Title: Methods to examine trait evolution on trees: advancing the field

Brian O'Meara, bcomeara@ucdavis.edu, *http://www.brianomeara.info*
Center for Population Biology, One Shields Ave, U. of California, Davis, CA, 95616

Project Summary: Examining the evolution of morphological, behavioral, geographical, genetic, biochemical, and physiological traits using methods that incorporate phylogenetic information has been an important area of growth in evolutionary biology for over two decades. As more methods are made available, the ability to test hypotheses and infer patterns of trait evolution becomes ever greater. However, the wide variety of methods in the literature and the limited communication between sub-disciplines in biology hampers method use and development. Broadly-applicable methods developed in one sub-discipline go unused in other sub-disciplines; biologists may not test certain hypotheses due to a perceived lack of relevant methods; and those who would develop methods to meet current needs have difficulty finding which areas require such work. I intend to develop and implement methods to help understand trait evolution, first identifying and communicating areas of need.

This project has four main goals. 1) Describing extant methods: A database of existing phylogenetic methods for examining trait evolution, from fields of genomics, paleontology, and comparative biology, will be creating using information from the literature. 2) Identifying underdeveloped areas: use database above. These areas will be communicated through a publication targeted to a general audience and through a website featuring the database and searchable by input data type, software availability, significance criterion, and other factors. This will allow empiricists to find appropriate methods quickly and permit theoreticians, especially those from fields outside biology, to identify areas still needing development. 3) Develop new methods in some of these areas. 4) Implement these new methods in existing software and integrate this software into *Mesquite* and CIPRES/Kepler. By the end of this project, empiricists will more easily identify relevant techniques and software to answer their questions, methods still needing development will be publicized, and a subset of the areas needing development will have methods created and implemented.

Introduction and Goals: Understanding how morphological, behavioral, and geographical traits have evolved has been a main goal of evolutionary biology since its inception. As biology has matured, traits of interest now include genetic, biochemical, and physiological traits. Methods using phylogenetic information have become powerful tools for addressing the evolution of such traits. Once an adequate method is developed and implemented, hypotheses previously untestable become tractable, spurring a flurry of research. The potential of phylogenetic methods for understanding trait evolution has created a large population of empirical biologists eager to use such methods and an active population of theoretical biologists seeking to build and implement new methods. As new questions arise, new methods need to be developed. However, the wave of interest in methods has actually made this difficult: information on the relevant literature and software is widely scattered and often segregated into different sub-disciplines, leading to empiricists choosing to use suboptimal methods or giving up on certain types of questions and to theoreticians failing to develop models where needed and occasionally duplicating the work of others. For the field to continue to progress, there is a need to 1)

categorize existing methods of examining trait evolution; 2) identify areas where new methods need to be developed; 3) develop methods in these targeted areas; and 4) implement these methods in user-friendly, open-source, cross-platform, fast software.

1) Categorize existing methods of examining trait evolution: Information about available methods is broadly scattered. This problem is made more acute due to lack of communication between sub-disciplines. Paleontologists, comparative biologists, and researchers in the various new "omics" sub-disciplines (genomics, metabolomics, etc.) often do not draw on the same methods. For example, Felsenstein's independent contrasts approach (Felsenstein, 1985) has formed the basis for many comparative studies, but is far less common in other areas of evolutionary biology. From January 2000 to November 2006, it was cited by 112 papers in the journal *Evolution* alone, just 9 times in two major paleobiology journals, and just 8 times in seven journals dealing largely with genomic studies (ISI Web of Knowledge). Methods are created redundantly and sometimes crudely. One such case was that of Jordan et al. (2005). In this *Nature* paper, the authors look for directional trends in amino acid substitutions by dividing a fifteen-taxon tree into potentially overlapping triplets and then infer direction of change by assuming the state possessed by the outgroup and one of the other two species is ancestral. However, there are at least two existing methods for doing this without needing to divide the tree: simple parsimony reconstruction of state changes, and a non-time-reversible likelihood method that allows for multiple changes (Pagel, 1994) implemented in the programs *Multistate* (Pagel) and *BayesTraits* (Meade & Pagel). The authors instead developed a novel method without any discussion of why this was needed or how this new method might be superior to existing methods. Had an existing likelihood method been used, the authors could have made many more inferences about relative transition rates, whether a model of directional change was a better fit to the data, and even what the ancestral amino acid frequency may have been. While this is an example of genomicists ignoring methods from phylogenetics, one could just as easily point to examples of phylogeneticists ignoring methods from paleontology, paleontologists ignoring methods from genomics (a method for understanding evolution of continuous characters on a tree may work equally well for molar height as for genome size), and so forth. A reference and an online database could help prevent this duplication of effort and missed opportunities.

2) Identify areas where new methods need to be developed: As both an empirical and theoretical biologist, I have been confronted by questions posed by my study system for which no tools are available and have then been forced to develop such tools. For example, there are methods for investigating correlations between states of two discrete characters (Maddison, 1990), rates of two discrete characters (Pagel, 1994), states of two continuous characters (Felsenstein, 1985), and, arguably, rates of two continuous characters (Garland, 1992). Attempts to correlate *state* of a discrete character (foraging period in my study organism — ants) with *state* of a continuous character (eye size) and *state changes* of a discrete character (changes in foraging period) with *rates* of a continuous character (rate of leg length evolution) in my study system have pointed to a need to develop such models (I have since developed and implemented these models). However, simply writing out the known methods, as above, would have directly suggested the need for such models, as well as additional models, such as one correlating rate of a discrete character with rate of a continuous character. Similarly, there are numerous methods that optimize the likelihood of a tree by applying one or more

stretching parameters to the entire tree — the optimized value of this parameter may provide information on whether changes occur early or late in evolution (Blomberg et al., 2003) or whether trait changes are associated with speciation events (Pagel, 1994). Comparing these methods with methods that allow different parameters on different parts of a tree suggests a way to further extend the methods: allow different stretching on different parts of a tree. New questions that could be answered could include, "Are trait changes more closely associated with speciation events when a lineage moves into an environment where sister species regularly contact one another?" or "Does temporal niche partitioning slow the rate at which habitat niches fill?" Identifying model similarities can also point to new models: Butler and King (2004) allow attraction strength and attraction mean values of an Ornstein-Uhlenbeck model to vary on a tree (but each branch may have only one value) but force the stochastic parameter to be constant; models of O'Meara et al. (2006) and Thomas et al. (2006) allow the stochastic parameter to vary between or even within branches but force the attraction parameter to be zero. This points to a general model, as yet unimplemented, that allows all three parameters to vary between and along branches.

The potential models outlined above suggest some of the benefits that could accrue from thorough comparison of existing models of character evolution; more such advances would come once even more models are compared in an organized manner.

3) Develop methods in targeted areas: Once a need is identified, a method to address this need must be developed. Fortunately, the pool of potential theoreticians has expanded from empirical biologists and a few dedicated theoreticians to also include researchers from fields of mathematics, computer science, and statistics. Researchers from these fields outside of biology have developed interesting new approaches in areas such as tree search (divide-and-conquer methods) and supertree construction; once clearly-delineated problems in trait evolution are described, they, along with traditional workers, may develop similarly promising solutions in these areas.

4) Implement new methods in user-friendly, open-source, cross-platform, fast software: For a software program to be useful to empirical biologists, it must take standard file formats, be well-documented, and be available on major platforms. As methods become more complex, and especially as integration of results across potentially thousands of trees (e.g., (Huelsenbeck and Rannala, 2003; Pagel and Meade, 2006)) becomes more popular, software speed also becomes an issue. It is also important for scientific software to be open-source. This allows inspection of code for bugs, other scientists to build on the work of others in the case of permissive licensing, and long-term maintenance and porting of software.

Proposed Activities: This project will have two components. One will be a synthetic literature and software review of available methods from paleontology, phylogenetics, genomics, and related disciplines that seek to understand character evolution using phylogenies, thus addressing Goals 1 and 2. These methods will be characterized for required data (character types, topologies or trees with branch lengths, measurement of intrataxon variation or single values per taxon, etc.), input format (Nexus, etc.), method type (nonparametric, likelihood, Bayesian, etc.), means for evaluating significance/meaningfulness of results (likelihood ratio tests, information theoretic methods, Bayes factors, etc.), software platform, software license (GNU Public License,

closed source, etc.), software distribution (by request only, free download, etc.), and frequency of use (measured by citations, weighted by year). This will allow users to identify the software they need to answer questions, especially approaches from other fields. Theoreticians and software developers will be able to identify what methods still need to be created and what methods still need implementation. This information will be compiled in an online MySQL database and also described in a publication. By assembling this information across sub-disciplines, empirical workers with different backgrounds may find an approach in another sub-discipline that will help solve a problem. Theoreticians will be able to identify missing tools in the empiricist's methodological toolbox and work to fix this.

Second, I will use information from this analysis to develop and implement new methods in a subset of the areas where needs are identified (Goals 3 and 4). This will be based on my existing C++, cross-platform, open source software *Brownie*, which uses open-source code from Rod Page, Paul Lewis, and the GNU Scientific Library. While the released version of this program can test for different rates of evolution on different parts of a tree, an unreleased version can now also test for change of the rate of a continuous character based on the state or the change of a discrete character, test for an Ornstein-Uhlenbeck model, estimate the ACDC parameter of Blomberg et al (2003), and incorporate measurement uncertainty in terminal taxa. New models to be developed and implemented include applying tree-stretching models (for example, models that test for association of speciation events with character change) to user-specified, character-mapped, or time-sliced branches (i.e., by a user specifying clades to examine, or by using reconstructed character states to assign models to branches, or by assigning different models to branches before and after a time event).

Centralization of tools into a small set of broadly-used and user-friendly applications will reduce learning time for empirical biologists (who will have to learn how to use just one or a handful of interfaces) and reduce development time for theoreticians (who will only have to build the core of the method and rely on a larger program for input and output). *Brownie* already has some of these advantages, as Paul Lewis' Nexus Class Library allows the interface to be similar to that of popular Nexus-input programs *PAUP* and *MrBayes*. The NSF-funded CIPRES project (*http://www.phylo.org*) is developing libraries that will enable even more centralization. Once these libraries are released (scheduled to be released in 2006), the general phylogenetic program *Mesquite* as well as the CIPRES/Kepler workflow tool will be able to call external programs. This means that *Mesquite*, which is user-friendly and cross platform but written in the relatively slow-running language Java (Vivanco and Pizzi, 2005) will be able to call the fast routines for trait analysis in *Brownie* without the user needing to make even the minimal effort to learn the *Brownie* interface.

Rationale for NESCent support: Many methodological advances are prompted by empirical biologists' need to understand their study systems. NESCent, with its assortment of postdocs and visiting faculty, plus its position in the Research Triangle, will provide a variety of biologists attempting to answer important questions who will know their methodological needs. These biologists will be of particular help when addressing Goals 1 and 2 of this proposal. NESCent's support for open source software development, goal of creating links across sub-disciplines in evolutionary biology, and

independence of postdocs creates an effective environment in which to conduct this work.

Proposed Timetable

Sept. 2007: Begin literature search, compiling database of methods and software. In parallel, continue implementation of available methods in *Brownie* software and incorporating this software into *Mesquite* and the CIPRES/Kepler workflow.

January 2008: Release a version of the software containing new methods.

May 2008: Finish database. Write manuscript and web interface for database.

June 2008: Submit review article manuscript. Before or with publication of the manuscript, make database web site live.

June 2008-Aug. 2009: Develop and implement new models for investigating character evolution; continue periodic software releases and website updates.

Anticipated Results

Papers: There will be one review paper examining available methods of evaluating character evolution from phylogenetics, paleontology, and genomics. The paper will be targeted for a journal with a readership from all the above fields and will highlight methods used primarily in one sub-discipline but applicable to others and areas where methods have yet to be developed at all. There will be 2-5 additional papers introducing new methods of investigating character evolution.

Software: An existing software program (*Brownie*) will be greatly extended. At least every eight months there will be a new release, compiled for Mac OS X and Windows and with source and a makefile for other systems. By the end of the first year, this software will be connected as a service for *Mesquite* and the CIPRES-Kepler project.

Web: The generated methods database will be posted online on a website; the database will be searchable based on multiple criteria and will be downloadable.

References

Blomberg S.P., Garland T., and Ives A.R. (2003). Evolution. *57(4)*, 717-745.

Butler M.A., and King A.A. (2004). American Naturalist. *164(6)*, 683-695.

Felsenstein J. (1985). American Naturalist. *125(1)*, 1-15.

Garland T. (1992). American Naturalist. *140(3)*, 509-519.

Huelsenbeck J.P., and Rannala B. (2003). Evolution. *57(6)*, 1237-1247.

Jordan I.K., Kondrashov F.A., Adzhubei I.A., Wolf Y.I., Koonin E.V., Kondrashov A.S., and Sunyaev S. (2005). Nature. *433(7026)*, 633-638.

Maddison W.P. (1990). Evolution. *44(3)*, 539-557.

O'Meara B.C., Ane C., Sanderson M.J., and Wainwright P.C. (2006). Evolution. *60(5)*, 922-933.

Pagel M. (1994). Proc Roy. Soc London B - Biological Sciences. *255(1342)*, 37-45.

Pagel M., and Meade A. (2006). American Naturalist *167(6),* 808-825.

Thomas G.H., Freckleton R.P., and Szekely T. (2006). Proc Roy. Soc London B - Biological Sciences. *273(1594)*, 1619-1624.

Vivanco R.A., and Pizzi N.J. (2005). Software-Practice & Experience. *35(3)*, 237-254.