Tutorial for the R package `seraphim` 1.0

Simulating relaxed random walk (RRW) diffusion along phylogenetic trees

Simon Dellicour, Philippe Lemey

June 17, 2019

The present tutorial describes how to use the "simulatorRRW1" function of the package "`seraphim`" [1] to simulate a relaxed random walk (RRW) diffusion process based on parameters previously estimated with the continuous phylogeographic model implemented in BEAST [2]. This tutorial explains how to simulate a RRW along a phylogenetic tree sampled in a posterior distribution of trees obtained from a continuous phylogeographic analysis of an Ebola virus dataset from the recent West Africa outbreak (2013-16)[3]. The present data set only contains phylogenetic branches associated with a dispersal distance smaller than 250 km. See also the package manual for further details. The first step is to download the package (`http://evolve.zoo.ox.ac.uk/Evolve/Software.html`) and place the "seraphim_1.0.tar.gz" file in a R workspace directory. The package can then be installed from this archive file using the following R command:

```
> install.packages("seraphim_1.0.tar.gz", repos=NULL, type="source")
```

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

```
> library(seraphim)
> library(OutbreakTools)
```

This tutorial requires the EBOV example files also available at `http://evolve.zoo.ox.ac.uk/Evolve/Software.html`. These files are "EBOV_cauchy.tree", a file containing the phylogenetic tree mentioned above, "EBOV_cauchy.log", a corresponding BEAST log file associated with this tree and that contains the RRW parameters estimated by BEAST, and "Empty_raster.asc", a raster file corresponding to the study area and with uniform cell values equal to "1" and "NaN" values on sea/non accessible areas.

Step 1: extraction of the spatio-temporal information contained in the tree

The first step is to extract the spatio-temporal information contained in the annotated phylogenetic tree. We will here use the "treeExtractions" function to extract the information contained in this post burin-in tree sampled in this posterior distribution. 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 the 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), "randomSampling" (boolean variable specifying if the trees have to be randomly sampled or sampled at the largest possible regular interval), "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 = "."
> allTrees = scan(file="EBOV_cauchy.tree", what=", sep="\n", quiet=TRUE)
> burnIn = 0
> randomSampling = FALSE
> nberOfTreesToSample = 1
> mostRecentSamplingDatum = 2015.696
> coordinateAttributeNam = "coordinates"
```

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

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

Step 2: simulations of a RRW diffusion process along the tree

The second step of this tutorial consists in using the "simulatorRRW1" function to simulate a RRW diffusion along the branches of the tree for which RRW parameters were previously estimated by BEAST. The "simulatorRRW1" function requires the user to specify (i) the "tree" along which the RRW diffusion process has to be simulated, (ii) the "rates" vector of trait evolutionary rates estimated by BEAST and associated with each branch, (iii) the vector of scale parameters "sigmas" for the latitude and longitude (retrieved from BEAST output; see below), (iv) the observed correlation "cor" between latitude and longitude (estimated from BEAST output; see below), (v) the "envVariables" list of environmental rasters that will determine the accessible area (simulated node positions will not fall into raster cells with "NaN" values), (vi) the most recent datum of sampling in decimal format, (vii) the vector of geographic coordinates "ancestPosition" (longitude, latitude) of the most ancestral node position (i.e. the starting position of

the RRW simulation), (viii) the "reciprocalRates" boolean variable corresponding to the BEAST parameter of the continuous diffusion model and specifying if branch lengths have to divided (TRUE) or multiplied (FALSE) by the trait evolutionary rates, (ix) the maximum number "n1" of branch rotations allowed per simulated node position (see details below), (x) the maximum number "n2" of re-simulations allowed per node position, (xi) the "showingPlots" boolean variable specifying if the different plots have to be displayed or not, and (xii) the "newPlot" boolean variable specifying if a new plot window has to be opened or if the simulated branches/nodes have to be plotted on a previously opened plot window e.g. with a specific raster and/or polygons already mapped on it (graphical parameter that is only useful when showingPlots=TRUE).

If a simulated node position falls into an inaccessible area (i.e. a raster cell with a NaN value), the simulator can rotate the branch around its oldest node position and while maintaining the geographic distance travelled by the branch until its youngest node position falls into an accessible area. In that situation, "n1" defines the maximum number of times that this rotation trial can be attempted. After "n1" rotation trials, the branch is not rotated anymore and the youngest node position and, if "n2" is higher than "0", the youngest node position is re-simulated (and hence its actual distance from the oldest node of the branch). In that situation, "n2" thus defines the maximum number of times that a new position is re-simulated (and then also rotated up to n1 times) before restarting the entire RRW simulation from the root of the tree. If "n2" is set to "0", the entire RRW simulation is immediately restarted from the root of the tree.

The script below includes several steps: (i) the extraction of the "rates" values from the tree file with the "read.annotated.nexus" function of the "OutbreakTools" package, (ii) the estimation of the correlation value based on the "TreeExtraction" file generated above, and (iii) the computation of "sigmas" values based on RRW parameters reported in the BEAST log file. Once all the parameter values are specified, the "simulatorRRW1" can be called and its output saved in a ".csv" file. Note the R file corresponding to this tutorial also includes a script to display the simulation outputs and compared the inferred and simulated branch positions on a map.

```
> tree = read.annotated.nexus("EBOV_cauchy.tree"); rates = c()
> for (i in 1:length(tree$annotations)) rates = c(rates, tree$annotations[[i]]$rate)
> tab = read.csv("TreeExtractions_1.csv",sep=",")
> log = read.table("EBOV_cauchy.log", header=T)
> col11 = log[1,"treeLengthPrecision1"]
> col12 = log[1,"treeLengthPrecision3"]
> col22 = log[1,"treeLengthPrecision2"]
> my_prec = c(col11, col12, col12, col22)
> my_var = solve(matrix(my_prec,nrow=2))
> sigma1 = sqrt(my_var[1,1])
> sigma2 = sqrt(my_var[2,2])
> sigmas = c(sigma1, sigma2)
> cor = my_var[1,2]/(sqrt(my_var[1,1])*sqrt(my_var[2,2]))
> envVariables = list(raster("Empty_raster.asc"))
> ancestID = which(!tab[,"node1"]\%in\%tab[,"node2"])[1]
> ancestPosition = c(tab[ancestID,"startLon"], tab[ancestID,"startLat"])
```

```
> reciprocalRates = FALSE
> n1 = 100
> n2 = 0
> showingPlots = TRUE
> newPlot = TRUE

> sim = simulatorRRW1(tree, rates, sigmas, cor, envVariables,
mostRecentSamplingDatum, ancestPosition, reciprocalRates,
showingPlots, newPlot)
> write.csv(sim, "TreeSimulation_1.txt", row.names=F, quote=F)
```


Acknowledgments

We are grateful to A. Rambaut and M.A. Suchard for their contribution to the original RRW paper [2], and John Welch in particular for contributing R scripting.

References

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

[2] 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.

[3] Dellicour S, Baele G, Dudas G, Faria NR, Pybus OG, Suchard MA, Rambaut A, Lemey P (2018). Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nature Communications* 9: 2222.