Tutorial for the R package `seraphim` 1.0

Simulating a relaxed random walk (RRW) diffusion process with a dispersal velocity impacted by an environmental raster

Simon Dellicour, Philippe Lemey

December 31, 2024

The present tutorial describes how to use the "simulatorRRW3()" function of the R package "`seraphim`" [1, 2] to conduct phylogeographic simulations based on a birth-death process and a relaxed random walk (RRW) diffusion process during which the dispersal velocity of lineages is impacted by an environmental raster. This function thus conduct forward-in-time joint simulations of both time-scaled phylogenies and the dispersal history of their branches on an underlying geo-referenced grid (raster) whose environmental values impact the dispersal velocity of lineages. At each time step of those RRW simulations, both the longitudinal and latitudinal displacements of evolving lineages are randomly drawn from a Gaussian distribution whose standard deviation is proportional to the underneath raster cell value while preventing lineage dispersal in inaccessible raster cells. Specifically, this tutorial explains how to simulate a RRW diffusion process [3] with simulation parameters (starting time and position, sampling time window) based on the study of a rabies virus spread in a North American raccoon population [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 "Elevation_rast.asc" raster file also available on the GitHub repository of the package (`https://github.com/sdellicour/seraphim/tree/master/tutorials`).

Step 1: preparation of the environmental rasters

The first step is to load the environmental raster i.e., in this case, the elevation layer for the study area. Once loaded, two transformations are performed on this raster: (i) negative values are set to "0" and (ii) all the positive values are rescaled between "0" and "10". The rescaling step is not compulsory but helps to figure out the relative impact of each cell value on the dispersal velocity.

```
> envVariable = raster("Elevation_rast.asc")
> envVariable[(envVariable[]<0)] = 0
> envVariable[] = (envVariable[]/max(envVariable[],na.rm=T))*10
```

Step 2: simulation of a RRW diffusion process on the rescaled raster

The "simulatorRRW3()" function requires the user to specify (i) the "envVariable" environmental raster on which the RRW diffusion process has to be simulated, (ii) the "resistance" boolean value specifying if the environmental raster as to be treated as a resistance or conductance factor, (iii) the "scalingValue" value used to rescale the standard deviation of the two normal distributions with mean = 0 in which longitudinal and latitudinal displacement values are randomly picked (see the package manual for further details), (iv) "ancestPosition" vector of geographic coordinates (longitude, latitude) of the most ancestral node position, i.e. the starting position of the RRW diffusion simulation, (v) the "birthRate" value defining the rate at which lineages are splitted in two new lineages (expressed in events per lineage per time unit), (vi) the "samplingRate" value defining the rate at which lineage are sampled and thus stopped moving and splitting on the raster (expressed in events per lineage per time unit), (vii) the "startingYear" year (or time in different units) at which the simulation begins from the ancestral node position, (viii) the "samplingWindow" vector of time values (in the same time unit as the "startingYear") defining the starting and ending times of the sampling period, i.e. the period during which the sampling rate will be effective, (ix) the "timeSlice" time interval (in the same time units as the "startingYear") at which unsampled lineages perform a new movement on the raster, (x) the "timeIntervale" time interval (in the same time units as the "startingYear") at which the plot displaying the simulation is updated (only taken into account when "showingPlots=TRUE"), (xi) the "showingPlots" boolean variable specifying if the different plots have to be displayed or not, (xii) the "extractionOfValuesOnMatrix" boolean variable specifying if the raster object has to be preliminary transformed into a matrix object. This operation allows a significantly faster simulation but has to be avoided if possible. Indeed, with this transformation, the raster and the ancestral position are projected on a flat grid and thus lose their initial projection during the whole simulation. This aspect can, in some cases, lead to over-simplistic situations. Note that if this option is selected, coordinates of simulated nodes are re-projected on the initial raster at the end of the simulation.

Note that in this example, "ancestPosition" and "startingYear" values were estimated from 100 trees sampled from the posterior distribution of trees inferred for the real "rabies

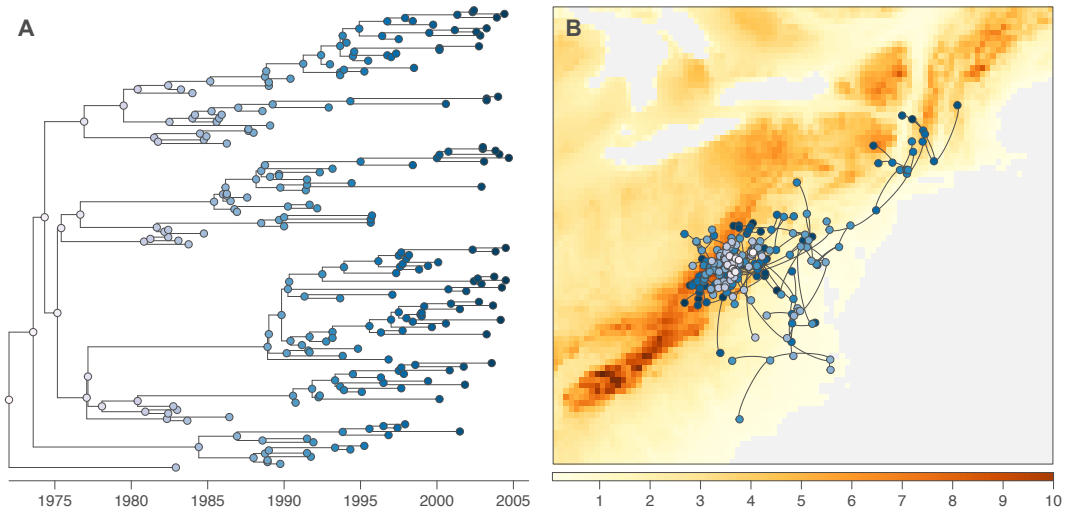

Figure 1: simulated time-scaled phylogenetic tree (**A**) also displayed on the environmental raster used for the RRW diffusion simulation (**B**).

raccoon" dataset (see the related tutorial for further details). In addition, the "samplingWindow" borders have set according to the real values observed in the original dataset.

```
> resistance = TRUE
> scalingValue = 2
> ancestPosition = c(-78.60,39.02)
> birthRate = 0.2
> samplingRate = 0.2
> startingYear = 1972
> samplingWindow = cbind(1982.2,2004.7)
> timeSlice = 1/12
> timeIntervale = 1/12
> showingPlots = FALSE
> extractionOfValuesOnMatrix = FALSE
```

Once all these parameters are specified, the "simulatorRRW3()" function can be called and both its outputs saved as a ".cs" and a "tree" file:

```
> simulation = simulatorRRW3(envVariable, resistance, scalingValue,
ancestPosition, birthRate, samplingRate, startingYear, samplingWindow,
timeSlice, timeIntervale, showingPlots, extractionOfValuesOnMatrix)

> write.csv(simulation[[1]], "RRW_simulation.csv", quote=F, row.names=F)
> write.tree(simulation[[2]], "RRW_simulation.tree")
```

Both outputs of the "simulatorRRW3()" function are displayed in Figure 1.

References

[1] Dellicour S, Rose R, Pybus OG (2016b). 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 (2016a). 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] Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA (2007). A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *PNAS* 104: 7993-7998.