Phylogenetics and Population Genetics in R (a primer)

Luis Amador

2023-03-02

Overview

In this tutorial students will learn basic phylogenetic and population genetics analyses using DNA sequences in the software R. The analytical software R is highly used for basic statistical programming, is easily accessible (it's free), has cross-platform compatibility (Windows, Mac, and Linux) (Láruson and Rees, 2021).

We will use scripts in R to build phylogenies and population structure plots based on a data set from DNA sequences of the brown anole *Anolis sagrei*. Scripts used in this tutorial follow or are modified from Cadotte and Davies (2016).

Required software

For this tutorial we need to install several programs on our computers. We can go to the posit website https://posit.co/download/rstudio-desktop/, where we can download everything we'll need to use R (R and RStudio for free).

Another software that we will use and need is Aliview (Larsson, 2014) which is an alignment viewer and editor. Aliview can be download at: https://ormbunkar.se/aliview/downloads/.

Installing R packages

The most frequently used package for phylogenetic analyses is *ape* (Analysis of Phylogenetics and Evolution; Paradis et al., 2004), and population genetics in R is the package *adegenet* (Exploratory Analysis of Genetic and Genomic Data https://github.com/thibautjombart/adegenet).

To install the packages first we need to open RStudio. Look for the RStudio application in your Downloads folder, or in the folder where you save the program. Once open RStudio we need to write the folllowing code:

```
#install the packages
install.packages(c("ape", "adegenet", "bios2mds", "pegas", "haplotypes", "phangorn"))
#Load the packages
library(ape)
library(adegenet)
library(bios2mds)
library(pegas)
library(haplotypes)
library(phangorn)
```

Brown anole data set

We will use the 321 sequences from the paper: "Genetic variation increases during biological invasion by a Cuban lizard", published by Kolbe et al. in the journal Nature. The sequences belong to the ND2 mitochondrial gene and are from several *Anolis sagrei* populations in the native and invasive range. Sequences from other *Anolis* species were used as an outgroup.

First, I extracted the GenBank Accession numbers from the supplementary material of Kolbe et al. (2014). See the documents in the folder called Anolis_taller. We read and obtained sequences with the function read.GenBank() of the package *ape*.

```
#Create a vector and get sequences from GenBank accession numbers used in Kolbe et al., 2004
Anolis_spp <- read.GenBank(c("AY655164", "AY655165", "AY655166", "AY655167", "AY655168", "AY655169", "A
class(Anolis_spp) #saved as a single DNAbin object
Anolis_spp #a brief summary
attributes(Anolis_spp) #list of attributes
names(Anolis_spp) #GenBank accession numbers
attr(Anolis_spp, "species") #get the names of species as a list
```

Next step is write our DNAbin object (Anolis_spp) as an alignment in the FASTA file format, this is one of the formats to use for building phylogenies, but first we have to visualize and edit our alignment.

```
#Write a fasta file
write.dna(Anolis_spp, file = "Anolis_spp.fasta", format = "fasta", append = FALSE, nbcol = 6, colsep =
```

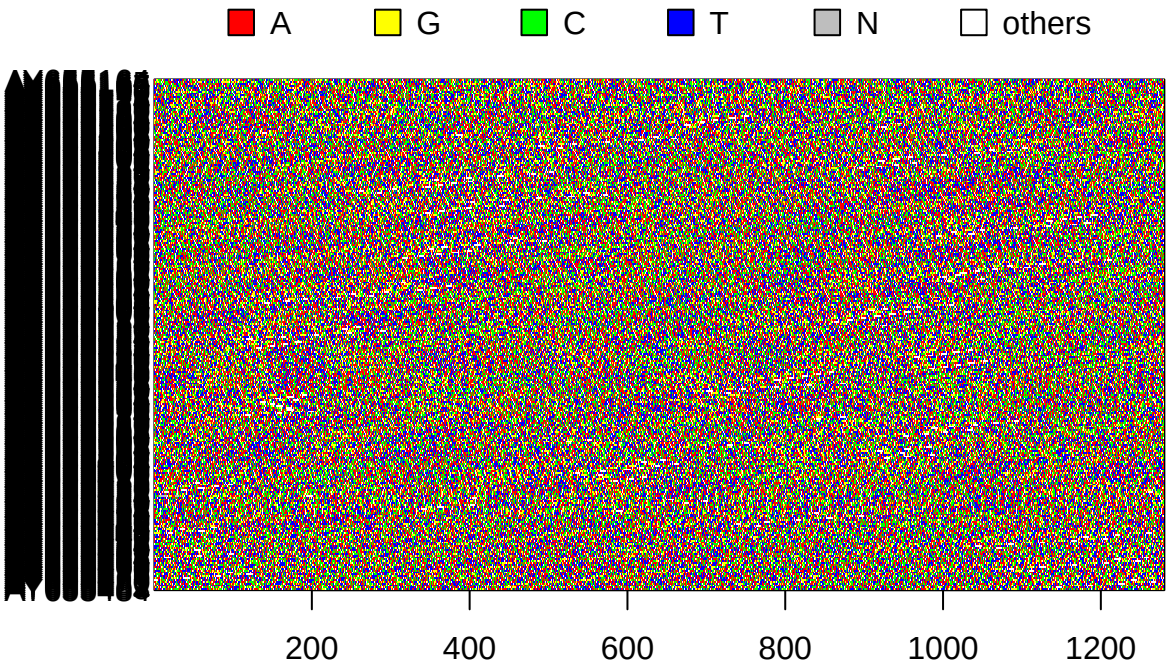
Once Anolis_spp.fasta was created, we have to open it in the software Aliview. Here we can visualize our alignment and edit it if it is necessary. In Aliview we edit and align our FASTA file and we save it as a new file in PHYLIP format.

```
#Read our new PHYLIP file and create a new DNAbin object for phylogentic analyses
Anolis_spp_dna <- read.dna(file = "Anolis_spp.phy")
class(Anolis_spp_dna)
```

Optional

We can also visualize our alignment in *ape* using the function image.DNAbin(), we need to convert our FASTA file to a new DNAbin object. However, to image.DNAbin() doesn't like sequences with different lenghts.

```
Anolis.DNAbin <- fasta2DNAbin("Anolis_spp.fasta", chunkSize = 10)
image.DNAbin(Anolis.DNAbin)
```



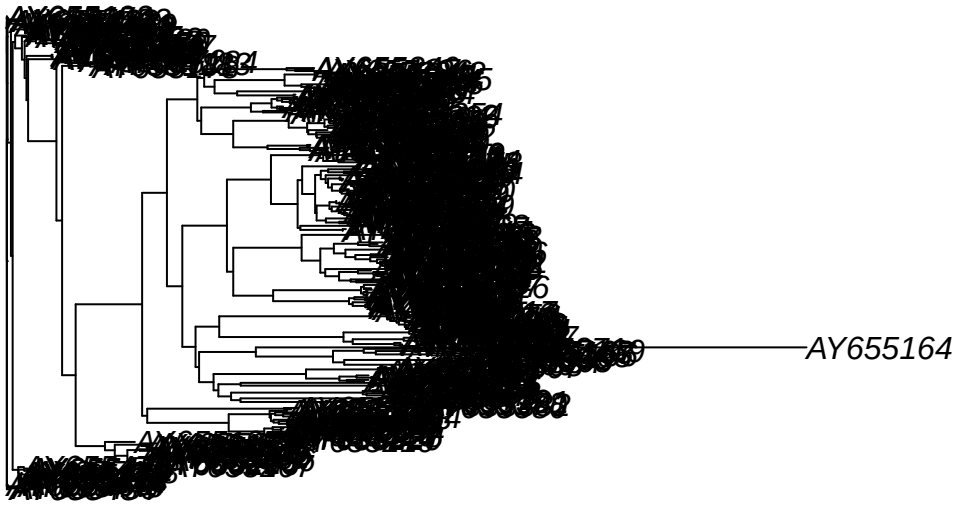
Phylogeny Reconstruction

To build phylogenies with our data set we will use several methods (or algorithms) that can be easily implemented in R.

Distance-Based

Distance-based methods generate a phylogenetic tree from a distance matrix calculated from the aligned sequences usinga neighbor-joining (NJ) method. Here wd use the njs() function in *ape*, on the DNA distance matrix estimated from our DNAbin object generated from our PHYLIP file obtained in Aliview.

```
Anolis.nj.tree <- njs(dist.dna(Anolis_spp_dna, model = "K80"))
plot(Anolis.nj.tree)
```

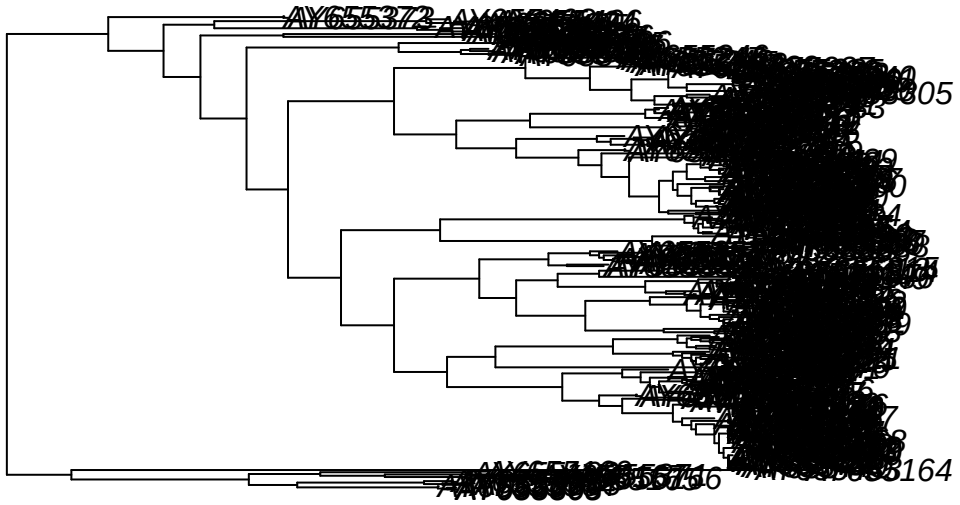



```
#Save the tree in our working directory
write.tree(Anolis.nj.tree, file = "Anolis_NJ.tree")
```

Parsimony

The principle of phylogeny reconstruction using maximum parsimony is to find the tree topology that minimizes the number of evolutionary steps required to explain the data (alignment of DNA sequences). To help the analysis along we provide a starting tree, in this case our NJ tree (Anolis.nj.tree), and the parsimony algorithm then proceeds by rearranging branches on the tree.

```
Anolis <- read.phyDat("Anolis_spp.phy", format = "phylip", type = "DNA")
Anolis.pars.tree <- optim.parsimony(Anolis.nj.tree, Anolis, rearrangements = "SPR")
#we can evaluate the length of the tree (number of evolutionary substitutions)
parsimony(Anolis.pars.tree, Anolis)
#Plot the tree.
Anolis.pars.tree <- acctran(Anolis.pars.tree, Anolis) #Using the ACCTRAN (accelerated transformation) o
plot(midpoint(Anolis.pars.tree))
```

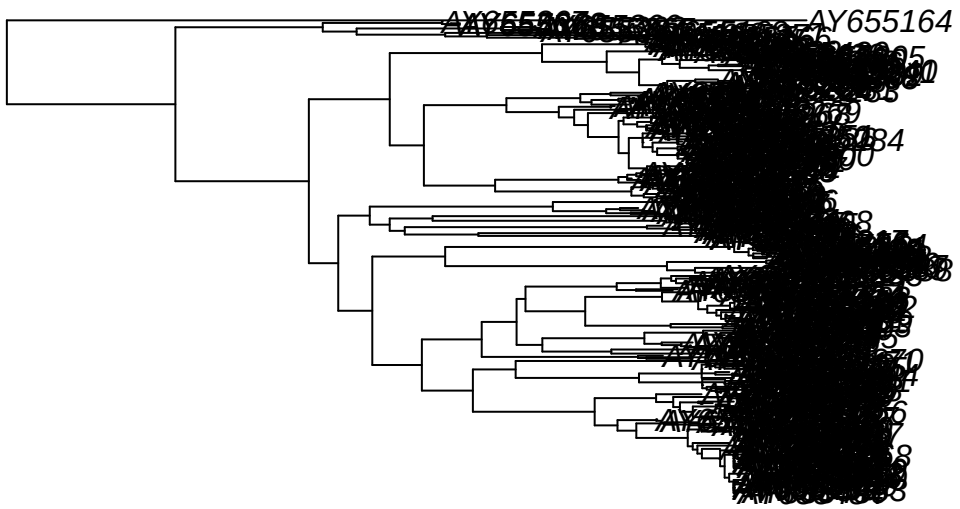


```
#Save the tree in our working directory
write.tree(Anolis.pars.tree, file = "Anolis_pars.tree")
```

Maximum Likelihood

Likelihood methods evaluate the likelihood of observing the data (aligned sequences) given a tree and a model of evolution. We again use the NJ tree as a starting tree.

```
#the likelihood of the tree given the data:
ml.model <- pml(Anolis.nj.tree, Anolis)
ml.model
#Optimize the tree
Anolis.ml.tree <- optim.pml(ml.model, model = "K80", optNni = TRUE)
Anolis.ml.tree
#We can plot our ML tree and compare it with the trees using NJ and parsimony
plot(midpoint(Anolis.ml.tree$tree))
```




```
#Save the tree in our working directory
write.tree(Anolis.ml.tree$tree, file = "Anolis_ML.tree")
```

Optional

We can visualize the trees in a program called FigTree, if you want download, you can do it from this GitHub repository: https://github.com/rambaut/figtree/releases where you can find compiled binaries for Mac, Windows and Linux.

References

Cadotte, M. W., & Davies, T. J. (2016). Phylogenies in Ecology: A Guide to Concepts and Methods. Princeton University Press. https://doi.org/10.23943/princeton/9780691157689.001.0001

Kolbe, J. J., Glor, R. E., Rodríguez Schettino, L., Lara, A. C., Larson, A., & Losos, J. B. (2004). Genetic variation increases during biological invasion by a Cuban lizard. Nature, 431(7005), 177–181. https://doi.org/10.1038/nature02807

Larsson A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics (Oxford, England), 30(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Láruson, A. J., Reed, F. A. (2021). Population Genetics with R: An Introduction for Life Scientists. Oxford University Press.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics, 20(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412