Project Description

Introduction

Phylogenetic trees are essential for scientists making evolutionary inferences. As hypotheses on inter-species relationships and history, phylogenies provide insight about the evolution of characters through time, about relationships among organisms and are used as an organizing framework for the study of molecular evolution. Time calibration of phylogenetic trees allows researchers to infer the timing of divergence events, and the rate of divergence accumulation, on the trees. Many modern comparative methods — methods that allow researchers to infer macroevolutionary patterns, such as speciation and extinction dynamics — require trees to be calibrated to an absolute time scale (e.g., years or millions of years). Because of the prime importance of time-calibrated trees in biology, the development of accurate methods for estimating of trees and divergence dates is a critical research area of modern evolutionary biology.

Time-scaling a phylogenetic tree is the process by which rate and time, which are conflated in the branch lengths of a phylogenetic tree, are disentangled. Time calibration of phylogenetic trees has traditionally been accomplished through the use of node-based calibrations. Node calibrations are independent pieces of information, such as the occurrence date of a fossil, that can be used to constrain the age of a particular node on the tree. This information is typically used to parameterize a probability distribution on the age of a node. In practice, parameterizing these distributions is often based on little more than intuition [Heath et al., 2012, 2014]. Due to the arbitrary nature of calibrations, uncertainty and conflict among results is to be expected.

Generally, not every node will have a calibration available. Traditional node calibrations require certainty in placing a specimen on a tree before it can be used. Uncertainty in placing some specimens means not all available fossils have been used. To time calibrate every node on a phylogeny, including those with no fossils, rates must be extrapolated across the whole tree. Clock and tree models are used for this purpose. Clock models describe how the variation in rates is distributed across the tree. Tree models are used in conjunction with clock models to extrapolate rates and times. Where clock models describe how branch-wise substitution rates are distributed across the tree, tree models describe how speciation events are distributed through time. The inherent subjectivity of parameterizing node calibrations, combined with not making use of all the available data can lead to conflicts among estimated ages between studies.

Stochastic-branching processes with sampling through time have been described as tree priors to account for fossil observations represented in the phylogeny [Heath et al., 2014, Stadler, 2010, Didier et al., 2012, Gavryushkina et al., 2015, Zhang et al., 2015]. These fossilized birth-death processes (FBD; [Heath et al., 2014]) are alternatives to traditional node-calibration approaches. Treating fossils and molecular data as part of the same generating process, these models condition the tree topology and branching times on several key parameters including the origin time of the process x_0, the speciation rate λ, the extinction rate μ, the probability of sampling extant lineages ρ, and the rate of sampling lineages through time ψ [Heath et al., 2014, Stadler, 2010]. Rather than assuming a tree-wide birth-death process and parametric prior densities for calibrated nodes, a single diversification process describing speciation, extinction, fossil recovery, and extant lineage sampling is used to calibrate the tree. The FBD process is useful as a tree model in any divergence-time analysis that includes extant and fossil taxa. FBD eliminates the need for node calibrations that may have little relationship to actual biology and provides coherent measures of statistical uncertainty in speciation times.

In addition to analytical simplicity, FBD models are advantageous for researchers who want to use morphological data more completely [Gavryushkina et al., 2015, Zhang et al., 2015]. Node dating using calibration densities discards morphological information and other data from the fossil record because calibration priors are often distilled down to their age information to obtain the minimum age for a clade, and use the oldest fossil in a clade. Under the FBD approach, all fossils belonging to a clade can be added to the analysis. Fossils are assumed to be part of the same process that generated the extant data (usually in the form of DNA sequences and morphological characters), though exact phylogenetic placement of the fossils may be uncertain. Allowing for uncertain placement of fossils with a clade allows fossils for which there is not a full matrix of morphological data to still be used in the analysis.

I will use ants as a model system to test new methods for phylogenetic divergence time estimation, and to develop a more precise estimate of ant phylogenetic relationships. Ants are an ideal study system to accomplish these goals because of the vast amount of genetic data and well-established collections of

ant fossils at the Field Museum. Making use of these combined resources will allow me to estimate a more-precise time-scaled tree, and to develop best-practice recommendations for scientists who wish to integrate molecular and morphological data in dating analyses.

Ants are a diverse group of organisms with wide ranging-impacts on plant, animal and human ecology. Current divergence estimates suggest that ants are about 145 million years old, and there are abundant fossils from the most recent 100 million years. Because commonly-used divergence dating methods are constrained to only use the oldest members of a taxon, most researchers have used the same set of fossils as node calibrations, but have applied different methods and assumptions to the data (e.g. [Brady et al., 2006, Moreau and Bell, 2013, Ward, 2014]). This has resulted in conflict about the age of the ant group, and this has been propagated up the tree, leading to conflict among age estimates within the ant group from different authors. I will make more complete use of the true richness of the ant fossil record to estimate a dated tree which will provide evidence to settle some of the controversies.

There are around 13,000 extant ant species, many of which have archived genetic data stored in GenBank. There are about 700 species of fossilized ants, with some represented by dozens or hundreds of individuals [Perrichot, 2014], with fossils present in formations around the globe dating from 100 million years ago through to recent times [Perrichot, 2014], and published data matrices used to build phylogenetic trees from morphological data [Urbani et al., 1992, Grimaldi et al., 1997, Brady and Ward, 2005]. I am proposing to work with the extensive ant (Formicidae) collections at the Field Museum to develop a time-scaled phylogenetic tree of ants and to use the data collected in building this tree to advance our understanding of the performance of FBD models. This project will generate many opportunities for the involvement of undergraduates in collections-based research, computational biology and evolutionary biology. The results of the proposed projects will be helpful to comparative and ant biologists, as well as to empiricists seeking best practices for their analyses. I am a computational evolutionary biologist, and this work will provide a platform for my future work by demonstrating the limits of FBD methods. Identifying the limits of the current methods will allow me to see where improvements can be made or entirely new models need to be developed.

Research Objectives, Methods and Significance

Hypothesis One: Increased sampling of morphological data will increase precision of divergence time estimation among ant species

Ants are one of the most ecologically successful terrestrial animal groups, and also engage in symbioses with a diversity of other organisms. A well-resolved and accurately dated phylogeny has implications for our understanding of not only the evolutionary history of this insect group, but for the other organisms that interact with ants including plants, fungi, bacteria, and other animals. The first objective I have is to reconstruct a well-resolved and time-calibrated phylogenetic tree of ants. This is an eminently achievable goal from my present skill set, and the tree will provide the organizing framework for the rest of the project. About half of the 13000 currently described ant species are present in GenBank. These data will supplement the matrices from previously-published studies [Brady et al., 2006, Moreau et al., 2006, Moreau and Bell, 2013], which were used for traditional node-calibration approaches to dating. Previous approaches have not made extensive use of the fossil record of ants, nor of the extensive collections of ants in many museums. Updating trees with this information will lead to new results and more accurate divergence time estimates. Time-calibrated trees are useful for a variety of comparative methods. This tree will be archived in open resources, such as TreeBase and Open Tree of Life, for reuse. Due to the size of the tree, I will have opportunities to subsample my data to examine various biases in divergence dating. The empirical practice of estimating this tree will allow me to provide many best-practice recommendations for other researchers working across the tree of life.

Methods

I will use publicly-archived DNA sequence data to augment previously-published matrices. There are nearly 300,000 ant DNA sequences archived in GenBank, covering approximately half of all ant species. Misidentified sequences are a concern in using publicly-archived sequences. To remove these sequences, I will use BLAST [Altschul et al., 1990] to search GenBank, using the downloaded sequences as a query. Sequences that return only the original sequence as a match, or clustering with sequences outside their

putative genus will be removed as potential mislabels. The most complete dataset that can be assembled from these data will be the backbone of the phylogenetic tree.

The Field Museum has an excellent collection of amber-preserved ant specimens, which will augment the existing amber and compression/rock fossil data available on AntWeb.org [Perrichot, 2014]. These data will be leveraged in two ways. I will collect previously-published morphological matrices [Grimaldi et al., 1997, Brady and Ward, 2005]. If a fossil has sufficient data to be added to pre-existing morphological matrices, I will score characters for that specimen and add it to published matrices. If there are not sufficient characters to add the specimen to the matrix, I will place previously-identified specimens on the tree and use their age information to date the bipartitions in the tree. One benefit of the FBD model for morphologists is that the exact placement of a fossil on the tree can be integrated out via Markov chain Monte Carlo (MCMC) sampling. This means that even highly-incomplete fossils can be used, as the fossils do not need to be added into an entire morphological matrix. Many unknown and unsorted ant specimens also exist in this collection. These specimens will be sorted, identified and databased, and added to the matrix or distilled to age and taxonomic information as appropriate.

When I have assembled my combined molecular-morphological dataset, I will use the FBD model to estimate a time-scaled tree. Different parameterizations of the FBD model exist. I will use Bayesian model-fitting to choose the optimal model parameters to describe the data [Baele et al., 2012]. Abundant fossil occurrence data will permit me to use selective subsamplings to look at the effects of biased or poor samples of fossil specimens. The end result of these various experiments will be a time-scaled phylogenetic tree with estimates of uncertainty. Because the FBD model also estimates speciation and extinction parameters, we will be able to use these results to discuss the diversification dynamics of the ant group.

Intellectual Merit

Performing an FBD dating analysis of ants will furnish a useful point-of-comparison to the existing molecular phylogeny of ants. Use of morphological information in FBD analyses differs from traditional node calibration in two important ways: Firstly, multiple fossils can be used to date one node, which allows us to use more specimens. Using more specimens increases the information available to date a bifurcation on the tree. Secondly, because the exact attachment (i.e., the precise phylogenetic placement of the specimen) point of a fossil to the tree is left vague, specimens of uncertain classification can be used. I predict, for these two reasons, that this approach will yield more precise dates for nodes in the ant tree by incorporating more completely the richness of the ant fossil record. I expect to observe this effect via smaller credible intervals on node dates. Particularly, this effect will be observed at the base of the group, as we will be able to incorporate more specimens from the outgroup, and more specimens known to be in the ingroup, but which are unable to be placed into specific clades with certainty. The age of the ant group has been controversial in the past, and my results will provide additional evidence as to the true age.

Divergence-time estimation methods under the FBD model are nascent and little literature on best practices for these analyses exists. Experimenting with these analyses in a well-sampled group will yield useful results about how this method compares to previously published node calibration approaches. This will provide useful results for researchers using primarily morphological data to co-estimate phylogenetic trees and divergence dates. This objective will also yield a set of genetic data and fossil occurances that will be used in testing Hypothesis Two.

Hypothesis Two: Differential preservation due to geographical and life history traits will decrease accuracy in divergence time estimation among ant species

The FBD model has a parameter, ψ, which describes the rate of fossil recovery over time. Fossil recovery events occur along a branch according to a Poisson process. Current implementations of the model assume that fossilization rates are constant throughout time, or are variable depending on time, and that recovery of specimens is not biased. In practice, we might expect that fossilization may be biologically or geographically biased. Quantifying the effects of these geographical and life history biases is a useful exercise. An understanding of the effects of these biases will help researchers using these methods assemble datasets that fit the assumptions and strengths of the model more effectively. For cases in which more, or different data cannot be gathered, an understanding of the effects of biases allows researchers to understand their results better and to make cautious scientific claims about their results. This work will also be useful to theoretical biologists and methods developers for understanding the biological realities faced by empirical researchers.

There are known geographical biases to where extinct ants have been recovered, such as poor fossil preservation in tropical regions [LaPolla et al., 2013]. Likewise, life history traits may also affect the sample of fossil ants we recover. For example, tree-dwelling ants are more likely to be preserved in amber, allowing them to be placed taxonomically with a greater degree of certainty than imprint fossils. These biases are quantifiable from the specimens at the Field Museum, which are cataloged with location and age information. The Field Museum, having a predominantly amber collection, is inherently biased towards ants that have a tree-habit lifestyle, but these data will be augmented with specimen information available on AntWeb.org.

Methods

I will simulate dataset of fossil data and extant nucleotide data. Using the ant time-tree generated in Objective One as a guide to the geographical and biological biases present in the real data, I will subsample my simulated data to match the geographical and life history biases seen in real data. This will entail writing a novel simulator, to be implemented in the open-source Bayesian phylogenetics package RevBayes. From the simulated data, I will estimate time-calibrated trees using both FBD and traditional node calibration. I will then quantify the mismatch between the estimated time-trees in terms of topological structure and estimated ages of nodes.

Intellectual Merit

Failing to adequately model aspects of the underlying divergence process is a form of underparameterization. Underparameterization of phylogenetic models is known to lead to inaccuracies in estimation of model parameters. I, therefore, expect that I will find that failing to account for the complexity of ant evolutionary history will lead to decreased accuracy and precision in estimation of phylogenetic trees and other model parameters from the observed data. I expect that this effect will be observed in several ways. Firstly, I expect that biases that prevent certain types of ants from being placed on the tree confidently will lead to a greater decrease in accuracy (as measured by fewer nodes being estimated and dated correctly) in node-dating than tip dating methods. This is expected because node-dating requires more taxonomic information before a specimen can be utilized. Uncertainty in placement will lead to a reduction in available information for node calibration, but not FBD dating. Secondly, I expect that I will observe a great deal of geographic bias in ant samples, as many ancient ants are known from European amber samples [Perrichot et al., 2008]. In simulation, we can examine what the effects would be if this is a mere artifact of preservation, not an indication of their true paleogeography. I expect that an artifact such as this will yield an inaccurate, but well-supported, estimate of the dates on the tree. The strength of support for the erroneous date will depend on the strength of the artifact: if nearly all non-amber fossils are excluded, there will be no conflict, resulting in a well-supported date. If some of the true biogeographical signal remains, this will weaken the support for the erroneous date.

This experiment will yield results showing which types of biological and geographical bias cause current methods of phylogeny and divergence date co-estimation to err. By identifying the important biases that affect our ability to accurately estimate species divergence times, I can generate recommendation best practices for these methods. This work will also expose the limits of these methods as currently implemented. Understanding the limitations and weaknesses of these models will help me develop my independent career in developing extensions of FBD models, developing new models and modifications to specific implementations.

Training Objectives and Career Enhancement

The projects outlined will build on and expand my current skill set. My work to date has focused on the estimation of undated phylogenetic trees. Using simulated and empirical approaches, I have explored the efficacy of phylogenetic models for estimating trees from morphological data [Wright and Hillis, 2014] and with comparative methods [Wright et al., 2015, Meirelles et al., 2015] for using phylogenetic trees to make evolutionary inferences. The proposed research will give me experience with divergence date estimation and with mathematical modeling that can be applied to a variety of further biological questions, particularly those relating to evolutionary inferences using phylogenetic trees. While I have applied models to data, I have not previously worked with writing or extending mathematical models for biological inference. The work I have proposed will give me experience with modeling of this type. My research thus far has required me to develop computational biology skills, my proposed work will allow me to expand those skills to the development of actual software. I will gain experience with C/C++ programming, which remains essential

in software approaches for phylogenetic inference. My work to date has largely been solitary; with the proposed projects, I will also gain experience working as part of a team for collaborative development of software (specifically RevBayes).

Presently, I do not have experience with collections-based research. While I have the skills that are needed to administer the computational side of museum collections, I have little exposure to the material side of collections. Working in the collections will give me hands-on experience with sorting and cataloging specimens, as well as data collection from fragile or unique organisms.

FBD methods will form a large part of my future independent career. I would like to continue working in the field of evolution research, particularly as a faculty member at a undergraduate-serving institution. I have not yet worked with undergraduate researchers. Scoring characters and using collections to quantify biases in the fossil record is not a one-person task. To complete this work, I will have to work with undergraduate assistants. This work will also provide the opportunity to work with computationally-inclined undergraduates in the development of the RevBayes software. These two projects all present unique opportunities to do cutting-edge research while mentoring a diverse group of students, as I will have to do as a faculty member.

Justification of Sponsoring Scientists

My proposed primary mentor, Dr. Tracy Heath, is an assistant professor at Iowa State University in the Department of Ecology, Evolution and Organismal Biology, and has made substantial contributions to the practice of divergence dating, in both theoretical and practical realms. Dr. Heath has authored a number of papers critically examining the role of fossils and lineage sampling in these methods. She is an ideal sponsor for this reason. Dr. Heath can also provide the computational support this project will require. She has developed a stand-alone software implementation of one of her models for divergence estimation, DPPDiv. She is also a core developer of RevBayes, the open source Bayesian phylogenetics software package in which I plan to implement my simulator. ISU has a supercomputing center that I will have access to for development and testing of software.

Dr. Corrie Moreau, Associate Curator of Insects at the Field Museum in Chicago, leads a thriving research program investigating ant evolution and systematics. She can provide needed access to collections, and to training in the handling and care of the specimens in the collection. The Field Museum also has a number of intern and undergraduate work programs through which I can obtain the assistance I will need to complete the work. Her empirical expertise and knowledge of the ant literature is necessary for this research. Dr. Moreau is also a curator with AntWeb, a resource with which I will have to become familiar to archive and manage specimen data appropriately for use by other researchers. The Field Museum and ISU are also within driving distance of one another to facilitate collaboration, both between myself and Dr. Moreau, and between Dr. Heath and Dr. Moreau.

Broader Impacts

The Field Museum environment also provides unique opportunities for engagement of the public, having many established weekly and monthly events for a variety of learners. The process of generating and subsampling the data can also be used to make a learning module on fossil ants for AIM-UP (Advancing Integration of Museums into Undergraduate Programs; http://aimup.unm.edu/), an educational effort in which Dr. Heath takes part [Cook et al., 2014].

I have a long record of service in computational education and outreach. I am an instructor with the Software Carpentry Foundation, a non-profit organization that sends volunteer computational scientists to universities and research facilities to train workers in computational best practices. I also developed, with a co-instructor, a free course for biologists at the University of Texas at Austin to learn computational biology. This course is beginning its fourth iteration in January. Materials for this course are Creative Commons licensed and freely available on my GitHub page (https://github.com/wrightaprilm). Dr. Heath and Dr. Moreau both have extensive records of outreach, particularly to women and minorities. I made it very clear from my first interactions with both mentors on the subject of postdoctoral research that I consider workshop teaching and outreach to be fundamental parts of my career. I am currently involved with organizing and teaching workshops and study groups at Iowa State University on computational topics and data management, I have already began organizing workshops and assisting in study groups at ISU.

Date	Item	Notes
June - August 2016	Primary data collection at Field Museum	Build a database of available fossils, occurrence time, taxonomy and fossil type. Field Museum undergraduates can be involved in this process.
Fall & Winter 2016	Phylogenetic estimation from the full dataset and the subsampled datasets will occur	ISU undergraduates can be involved in the computational aspects of tree estimation
Spring 2017	Writing and submission of the publication for Objective One at both ISU and the Field Museum	New calibrations will be archived publicly in the Paelontologica Electronica Fossil Calibration database and the tree of extant species will be submitted to Open Tree of Life's synthesis tree
Spring-Summer 2017	Simulations for Objective Two will be conducted	Simulation software will be released in the RevBayes software package when it is complete
Fall 2017	Data analysis for Objective Two	ISU undergraduates interested in biology and computer science can be involved in this and the previous steps.
Winter 2017 and Spring 2018	Writing and submission of the publication for Objective Two at both ISU and the Field Museum	Dr. Heath has committed funds for open access publishing

Table 1: Timetable.

I will develop new workshops at Field Museum for collections researchers who want to learn more about computational biological methods, and in conjunction with the Women in Science group run by Dr. Moreau.

The research proposed will require undergraduate assistants. These undergraduates will have to master a variety of skills. For building a time-calibrated tree, these students will be programmatically accessing databases (GenBank and AntWeb) and parsing downloaded data. To quantify biases in the ant fossil record, these students will be getting hands-on experience with the materials side of biological collections, but also the computational skills of cataloging specimens. To assist me in Objective Two, students will gain skills in data simulation and analysis. My projects provide students with training in the computational techniques increasingly important in academia, industry and the public sector, with a strong biological basis. Dr. Heath is involved with the Society for Advancement of Hispanics/Chicanos and Native Americans in Science, and Dr. Moreau is involved with the Field Museum's Women in Science program. I have a history of recruiting women instructors in my own courses and workshops and of representing the interests of women and underrepresented minorities on the diversity boards for the Software Carpentry Foundation and the SciPy 2015 meeting diversity committee. I would like to continue this trend by recruiting undergraduates using the professional networks afforded to me by my co-supervisors.

Lastly, I am an open-source programmer. All the code associated with my papers and dissertation work is hosted freely on the internet and archived in appropriate repositories. Software for the outlined projects will be released via the RevBayes project (github.com/revbayes/revbayes). RevBayes is an open-source, generously-licensed software package with contributors from six US-based and international universities. The software is coded in an object-oriented framework to allow users more flexibility in analyses, but also to lessen the burden of project maintenance and expansion. New fossil age estimates or new classifications will be archived in appropriate databases. Dr. Heath has committed funds for open-access publication. Dr. Moreau is a curator for AntWeb (an online, open-access database of specimen records and images for ants), and fossil ocurrances and any related digitized material will be housed in AntWeb and iDigBio.