BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 20 2011, pages 2910–2912
doi:10.1093/bioinformatics/btr481

Phylogenetics

Advance Access publication September 11, 2011

SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics

Filip Bielejec[1,*], Andrew Rambaut[2,3], Marc A. Suchard[4,5,6] and Philippe Lemey[1]

[1]Rega Institute for Medical Research, Clinical and Epidemiological Virology Section, Katholieke Universiteit Leuven, Leuven, Belgium, [2]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK, [3]Fogarty International Center, National Institutes of Health, Bethesda, MD, [4]Department of Biomathematics, [5]Department of Biostatistics and [6]Department of Human Genetics, University of California, Los Angeles, USA

Associate Editor: David Posada

ABSTRACT

Summary: SPREAD is a user-friendly, cross-platform application to analyze and visualize Bayesian phylogeographic reconstructions incorporating spatial–temporal diffusion. The software maps phylogenies annotated with both discrete and continuous spatial information and can export high-dimensional posterior summaries to keyhole markup language (KML) for animation of the spatial diffusion through time in virtual globe software. In addition, SPREAD implements Bayes factor calculation to evaluate the support for hypotheses of historical diffusion among pairs of discrete locations based on Bayesian stochastic search variable selection estimates. SPREAD takes advantage of multicore architectures to process large joint posterior distributions of phylogenies and their spatial diffusion and produces visualizations as compelling and interpretable statistical summaries for the different spatial projections.

Availability: SPREAD is licensed under the GNU Lesser GPL and its source code is freely available as a GitHub repository: https://github.com/phylogeography/SPREAD

Contact: filip.bielejec@rega.kuleuven.be

Received on June 17, 2011; revised on August 11, 2011; accepted on August 13, 2011

1 INTRODUCTION

The advent of powerful and flexible Geographic Information Systems (GIS) has fostered an increasing interest in incorporating geographical information into molecular phylogenetic methods. Spatial phylogenetic projections in a cartographic background play an important role in these developments (Kidd and Ritchie, 2006; Parks *et al*., 2009), but most applications remain limited to mapping phylogenetic tip taxa to their geographical coordinates. Mapping phylogeographic histories of geo-referenced taxa requires a robust statistical estimate of the geographic locations at the ancestral nodes of the tree. To obtain such estimates under stochastic process-driven models, we recently proposed a suite of Bayesian inference approaches for the joint reconstruction of evolutionary and geographic history (Bloomquist *et al*., 2010; Lemey *et al*., 2009, 2010). Models that accommodate spatial diffusion in discrete and continuous space have been implemented in a flexible Bayesian statistical framework (BEAST, Drummond *et al*., 2007; http://beast.bio.ed.ac.uk) for hypothesis testing based on time-measured evolutionary histories. These approaches produce statistical distributions of temporal-spatial phylogenies and this has created new challenges for statistical phylogenetics in producing informative and compelling visualizations. Here, we present software to fully exploit spatial–temporal annotations on phylogenies by providing flexible visual summaries that can be further examined in an interactive manner using GIS or virtual globe software.

*To whom correspondence should be addressed.

2 FEATURES

SPREAD provides four templates to analyze and visualize different aspects of phylogeographic diffusion, labeled: Discrete Tree, Discrete Bayes Factors, Continuous Tree and Time Slicer. The discrete and continuous tree templates typically provide posterior summaries of diffusion from a Bayesian analysis along a high-probability tree, such as the maximum clade credibility (MCC) tree of BEAST. However, SPREAD is not necessarily limited to this input, as it employs a general, trait-annotated NEXUS format; we provide several tree file examples for both discrete or continuous annotations (Compiled, runnable package and supplementary data are hosted at: http://www.phylogeography.org/SPREAD.html). SPREAD supports customized visualization of spatially mapped trees, including branch coloring according to time and branch width manipulation.

Discrete tree associates geographic coordinates with the discrete location states annotated to tree nodes and projects branches that accommodate location changes on a map. Branches that maintain a location state are visualized using customizable circular polygons.

Discrete Bayes factors summarizes the data support for each pairwise rate of diffusion between locations based on Bayesian stochastic search variable selection estimates inferred using BEAST. This template takes as input a posterior sample of rate indicators from an augmented continuous-time Markov chain model and the Poisson prior specifications for the total number of non-zero rates. Lemey *et al*. (2009) describe the Bayes factor calculations in more detail.

Continuous tree maps all branches of a continuous diffusion phylogeography and allows plotting the uncertainty of geographic coordinates at the internal nodes through their annotated highest posterior density contours.

Time slicer supplements the visualization of geographic locations estimated using continuous diffusion by summarizing

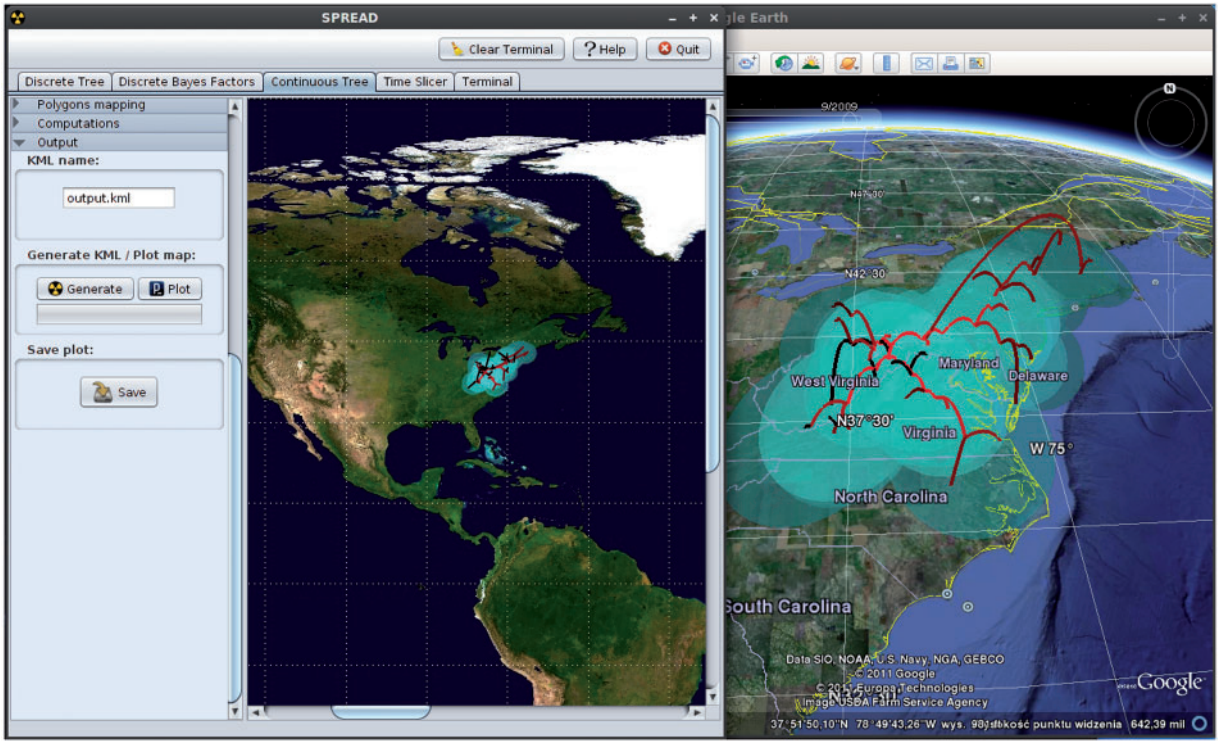

Fig. 1. Screenshot of SPREAD. This example visualizes the phylogeographic history of rabies among raccoons along the Eastern US seaboard under a continuous diffusion model. Such visualizations allow users to quickly inspect key evolutionary changes in their geographic context. Further generation of KML output enables interactive exploration in the time dimension as well in freely available virtual globe software, such as GOOGLE EARTH (on the right).

the rate and degree of geographical movement over the complete posterior distribution of trees. To obtain such ranges, we slice each rooted tree in the posterior sample at a number of points within a time interval, usually defined by the length of a summary (MCC) tree, and impute the unobserved ancestral locations for each branch that intersects those time points. To fully accommodate the uncertainty of the original inference, the imputation involves building Brownian bridges that can take into account branch-specific scaling factors of diffusion rates under relaxed random walks (Lemey *et al.*, 2010). For each time point, we collect the imputed locations across the posterior distribution and use bivariate kernel density estimates to plot the highest posterior density contours. The kernel density estimation follows (Snyder, 1978) and uses bivariate normal kernels with a diagonal bandwidth with bandwidths based on Silverman's 'rule-of-thumb' plug-in value (Silverman, 1978).

3 EXAMPLE AND PERSPECTIVES

For all four template analyses, SPREAD offers direct visualization and also can export the mapped objects to a KML file suitable for viewing with virtual globe software, such as Google Earth (http://earth.google.com). A limited example on raccoon rabies diffusion (Lemey *et al.*, 2010) finds itself in Figure 1; dynamic visualizations of this example as well as others are provided at http://www.phylogeography.org/. KML files can be imported and visualized by many GIS software packages, including ARCGIS and Cartographica.

SPREAD is generally not the run time limiting analysis, compared with fitting the original phylogenetic model, and readily accommodates larger problems. Even for Bayesian phylogenetic analyses, Thousands of pathogen sequences can be accommodated, e.g. (Rambaut *et al.*, 2008), and new computational technologies are actively stretching these limits (Suchard and Rambaut, 2009). Future developments of SPREAD are aimed at extending built-in rendering functionality such as parsing custom base maps and adding mouse-driven camera support to the embedded renderings.

ACKNOWLEDGEMENTS

We thank Nuno Faria and Bram Vrancken for their critical insight and testing of the software and Guy Baele for his Java expertise.

Funding: European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement no. 260864; National Institutes of Health (R01 GM086887); The National Evolutionary Synthesis Center (NESCent) catalysed this collaboration through a working group (NSF EF-0423641).

Conflict of Interest: none declared.

REFERENCES

Bloomquist,E.W. *et al.* (2010) Three roads diverged? Routes to phylogeographic inference. *Trends Ecol. Evol.*, **25**, 626–632.

Drummond,A.J. and Rambaut,A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, **7**, 214.

Kidd,D.M. and Ritchie,M.G. (2006) Phylogeographic information systems: putting the geography into phylogeography. *J. Biogeogr.*, **33**, 1851–1865.

Lemey,P. *et al*. (2009) Bayesian phylogeography finds its roots. *PLoS Comput. Biol.*, **5**, e1000520.

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

Parks,D.H. *et al*. (2009) GenGIS: a geospatial information system for genomic data. *Genome Res.*, **19**, 1896–1904.

Rambaut,A. *et al*. (2008) The genomic and epidemiological dynamics of human influenza a virus. *Nature*, **453**, 615–619.

Silverman,B.W. (1978) *Density Estimation*. Chapman and Hall, London.

Snyder,W.V. (1978) Algorithm 531: contour plotting [J6]. *ACM Trans. Math. Softw.*, **4**, 290–294.

Suchard,M.A. and Rambaut,A. (2009) Many-core algorithms for statistical phylogenetics. *Bioinformatics*, **25**, 1370–1376.