Plans for the Phylogenetic Comparative Methods Benchmark Database (PhyCoMB)

August 30, 2017

CONTENTS

1 INTRODUCTION

This document is intended to be a high-level description of the functionality and structure of what PhyCoMB will eventually look like. Ultra-briefly: The goal is to allow users to compare the performance of different phylogenetic methods. There will be a web interface that allows for browsing those performance results, and for contributing new results. Behind that will be a database of the testing datasets, methods, and results.

2 PROPOSAL CONTENT

Here is what was originally proposed to NSF/BSF.

PhyCoMB will challenge new methods in a standardized manner, revealing their strengths and weaknesses early in their lifecycle. It will focus on testing if discrete traits affect rates of speciation and extinction, but the work will also support questions of trait evolution and lineage diversification separately.

2.1 Community resource for benchmark tests of model performance

A paper introducing a new phylogenetic comparative method typically includes essential simulations that examine its power and bias and reveal the parameter space in which it will be most beneficial. These simulations typically follow the assumptions of the underlying model. Poor behavior, however, may arise when those assumptions are not met. Because all empirical datasets have been shaped by processes outside the assumptions of any model, it seems self-evident that methods should routinely be tested in many situations beyond their specific focus. Why is this not standard practice for comparative methods developers? It is time-consuming or impossible for a single developer to craft diverse testing datasets that encompass the biological phenomena likely to 'break' his/her new method. Furthermore, there is not a culture—during development, peer review, and empirical application—of valuing and hence requiring robustness testing in this field. Thus, phylogenetic comparative methods papers typically discuss possible artifacts but fall short of providing concrete guidance about whether a method can be reliably applied to data at hand.

We aim to break these barriers to the routine, rigorous testing of new phylogenetic comparative methods by making robustness testing more straightforward. We will develop a suite of tests that can easily be deployed to assess the performance of a new method and compare its behavior against other methods designed for the same questions. The product will be called **PhyCoMB: Phylogenetic Comparative Methods Benchmarking** (Fig. 1). Benchmark tests

are standard practice in many fields, from computer hardware to bioinformatics. For example, BAliBASE (the Benchmark Alignment dataBASE; Thompson et al. 1999, 2005) has been widely adopted for assessing the performance of multiple sequence alignment algorithms. Further examples of standardized benchmark tests are Assemblethon and BUSCO for genome assembly (Bradnam et al. 2013; Simão et al. 2015) and T. Warnow's resources for phylogeny estimation (*http://www.cs.utexas.edu/~phylo/datasets*). PhyCoMB's structure will mirror other benchmark suites:

◦ *Tasks* are specific questions within the domain of phylogenetic comparative methods. They include tests of whether evolving traits affect rates of speciation and extinction (our focal task here), clade-specific diversification rate shifts, irreversible evolution, and discrete trait correlations.

◦ *Methods* are procedures designed to accomplish a task. They consist of a model or other technique (e.g, BiSSE, sister clades), plus a statistical inference framework (e.g., AIC, model averaging, sign test).

◦ *Elements* are collections of trees and optionally traits, all with the same properties. They may arise from empirical or simulated data. A method is applied to each tree/trait item within an element.

◦ *Reference sets* are chosen to present particular challenges to a method, within the context of a task. Each is a group of elements. For example, the 'power' reference set includes elements with small and large trees. The 'pseudoreplication' reference set contains trees in which traits change rarely and are accidentally associated with diversification shifts. The 'diversification heterogeneity' reference set contains trees with complex shifts in speciation, on which neutral traits are evolved.

◦ *Curated benchmarks* are the heart of PhyCoMB. The underlying database will eventually contain dozens of reference sets with many thousands of elements, but we will carefully select a subset to be the benchmark test for a task. For example, the benchmark for trait-dependent diversification might consist of 16 elements forming a progression of increasingly challenging power tests, 24 elements with different forms of pseudoreplication, and 36 elements with different forms of diversification heterogeneity and neutral trait evolution. Manual curation will maximize the diversity of challenges faced by a method while keeping the amount of testing manageable. Clearly-defined benchmarks will allow the performance of different methods to be easily compared by those developing new methods and those applying them to empirical systems.

◦ *Annotations* will make the results obtained from PhyCoMB easier to explore and interpret. Each element will be labelled with, e.g., the number of tips and sampling completeness, the type of diversification process for simulated trees, construction methods for empirical trees, and the model of evolution for simulated traits. Contributors will note the origin of the element's data (literature citation or generating script).

◦ *Reports* will allow developers to see how a new method performs relative to others, and they will allow users to understand the strengths and weaknesses of a variety of methods for a given task.

For an empiricist, PhyCoMB will provide not only structured information about methods, but also the means for directed testing. This includes scripts for creating test data that reflect properties of real data, and for running methods on them. It can thus enhance the impact of empirical work by aiding demonstrations of the robustness of a method for the data and question at hand. Such improved connections between methods development and empirical use will bring greater stability and transparency to the field (Cooper et al. 2016).

The first life-stage of PhyCoMB will be initiated by this proposal. We will generate a collection of elements and benchmarks (soliciting ideas during the workshop; see Broader Impacts), and we will build a web-based interface for developers and end users. This first stage will demonstrate substantial benefits to the phylogenetics community: straightforward comparisons of old and new approaches, reproducibility of results, early warnings of methodological weaknesses, and highlights of when methods are particularly powerful and robust. The second stage that we envision for PhyCoMB is gradual community adoption beyond the timeframe of the proposed work. PhyCoMB will evolve through contributions of new elements, as future work uncovers complex scenarios that challenge the assumptions of our methods.

2.2 References

Beaulieu, J.M., B.C. O'Meara, 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Systematic Biology 65:583–601.

Bradnam, K.R., J.N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J.A. Chapman, G. Chapuis, R. Chikhi, et al., 2013. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2:1.

Cooper, N., G.H. Thomas, R.G. FitzJohn, 2016. Shedding light on the dark side of phylogenetic comparative methods. Methods in Ecology and Evolution 7:693–699.

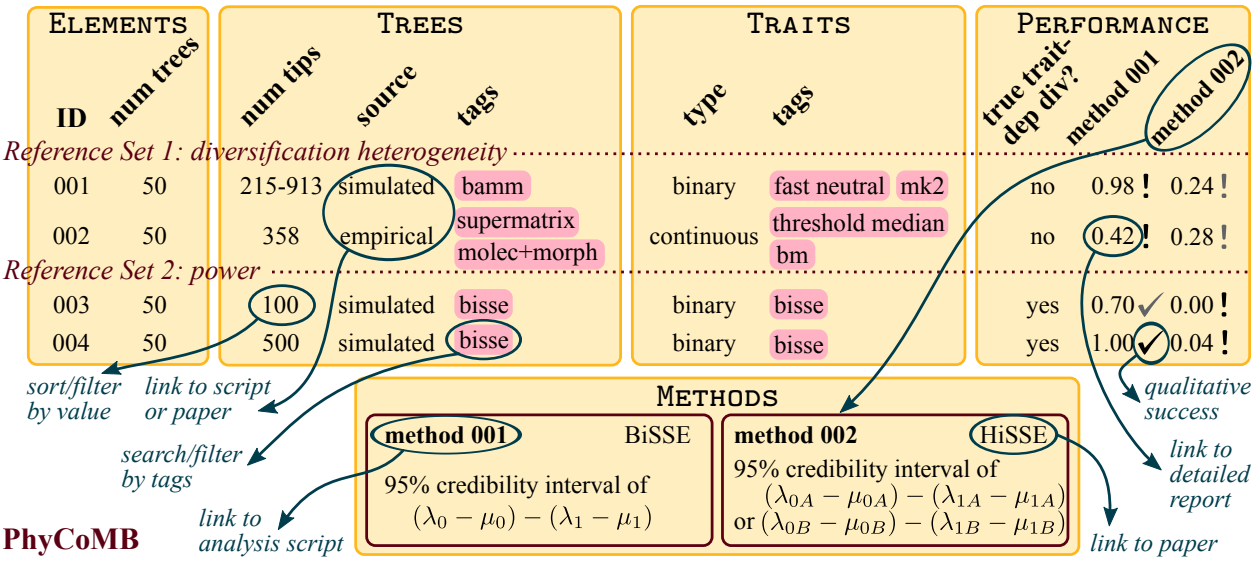

Figure 1: Vision for PhyCoMB. The Phylogenetic Comparative Methods Benchmark database will provide test datasets and reports of method performance. Example annotations are shown for four elements (comprising trees and traits), and interactive behavior of the web interface is noted. Compared are two methods, which test for trait-dependent diversification by assessing the difference in estimated net diversification rates. Performance is reported as the proportion of trees on which significant trait-dependent diversification is inferred. Reference set 1 reveals the BiSSE-based method is prone to incorrectly associate two kinds of neutral traits with diversification rate, while the HiSSE-based method is less so. Reference set 2 reveals that HiSSE has much lower power than BiSSE in this test. (Note that HiSSE is much more effective under model averaging; Beaulieu & O'Meara 2016.)

Simão, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212.

Thompson, J.D., P. Koehl, R. Ripp, O. Poch, 2005. BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark. Proteins: Structure, Function, and Bioinformatics 61:127–136.

Thompson, J.D., F. Plewniak, O. Poch, 1999. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15:87–88.

2.3 Later thoughts

This subsection wasn't part of the proposal. But it explains some ways in which the current plan differs from what was proposed.

Task needs to be defined a bit more carefully. For example, even within the realm of state-dependent diversification, one approach/question is hypothesis testing and another approach/question is parameter estimation. Results for those questions would be reported differently. And even a given method on a given element could perform better on one question than the other. Below, *Task* is assumed to include this refinement. But we could instead have a separate *Question* component that intersects with *Task* to specify the problem.

If we are going to include methods that co-infer the phylogeny along with answering the comparative methods question, we'd need to allow for input data to be a sequence alignment rather than a tree. That is a potentially promising direction for methods development, but it's beyond the immediate scope of PhyCoMB.