Tutorial for the R package `seraphim` 1.0

Plotting the dispersal history of viral lineages

Simon Dellicour

December 26, 2024

The present tutorial describes how to use the R package "`seraphim`" [1, 2] to exploit phylogenetically informed movement data obtained through continuous phylogeographic reconstruction [3] in order to plot the dispersal history of Yellow fever viral lineages in Brazil [4]. See also the package manual for further detail on its different functions.

The R package "`seraphim`" is hosted on GitHub (`https://github.com/sdellicour/seraphim`) and the first step is to install it using the "install_github()" function of the "`devtools`" package:

```
> install.packages("devtools"); library(devtools)
> install_github("sdellicour/seraphim/unix_OS") # for Unix systems
> install_github("sdellicour/seraphim/windows") # for Windows systems
```

Note that the installation of "`seraphim`" requires the preliminary installation of the following R packages: "ape", "doMC" (only available for Unix systems), "fields", "gdistance", "HDInterval", "ks", "phytools", "raster", "RColorBrewer", "rgeos", and "vegan". Once installed, the package has to be loaded as follows:

```
> library(seraphim)
```

This tutorial requires the Yellow fever virus (YFV) example files also available on the GitHub repository of the package (`https://github.com/sdellicour/seraphim/tree/master/tutorials`): (i) "YFV_1000.trees", a file containing 1,000 phylogenetic trees sampled from the post-burn-in posterior distribution of trees inferred for the YFV data set using the method of Lemey *et al.* [3] by Faria and colleagues [4]; (ii) "YFV_MCC.tree", a maximum consensus tree (MCC) tree file estimated from these 1,000 sampled trees; and (iii) "YFV_studyArea.asc", a raster file corresponding to the study area. This is out of the scope of the present tutorial but a detailed procedure on how to prepare and conduct a continuous phylogeographic reconstruction using the relaxed random walk (RRW) diffusion model [3] implemented in the software package BEAST 1.10 [5] is available

on the BEAST community website (`https://beast.community/workshop_continuous_diffusion_yfv`).

In the context of this tutorial, we will also need the additional package "`diagram`". This package can be installed and loaded using the following command lines in R:

```
> install.packages("diagram")
> library(diagram)
```

Step 1: extracting spatio-temporal information embedded in posterior trees

The first step is to extract the spatio-temporal information embedded in annotated phylogenetic trees sampled from the posterior distribution of a continuous phylogeographic analysis. The tree file "YFV_1000.trees" contains 1,000 trees sampled in the post-burn-in posterior distribution of trees. We will here use the "treeExtractions()" function to extract the information embedded in these 1,000 posterior trees. The "treeExtractions()" function first requires the definition of the following parameters: "localTreesDirectory" (name of the directory to create and where spatio-temporal information contained in each tree will be saved), "allTrees" (name of the tree file), "burnIn" (number of trees to discard as burn-in, i.e. a number defining a series of first trees in which no tree will be sampled – has to be set to "0" as burn-in trees are, in the present case, already discarded), "randomSampling" (boolean variable specifying if the trees have to be randomly sampled or sampled at the largest possible regular interval – not relevant in the present case as the trees have already been sampled), "nberOfTreesToSample" (number of trees to sample), "mostRecentSamplingDatum" (most recent sampling date in a decimal format) and "coordinateAttributeName" (attribute name used to indicate the geographic coordinates within the tree file).

```
> localTreesDirectory = "Extracted_trees"
> allTrees = scan(file="YFV_1000.trees", what="", sep="\n", quiet=TRUE)
> burnIn = 0
> randomSampling = FALSE
> nberOfTreesToSample = 1000
> mostRecentSamplingDatum = 2017.304
> coordinateAttributeNam = "location"
```

Once all these parameters have been specifying, the "treeExtractions()" function can be launched as follows:

```
> treeExtractions(localTreesDirectory, allTrees, burnIn, randomSampling,
nberOfTreesToSample, mostRecentSamplingDatum, coordinateAttributeName)
```

Note that in the R script of this tutorial, we also describe how to use the "postTreeExtractions()" function of the package to extract the spatio-temporal information embedded in posterior trees inferred by such a Bayesian continuous phylogeographic analysis (and the

use of this new function is actually recommended because, depending on the continuous phylogeographic inference settings, the "treeExtractions()" function can sometimes fail to perform the extractions, returning an error message).

Step 2: extracting spatio-temporal information embedded in the MCC tree

Spatio-temporal information embedded in a maximum clade consensus (MCC) tree can be extracted using the "readAnnotatedNexus()" function that was previously available in the package "`OutbreakTools`" [6]:

```
> mcc_tre = readAnnotatedNexus("YFV_MCC.tree")
```

We can then organise and save this information in a matrix were each row gathers the data related to a distinct phylogeny branch. We will do that with the "mccExtractions()" function applied as follows:

```
> mcc_tab = mccTreeExtraction(mcc_tre, mostRecentSamplingDatum)
> write.csv(mcc_tab, "YFV_MCC.csv", row.names=F, quote=F)
> mcc_tab = read.csv("YFV_MCC.csv", head=T)
```

Step 3: estimating the HPD region for each time slice

The functions "spreadGraphic1()" and "spreadGraphic2()" in "`seraphim`" can be used to estimate the HPD regions. The function "spreadGraphic1()" generates a raster made by the superimposition of distinct layers corresponding to successive time slices. Each time slice layer is build by estimating the HPD (highest posterior density) region based on all the ending positions of phylogenetic branches whose ending time falls within the considered time slice. Alternatively, instead of a unique raster, "spreadGraphic2" produces a list of distinct spatial polygon data frames, with one data frame for each time slice. See also, for instance, the programs SPREAD 4 [7] and spread.gl [8] for a very similar approach. Here, we use the "spreadGraphic2" to estimate and display the uncertainty related to the continuous phylogeographic inference. The different parameters of the "spreadGraphic2()" function have to be specified as follows:

```
> nberOfExtractionFiles = nberOfTreesToSample
> prob = 0.95
> precision = 0.025
> startDatum = min(mcc_tab[,"startYear"])
```

where "nberOfExtractionFiles" is the number of extraction of files to use (this number cannot be higher that the number of extractions specified by "nberOfTreesToSample" in the first step), "prob" is the probability that will be used to estimate the HPD (highest posterior density) regions, "precision" is the time interval that will be used to define

the successive time slices, and "startDatum" is the number in a decimal format defining the beginning of the dispersal history. Once these parameters are specified, the "spreadGraphic2()" function can be called as follows:

```
> polygons = suppressWarnings(spreadGraphic2(localTreesDirectory,
nberOfExtractionFiles, prob, startDatum, precision))
```

Step 4: defining the different colour scales to use

Here, we will define two distinct sets of colours based on the same colour scale "RdYlGn" generated by the "RColorBrewer" package: (i) "endYears_colours" defining the colour of each node in the MCC tree (i.e. the colour assigned to the descendent node of each branch summarised in "mcc_tab"), and (ii) "polygons_colours" defining the colour of each HPD polygon in the "polygons" list of spatial polygon data frames.

```
> colour_scale = colorRampPalette(brewer.pal(11,"RdYlGn"))(141)[21:121]
> minYear = min(mcc_tab[,"startYear"]); maxYear = max(mcc_tab[,"endYear"])
> endYears_indices = (((mcc_tab[,"endYear"]-minYear)/(maxYear-minYear))*100)+1
> endYears_colours = colour_scale[endYears_indices]
> polygons_colours = rep(NA, length(polygons))
> for (i in 1:length(polygons)) {
>        date = as.numeric(names(polygons[[i]]))
>        polygon_index = round((((date-minYear)/(maxYear-minYear))*100)+1)
>        polygons_colours[i] = paste0(colour_scale[polygon_index],"40")
> }
```

Step 5: co-plotting the HPD regions and MCC tree

For this final step, we first need to load a template raster that will serve as background to plot the dispersal history of viral lineages, as well as the subnational admin-1 borders of Brazil. These administrative borders are here downloaded from the GADM database (https://gadm.org) and directly cropped using the extent of the template raster:

```
> template_raster = raster("YFV_studyArea.asc")
> borders = crop(getData("GADM", country="BRA", level=1), extent(template_raster))
```

HPD regions and MCC tree can then be co-plotted using the script below. This script uses the "curvedarrow()" function from the package "diagram" to plot phylogeny branches as curves (and you can edit its "curve" parameter to play with the strength of curves, "curve" set to "0" returning straight line segments). Note that the script below shows how to plot tree nodes and HPD regions following the same colour scale. The result is displayed in Figure 1.

```
> dev.new(width=6, height=6.3)
> par(mar=c(0,0,0,0), oma=c(1.2,3.5,1,0), mgp=c(0,0.4,0), lwd=0.2, bty="o")
> plot(template_raster, col="white", box=F, axes=F, colNA="grey90", legend=F)
> for (i in 1:length(polygons)) {
```

```
>       plot(polygons[[i]], axes=F, col=polygons_colours[i], add=T, border=NA)
> }
> plot(borders, add=T, lwd=0.1, border="gray10")
> for (i in 1:dim(mcc_tab)[1]) {
>       curvedarrow(cbind(mcc_tab[i,"startLon"],mcc_tab[i,"startLat"]),
>                   cbind(mcc_tab[i,"endLon"],mcc_tab[i,"endLat"]), arr.length=0,
>                   arr.width=0, lwd=0.2, lty=1, lcol="gray10", arr.col=NA,
>                   arr.pos=F, curve=0.1, dr=NA, endhead=F)
> }
> for (i in dim(mcc_tab)[1]:1) {
>       if (i == 1) {
>            points(mcc_tab[i,"startLon"], mcc_tab[i,"startLat"], pch=16,
>                   col=colour_scale[1], cex=0.8)
>            points(mcc_tab[i,"startLon"], mcc_tab[i,"startLat"], pch=1,
>                   col="gray10", cex=0.8)
>       }
>       points(mcc_tab[i,"endLon"], mcc_tab[i,"endLat"], pch=16,
>              col=endYears_colours[i], cex=0.8)
>       points(mcc_tab[i,"endLon"], mcc_tab[i,"endLat"], pch=1,
>              col="gray10", cex=0.8)
> }
> rect(xmin(template_raster), ymin(template_raster), xmax(template_raster),
>      ymax(template_raster), xpd=T, lwd=0.2)
> axis(1, c(ceiling(xmin(template_raster)), floor(xmax(template_raster))),
       pos=ymin(template_raster), mgp=c(0,0.2,0), cex.axis=0.5, lwd=0, lwd.tick=0.2,
       padj=-0.8, tck=-0.01, col.axis="gray30")
> axis(2, c(ceiling(ymin(template_raster)), floor(ymax(template_raster))),
       pos=xmin(template_raster), mgp=c(0,0.5,0), cex.axis=0.5, lwd=0, lwd.tick=0.2,
       padj=1, tck=-0.01, col.axis="gray30")
> rast = raster(matrix(nrow=1, ncol=2))
> rast[1] = min(mcc_tab[,"startYear"])
> rast[2] = max(mcc_tab[,"endYear"])
> plot(rast, legend.only=T, add=T, col=colour_scale, legend.width=0.5,
       legend.shrink=0.3, smallplot=c(0.40,0.80,0.14,0.155), legend.args=list(text="",
       cex=0.7, line=0.3, col="gray30"), horizontal=T, axis.args=list(cex.axis=0.6,
       lwd=0, lwd.tick=0.2, tck=-0.5, col.axis="gray30", line=0, mgp=c(0,-0.02,0),
       at=seq(2016.4,2017.2,0.2)))
```

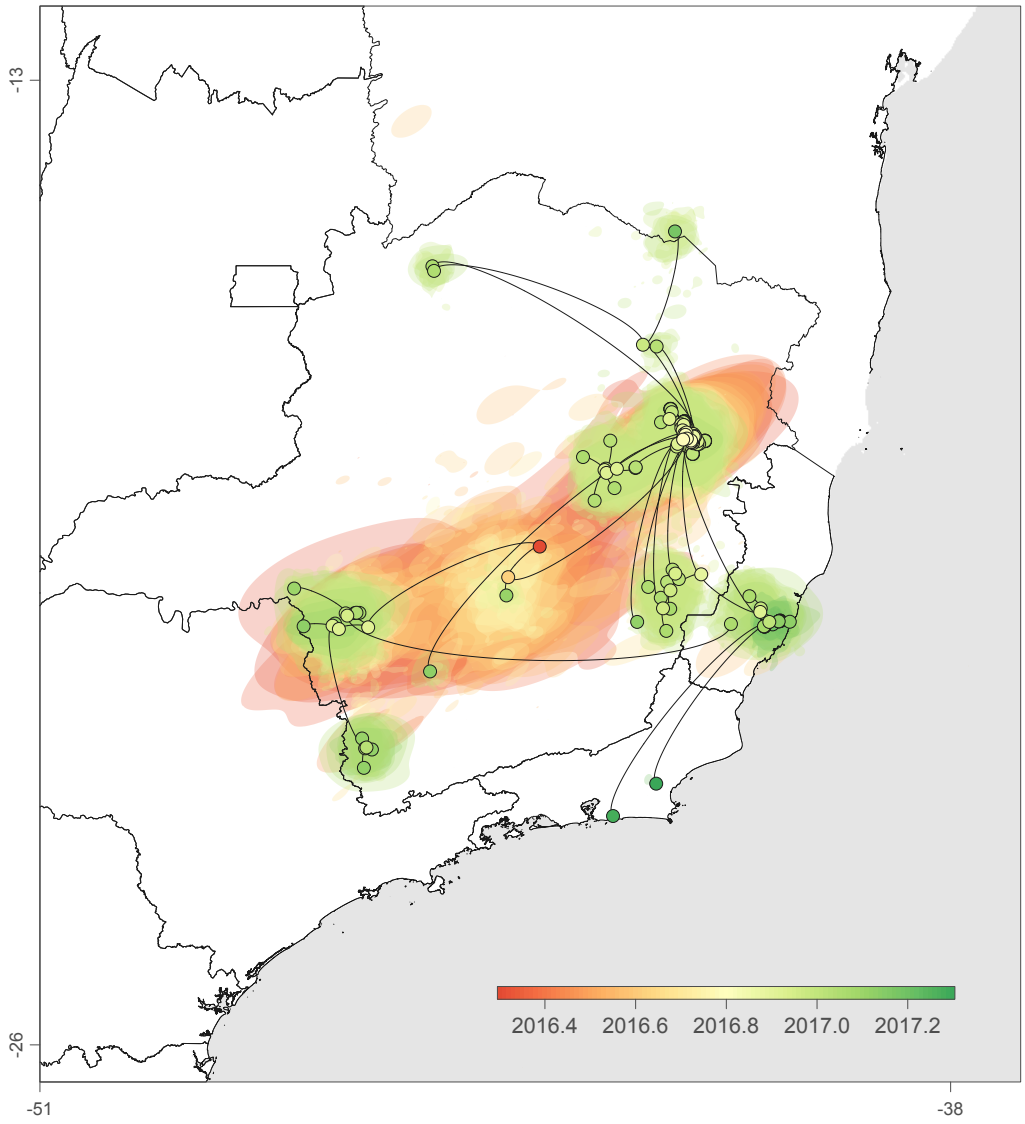

Figure 1: reconstructed spatiotemporal dispersal history of YFV lineages in Brazil based on the data set published by Faria *et al.* [4]: maximum clade credibility (MCC) trees and 80% HPD regions based on 1,000 trees subsampled from the posterior distribution of a continuous phylogeographic analysis. Nodes of the MCC tree are coloured according to their time of occurrence. 80% HPD regions were computed for successive time layers and then superimposed using the same colour scale reflecting time.

References

[1] Dellicour S, Rose R, Pybus OG (2016a). Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics* 17: 82.

[2] Dellicour S, Rose R, Faria N, Lemey P, Pybus OG (2016b). SERAPHIM: studying environmental rasters and phylogenetically-informed movements. *Bioinformatics* 32 (20): 3204-3206.

[3] Lemey P, Rambaut A, Welch JJ, Suchard MA (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology & Evolution* 27: 1877-1885.

[4] Faria NR, Kraemer MUG, Hill SC, Goes de Jesus J, de Aguiar RS, Iani FCM, Xavier J, Quick J, du Plessis L, Dellicour S, Thézé J, Carvalho RDO, Baele G, Wu CH, Silveira PP, Arruda MB, Pereira, MA, Pereira GC, Lourenço J, Obolski U, Abade L, Vasylyeva TI, Giovanetti M, Yi D, Weiss DJ, Wint GRW, Shearer FM, Funk S, Nikolai B, Adelino TER, Oliveira MAA, Silva MVF, Sacchetto L, Figueiredo PO, Rezende IM, Mello EM, Said RFC, Santos DA, Ferraz ML, Brito MG, Santana LF, Menezes MT, Brindeiro RM, Tanuri A, dos Santos FCP, Cunha MS, Nogueira JS, Rocco IM, da Costa AC, Komninakis SCV, Azevedo V, Chieppe AO, Araujo ESM, Mendonça MCL, dos Santos CC, dos Santos CD, Mares-Guia AM, Nogueira RMR, Sequeira PC, Abreu RG, Garcia MHO, Alves RV, Abreu AL, Okumoto O, Kroon EG, de Albuquerque CFC, Lewandowski K, Pullan ST, Carroll M, Sabino EC, Souza RP, Suchard MA, Lemey P, Trindade GS, Drumond BP, Filippis AMB, Loman NJ, Cauchemez S, Alcantara LCJ, Pybus OG (2018). Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* 361: 894-899.

[5] Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* 4: vey016.

[6] Jombart T, Aanensen DM, Baguelin M, Birrell P, Cauchemez S, Camacho A, Colijn C, Collins C, Cori A, Didelot X, Fraser C, Frost S, Hens N, Hugues J, Höhle M, Opatowski L, Rambaut A, Ratmann O, Soubeyrand S, Suchard MA, Wallinga J, Ypma R, Ferguson N (2014). OutbreakTools: A new platform for disease outbreak analysis using the R software. *Epidemics* 7: 28-34.

[7] Nahata K, Bielejec F, Monetta J, Dellicour S, Rambaut A, Suchard MS, Lemey P (2022). SPREAD 4: online visualization of pathogen phylogeographic reconstructions. *Virus Evolution* 8: veac088.

[8] Li Y, Bollen N, Hong SL, Brusselmans M, Gambaro F, Suchard MA, Rambaut A, Lemey P, Dellicour S, Baele G (2024). Spread.gl: visualising pathogen dispersal in a high-performance browser application. *Bioinformatics* 40: btae721.