Project Narrative: Improving Modeling of Process Heterogeneity on Fossil-Based Phylogenetics

Wright

Today

1 Rational of the Project

1.1 Potential for National Competitiveness

My prior work has been supported by the National Science Foundation. My graduate work was supported on a National Science Foundation Doctoral Dissertation Improvement Grant (grant no). I was awarded a National Science Foundation Postdoctoral Research Fellowship in 2016 (grant no). The work funded by my NSF PDF is directly informing the research I propose in this grant. However, the PDF is a mostly solitary research fellowship, with awards not funding research assistants, and broader impact activities being tightly constrained by the small allowance for such activities. These funds have been very useful for developing my own ideas, and technical competancies to carry out mathematical and computational work.

However, nationally-competitive funds for early career researchers, such as the NSF's CAREER fund, require evidence of the integration between teaching and research. In order to achieve competitiveness for these types of funds, I need to be able to demonstrate significant overlap between my educational activities and my research activities - i.e., that my classroom activities and outreach activities are informed by the research I perform. National research foundations, such as NSF and National Institutes of Health, are also emphasizing assessment of educational and outreach work. Therefore, in the context of the BOR RCF award, it will be important for me to establish not simply a research program that employs student workers, but a program of outreach and education, and to be performing regular assessment of efficacy for these activities.

1.2 Plan for Achieving Competitiveness

As described above, my ideas have been invested in consistently by the NSF. However, most of the funding has been investment in my development, and the development of theory and software with which I am working. In order to

pivot the funding that I have been successful in achieving to be competitive for non-training grants, I propose to spend the majority of funds through this opportunity on development of student training pipelines, and broader impacts with assessment.

1.3 Competitive Area One: Student Training

Many national awards, such as the NSF CAREER award and NSF ABI awards, rate broader impacts at 50% of the proposal weight. Included in these broader impacts is student mentorship. During my time as an NSF PDF, I developed a set of training materials to onboard research assistants into research on my project. These training materials covered introductory bioinformatics and statistics. I worked with a total of 3 students during my postdoctoral fellowship. I would like to use funding from this project to pay student salaries over the summer, as mastering computational science is not possible in one semester. I would also like to set up a more formal system of assessing student knowledge, and tracking what students go on to do after leaving my laboratory. This type of assessment is prized by most NSF awards, and the interaction of research and mentorship is particularly valued by the CAREER program.

1.4 Competitive Area Two: Outreach and Assessment

My previous NSF funding has allowed for limited interaction with non-academic broader impacts. For competitiveness in national-level funding, I need to be able to demonstrate a clear connection between my research work and outreach. In Louisiana, it is legal to teach creationism in schools. Being an evolutionary biologist, I propose a teacher's workshop at the Turtle Cove facility operated by Southeastern Louisiana University. This workshop, in its first year, will allow teachers to come meet evolutionary biologists, and leave with several classroom activities about evolution. During the first year, I will also hold a listening session, where I can learn about the realities faced by public school teachers in Louisiana - what resources they have in the schools and how they feel the teaching of evolution is received.

In the second year, the same teachers will be invited back to talk about how the activities were received, to compare notes and to refine the activities. The same teachers will be invited to stay for a second workshop day, in which they will help me demonstrate the refined activities to a new set of public school teachers, with the aim of using peer networks to disseminate successful activities among colleagues. We will also introduce bioinformatic activities on evolution for teachers with access to a computer lab.

In the third year of funding, the activities will again be refined, and teachers from the previous two years will be invited to train a third cohort of teachers on how to run the activities. The bioinformatics materials will also be workshopped, and a second cohort of teachers trained in how to use them. Data will again be collected, both in terms of focus group discussions and written surveys to allow me to assess the dissemination and usefulness of the activities.

The multi-year workshop set up will allow me to collect long-term data on how and if materials are being adopted in the classroom. This data will be formally collected and turned into a publication on teaching evolution in a K12 setting. By collecting data on the actual adoption of materials and activities, I will demonstrate that I am capable of performing long-term assessment of outreach activities, which is crucial for awards such as the CAREER.

1.5 Previous Proposals

2 Introduction

2.1 Bayesian Divergence Time Estimation

The estimation of phylogenetic trees has established itself as a chief interest of biologists, and a key step in many different types of biological data analysis. Phylogenetic trees are often interesting in and of themselves, demonstrating the relationships between organisms. These representations of evolutionary history are often also combined with phenotypic data to allow for inferences about trait evolution. Since the late 1970's, it has been appreciated that not just the topology of a phylogenetic tree, but its branch lengths (a representation of the amount of evolutionary change between a common ancestor and its descendants) can be informative for understanding the evolution of organisms and their traits. In recent years, methods for estimating more complex models of evolution for phenotypic traits have arisen. These models typically require a phylogenetic trees that have been scaled to absolute time.

Scaling a phylogenetic tree to absolute time is challenging. Sequence data does not contain direct information about the timing of evolutionary events. For example, two nucleotide sequences that are very different could be very different because they diverged recently and have a high rate of evolution (Fig 1a), or because they diverged a long time ago and have been evolving slowly (Fig 1b). Initial attempts at estimating phylogenetic trees with absolute time dates solved this issue by assuming a single rate of molecular evolution across the tree (Zuckerkandl and Pauling 1969). Further refining of this method allowed researchers to use different rates for each gene in a dataset - but still assumed strongly clock-like modes of evolution.

This has long been recognized as biologically unrealistic. Rates of molecular evolution may vary across organisms due to a variety of biological factors including reproductive rate and mutation rate. Due to these inadequacies, researchers began to incorporate more biological and paleontological data to inform their divergence time estimation. One way in which this has been performed is through the use of node calibrations. Node calibrations use information from other sources, most commonly fossils to constrain the age of a node. This is typically performed by a researcher ascertaining what the oldest fossils in their group of interest are, then using these fossils to place a distribution on the node. This distribution corresponds to a range of ages that the node could possibly have, given the fossil information. These ages typically represent a minimum age: the

lineage subtending the node must be *at least* the age of the fossils belonging to it.

Node calibrations represented a leap forward in the realism of divergence dating analaysis by allowing researchers to incorporate more of the wealth of biological data available to them. But they still have problems: specifying a distribution of ages on a particular node is often subjective. In distilling down the full fossil record to just one or a couple calibrations, much data is thrown away. In 2012, methods began to appear to more completely integrate fossil data with molecular data for dating phylogenetic trees. These methods allow for fossils to be terminal taxa on a phylogenetic tree, with their placements estimated from data, as opposed to assigned by an expert. Performing divergence dating analysis in this way removes the subjectivity of forcing researchers to quantify how old they think a node is, and rather to infer that age.

In 2014, the Fossilized Birth Death model for divergence time estimation was implemented by Heath, Huelsenbeck and Stadler. This model proposes a unified framework in which all taxa on the tree, including whether they are extinct or extant, represented by fossil or molecular data, are assumed to be part of the same process of lineage creation and extinction. In effect, the FBD allows researchers to include as many fossils as they have in their datasets, rather than picking a few to estimate their node calibrations. In modeling the whole process of lineage diversification through time, in addition to estimating a time-calibrated phylogeny, these models also estimate other interesting parameters, such as speciation, extinction and rates of evolution across the tree. An additional benefit to performing divergence dating in a model-based framework is that there is a clear mathematical context to improve the model. For example, if one desired, multiple FBD models can be fit across the tree. If one thought the process of speciation and extinction was different over various time points (for example, before and after the Permian mass extinction), this can be written out as a model extension. Likewise, if one thought that the FBD process is different in one lineage than another lineage, this can also be written out as a model extension.

In this proposal, I propose three goals. These goals build on my National Science Foundation Postdoctoral fellowship, and are designed to help me take that body of work, and pivot that previously solo work into a project on which a full laboratory can collaborate. My first goal is build a time-scaled phylogenetic tree of ants using the FBD model. Ants have a rich fossil record, including many precisely dated specimens. Because of their interest to agriculture and ecology, there are also many genetic resources available for ants. Together, these factors make ants an ideal system for studying, understanding, and developing divergence time estimation methods.

My second goal is to examine adding realism to the modeling of ant divergence dates by allowing for the parameters of the FBD to vary throughout time. Because of their rich fossil record, we know that there are periods in which we have many fossils, and periods in which we have few. Allowing the dynamics of speciation and extinction to vary over time across the phylogeny will likely be a better fit for these data than assuming that the same speciation and extinction

rates apply to the whole tree, in all timepoints. Particularly, I am interested in comparing time-stratified FBD models to FBD models that allow rates of speciation and extinction to vary across clades on the tree for this dataset. Lastly, our understanding of the performance of FBD models is still developing. Therefore, I would like to perform simulations to assess how many data are required to distinguish between time-stratified and cladewise FBD models. Time-stratified models are simpler, and paleontological datasets are often strongly limited in terms of how much more data can be collected. Establishing how researchers can choose between these two model types will be a useful theoretical entry to this body of work.

- Phylogenetic trees and their importance in biology (establish competitiveness)
- A historical overview: From Zuckerkandl and Pauling to Node-Based Calibration
- From 2012 Onward: A more serious role for morphology
- 2014-Now: The rise of explicitly mechanistic models of evolution

  Address competitiveness
-

2.2 Bayesian Modeling of Morphology

- 2001-2014: Paul Lewis and the Mk Model
- From 2014 Onward: April Wright and increasing the biological realism of morphological models

2.3 The Proposed Work

- Hypothesis One: A mechanistic model of evolution will improve divergence time estimation in the Formicidae
- Hypothesis Two: Skyline models for Bayesian divergence time estimation will perform better than clade-specific models for Formicid data
- Hypothesis Three: Simulation study - clade bias in speciation and extinction rates has to be very large for a more-complex clade-wise model to outperform a simpler skyline model.

3 Research Plan

3.1 Hypothesis One: A mechanistic model of evolution will improve divergence time estimation in the Formicidae

3.1.1 Problem

3.1.2 Methods

3.1.3 Expected outcomes

3.2 Hypothesis Two: Skyline models for Bayesian divergence time estimation will perform better than clade-specific models for Formicid data

3.2.1 Problem

3.2.2 Methods

3.2.3 Expected outcomes

3.3 Hypothesis Three: Simulation study - clade bias in speciation and extinction rates has to be very large for a more-complex clade-wise model to outperform a simpler skyline model

3.3.1 Problem

3.3.2 Methods

3.3.3 Expected outcomes

4 Personnel

4.1 Me

4.2 Undergraduate Involvement

5 Facilities

5.1 Compute Resources

6 Timetable