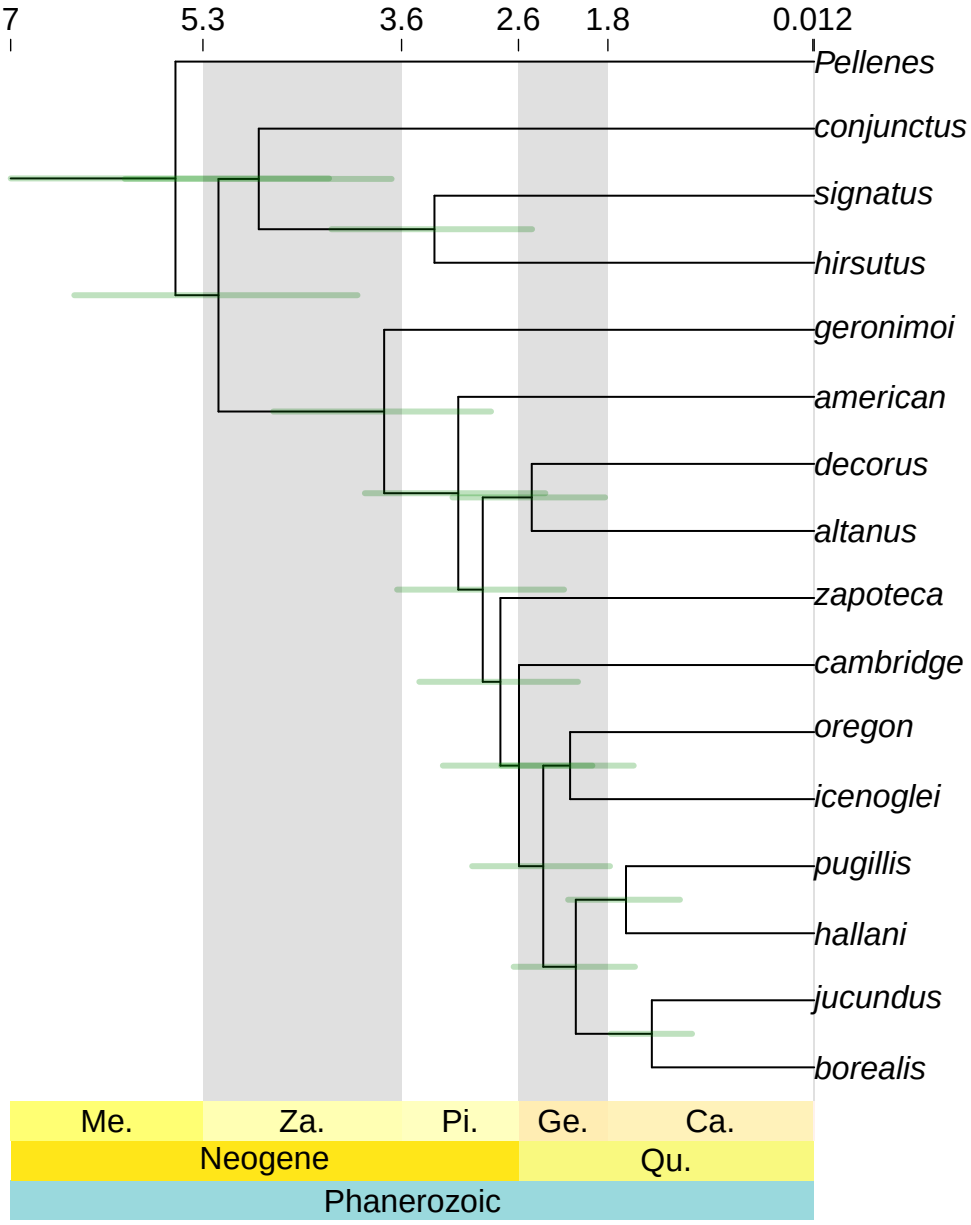
Once MrBayes finishes, go ahead and download the tree file to plot in R. You do not need to worry about checking the convergence diagnostics this time; I know from my own tests that our settings and number of iterations should be sufficient for good mixing.

Use the MCMCtreeR package to plot the scaled tree with divergence times. You may notice that our divergence times for this exercise are all much more recent than the times we had in the practice, so I suggest you alter the scale.res() settings in the plot function to be more informative. You can consult the MCMCtreeR help page for details about your options.

Question 3:

Provide your plot of the Habronattus phylogeny with divergence time estimates. (4 points)

```
tree.mb = 'SpidersConcat.nex.con.tre'
MCMC.tree.plot(analysis.type='mrbayes',
               directory.files=tree.mb,
               plot.type="phylogram", lwd.bar=3,
               time.correction=100,
               scale.res=c("Eon", "Period", "Age"),
               node.method='bar',
               col.age="#008b0040",
               no.margin=TRUE)
```



Used: [https://cran.r-project.org/web/packages/MCMCtreeR/vignettes/MCMCtree_plot.html]

Diversification Rates

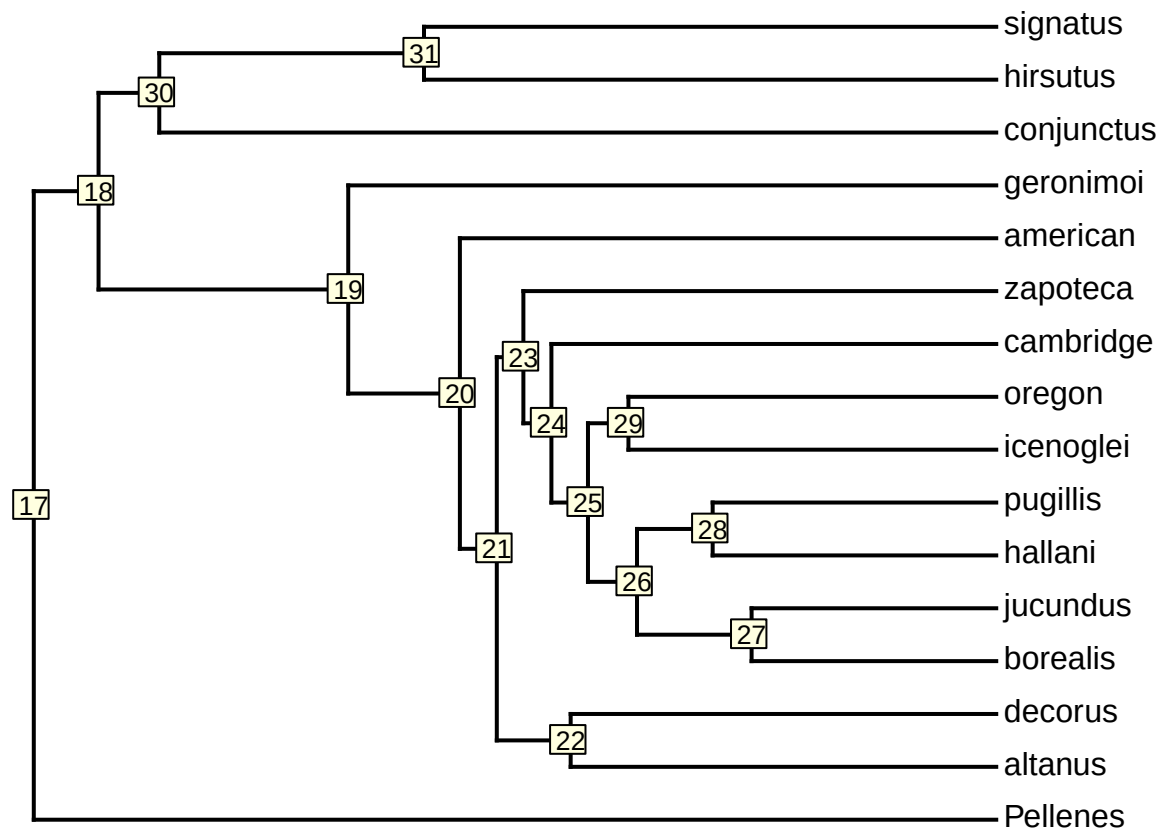
Most jumping spider species come from North America, and many of the species in our dataset live in the western United States, where much of the habitat is arid desert. However, it didn't used to be this way, and at different points during the last Ice Age this part of the country contained many large lakes and lush forests. One theory about the evolution of Habronattus is that the loss of of these wet habitats led to

fragmented populations which later became many new species. We can use our time-calibrated tree to test this theory by seeing if our branch times fit a continous birth/death model, or if there is a point in time where speciation seems to accelerate.

Let's start by just testing whether or not our data fit the null expectation for diversification rates. Use the read.nexus() function to read your MrBayes tree file into R. Then use the following code to perform a goodness-of-fit test for the null speciation model:

You can also embed plots, for example:

```
bayesTree = read.nexus("SpidersConcat.nex.con.tre")
plotTree(bayesTree, edge.width=2, font=1)
nodelabels(bayesTree$node.label, cex=0.8, bg = "lightyellow")
```




```
bt = branching.times(bayesTree)
bt = sort(bt)
diversi.gof(bt)
```

```
##
## Tests of Constant Diversification Rates
##
## Data: bt
## Number of branching times: 15
## Null model: exponential
##
## Cramer-von Mises test: W2 = 1.768   P < 0.01
## Anderson-Darling test: A2 = 2.825   P < 0.01
```

Question 4:

Based on the results of your test, does it appear that rates are not constant in this clade? (1 point)

Answer: The data does not follow normal distribution and it does appear rates are constant mainly.

We can also use a method called MEDUSA to pinpoint if there is a particular time in our tree where diversification rates appear to change. First, you need to install the geiger package.

Next, we need to create a species richness table to define how many taxa from each clade are not present in our tree. If the tip labels of your tree are in alphabetical order, you can use this code:

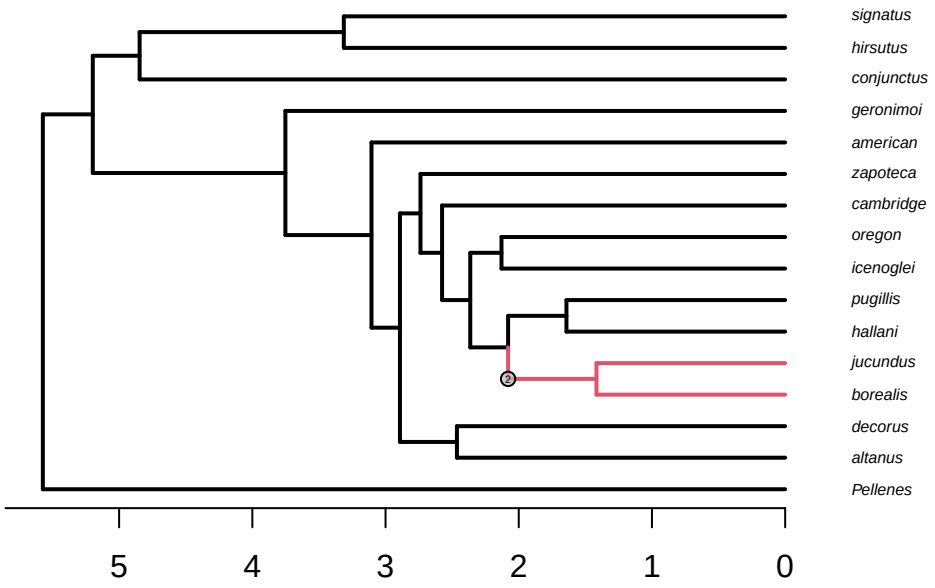
```
n.taxa = c(2,9,14,2,5,1,2,1,1,2,3,1,1,2,2,2)
richness = data.frame(taxon=bayesTree$tip.label, n.taxa=n.taxa)
```

Now run MEDUSA and plot the results:

```
medusa.res = medusa(bayesTree, richness=richness, warnings=FALSE)
```

```
## Appropriate  aicc-threshold for a tree of 16 tips is: 0.
##
## Step 1: lnLik=-49.83671; aicc=101.8113; model=yule
## Step 2: lnLik=-44.26669; aicc=95.42227; shift at node 27; model=yule; cut=stem; # shifts=1
##
## No significant increase in aicc score. Disregarding subsequent piecewise models.
##
## Calculating profile likelihoods on parameter values.
##
##   Model.ID Shift.Node Cut.At Model Ln.Lik.part       r epsilon     r.low
## 1        1         17   node  yule   -37.51955 0.44659      NA 0.3040539
## 2        2         27   stem  yule   -6.747143 1.51046      NA 0.8240196
##      r.high
## 1 0.6327324
## 2 2.6578433
```

```
plot(medusa.res, label.offset=0.5, edge.width=2)
```



Question 5:

Compare your plot from MEDUSA with your original plot with divergence times. Is there a particular point in time where diversification rates appear to increase? If so, what geological era or age does this coincide with? Does the ice age theory seem to be supported? Why or why not? (4 points)

Answer: Looking at the MEDUSA plot, it seems that there is a frequency shift at the common ancestor of the common ancestor of borealis and jucundus. It is in the Phanerozoic eon, cenozoic era, quaternary period and calabrian age. This is not the ice age period so the theory is not supported.

Testing for Gene Flow

For the last part of this assignment, you will be testing for the possibility of gene flow among a subset of more closely-related Habronattus species using the ABBA-BABA test.

One of the species from our MrBayes analysis, H. americanus, is known to live in very close proximity to 2 other jumping spider species: H. ophrys and H. tarsalis, and there is potential for hybridization between members of different species. To test for the presence and the direction of gene flow, we will be using 50,000 SNP markers sequenced in americanus, ophrys, tarsalis and an outgroup: H. signatus. These data are in nexus format

Use SVDQuartets to Get a Tree Topology

In order to perform the ABBA-BABA test, we need a starting tree topology. Since we only have SNP data for this group, SVDQuartets will be a good option for getting a tree. Upload the nexus file to the cluster, and run SVDQuartets through the PAUP package. You can refer to this practice page for instructions. After reading in your data but before running the analysis, type in "outgroup 4" to set the 4th taxon (signatus) as the outgroup. Do NOT run bootstrapping.

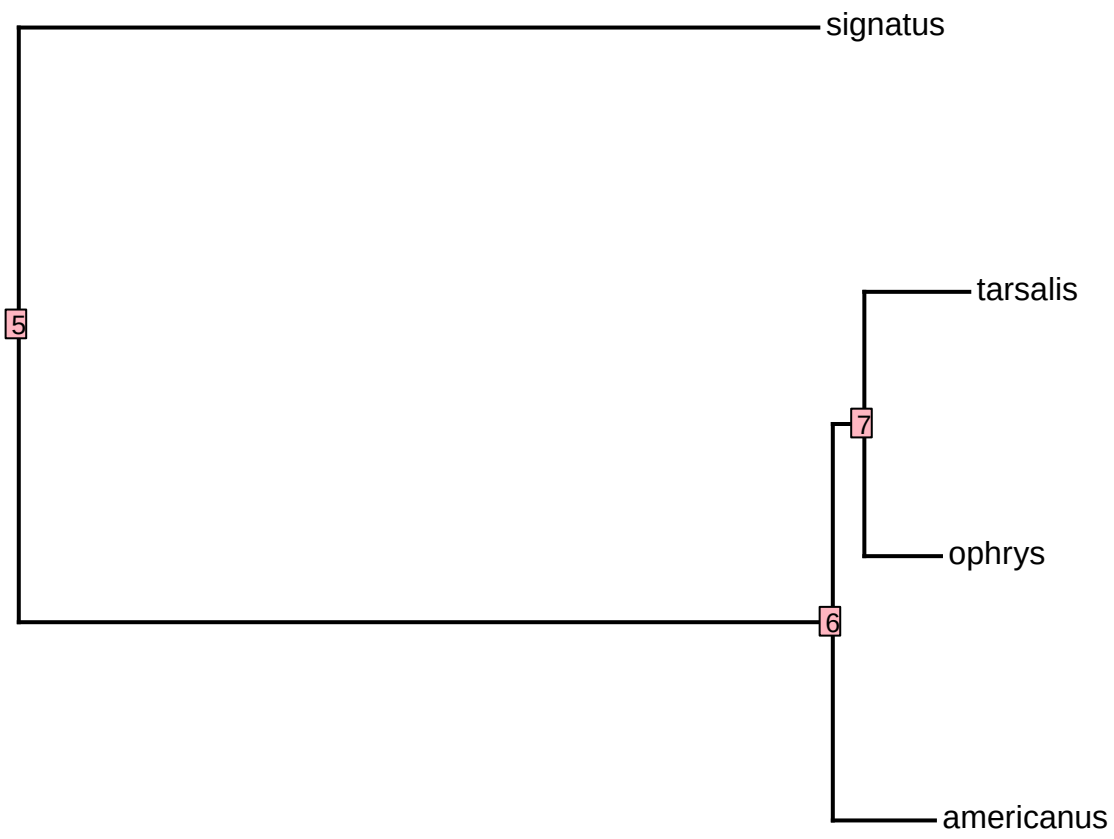
```
/projects/class/binf6205_001/paup4a168_centos64

paup> log file=paup_test.log
paup> exec spider-snps.nex
paup> outgroup 4
paup> svdq eval=all
paup> savetrees file=spider_test_tree.nex brlens=yes supportValues=Both
```

Save your tree, download it and plot it in R.

Question 6:

Provide the plot of your SVDQuartets tree. (1 point)

```
spider_best_tree = read.nexus("spider_test_tree.nex")
spider_best_tree2 = root(spider_best_tree, outgroup=c("signatus"), resolve.root=TRUE)
plotTree(spider_best_tree2, edge.width=2, font=1)
nodelabels(spider_best_tree2$node.label, cex=0.8, bg = "lightpink")
```




```
svd_snps.tree = read.nexus.data("spider-snps.nex")
```

Below is the general structure of the topology we use for the ABBA-BABA test. Based on your SVDQuartets tree, decide which of your taxa should be designated P1, P2, P3 and Out.

Question 7:

What are your species assignments for P1, P2, and P3? (2 points)

Answer: Looking at the reference in the assignment, P1 and P2 would be tarsalis and ophrys because they are the most related. P3 would be americanus because it is less related to tarsalis then ophrys is (and vice versa), but is more related to the two than signatus (the outgroup) is. Signatus is the outgroup do to it being the least related to the other three.

Perform the ABBA-BABA Test

We can use the R package evobiR to both calculate the D-statistic and test for significance.

The function we want to use requires use to have a FASTA file of our sequences, and they have to be in the order of P1, P2, P3, Out. To convert the SNP nexus file into this format, start by using the read.nexus.data() function from the phangorn or ape package. Then use the as.DNAbin function to convert the data to a DNAbin object.

Create a vector of taxa names in the order that you want to get your data into (the order will depend on what you've concluded about which species are P1, P2, P3, and Out.) The names need to match the labels that are in the tree.

```
YOURDNAbinObject = as.DNAbin(svd_snps.tree)
ordered.names = c("tarsalis", "ophrys", "americanus", "signatus")
```

Next, match up the vector of names in the order that you want with the current sequence labels, and then re-order the DNAbin object:

```
YOURDNAbinObject = as.DNAbin(svd_snps.tree)
m = match(ordered.names, labels(YOURDNAbinObject))
new.dna.bin = YOURDNAbinObject[m]
write.FASTA(new.dna.bin, file="spidersFA.fa")
```

From there, it is just one simple command to perform the test:

```
CalcD(alignment="spidersFA.fa", sig.test="B")
```

```
##
## performing bootstrap.....................................................................................
## Sites in alignment = 50000
## Number of sites with ABBA pattern = 554
## Number of sites with BABA pattern = 321
##
## D raw statistic / Z-score =  0.2662857  /  8.125676
##
## Results from  1000 bootstraps
## SD D statistic = 0.0327709
## P-value (that D=0) =  4.440892e-16
```

Question 8:

Is there evidence of gene flow between any of these species? If yes, then which pair of species is exchanging genes? (2 points)

Answer: Because of the positive D statistic and the fact that there are more ABBA sites than BABA sites, it can be concluded that there is gene flow. Therefor, P2 (ophrys) and P3 (americanus) are exchanging genes.
