Tutorial for the R package `seraphim` 1.0

Using phylogenetically informed movement data to study the association between an environmental variable and the dispersal location of viral lineages.

Simon Dellicour

July 7, 2025

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 investigate the association between an environmental variable and the dispersal location of viral lineages [4]. Specifically, we here illustrate how to test wether rabies virus (RABV) lineages having spread in the North American raccoon population [1, 5] tended to remain in and/or to disperse to lower altitude areas.

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 following example files also available on the GitHub repository of the package (`https://github.com/sdellicour/seraphim/tree/master/tutorials`): (i) "RABV_gamma.trees", a file containing annotated phylogenetic trees sampled from a posterior distribution of trees inferred for the RABV dataset using the continuous phylogeographic approach developed by Lemey and colleagues [3]; and (ii) "Elevation_rast.tif", the environmental raster "elevation" encompassing 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 [6] is available on the BEAST community website (`https://beast.community/workshop_continuous_diffusion_yfv`).

Step 1: extracting spatio-temporal information in 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 "RABV_gamma.trees" contains 5,001 trees sampled by the MCMC chain. We will here use the "treeExtractions()" function to extract the spatio-temporal information embedded in 100 post-burin-in trees 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 embedded in each tree will be saved), "allTrees" (all the posterior trees), "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 posterior 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="RABV_gamma.trees", what="", sep="\n", quiet=T)
> burnIn = 1001
> randomSampling = FALSE
> nberOfTreesToSample = 100
> mostRecentSamplingDatum = 2004.7
> coordinateAttributeName = "location"
```

Once all these parameters have been specifyied, 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). After this extraction step, each phylogenetic branch can be considered as a vector defined by its start and end location (latitude and longitude), and its start and end dates (in decimal units). Each phylogeny branch therefore represents a conditionally independent viral lineage dispersal event [7].

Step 2: investigating the association between the altitude and the dispersal location of viral lineages

We then use these extraction files to investigate if RABV lineages tended to remain in and/or to disperse towards areas of higher human population density. For that purpose, we will use the "spreadFactors()" function that is also used in the tutorial "impact on dispersal velocity" but this time, we will specify that we do not want to use a path model ("pathModel = 0"). By doing so, we specify that we do not want to use a path model to compute environmental distances and analyse their correlation with branch dispersal durations. In this example, we will also specify that we want to test the environmental raster as a potential "conductance" factor ("resistances = list(FALSE)"). Indeed, as stated above, we here want to investigate if areas associated with higher altitude tended to "repulse' dispersal events of RABV lineages. It is important to note that the "resistance/conductance" terminology is based on the path model specification. In the present context where we do not use/set a path model, "resistance factor" or "conductance factor" mean that we test the corresponding environmental variable as a factor repulsing or attracting tree nodes, respectively.

The other parameters of the "spreadFactors()" function have to be specified as follows:

```
> localTreesDirectory = "Extracted_trees"
> nberOfExtractionFiles = 100
> envVariables = list(raster("Elevation_rast.tif"))
> pathModel = 0
> resistances = list(TRUE)
> avgResistances = list(TRUE)
> fourCells = FALSE
> nberOfRandomisations = 1
> randomProcedure = 3
> outputName = "Elevation"
> showingPlots = FALSE
```

Although we focus on only one raster file in this case, the "envVariables" object has to be a list of raster files and the "resistances" object has to be a vector of boolean variables specifying if each raster has to be treated as a resistance ("TRUE") or a conductance ("FALSE") variable. The "avgResistance" and "fourCells" parameters are not at all used for this analysis but cannot be left unspecified. As for the "outputName" string, it will be used as a prefix to name the different outputs of the function. If the boolean parameter "showingPlots" equals "TRUE", the function will generate and save several graphs (but in that case, the function will run much slower).

The function can then be launched using the following command:

```
> spreadFactors(localTreesDirectory, nberOfExtractionFiles, envVariables,
  pathModel, resistances,  avgResistances, fourCells, nberOfRandomisations,
  randomProcedure, outputName, showingPlots)
```

With "pathModel" set to "0", the "spreadFactors()" function will compute (i) the value E, which is the mean of the environmental values extracted at the tree nodes' position, and (ii) the ratio R defined as the proportion of branches for which the environmental value recorded at the oldest node position is higher than the environmental value recorded at the youngest node position. While E measures the tendency of tree nodes to remain located in lower/higher environmental values, R rather measures the tendency of lineages to disperse towards lower/higher environmental values. These two metrics are computed for each tree sampled from a posterior distribution, and we therefore obtain a posterior distribution for E and R. Finally, each of these two posterior distributions is compared to a null distribution of the same metric computed after having randomised phylogenetic node positions within the study area, under the constraint that branch lengths, tree topology and root position are unchanged ("randomProcedure = 3"; Fig. 1). This approach only requires one randomisation per sampled tree and leads to the approximation of a Bayes factor (BF) support for each statistic. For a particular environmental factor e tested as a factor attracting lineages, the Bayes factor BF_e associated with the statistic E is approximated by the posterior odds that $E_{estimated} > E_{randomised}$ divided by the equivalent prior odds (the prior probability for $E_{estimated} > E_{randomised}$ is considered to be 0.5):

$$BF_e = \frac{p_e}{1 - p_e} / \frac{0.5}{1 - 0.5} = \frac{p_e}{1 - p_e}$$

where p_e is the posterior probability that $E_{estimated} > E_{randomised}$, i.e. the frequency at which $E_{estimated} > E_{randomised}$ in the samples from the posterior distribution. The prior odds is 1 because we have an equal prior expectation for $E_{estimated}$ and $E_{randomised}$. The formal estimate of posterior predictive odds is analogous to approximating BFs in case two alternative hypotheses exist, such for the inclusion of rate parameters or predictors in BSSVS procedures (Bayesian stochastic search variable selection [8, 9]). Bayes factor are automatically approximated by the "spreadFactors()" function when the "nberOfRandomisations" is at least set to "1". As output, the function will generate a text file reporting the different BF values. Alternatively, if the environmental factor was tested as a factor repulsing lineages, BF_e would be approximated by the posterior odds that $E_{estimated} < E_{randomised}$ divided by the equivalent prior odds.

The same approach is used to approximate Bayes factor supports for the statistic R. Whether the environmental factor is tested as a factor attracting or repulsing lineages, the posterior BF_e for R is approximated by the posterior odds that $R_{estimated} < R_{randomised}$ (attracting lineages) or that $R_{estimated} > R_{randomised}$ (repulsing lineages) divided by the equivalent prior odds.

The function will create four distinct output files: (i) one text file reporting the Bayes factor supports associated with each tested environmental factor for the statistic E ("Elevation_position_E_BF_results.txt"), (ii) one text file reporting the Bayes factor supports associated with each tested environmental factor for the statistic R ("Elevation_position_R_BF_results.txt"), (iii) one ".csv" file reporting for each posterior tree the mean environmental value extracted at tree node locations, and (iv) one ".csv" file reporting for each randomised tree the mean environmental value extracted at tree node locations. In the case of the elevation raster tested as a factor repulsing RABV lineage dispersal events, the BF support is lower than 3, which indicates the absence of statistical support [10],

i.e. no support found for a tendency of these inferred RABV lineages to preferentially circulate in low altitude areas).

As these tests are directly based on the environmental values extracted at internal and tip node positions, their outcome can be particularly impacted by the nature of sampling [11]. Indeed, half of the node positions, i.e., the tip node positions, are directly determined by the sampling. To assess the sensitivity of the tests to heterogeneous sampling, one could, e.g., repeat these tests while only considering internal tree nodes. Since internal nodes are phylogeographically linked to tip nodes, discarding tip branches would, however, only mitigate the direct impact of the sampling pattern on the outcome of the analysis. Overall, those tests remain influenced by sampling effort and pattern, and should then be considered more as a description of the environmental context of inferred virus lineage dispersal rather than a robust test of the impact of those conditions on dispersal [4].

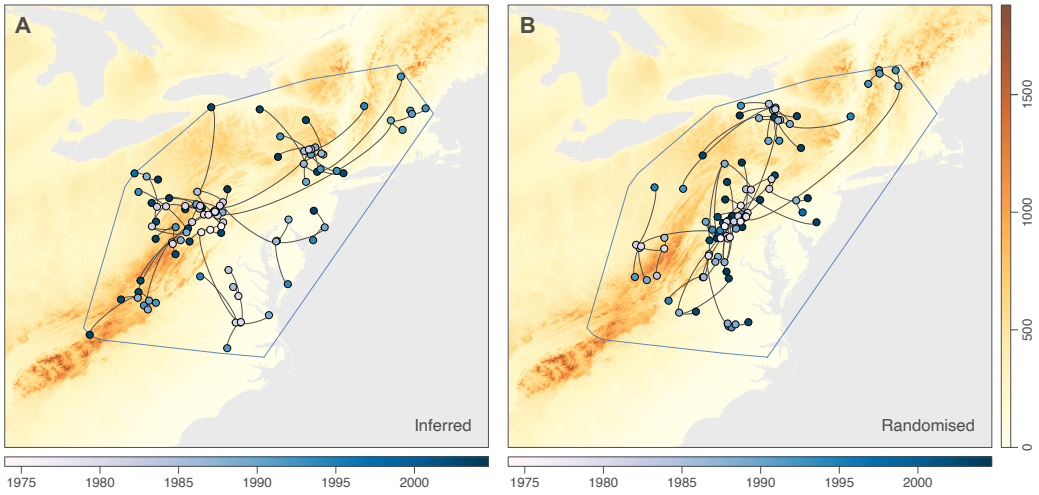

Figure 1: sampled and randomised trees mapped on the elevation raster. **A**: the original environmental raster (representing, in this case, elevation) upon which is superimposed the lineage dispersal events extracted from one annotated tree sampled from the posterior distribution of trees obtained through continuous phylogeographic inference. **B**: the result of one randomisation of branch positions. This randomisation procedure is performed within a minimum convex hull (shown in blue), which is defined by the node locations of all selected phylogenies.

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] Dellicour S, Troupin C, Jahanbakhsh F, Salama A, Massoudi S, Moghaddam MK, Baele G, Lemey P, Gholami A, Bourhy H (2019). Using phylogeographic approaches to analyse the dispersal history, velocity, and direction of viral lineages – application to rabies virus spread in Iran. *Molecular Ecology* 28: 4335-4350

[5] 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. *Proceedings of the National Academy of Sciences of the USA* 104: 7993-7998.

[6] 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.

[7] Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, Gray RR, Arinaminpathy N, Stramer SL, Busch MP, Delwart EL (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences of the USA* 109: 15066-15071.

[8] Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology* 5.

[9] Dellicour S, Rose R, Faria NR, Vieira LFP, Bourhy H, Gilbert M, Lemey P, Pybus OG (2017). Using viral gene sequences to compare and explain the heterogeneous spatial dynamics of virus epidemics. *Molecular Biology & Evolution* 34: 2563-2571.

[10] Kass RE, Raftery AE (1995). Bayes Factors. *Journal of the American Statistical Association* 90: 791.

[11] Dellicour S, Lequime S, Vrancken B, Gill MS, Bastide P, Gangavarapu K, Matteson NL, Tan Y, du Plessis L, Fisher AA, Nelson MI, Gilbert M, Suchard MA, Andersen KG, Grubaugh ND, Pybus OG, Lemey P (2020). Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nature Communications* 11: 5620.