Supporting Synthetic Science by Integrating Evolutionary Data in the EcoData Retriever

Ethan P. White, Department of Biology and the Ecology Center, Utah State University, 5305 Old Main Hill, Logan, UT 84341, Email: ethan.white@usu.edu, Phone: 435-797-2097

Project Summary

Large amounts of biological data are available for analysis and synthesis. One of the primary challenges for synthetic research is the process of making individual datasets usable and then integrating the different datasets. This process is difficult because most biological data does not adhere to standard formats and violates basic rules of data structure. In addition, linking datasets is challenging due to differences in structure and problems with identifiers for linking data. This means that compiling and combining data to answer synthetic questions is difficult, and requires computational skills and knowledge that many biologists lack. My lab developed software to reduce these barriers to synthetic science in ecology. This software automatically downloads data, processes it into proper structure, and stores it in a variety of common formats for easy analysis. The proposed research would support a major expansion of this software to: 1) Include evolutionary, taxonomic, and trait information of interest to evolutionary biologists; and 2) Automatically generate synthetic datasets combining data across datasets to allow researchers to rapidly conduct synthetic analyses. These efforts will focus on facilitating eco-evolutionary synthesis, which is particularly difficult due to differences in data structure among disciplines and the fact that most researchers are not experts in both evolutionary and ecological data.

Public Summary

Large amounts of biological data are being collected every year and made available for analysis. However, this data is collected by many different groups and in many different ways, which makes combining the data to answer synthetic questions difficult. To allow scientists to quickly and easily leverage this data, we have started developing a software package that automatically downloads, cleans up, and installs many of the most important datasets in synthetic ecology. The proposed sabbatical will allow the extension of this software to evolutionary biology, and will allow it to produce complete synthetic datasets that combine many datasets into one. This will support evolutionary and eco-evolutionary synthesis by making it easier to assemble, clean, and combine the necessary data, thus leaving scientists with more time to focus on doing science.

Introduction and Goals

Increasingly large quantities of biological data are being generated a combination of: 1) large coordinated sampling efforts such as the National Ecological Observatory Network (http://neoninc.org); 2) compilations of results into standardized repositories such as GenBank (http://www.ncbi.nlm.nih.gov/genbank/) or the Tree of Life (http://tolweb.org/tree); and 3) the publication of individual datasets in repositories like Dryad (http://www.datadryad.org; Whitlock et al. 2010). As a result of this rapidly expanding availability of data, synthetic research in biology is increasingly limited by the rate at which available data can be acquired, organized, and analyzed. Unfortunately, this process of assembling data for synthesis is time consuming and error prone, because most biological datasets do not adhere to any agreed-upon standards in

format, data structure or method of access (e.g., Jones 2006, Reichman et al. 2011). Even when individual datasets can be relatively easily acquired, combining them to address synthetic questions is difficult and time consuming even for experienced data scientists.

These problems mean that scientists spend a lot of time simply assembling, cleaning, and combining datasets. In addition, many biologists lack the skills and tools required to assemble this type of data. This means that the difficulty of assembling data can serve as a major barrier, preventing many scientists from conducting synthetic research at all. These challenges are magnified when trying to work across disciplinary boundaries, because of the lack of researchers with expertise in the availability and structure of data in their across multiple disciplines.

My lab has begun to address these challenges by developing software (the EcoData Retriever; http://ecodataretriever.org) to make it easier to conduct synthetic science in ecology. This software automatically downloads data for individual datasets, cleans it, processes it into proper structure, and stores it in a variety of common formats for easy analysis. While this software has proved useful to us and a growing number of other researchers, it is limited by a lack of evolutionary data and an inability to automatically integrate evolutionary and ecological datasets.

With the leadership of NESCent over the past decade, evolutionary synthesis has now come to the forefront of synthetic research in biology. In addition, it is increasingly recognized that evolution and ecology are inextricably linked. This means that in order to best facilitate synthesis in both evolution and ecology the Retriever needs to be expanded to include evolutionary data. It also needs to move beyond helping researchers at the level of the individual dataset and begin facilitating the combination of datasets to automate the process of producing data for synthetic projects. I propose to address these critical needs by addressing three major goals:

Goal 1: Incorporate evolutionary and taxonomic data. I will add evolutionary data to the Retriever including phylogenetic, taxonomic, trait, and other useful data identified through interactions with the NESCent community. This work will utilize existing informatics resources to avoid duplication of effort.

Goal 2: Automatically build synthetic datasets. Once clean, well formatted, data exists for individual datasets, the next major challenge is combining those datasets for analysis. I will add the ability to combine datasets for synthesis by adding core architecture to address this problem and by modifying our existing approach to allowing users without strong computational backgrounds to use the Retriever to make it possible for them to produce these datasets.

Goal 3: Provenance baked in. It is increasingly recognized that for science to be reproducible that the history of the entire process from data collection through analysis needs to be documented. However, rigorous documentation often begins after data acquisition and cleaning has occurred. I will add automated provenance recording and data archiving to the Retriever to make recording the details of data acquisition and processing as simple as pushing a button.

Addressing these goals will result in a key tool for immediately supporting evolutionary and eco-evolutionary synthesis using published data in its existing format, while broader initiatives (e.g., Parr et al. 2012, DataONE) work towards developing more complex and inclusive solutions based on new standards and infrastructure for data publication.

Proposed Activities

Goal 1: Incorporate evolutionary and taxonomic data

A broad array of data relevant to evolutionary and eco-evolutionary synthesis will be added to the Retriever. Thanks to a number of ongoing efforts by other groups (including some based at NESCent) much of this work will simply involve accessing web services to allow their data to be accessed through the Retriever and readily integrated with other data sources. This work will include integration with existing taxonomic and phylogenetic informatics efforts (see below) and the addition trait based datasets as well as other important datasets for evolutionary synthesis determined through interactions with members of the NESCent community.

Integration of taxonomic tools: One of the major challenges in linking datasets in biology is differences in in taxonomies. I will integrate the Retriever with existing resources for standardizing taxonomic information (e.g., iPlant and Phylotastic's Taxonomic Name Resolution Services, http://tnrs.iplantcollaborative.org/, http://www.evoio.org/wiki/Phylotastic/TNRS; Integrated Taxonomic Information Service, http://www.itis.gov/). When importing datasets the Retriever will (optionally) clean up and standardize the taxonomic information using these services. Logs will be generated of all taxonomic changes (either as log files or as tables added to the database) to allow for full back tracking to the original data source. In addition to improving analysis and reporting of individual datasets, this integration of taxonomic tools will facilitate **Goal 2**.

Integration of phylogenetic tools: One of the key components to integrating evolutionary and ecological data is the phylogenetic tree for the species involved. I will leverage existing informatics efforts that are compiling and providing access to phylogenetic data (e.g., Phylotastic, http://phylotastic.org; Tree of Life; http://tolweb.org; TreeBASE, http://treebase.org/) through web services to integrate phylogenies with other biological data.

Trait and life history data: Publicly available trait and life history data is currently distributed across a wide array of different sources (e.g., Freshwater Biological Traits Database, http://www.epa.gov/ncea/global/traits/; AnAge, http://genomics.senescence.info/species/; USDA plants, http://plants.usda.gov/). The available data will be added to the Retriever, and restructured to allow all of the different trait databases to be consistently integrated with other datasets.

Goal 2: Automatically build synthetic datasets

One of the biggest challenges in synthetic research is combining datasets. I will develop tools in the Retriever to facilitate two different kinds of dataset combination. First, I will build functionality to combine similar data from datasets with different structures into single synthetic datasets. For example, in White et al. 2012 we combined data from six different datasets, with widely differing structures, to analyze species abundance distributions across ecosystems and taxonomic groups. However, we did this work by writing individual sets of queries for each dataset, resulting in over 300 lines of Python + SQL that is fragile, scales poorly, and is not generalizable. The updated Retriever will allow the specification of fields in different datasets that should be combined into a single core field in a synthetic dataset. The ability to use queries

of the original dataset to generate the table with the necessary fields will also be included. Second, I will build functionality to allow different types of datasets to be combined based on a set of common linking fields. This is again work that my group has done in the past (e.g., Thibault et al. 2011 combines trait and ecological data to analyze patterns of avian body size), but in very project specific ways. I will generalize the Retriever to allow taxonomic, phylogenetic, trait, and ecological data to be combined into synthetic datasets for rapid analysis.

Because the combinations of data for synthetic projects vary substantially, I will also improve and expand our existing system to allow users with limited computational backgrounds to combine datasets in customized ways through a simple web interface. In collaboration with Ben Morris, I will improve our existing approach to adding datasets to the Retriever by transitioning from our current custom scripting system for describing the structure of individual datasets to formal semantic web standards including the Resource Description Framework (RDF) and Web Ontology Language(OWL). This will make it easier to describe the relationships between different datasets and serve as a building block for integrating the Retriever more thoroughly into the growing eco-evolutionary informatics ecosystem.

Goal 3: Provenance baked in

One of the current challenges with synthetic data analysis is tracking the many steps involved in assembling the data. While workflows of various forms increasingly document analyses, this process often starts after the data download and processing has already occurred. I will add functionality to the Retriever to store all of the information necessary for complete provenance with the resulting data. Data sources, download dates, and the version of the Retriever used to process the data, will all be recorded. In combination with the Retriever's open source, version controlled, code base this will allow the entire process of data acquisition and processing to be recorded. This metadata will be stored in the appropriate table metadata systems for the database management systems, and in structured comments in csv files. I will also add an archive option to the Retriever that will store this data in a log file, and will store both the raw downloaded data and the processed form as exports/dumps from the chosen database management system.

Rationale For NESCent Support

NESCent is a major center of evolutionary informatics and is therefore the ideal location for this work. My efforts will use existing initiatives, to avoid duplicating effort, and interactions with NESCent's informatics team will help facilitate the integration of the Retriever with existing initiatives. In addition, NESCent hosts many researchers addressing synthetic evolutionary questions. Interacting with these researchers will allow me to learn more about what kinds of datasets and combinations of datasets are most needed by the evolutionary synthesis community. Finally, NESCent's involvement in the development of DRYAD (one of the primary repositories for evolutionary data) will facilitate the integration of DRYAD data into the Retriever.

Collaborations

Ben Morris, the undergraduate student in my lab who lead the development of the Retriever, is now in graduate school at the University of North Carolina and part of the NESCent informatics group. Ben has agreed to collaborate on this expansion of the Retriever, and being in the same

location as will facilitate this effort. In addition, I hope to meet regularly with members of the NESCent IT group to facilitate the integration of evolutionary tools into the Retriever.

Proposed Timetable

I propose a 12-month sabbatical from August 15, 2013 to August 15, 2014. **Months 1-2:** Improve fundamentals of the Retriever to make adding new features easier and the code more maintainable. Improvements to the testing framework and documentation, and addition of continuous integration. Interact actively with the NESCent informatics and scientific communities to identify the best areas to focus effort in Goals 1 & 2. **Months 3-6:** Goal 1. **Months 7-10:** Goal 2. **Months 11-12:** Goal 3 and wrap up of remaining tasks from Goals 1 & 2. Releases will occur throughout the year as features are completed.

Anticipated IT Needs & Plan For Making Data/Software Available

I will be developing software as part of this proposal, but I will be handling all of the development myself and therefore will not require IT support (though I greatly look forward to the possibility of interacting with this group). The software developed at NESCent will be released under an MIT License (an approved OSI open source license). I and my lab have a long history of open science activities including the publication of data and software under open source licenses. In fact, this entire proposal was developed in the open on GitHub under a CC-BY license (https://github.com/ethanwhite/nescent-sabbatical-proposal).

Anticipated Results

This proposal will result in 2-3 major releases of the Retriever and substantial expansions and improvements of the website (http://ecodataretriever.org) and associate documentation. In addition, the rOpenSci team and I are planning to wrap the Retriever's command line interface to allow it to be used from inside of R. This will result in a new R package published on CRAN.

References Cited

Jones, M. B. et al. 2006. The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. Annual Review of Ecology, Evolution, and Systematics 37: 519-544.

Parr, C.S., R. Guralnick, N. Cellinese, and R.D.M. Page. 2012. Evolutionary informatics: unifying knowledge about the diversity of life. Trends in Ecology & Evolution 27:94 - 103.

Reichman, O.J., M.B. Jones, and M.P. Schildhauer. 2011. “Challenges and Opportunities of Open Data in Ecology.” Science 331, 703–705.

Thibault, K.M., E.P. White, A.H. Hurlbert, and S.K.M. Ernest. 2011. Multimodality in the individual size distribution of bird communities. Global Ecology and Biogeography 20:145-153.

White, E.P., K.M. Thibault, and X. Xiao. 2012. Characterizing species-abundance distributions across taxa and ecosystems using a simple maximum entropy model. Ecology 93:1772–1778.

Whitlock, M.C., M.A. McPeek, M.D. Rausher, L. Rieseberg, and A.J. Moore. 2010. Data Archiving. American Naturalist 175:145-146.