# Adding Phylogenies to QGIS and Lifemapper for Evolutionary Studies of Species Diversity

Jeffery A. Cavner
*University of Kansas (USA)*

Aimee M. Stewart

Charles J. Grady

James H. Beach

# Adding Phylogenies to QGIS and Lifemapper

**for Evolutionary Studies of Species Diversity**

*by Jeffery A. Cavner, Aimee M. Stewart, Charles J. Grady, James H. Beach*

*University of Kansas (USA).* jcavner@ku.edu

## Abstract

Phylogenetic data from the "Tree of Life" have explicit spatial and temporal components when paired with species distribution and ecological data for testing contributions to biological community assembly at different geographic scales of species interaction. Important questions in biology about the degree of niche suitability and whether the history of a community's assembly for an area can affect whether the species in a community are more or less phylogenetically related can be answered using several different spatially-filtered measures of phylogenetic diversity. Phylogenetic analyses which support the description of ecological processes are usually achieved in a handful of software libraries that are narrowly focused on a single set of tasks. Very few applications scale to large datasets and most do not have an explicit spatial component without relying on external visualization packages. This prompted us to explore bringing phylogenetic data into an open-source GIS environment. The Lifemapper Macroecology/Range & Diversity QGIS plug-in is a custom plug-in which we use to calculate and map biodiversity indices that describe range-diversity relationships derived from large multi-species datasets. We describe extensions to that plug-in which expand the Lifemapper set of ecological tools to link phylogenies to spatially-derived 'diversity field' statistics that describe the phylogenetic composition of natural communities.

**Keywords:** QGIS, WPS, Distributed Computing, Biogeography, Range and Diversity, Lifemapper, Macroecology, Phylogenetics.

## 1. Background

Community phylogenetics, the focus on how species relatedness and species traits are associated with how evolution extends into ecological processes and spatial patterns, and biogeography or meta-community ecology, largely focused on the spatial regulation of species distributions, should assay the spatial variation of phylogenies by mapping phylogenetic community values across space and time at different scales using advances in GIS techniques. One such approach would to be bring phylogenetic data into a GIS environment. We have begun to develop such an approach as an addition to the Lifemapper project (www.lifemapper.org) in a Lifemapper Range & Diversity (LmRAD) QGIS plug-in (Cavner et al. 2014) that provides phylogenetic visualization and analysis tools for spatially linked range-diversity relationships derived from presence-absence matrices (PAMs). We developed the tool also hoping to expand it to include historical biogeography meta-community analyses and community assembly analyses focused on phylogenetic-diversity area relationships where analysis across geographic scale leads some of the most important questions in biodiversity.

The LmRAD QGIS plug-in creates, maps and analyzes presence-absence matrices or PAMs, one of the core data structures for macroecological research. It links the resulting data to phylogenetic and spatial views of a set of range-diversity statistics derived from the PAM. The PAM or incidence matrix is a 2-dimensional Boolean matrix constructed from a spatially defined grid of regular polygons where the presence or absence of each species of hundreds or thousands of species are recorded for each cell. One axes of the matrix represents species and the orthogonal axis represents geographic localities described by the regular polygons. Each geographic site is coded for the presence (1) or absence (0) of each species. It summarizes the two fundamental units of biogeography, the distributional range of a species (both their position and size, range size simply equals the total of the species axes across sites) and the species diversity of sites or the number of different species in each site as summarized by site axes totals.

Several mathematical and biological relationships obtain across the PAM that link spatially derived statistics with species based statistics. Of interest for phylogenetic relationships are the species based statistics calculated from the PAM that measure the "diversity field" of a species (Arita et al. 2008). The diversity field is the set of diversity values of sites in which a species occurs. For example, the

diversity field volume, i.e. the summation of those species diversity values within a species' range divided by the range size of the species allows us to calculate the average species diversity within the range of that species. We represent that volume as a proportion of the total number of species in the study area. Including the total area of the study area allows us to illustrate the proportion of the sites in which two species co-occur. The average association of a species with all of the species in the study area allows us to illustrate that there is an inverse relationship between the proportional range of a species and the difference between the mean proportional diversity within its range and the average proportional diversity in the study area (Arita et al. 2008). The mathematical reciprocal of the average proportional diversity of the study area is a well-studied measure of species turnover called Whittaker's beta diversity. It is a measure of the ratio between the overall diversity of the study area and the average local diversity (Arita et al. 2008). There are closely associated beta measures of diversity for several different types of diversity. Different approaches to species diversity such as phylogenetic diversity – the degree of relatedness of species in a community based on their evolutionary history – abundance and ecosystem function measures of diversity all can be decomposed into measures of local and regional diversity ratios that are highly dependent on scale.

Analyzing the diversity field within the range of a species is equivalent to studying it's covariance with all the species in a study, i.e. the degree of association of species within their ranges. We plot this association in QGIS through the plug-in in a "range-diversity" plot. Curves on the plot for species follow a line defined by the inverse relationship between the range of a species and the difference between the two diversity statistics. When plotting the species in this way, species with equal degrees of association with one another arrange themselves along lines of isocovariance. The Lifemapper plug-in allows the user to "brush" data points along those curves in the interactive range-diversity plot which selects the individual species in the linked data space for the phylogenetic tree. In this way the spatially derived statistics for diversity from the PAM can be compared to the degree of phylogenetic relatedness within species communities.

The plug-in accomplishes this by using QGIS as a WPS client to Lifemapper web services (Stewart et al. 2014) and by using JavaScript based visualization technologies for large phylogenetic trees within the plug-in. Macroecology algorithms are exposed as Open Geospatial Consortium (OGC) Web Processing Services (WPS) (Open Geospatial Consortium, Inc. 2007b) so that larger distributed computing environments can be brought to bear on large datasets. The Lifemapper web services are organized as two modules, LmSDM, and LmRAD. The LmSDM module uses RESTful and OGC specifications to build species distribution models based on the predicted niche for a species using climate and species occurrence data. The LmRAD (Range and Diversity) is a multi-species platform for PAM based range and diversity calculations. Both modules can be accessed through the plug-in, and outputs from LmSDM can be piped into LmRAD as species inputs to PAMs. This paper will focus on the range and diversity capabilities of the plug-in and how the spatial component to phylogenetic data recently added to the plug-in can be used with the biodiversity indices calculated from the PAM and areas where phylogenetic data can be used to explore other types of diversity measures for species communities. This paper will begin by outlining use cases and common threads that connect them and how we have begun to address them with a focus on new interface capabilities for phylogenetic data and linked data spaces. Next we will describe how the Lifemapper plug-in and it's supporting web services were designed to take advantage of a client-server architecture in order to be able to use geographic processing standards on large datasets. This is followed by a comparison of related software with a focus on phylogenetic algorithms and scripts with a spatial component. We end by discussing findings, and future directions for the Lifemapper plug-in.

## 2. Use Cases and Capabilities

### 2.1 Range and Diversity Plots and Maps with Phylogenetic Trees

Phylogenetic based ecology is a growing field. Its practice both at small scales and larger biogeographic scales – it goes under several names: phylogeography, ecophylogenetics, or phylogenetic community ecology – share two obvious constraints for incorporating phylogenetic data into ecology research. First, many ecophylogenetic methods are not available as open-source software packages, and are therefore not extensible or customizable, and second; the tools are scattered across specialty software each with their own learning curve and with unique data formats (Kembel et al. 2010). When

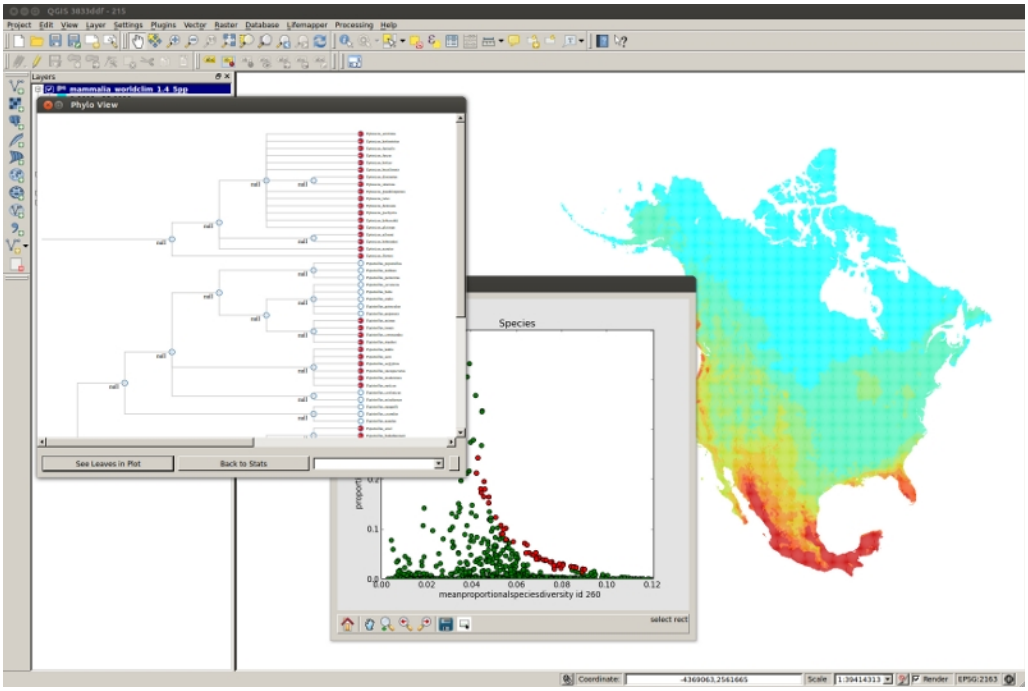


Figure 1: Lifemapper Range-Diversity Plug-in in QGIS with Range-Diversity Plot and Tree View.

we extended Lifemapper to include the multi-species range and diversity experiments on PAMs in the LmRAD module we encountered a third constraint. Untangling species associations at large scales remains an important research question (Arita et al 2008) and these scales require large taxa with numerous species in single experiments, e.g. approximately 3000 + species of birds in South America. Each of these species represents some type of geographical map which have to be intersected with large numbers of polygons resulting in PAM matrices of several million elements. These models must also be permuted, i.e. randomized to perform null model hypothesis testing, resulting in a large number of unwieldy data structures. Additionally these studies need to be done at several different spatial scales and across time. Totaling all of these dimensions one can readily see that this presents serious computational challenges. These factors informed the design of LmRAD as a module to the existing Lifemapper computational platform which uses co-located distributed high-through-put compute resources to calculate large multi-part ecological modeling jobs, and a thick client front-end in QGIS to those services.

Important biological relationships are expressed in the PAM as associations between species diversity in communities and the range size of those species. Two important correlations are between the species diversity of sites and mean range sizes of species occurring in the site; and between the range sizes of species and the mean species diversity within those ranges. The set of range statistics for species within a site are described by a 'dispersion field'. The analogue to that measure for species is the 'diversity field' which quantifies the species diversity of sites in which a species occurs (Arita et al. 2008). Range-diversity plots are produced in the Lifemapper plug-in that summarize these fields as indexes of site similarity and the degree of association of species adding to our knowledge of species communities. Data points in the plots are constrained by the association of species and site similarity and the proportional fill of the matrix, i.e. the ratio of presences of species to their absence. The plug-in allows a researcher to build several models across scale, experiment with fill, extent and resolution. The dispersion field and diversity field measures in the range-diversity plots are interactive and allow multiple pane selection. Visualization and data exploration are presented in both geographic and phylogenetic data spaces. The dispersion field statistics are viewed in an interactive "by-sites" range-diversity plot and are linked geographically to a map of the range statistics attached to the input grid for the PAM so that they can be overlaid with other geographic data. For species associations within communities the data derived from

the PAM are depicted in an interactive "by-species" range-diversity plot for the diversity field and are linked to a dendrogram that represents their phylogenetic relationship. All of the data spaces allow for 'brushing' of datasets by species or location across tree data space and geographic data space. Selecting species in the phylogenetic tree viewer select those species in the "by-species" range-diversity plot. In this way the trees can act as a data exploration tool against the diversity indices derived from the PAM providing insight into the phylogenetic composition of communities where species co-occur. (see figure 1.)

## 2.2. Across Space and Time: Scale Considerations

The indices currently calculated through the plug-in, including ecology staples, such as beta diversity, along with measures of nestedness – the degree to which diversity loss occurs by species, leaving isolated "islands" of diversity – are all effected by scale. The degree to which these indices are effected by scale and the mechanisms involved are important research questions (Arita et al. 2008, Lira-Noriega et al. 2007). Most analyses of scaling effects on diversity have been based on coarse input grids. For example Hawkins et al. (2003) based a diversity study comparing the effect of scale using 85 datasets with resolutions ranging from 103 to 105 km2 (Lira-Noriega et al. 2007). Lira et al. performed a study with finer PAM resolutions starting at 11.4 km2 and incrementally climbing to 2.93 x 103 km2 for an area of ˜ 138,200 km2. The Lifemapper plug-in has been used to construct PAMs for much larger areas, ˜ 24,709,000 km2 with slightly larger cell resolutions of 100 km2, but with the recent additions of data parallelization and portable instances of Lifemapper we expect to be able to produce PAMs with cell resolutions lower than 1/32o for the globe. We can also currently test scale related hypotheses about range size and diversity such as predictions that for the same kind of organism, organized by taxa, and their ability to disperse across the landscape, stronger negative correlations between range size and diversity should exist the greater the scale. Several questions that relate to spatial scale can also be asked of phylogenetic-diversity area relationships, and the extent to which speciation and adaptation contribute to community assembly with the incorporation of phylogenetic tree data into the plug-in.

Because biogeographers are increasingly interested in methods in phylogeography and community assembly, research questions addressed by both species richness based diversity measures, phylogenetic diversity and functional diversity need to benefit from relative findings and work together to complement one another (Cianciarus 2011). A common thread connecting different concepts of diversity are questions about the evolutionary and biogeographical history of a species and how temporal and spatial scales affect the evolutionary relatedness of species in a habitat and the degree that those assemblages are consistent with environmental filtering or competitive interaction (Emerson and Gillespie 2008). The species composition of natural communities is tied to questions of range contraction and local extirpation of species in relation to niche processes like climate change. The Lifemapper/QGIS plug-in allows the user to build PAMs that describe range and diversity relationships across time in relation to climate change by using predicted eco-niches based on climate scenarios, derived from LmSDM, as inputs to future PAMs.

Phylogenetic data has both spatial and temporal components. Patterns of co-occurrence of species in a spatially defined community is effected over different time and spatial scales by the similarity, and distance of other habitats, the degree that niches are filled with current inhabitants and the relative time available for colonization or adaptation (Emerson and Gillespie 2008). Patterns of community structure and co-occurrence of species can be summarized by two related statistics derived from phylogenies for a geographic area, phylogenetic clustering, and phylogenetic over-dispersion/evenness. Phylogenetic clustering occurs when co-occurring species are more closely related than can be expected by chance. Phylogenetic over-dispersion/evenness occurs when co-occurring species are more distantly related than can be expected by chance. With the tree viewer these phenomena are easily discernible for small trees with species selected that co-occur within a community. Both of these measures will need to be quantified for larger trees and both require that they be tested against null models generated from the tree and its spatial components. Lifemapper currently implements some very efficient bit-wise operations for randomizing null models from the PAM. To permute the tree data, we will in the future build out the architecture for encoding the tree topology from large phylogenies into matrices that will use similar methods for randomization.

Clade based analyses of traits related to niche occupancy helps us to understand the relative importance of environmental filtering. Using cross scale

comparisons in the plug-in with the phylogenetic trees could help to tease out effects of both temporal and spatial scale. Larger extents within an LmRAD experiment should show phylogenetic clustering due to environmental filters, and local areas which will naturally contain subsets of the same taxa used in the experiment should show local over-dispersion due to competitive interaction. For temporal scale, range and diversity measures from time-stepped PAMs achievable with the recent acquisition of paleontological climate layers, and the future climate scenario data currently in the plug-in, should allow us, with the use of the trees be able to look at colonization dynamics, and how over-dispersion and cladogenesis become more important over time for isolated niches and how species new to a habitat over large time frames, e.g. island migration, show shared common traits pre-adapted to a habitat (Emerson and Gillespie 2008).

# 3. Design and Architecture

## 3.1 Lifemapper Distributed Computational Services

The Lifemapper Range and Diversity (LmRAD) module is an analysis suite that extends the current Lifemapper (www.lifemapper.org) platform allowing us to leverage the computational power of distributed computing environments to execute the range-diversity analyses as distributed algorithms. The algorithms are exposed as Open Geospatial Consortium Web Processing Services (WPS) (Open Geospatial Consortium, Inc. 2007b), and RESTful web-services for simple data retrieval and viewing. The Lifemapper infrastructure is composed of a central management component, LmDbServer, which manages data and analysis operations with a "data pipeline" written in Python (www.python.org) and a PostgreSQL/PostGIS database; multiple instances of LmCompute that can be co-located across institutions, currently deployed at compute clusters at University of Kansas, University of Florida, and San Diego Supercomputer Center; a continuously updated species model and species occurrence set archive based on museum data for species from the Global Biodiversity Information Facility (GBIF); and LmWebServer which manages all communications between the components and client applications. (see Figure 2.) LmRAD specifically is a distributed mulit-species modeling module within this system with custom algorithms for working with presence-absence data, including matrix definition, construction, calculation, randomization for null models and preparation of visualization outputs, trees and maps.

As a job based infrastructure, LmRAD and LmSDM algorithms are environmentally agnostic and are portable across compute environments through instances of LmCompute that are deployable in several types of distributed compute environments. LmCompute is a pluggable, configurable, open source client that abstracts the details of the compute job away from the physical system. LmWebServer contains a Job Server tier that feeds jobs to any compute environment that can sponsor an instance of LmCompute. LmCompute is also generalizable, since LmCompute only interacts with the physical system through a mediator designed along the mediator and facade design patterns (Gamma et al. 1994) the compute plug-in expects just a few stock functions. A "request job" method call might just as easily get a local XML job definition or pull a job from the Lifemapper Job Server. An instance of LmCompute can use a job response to instantiate a Job Runner object and retrieve inputs to the methods requested. Each of these computational tasks or group of related tasks is a compute plug-in based on the template method and strategy design pattern (Gamma et al. 1994). The compute plug-in is wrapped in a "runner" class that depending on its run method can execute an external application or run custom algorithms like LmRAD algorithms. A compute plug-in receives its jobs through a job controller that acts as a hub for producing job outputs. Using the factory method pattern and command pattern (Gamma et al. 1994), the controller sits in front of a compute environment, requests data inputs for a job, and determines through Python "duck typing" which compute plug-in is appropriate for the computation. The pipeline and LmDbServer are responsible for presenting jobs to the Job Server on LmWebserver and moving jobs through the system. At different stages in a LmRAD experiment dependencies and statuses are updated by LmCompute which posts back to the Job Server during the process. LmRAD PAM operations specifically have been parallelized across processors on any compute environment that receives a PAM job. Data products for large PAMs at high resolutions (10 km) with upwards of 800 species can be constructed and analyzed in this way with reasonable response times. Results from the experiment are then posted back to the Job Server from the compute environment and are written to the database and file system shared by the LmDbServer and LmWebserver.

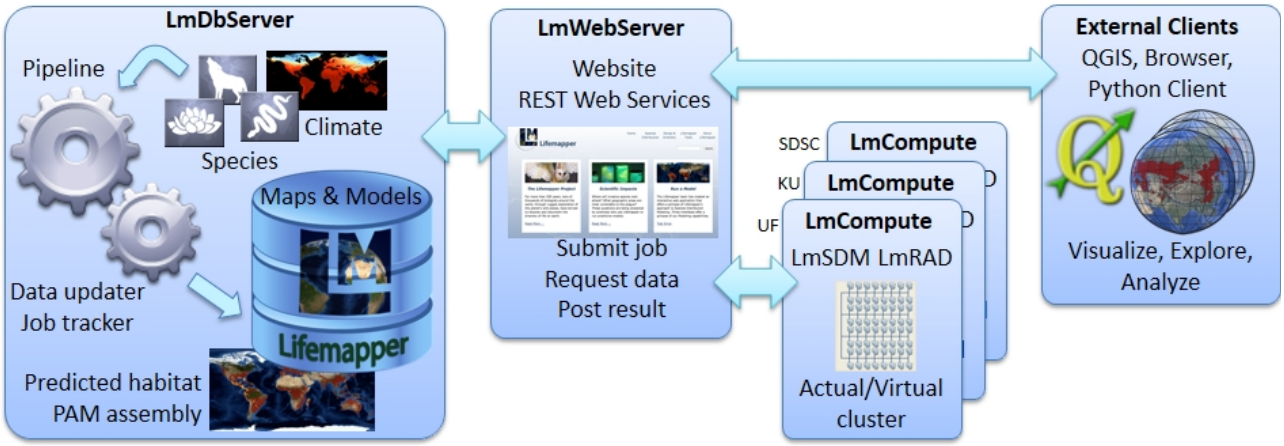


Figure 2: Lifemapper Components.

Data parallelization across multi-core architectures in each of the environments hosting LmCompute help to speed the PAM matrix construction, which uses a combination of Rtree (https://pypi.python.org/pypi/Rtree/) and matplotlib's nxutils and GDAL (GDAL 2013) for vector and raster based intersections, respectively. Calculations on the matrices use NumPy (Jones 2001) built with the Basic Linear Algebra Subprograms (BLAS). Permutations on the PAM matrices use methods that are specific to binary matrices, where row and column totals can both be kept intact while changing the mix of species in sites and the range size of each species. Data parallelization is not suited to these computations since the entire matrix needs to be taken into account. But since several hundred permutations may be required per experiment, the current job based parallelization across compute nodes works well for permuting several hundred matrices. Another method in LmRAD for permuting the matrix is perfectly suited for both types of parallelization. It uses a dye dispersion algorithm which is a 2-Dimensional geometric-constraints model that assumes range continuity (Jetz and Rahbek 2001). Since range allocations are reassembled individually for each species, those data can be split across cores on a single machine or across nodes.

## 3.2 QGIS Plug-in, WPS Client, JavaScript, Plots, Visualization

The computational constraints of operating against large matrices in current desktop software informed the design of LmRAD as a client-server architecture using web-services to off-load the heavy lifting for PAM operations to remote compute environments with the use of a thick client inside a feature-rich open source GIS environment. The Lifemapper plug-in for QGIS allows QGIS to operate as a web service client to the LmSDM services and as a WPS client to the LmRAD analyses, edit and submit data, parameterize inputs and request computations with the added feature of being able to pull down statistical results, geospatial outputs and work with phylogenetic tree data. It is able to do this by interacting with a multi-platform Python client library that abstracts the communication layer away from the user. The Lm client library can also be decoupled from the client so that developers can use it to program a variety of standards based clients. The added benefit of using the library within the QGIS plug-in is that all LmRAD functionality for dealing with PAM operations are wrapped in easy to use, point and click operations with results automatically downloaded to a managed workspace and presented in QGIS.

Viewing the phylogenetic trees required that a highly interactive and lightweight interface be built in the plug-in without library dependencies. Rather than deal with heavy Qt (http://qt-project.org/) solutions for graphics we decided to leverage recent advances in web based standards for visualization of phylogenetic trees in the QGIS plug-in using a document driven JavaScript framework. Tree data from the phylogenetic community can take any one of several forms, phyloXML (www.phyloxml.org), Newick (http://bit.ly/1n6ELcZ), Nexus (Maddison et al. 1997), and NeXML (http://nexml.org). All of these formats are easily translated into JSON which maps into Python dictionaries and works well with web standards based solutions for visualizations

based on JavaScript. Additionally tree providers, like Open Tree of Life (http://blog.opentreeoflife.org/) are developing NexSON, a badgerfish convention JSON translation of Newick as a data document for transport from web-services that provide trees served from graph databases. Data like these are perfect for producing a scene graph, can be made available from web-services, are easily transported back and forth from LmCompute for analysis and can be used directly in a document driven visualization framework.

The tree viewer presents the phylogenetic data as interactive SVG built dynamically from incrementally loaded JSON data. This is made possible with the JavaScript library D3.js (Data Driven Documents) (http://d3js.org/). D3 allows the JSON document to be dynamically bound to the Document Object Model so that data-driven transformations can be applied to the document with smooth transitions and fluid interaction. The data are directly mapped to visual elements in the DOM without an internal or intermediate representation or abstraction of the DOM. The document is the scene graph. This allows for much better performance since the focus is on transformation of the document (Bostock et al. 2011). Selections against the DOM are declarative in a functional programming style with predicates from the W3C Selectors API similar to jQuery allowing CSS properties to be specified as functions. Incoming data can create new nodes in the DOM, and outgoing data can remove nodes using Enter and Exit selections. This is especially useful when navigating large trees, since the large number of nodes and edges for large phylogenies have in the past been hurdles for visualizing tree data in a way that is responsive to user interaction conditioned to fast response times.

The D3 based interactive tree is rendered in the plug-in through a Qt dialog using QtWebKit. Communication between the tree and the rest of the plug-in is effected by QtWebKit Bridge. The bridge allows the JavaScript and PyQt objects to communicate with one another. The tree viewer is linked to the interactive range-diversity plots in matplotlib (Hunter 2007) by simple PyQt signals and slots. A similar method connects the range-diversity plots for site-based statistics to the maps in QGIS based on the PAM. Using JavaScript in PyQt dialogs for QGIS allowed us to achieve fluid visual representations of trees for large clades, e.g. one tree used in testing is the entire phylogeny for the Phylum Mollusca with over 85,000 nodes.

# 4. Comparison of Approaches

Several phylogenetic analysis software implementations exist, the number is too daunting to recount them all here and most are implemented in R scripts and free but not necessarily open C++ software. Very few integrated systems exist that address biogeography, species communities, ecological niche and phylogeny. With the growth in phylogenetic data, web-based solutions for viewing trees are popular, but those concentrate on data already analyzed for specific taxa and tend to illustrate simple clade-area relationships. Challenges for both analyzing and exploring large phylogenies exist both on the computation side and the visualization side. We mention some very powerful approaches that contain a spatial component in relation to phylogenetic analysis and compare them to our tool which aims at bringing phylogenetic data into a GIS based tool that is sustainable and extensible using an analysis, that until now has not been systematized, using PAMs and their inherent range-diversity relationships

GeoSSE (Geographic State Speciation and Extinction, Goldberg et al. 2011) is a geographic range/-phylogeny model. GeoSSE is an an extension of the BiSSE (binary state speciation-extinction) model that allows tests for relationships between speciation or extinction and geographic range. GeoSSE is a method for analyzing the reciprocal influence of character traits and speciation/extinction, where character states are defined by spatial distributions. Transitions between states are parametrized in terms of range expansion through dispersal and range contraction through local extirpation. The model has the liability of requiring fairly large phylogenies with one or two hundred species at the leaf nodes as a minimum. The increasing availability of larger trees shouldn't make this much of a problem in the future, but may potentially also require computational solutions addressed by a distributed or parallel implementation.

Picante (Kembel et al. 2010) is a comprehensive R package for calculating phylogenetic diversity of ecological communities. It contains functions for both local or alpha phylogenetic diversity and beta phylogenetic diversity. Local community diversity indexes include Faith's phylogenetic diversity (PD) (Faith 1992), taxonomic distinctness indexes, mean pairwise phylogenetic distance (MPD) and mean nearest taxon distance (MNTD) within communities. Clustering and evenness are represented by several measures calculated in Picante. Beta phylogenetic diversity is also addressed with MPD and MNTD be-

tween communities, Sorenson index and the UniFrac phylogenetic distance metric. Picante also has robust null model capability, performing numerous permutation procedures. Ecological correlation is also included with species-environmental regressions. Picante would be an extremely powerful addition to a workflow involving large matrices using parallel methods in R or a framework like Lifemapper. Picante's methods are staples and starting points for numerous different analyses that could be performed in QGIS, benefiting from an explicit spatial component especially in regards to its ecological links to phylogenetic statistics.

Landis, Matzke, Moore, Huelsenbeck (2013), recognize that the main constraints on using models to describe the geographic evolution of species ranges as processes of dispersal and extinction is the computational limit on the number of areas that can specified. Where Lifemapper choose to leverage distributed computational resources to solve similar scale problems for large numbers of sites the Landis et al. method uses a Bayesian approach for inferring biogeographic history that allows more realistic problems involving large numbers of geographic sites implemented in BayArea, a free C++ command-line program that uses PAMs and phylogenetic data in the Newick format as inputs. Its outputs can be visualized as tree/map animations in an external JavaScript web service for filtering phylogenetic reconstructions and mapping them.

Biodiverse (Laffan, Lubarsky, Rosauer 2010), an open-source project similar to the Lifemapper plugin, provides linked visualization across different data spaces. Biodiverse links species distributions in geographic, phylogenetic, taxonomic and matrix space. One advantage of Biodiverse similar to Lifemapper is that scale comparison are achieved through a window analysis for endemism, phylogenetic diversity, and beta diversity. By varying the size of the windows one can start to understand the effects of scale on those statistics. Currently the Lifemapper plugin uses a multi-grid approach where several subsets at different cell resolution can be built out within the same experiment allowing comparisons across scale for the range and diversity statistics including beta diversity.

# 5. Future Directions and Conclusion

## 5.1 Incorporation of R for ad-hoc phylogenetic diversity-area measures against a PAM archive

The Lifemapper Project is exploring mapping it's algorithms into a MapReduce paradigm using an Apache Hadoop-based Architecture (HBA) and software-defined systems (SDS) and Multiple-Domain Distribution/Replication (MDD) of Lifemapper itself as part of a push for investment in sustainable biodiversity cyberinfrastructre. Allowing Lifemapper to live at other institutions through MDD will allow platform owners to define the types of analyses supported by Lifemapper meeting an ever growing need for more flexible and ad-hoc algorithm deployment. Researchers in the areas of bioinformatics that Lifemapper supports live in a world dominated by R scripting. Parallelizing R for Hadoop, using one of several well established methods for this, like R+Hadoop or RHIPE may allow us to calculate larger jobs in a finer grained manner, allowing code reuse, and uncoupling analyses from siloed stacks in Python on LmCompute.

A useful application of this would be the calculation of phylogenetic-diversity, over-dispersion/evenness and clustering for user defined subsets of a PAM archive or Global PAM (GPAM). With the GPAM, PAM construction could be pipelined and a continuously updated PAM archive for all the world's terrestrial species from GBIF could be sub-setted, both taxonomically and spatially, by a user for on-demand data needs. Phylogenetic trees would have to be resolved from tree provider services, now coming on-line, for the species in the PAM, and Lifemapper services could enable those data through a phylo-to-matrix module, that would abstract the phylogenetic topology into a series of matrices and provide permutations of the phylogenetic data for hypotheses testing. These products would have several over-linking uses across different types of analyses. Such a PAM archive and its computational architecture for distributing matrix math across compute resources could also support the quantitative evaluation of the joint effects of historic biogeographic events to test whether different species are more or less constrained by past biogeographic events. A meta-community analysis like this is outlined by Leibold et al. 2010, where the degree of contingent historical constraint is compared to

environmental suitability across a phylogeny using correlation matrices derived from several types of data. The authors of this method point to the need for addressing issues of range shifts and phylogenetic adaptation in meta-communities across several clades requiring extensive phylogenetic information (Leibold et al. 2010). Adding more robust phylogenetic based analyses to models in Lifemapper in combination with the niche models in its archive would be a valuable resource for such an analysis.

### 5.2 Conclusion

We have summarized an on-going effort to incorporate phylogenetic data into a flexible computational platform for multi-species range and diversity modeling in order to bring a more complete history of the diversity patterns of species' communities into focus. Concentrating on range-diversity relationships and a species 'diversity field' derived from calculations on large matrices presented to a thick GIS client in QGIS as web-services allowed us to build a set of robust tools that leveraged open-software, and exposed those analyses to a larger audience, enabling transformative new science. The addition of phylogenetic data to the range-diversity plots and maps allows a user to explore community assembly of species habitats and answer questions about dispersal, competition and adaptation to the environment.

With the explosion of data across all areas of ecology and especially in the phylogenetic community, the need for scalable software solutions for dealing with computationally intensive calculations on large datasets is increasingly clear. Common to most of the methods discussed for analyzing phylogenies is the wish to combine them with environmental data and species range data. Macroecology and biogeography are becoming more cross-disciplinary and are incorporating more methods from community phylogenetics. As this happens phylogenetic datasets will need to reach across more of the tree of life. Spatially they will become biogeographical in scale requiring that researchers have access to computational resources not easily accessible to non-computer specialists. A set of phylogenetic community ecology algorithms that leverage those resources through a suite of web services with a thick client should be designed for maximum flexibility allowing code reuse, and definable by the end user freeing the researcher to concentrate on formulating and testing hypotheses in order to be able to describe the earth's diversity and answer important questions about the fate of our planet's health. Lifemapper is a computational platform that answers some of these challenges, it has implemented a suite of range-diversity statistics never before formalized in relation to phylogenetic data, with a unique interface which scales to large phylogenetic trees, embedded within a rich spatial GIS environment.

**Acknowledgements:** Authors were supported by NSF/BIO/AVAToL Award #1208472. We are grateful to our colleagues and collaborators, Jorge Soberon, Andres Lira-Noreiga and Rafe Brown.

## References

Arita, H. T., Christen, J. A., Rodriguez, P., & Soberón, J. (2008). 'Species diversity and distribution in presence-absence matrices: mathematical relationships and biological implications.' The American Naturalist, 172(4), 519-532.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). 'Dě data-driven documents.' Visualization and Computer Graphics, IEEE Transactions on, 17(12), 2301-2309.

Cavner, J.A., Beach, J.H. Stewart, A.M. Grady, C.J (2014) Lifemapper Macroecology Range and Diversity Tools v. 2.0.1 [QGIS plugin, Computer Software], Lawrence, KS: University of Kansas Biodiversity Institute. Available from http://plugins.qgis.org/plugins/lifemapperTools/

Cavner, J. A., Stewart, A. M., Grady, C. J., & Beach, J. H. (2012). 'An innovative Web Processing Services based GIS architecture for global biogeographic analyses of species distributions'. OSGeo Journal, 10(1), 11.

Cianciaruso, M. V. (2011). 'update: Beyond taxonomical space: large-scale ecology meets functional and phylogenetic diversity.' Frontiers of Biogeography, 3(3).

Emerson, B. C., & Gillespie, R. G. (2008). 'Phylogenetic analysis of community assembly and structure over space and time.' Trends in Ecology & Evolution, 23(11), 619-630.

Faith, D. P. (1992). 'Conservation evaluation and phylogenetic diversity'. Biological Conservation, 61(1), 1-10.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). Design patterns: elements of reusable object-oriented software. Pearson Education.

GDAL (2013) Geospatial Data Abstraction Library: version 1.10.1 Open Source Geospatial Foundation, http://gdal.osgeo.org

Goldberg, E. E., Lancaster, L. T., & Ree, R. H. (2011). 'Phylogenetic inference of reciprocal effects between geographic range evolution and diversification.' Systematic Biology, 60(4), 451-465.

Jetz, W. and Rahbek, C. (2001) 'Geometric constraints explain much of the species richness pattern in African birds.' Proc. Nat. Acad. Sci. USA 98:5661–5666

Jones, E. (2001) SciPy: Open Source Scientific Tools for Python Url: http://www.scipy.org/

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., ... & Webb, C. O. (2010). Picante: R tools for integrating phylogenies and ecology. Bioinformatics, 26(11), 1463-1464.

Laffan, S. W., Lubarsky, E., & Rosauer, D. F. (2010). 'Biodiverse, a tool for the spatial analysis of biological and related diversity. ' Ecography, 33(4), 643–647.

Landis, M. J., Matzke, N. J., Moore, B. R., & Huelsenbeck, J. P. (2013). 'Bayesian analysis of biogeography when the number of areas is large.' Systematic biology, 62(6), 789-804.

Leibold, M. A., Economo, E. P., & Peres-Neto, P. (2010). Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. Ecology Letters, 13(10), 1290-1299.

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). NEXUS: an extensible file format for systematic information. Systematic Biology, 46(4), 590-621.

OGC 2007b OpenGIS Web Processing Service, Version 1.0.0. Wayland, MA, OGC Document No 05-007r7

Scheiner, S. M. (2012). ' A metric of biodiversity that integrates abundance, phylogeny, and function.' Oikos, 121(8), 1191-1202.

Stewart, A.M., Beach, J.H., Grady, C.J., Cavner, J.A. (2014) Lifemapper [Computational platform services for species distribution modeling and continental-scale biodiversity pattern analyses] Web: www.lifemapper.org