00000100000
00001010000
00010001000
00100000100
01000000010
10000000001
Software notes

Ecography 33: 633–636, 2010
doi: 10.1111/j.1600-0587.2010.06312.x

Subject Editor: Thiago Rangel. Accepted 24 January 2010

GeoPhylo: an online tool for developing visualizations of phylogenetic trees in geographic space

Andrew W. Hill and Robert P. Guralnick

A. W. Hill (Andrew.Hill@colorado.edu) and R. P. Guralnick, Univ. of Colorado Museum of Natural History and Dept of Ecology and Evolutionary Biology, Univ. of Colorado Boulder, Boulder, CO 80309-0265, USA.

GeoPhylo is a scalable online service for developing 3-dimensional geographic visualizations of phylogenetic trees in the keyhole markup language (KML). These geographic phylogenies, geophylogenies, can then be viewed in Google Earth or Nasa's World Wind. Advanced features provide users the ability to change many aspects such as scaling and coloring of branches. The GeoPhylo engine has been deployed on the Google App Engine in order to be scalable, sustainable and easily updated, while providing long-term support for stable releases. These features will allow developers to use GeoPhylo as a service in their own applications without concerns of incompatible changes made in future updates.

"As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever-branching and beautiful ramifications". – Darwin, 1872

At the heart of biodiversity and biogeography is Darwin's great Tree of Life. Darwin's metaphor beautifully captures the importance of diversification and extinction, geography and time in determining the current arrangement of global biodiversity. Darwin's initial drawings are some of the first phylograms, documenting the connection between ancestral and descendent species. The metaphor of the Tree of Life has become directly codified in how we visualize historical relationships among entities, whether populations, species or larger clades. Such phylogenetic trees are key outputs from evolutionary analyses. Overlaying these phylogenies across past and present geographies is a central endeavor in historical biogeography and macroecology, allowing inferences about the tempo and mode of diversifications over time and space and providing information about the phylogenetic diversity found in a particular region. Geophylogenies are useful for quick and simple data exploration, and can also be be further used as spatial networks in more formal analysis.

In the age of digital data, phylogenetic trees have been liberated from the printed page and onto the Web so that anyone may traverse the full tree from the deepest branches to the shallowest twigs and tips of the tree <http://tolweb.org/tree/phylogeny.html>. The digitization of phylogenetic knowledge has been an essential task in completing Darwin's vision of a full Tree of Life, and is one that combines the skillset of evolutionary biologists, computer scientists and graphic designers. An obvious next step is projecting such digital trees onto present or past geographies in order to add an explicit spatiotemporal context for lineage diversification (Kidd and Ritchie 2006). Developing these geophylogenies (sensu Kidd and Xianhua 2008) provides added challenges and opportunities with the development of online GIS and virtual globe technologies. By utilizing such technologies, it is possible for much richer and deeper visualization of diversification events across the globe than ever before. But the devil is in the details about how to create these geophylogenies.

Below we present and discuss a web-based application for creating a geophylogeny that can be opened in a virtual globe platform such as Google Earth or WorldWind. GeoPhylo Engine is not the only such geophylogeny creation tool. Kidd and Xianhua (2008) developed one of the first tools, GeoPhyloBuilder, as an ArcGIS extension. Because GeoPhyloBuilder requires ArcGIS and does not directly produce KML files, it is for more specialized use. Supramap <www.supramap.osu.edu/> and Treebase <www.treebase.org/getrees/uploadTrees.html> have both developed web-based tools that create KML files as well. As we discuss in more detail below, the customizability, workbench and web-service based approach to GeoPhylo Engine sets it apart from others.

The key input data is simply an evolutionary tree and the geographic location of the ''tip'' taxa that were sampled in order to create the tree. We initially developed these visualizations to document the continuing spread of H5N1 avian influenza across the landscape, and to use the

visualization capacities afforded by virtual globes in order to show the spread of lineages with mutations that were thought to potentially increase transmission or virulence of the disease (Janies et al. 2007). The visualization in Hill et al. (2009) show the increasing rate of drug resistant strains over time in H5N1 avian influenza and then ties that back to potential explanation for the emergence and spread of resistant genotypes.

We have since enhanced these visualizations in order to provide cleaner and more detailed information about the ecology and evolution of the virus (Hill et al. 2009; Fig. 1). We have argued that such geophylogenetic visualizations provide important information to the scientific and policy and management communities, as well as the general public. They allow any user with a virtual globe to download the geophylogeny and explore how a pathogen such as H5N1 or H1N1 is both spreading geographically and evolving.

Since development of our geophylogenies is automated, it is easy to now provide a service allowing any user to create and store their own geophyogeny. The online, cloud-based application we have developed, called GeoPhylo Engine, is now available <http://geophylo.appspot.com/>. The rest of this paper is divided into sections that explain the methodology for creating geophylogenies, the development platform we have chosen for serving our application and the key features we have implemented in GeoPhylo. We close by discussing next steps we foresee for further increasing the utility of such virtual globe based geophylogenies. In particular, we discuss how to automatically integrate more knowledge from multiple online sources into geophylogenies and create broader mash-ups.

Methodology for creating geophylogenies

Geophlogenies can be generated in Keyhole Markup Language (KML) format from a phylogenetic tree and geographic coordinates for each tip taxa (leaf) within the tree. The GeoPhylo Engine parses the phylogenetic tree from a Newick formatted tree-file, strictly containing distances and leaf names. In future versions we will expand this feature to allow users to name internal nodes or upload different tree formats. Next, the GeoPhylo Engine parses geographic coordinates for each leaf from a comma seperated file (CSV). The format of the CSV is given at <http://geophylo.appspot.com/coords>. In addition to coordinates, users have the option to include temporal information and specify external icons for node representation in the final output geophylogeny through the CSV file. At the GeoPhylo Engine, users are given a set of advanced options that allow them to modify many of the specific elements of KML generation. Documentation and tutorials are available through our main website, <http://geophylo.appspot.com>. Below we will detail some of the main elements of KML generation at GeoPhylo Engine.

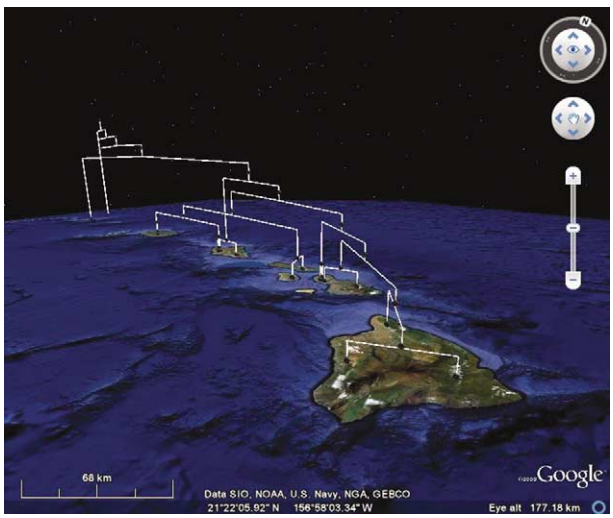

Figure 1. The phylogeny of *Banza* as reported in Shapiro et al. (2006) here presented as a geophylogeny similar to that of Page (2008) but built using the automated services of the GeoPhylo Engine. See the ''Examples'' section of the GeoPhylo Engine website for details on this geophylogeny.

Altitude growth

All leaf taxa are placed on the ground (altitude = 0 m); internal nodes are given an altitude in the sky such that the tree evenly grows through space for organized visual representation. We detailed this process in previous work (Janies et al. 2007), but at GeoPhylo Engine we give the user further control over this process through our Advanced options under ''Global Scaling Factor''. This will allow users to tailor KML generation for local geophylogenies (i.e. the tree should only grow at a minimum rate) or for global phylogenies (i.e. the tree should grow at a faster rate). In future versions, GeoPhylo Engine will allow greater user control over the projection of the phylogeny, including visualizing branch lengths on the geophylogeny.

Local tree representation

In Hill et al. (2009) we created geophylogenies for H5N1 influenza A where numerous (30–50) leaf taxa occupied the same latitude and longitude. To give the user greater ability to interpret the phylogeny in these locations, we developed a method that scatters the points with shared coordinates around a circle. We kept phylogenetically close taxa closer on the circle than those more distantly related. Next, we changed our branch drawing method to a rectangular cladogram. The combination of these three features makes local representations of the phylogenies much cleaner and more interpretable than earlier attempts. We have implemented algorithmic methods for each of these strategies in the GeoPhylo Engine code and they are automatically applied when generating new geophylogenies (Fig. 1).

Temporal information

Our previous research has focused on the evolution of influenza subtypes (Janies et al. 2007, Guralnick and Hill 2009, Hill et al. 2009). The influenza virus is rapidly evolving and the temporal information can be very informative when viewing a geophylogeny. Users can literally watch diversification events occur scaled to the

timeframe at which the influenza was sampled. The GeoPhylo Engine can handle temporal information in the ISO format, YYYY-MM-DDThh:mm:ss, to any level of detail. If the user chooses to add temporal information, each Placemark in the KML is given a TimeStamp. Ancestral nodes are given the date of their oldest descendant. Users can then view the expansion and movement of lineages over time using the built-in functions of Google Earth.

Pop up boxes

Pop up boxes serve as both sources of information and as navigation aids in KML based geophylogenies. In Hill et al. (2009) we demonstrated how both of these functions could be put to use. As valid XML, the KML based geophylogenies can be easily parsed by users to include any combination of information and navigation tools. We provide additional information on parsing KML on our website. The default content of pop up boxes in KML returned by the GeoPhylo Engine is minimal; containining hypertext links that will automatically fly the viewer to nodes closely positioned in the phylogeny (sister, ancestor, and descendant nodes). In future updates we will be allowing users to submit a third data file where users can include KML Description window content (including HTML and CSS formatted information).

Development platform for GeoPhylo Engine

GeoPhylo Engine was written in Python and hosted in the Google App Engine <http://code.google.com/appengine/>. We chose this method for three primary reasons. First, it allows us to offload the computation of geophlogenies to a cloud computing environment. This will make the GeoPhylo Engine highly scalable and ready to handle high levels of traffic and use. Second, it allows us to run old versions of GeoPhylo Engine side-by-side with our latest version with minimal overhead. This will allow developers of other projects to utilize our service without concern that a future version will render their project unusable. For example, if a developer wanted to submit data and retrieve a KML from GeoPhylo Engine, they would develop their project to utilize a specific version of our service, i.e. ver. 2.4. Following that, they could link their project directly to ver. 2.4 (e.g. <http://2-4.latest.geophylo.appspot.com/>) and then can update their project to use future implementations of GeoPhylo Engine at their own pace. Finally, we wanted to take advantage of the Google App Engine datastore to give users an option to persistantly store their KMLs for sharing and modification, discussed below.

GeoPhylo Engine as a web service

Users who wish to access the GeoPhylo Engine as a web service can do so through a simple REST interface. We provide an example of how to do so on the GeoPhylo website under ''Sample Code''. The user can submit very basic to detailed information regarding how each KML should look and then receive either the full KML itself or a link to the KML stored on the GeoPhylo website. Developers are encouraged to design for a given release of the GeoPhylo Engine for which we offer long-term support. Currently, the only version with planned long-term support is ver. 2.4. As new versions are released we will make the changes log available to users as well as documentation indicating which versions will have long-term support. We hope this encourages users and web-developers to include the GeoPhylo Engine as a resource for mashups containing phylogenetic information.

GeoPhylo Engine as a workbench

Users of the GeoPhylo Engine have the ability to generate and store their KMLs at the GeoPhylo website. To do so, a user must select ''Permalink'' in the Advanced options. When selected the user will be given a private-key which they can use to update their geophylogeny in the future. When the user submits their data, instead of a complete KML being returned, the user will recieve a KML containing only a NetworkLink to the KML stored at GeoPhylo. For large geophylogenies, this will give the user the ability to share it with many users without exchanging the large file over email. For users who want to distribute a KML and then update it in the future, this feature will allow them to manage that through the GeoPhylo Engine website.

To update a stored KML, the author needs to enter their private-key and the KML's public-key under the ''Update existing geophylogeny stored at GeoPhylo?'' section of the Advanced options. The public-key is provided in the description window of the NetworkLink KML, or in the file itself in the path to the KML stored at GeoPhylo Engine as, <http://geophylo.appspot.com/public-key/networklink.kml>. If the two match the details for a stored KML, the user will update that stored KML automatically when new data is submitted. Once updated, any user who has been given the NetworkLink KML will see the new KML the next time they open Google Earth.

Next steps for virtual-globe based geophylogenies

The challenge we have met here is making a serviceable and simplified method of generating geophylogenies. We anticipate that such a service will encourage a greater number of mashups and incorporation of geophylogenies into existing phyloinformatic workflows especially since KML is now an Open Geospatial Consortium (OGC) standard <www.opengeospatial.org/standards/kml>. With the recent release of the Google Earth browser plug-in <http://code.google.com/apis/earth/>, integration of outputs into web browsers is further simplified. In near future releases, we plan to support further upload capabilities for Descriptions windows as discussed above. We can also leverage other services providing biological data in order to automatically enrich user-generated geophylogenies. As an example of such a service, we plan to support acquisition of available species ranges in the form of polygons linked to each named taxon within the phylogeny. Geophylogenies offer a bridge between phylogeny and biodiversity through their common geographic components. Therefore, integration of further geospatial information will continue in the development of the GeoPhylo Engine service.

To cite GeoPhylo or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for ''version 0'':

Hill, A. W. and Guralnick, R. P. 2010. GeoPhylo: an online tool for developing visualizations of phylogenetic trees in geographic space. – Ecography 33: 633–636 (Version 0).

Acknowledgements – We thank the our reviewers for useful comments on the manuscript. Early work on geophylogenies was supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number W911NF-05-1-0271. We acknowledge recent support by NSF ABI grant 0930344.

References

Darwin, C. R. 1872. The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, 6th ed. – John Murray, London.

Guralnick, R. P. and Hill, A. W. 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. – Bioinformatics 25: 421–428.

Hill, A. W. et al. 2009. Evolution of drug-resistance in multiple distinct lineages of H5N1 avian influenza. – Infect. Genet. Evol. 9: 169–178.

Janies, D. et al. 2007. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). – Syst. Biol. 56: 321–329.

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

Kidd, D. K. and Xianhua, L. 2008. GEOPHYLOBUILDER 1.0: an ARCGIS extension for creating 'geophylogenies'. – Mol. Ecol. Resour. 8: 88–91.

Page, R. 2008. Towards realising Darwin's dream: setting the trees free. – Nature Precedings, doi:10.1038/npre.2008.2217.1.

Shapiro, L. H. et al. 2006. Molecular phylogeny of *Banza* (Orthoptera: Tettigoniidae), the endemic katydids of the Hawaiian Archipelago. – Mol. Phylogenet. Evol. 41: 53–63.