SERAPHIM: studying environmental rasters and phylogenetically-informed movements

Simon Dellicour[1],*, Rebecca Rose[2], Nuno R. Faria[3], Philippe Lemey[1], Oliver G. Pybus[3]*

1 Rega Institute for Medical Research, Clinical and Epidemiological Virology, Department of Microbiology and Immunology, KU Leuven—University of Leuven, Minderbroedersstaat 10, 3000 Leuven, Belgium.

2 BioInfoExperts, Thibodaux, Louisiana, USA.

3 Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom.

Abstract

Summary: SERAPHIM ("Studying Environmental Rasters and PHylogenetically Informed Movements") is a suite of computational methods developed to study phylogenetic reconstructions of spatial movement in an environmental context. SERAPHIM extracts the spatio-temporal information contained in estimated phylogenetic trees and uses this information to calculate summary statistics of spatial spread and to visualize dispersal history. Most importantly, SERAPHIM enables users to study the impact of customized environmental variables on the spread of the study organism. Specifically, given an environmental raster, SERAPHIM computes environmental "weights" for each phylogeny branch, which represent the degree to which the environmental variable impedes (or facilitates) lineage movement. Correlations between movement duration and these environmental weights are then assessed, and the statistical significances of these correlations are evaluated using null distributions generated by a randomization procedure. SERAPHIM can be applied to any phylogeny whose nodes are annotated with spatial and temporal information. At present, such phylogenies are most often found in the field of emerging infectious diseases, but will become increasingly common in other biological disciplines as population genomic data grows.

Availability and Implementation: SERAPHIM 1.0 is freely available from http://evolve.zoo.ox.ac.uk/. R package, source code, example files, tutorials and a manual are also available from this website.

Contact: simon.dellicour@kuleuven.be or oliver.pybus@zoo.ox.ac.uk

Introduction

Phylogenetic techniques are now a standard tool in the study of the spatial and demographic history of organisms (e.g. Lemey *et al.*, 2009, 2010; Lemmon and Lemmon, 2008; Sanmartín *et al.*, 2008). In the context of infectious diseases, for example, phylogenetic inference can be used to reconstruct epidemic history (e.g. Carroll *et al.*, 2015). Of most relevance here, phylogeographic methods, such as that implemented in the software program BEAST (Drummond *et al.*, 2012), enable the reconstruction of the dispersal history of a phylogeny of a given set of genomes sampled through time in continuous space (Lemey *et al.*, 2010). The current version of this approach uses a relaxed random walk model (Pybus *et al.*, 2012) to reconstruct historical dispersal. Phylogeographic methods model the geographic locations of nodes in a phylogeny, and can therefore be considered a type of phylogenetic character mapping approach. Trees inferred using this method, or related methods such as the structured coalescent model (De Maio *et al.*, 2015) can be visualized in a geographical context, such that each branch is interpreted as a vector that characterizes a distinct path through time and space. Thus, the tree represents a collection of phylogenetically informed movement events and contains useful information about the past dynamics of spatial spread.

Here we present SERAPHIM ("Studying Environmental Rasters and PHylogenetically Informed Movements"), a new suite of computational and statistical methods developed to study the environmental context of spatio-temporal phylogenies. SERAPHIM is available as a R package (R Core Team, 2016, http://www.R-project.org) that (i) extracts spatio-temporal information from a set of phylogenetic trees, (ii) estimates and plots dispersion statistics, and (iii) statistically test hypotheses regarding the relationship between dispersal velocity and environmental variables, such as altitude, population density and land cover.

Features

SERAPHIM first extracts the spatio-temporal information contained in a set of suitable phylogenetic trees from a Nexus file containing trees in a Newick format. Phylogenetic uncertainty will be incorporated into the analysis if a set of trees is used, for example a posterior distribution of trees obtained using a Bayesian phylogenetic approach such as that implemented in BEAST. The user can then compute a set of spatial statistics that summarise dispersal. These include the mean branch velocity, the diffusion coefficient as defined in Pybus *et al.* (2012), the weighted diffusion coefficient as defined in Trovão *et al.* (2015), and plots of the maximal distance of an invasion wavefront from the spatial origin against time. Wavefront plots are generated using two distance metrics: (i) the direct distance, which corresponds to the straight-line distance between the origin and the wavefront, and (ii) the patristic spatial distance, which is the sum of the spatial distances along all phylogeny branches between the tree root and the corresponding tree tip. These summary statistics can be useful to compare the mode and rate of spatial spread of different populations. Supplementary Figure S1 illustrates the plots obtained when SERAPHIM is applied to phylogenies that represent dispersal history of the West Nile virus (WNV) in North America. The phylogenies used were those reported in Pybus *et al.* (2012). SERAPHIM also allows users to generate graphs of spatial dispersal history. These plots are similar to those produced by the software package SPREAD (Bielejec *et al.*, 2011). Supplementary Figure S2 shows the reconstructed dispersal history of WNV in North America, again obtained by applying SERAPHIM to the data from Pybus *et al.* (2012). A detailed tutorial on how to undertake these analyses is available with the SERAPHIM package.

The main functionality of SERAPHIM is to statistically test the impact of environmental factors on dispersal velocity. SERAPHIM does this in three steps. First, it computes an environmental "weight" for each branch using a pre-specified environmental raster in an ASCII format. Second, it estimates the correlation between branch duration and branch weight. Third, it tests the statistical significance of these correlations using null distributions generated by a randomization procedure (see manual for details). SERAPHIM implements three different movement models for computing the environmental weight of each dispersal vector (i.e. each phylogeny branch): the (i) straight-line, (ii) least-cost (Dijkstra, 1959) and (iii) random walk models (McRae, 2006). The random walk dispersal model is based on circuit theory (McRae, 2006) and for this model, SERAPHIM calls CIRCUITSCAPE, an external Python package (McRae, 2006). These three models allow the user to explore and test different modes of dispersal when there is no prior information about which movement model might be most appropriate. Once the environmental weights are computed for each branch and environmental raster, the correlation between vector durations and environmental variables are calculated using either a univariate or multivariate regression analysis within a generalized linear model (GLM) framework. Following the GLM approach of Faria *et al.* (2013) and Lemey *et al.* (2014), all variables are log-transformed and standardized. SERAPHIM implements randomization procedures to create null distributions to test for statistical significance, for example by randomizing phylogeny node positions but maintaining branch lengths, tree topology and the location of the root node. An application of the present workflow is illustrated in Dellicour *et al.* (2016).

Perspectives

Although the framework presented here has been developed for the analysis of movement events extracted from spatiotemporal phylogenies (e.g. trees generated by BEAST), it can also be applied more generally to dispersal events that have been observed using non-phylogenetic methods, such as GPS collars or through capture-mark-recapture. Any method that produces a collection of movement vectors, each defined by a start and end location and a start and end time, will be compatible with the SERAPHIM framework.

Acknowledgement

We thank three anonymous reviewers for their useful comments.

Funding

This work was funded by the Wiener-Anspach Foundation. S.D. is a post-doctoral research fellow funded by the Fonds Wetenschappelijk Onderzoek (FWO, Flanders, Belgium). R.R. received funding from the Medical Research Council under a Methodology Research Fellowship grant agreement no. 99204, and from the European Union's Horizon 2020 research and innovation programme under grant agreement 634650-VIROGENESIS. P.L. acknowledges funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864-ViralPhylogeography. OGP received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 614725-PATHPHYLODYN.

Conflict of Interest: none declared.

References

Bielejec, F., et al. (2011) SPREAD: Spatial phylogenetic reconstruction of evolutionary dynamics, Bioinformatics, 27, 2910-2912.

Carroll, M.W., et al. (2015) Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa, Nature, 524, 97-101.

De Maio, N., et al. (2015) New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation, PLoS Genet., 11, e1005421.

Dellicour, S., et al. (2016) Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data, BMC Bioinform., 17, 1-12.

Dijkstra, E.W. (1959) A note on two problems in connexion with graphs, Numerische Mathematik, 1, 269-271.

Drummond, A.J., et al. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7, Mol. Biol. Evol., 29, 1969-1973.

Faria, N.R., et al. (2013) Simultaneously reconstructing viral crossspecies transmission history and identifying the underlying constraints, Phil. Trans. R. Soc. B, 368.

Lemey, P., et al. (2009) Bayesian phylogeography finds its roots, PLoS Comput. Biol., 5.

Lemey, P., et al. (2010) Phylogeography takes a relaxed random walk in continuous space and time, Mol. Biol. Evol., 27, 1877-1885.

Lemmon, A.R. and Lemmon, E.M. (2008) A likelihood framework for estimating phylogeographic history on a continuous landscape, Syst. Biol., 57, 544-561.

McRae, B.H. (2006) Isolation by resistance, Evolution, 60, 1551-1561.

Pybus, O.G., et al. (2012) Unifying the spatial epidemiology and molecular evolution of emerging epidemics, Proc. Natl. Acad. Sci. U.S.A., 109, 15066-15071.

R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing 2016, Vienna, Austria.

Sanmartín, I., et al. (2008) Inferring dispersal: A Bayesian approach to phylogenybased island biogeography, with special reference to the Canary Islands, J. Biogeogr., 35, 428-449.

Trovão, N.S., et al. (2015) Bayesian inference reveals host-specific contributions to the epidemic expansion of Influenza A H5N1, Mol. Biol. Evol., in press.