Title: Bayesian phylogenetic placement of fossil taxa from quantitative morphometric data
Authors: Caroline Parins-Fukuchi
Journal: Evolution?
Abstract: Stuff and things

Introduction: The role of fossil data in reconstructing phylogeny among living organisms has long been a central, yet contentious, topic in evolutionary biology. One view has historically suggested that imperfections in the preservation and identification of fossils should preclude their inclusion in phylogenetic inference, instead suggesting that fossils be placed into the stem lineages of trees representing extant species (Hennig 1965). Other researchers have argued that fossil information is fundamentally important when inferring evolutionary dynamics and relationships (Donoghue et al. 1989). Since this time, there has been significant interest in the simultaneous reconstruction of fossil and living organisms in a 'total evidence' framework. Approaches based upon probabilistic models of molecular and morphological character that incorporate fossil taxa into the reconstruction of phylogenetic relationships and divergence times have increased understanding of evolutionary patterns across large clades, and provide compelling evidence in favor of incorporating fossils in phylogenetic analyses (Pyron et al. 2011, Ronquist et al. 2012, Zhang et al. 2015).

A fundamental challenge when jointly estimating phylogeny between living and extinct organisms is the unavailability of molecular data in nearly all fossil taxa. As a result, there has been a need to explore the compatibility of molecular with morphological data to better understand the capability of fossil and extant species to reciprocally inform reconstruction of the other's evolutionary patterns. Previous work has sought to determine whether the inclusion of molecular data representing extant species can improve the reconstruction of relationships among fossils represented by morphology alone (Wiens 2009, 2010). The result of these studies suggest that the inclusion of morphological characters comprising living and fossil species does not have a tendency to decrease the accuracy of phylogenetic reconstructions, and can improve estimation of fossil placements in well-behaved datasets. Expanding upon these observations, Berger and Stamatakis (2010) have shown that methods placing fossils on fixed molecular phylogenies can yield accurate results by filtering through conflicting signal. This demonstrations that placing incomplete fossil data in a molecular context scaffolded by extant relationships can improve reconstruction among fossils.

Researchers' enthusiasm for reconstructing a comprehensive tree of life has encouraged the integration of fossils with living taxa in phylogenetic analyses. Improving integration between fossil and living taxa has the capability to benefit both paleo- and neontological studies. In addition to the results of Berger and Stamatakis discussed above, the inclusion of fossils improves the reconstruction of ancestral states using phylogenetic comparative methods (Slater et al. 2012). As another example, increasing the rigor with which fossils are placed on phylogenies is expected to improve the accuracy and treatment of uncertainty in divergence time estimation (Guindon 2018).

Despite the abundance of both clear and subtle benefits to improving the integration of fossil and living taxa in phylogenetics, challenges have arisen from conflicting and noisy information presented by morphological data. The fragmentary sampling of fossil data further exacerbates this problem, and can lead to erratic results and high uncertainty in posterior estimates (Ronquist et al. 2016). Another issue that is noted by Berger and Stamatakis (cited above) stems from the reality that morphological alignments commonly contain very few sites, often 50-500, compared to molecular datasets, which can contain hundreds of thousands of sites. This can cause the likelihood of molecular partitions to dwarf those of morphological partitions, limiting the influence of morphology in reconstructions of topology and branch lengths. For these reasons, Berger and Stamatakis advocated fixing the relationships of extant taxa *a priori* using molecular reconstructions, and using the resulting scaffold to identify conflicting signal in morphological data.

Compounding upon challenges associated with their combination with molecular data, morphological data themselves often exhibit major imperfections. These occur at multiple levels important to phylogenetic analysis. At one level, morphological data are frequently susceptible to displaying biased or misleading signal. This may often stem in part from the general practice of assigning discrete character states to taxa through qualitative assessment. The subjective nature of this process can cause major irreconcilable disagreement between results achieved from different researchers (cite examples?). As an added layer of potential bias, these matrices are also frequently filtered to include only characters that researchers consider before the fact to be accurately informative in phylogenetic reconstruction. At another level, the discrete character matrices most commonly employed in phylogenentics can often be difficult to adequately model. At present, researchers employing probabilistic methods generally use the so-called ‘Mk’ model (Lewis 2001). This is a generalization of the Jukes-Cantor model of nucleotide substitution that accommodates *k* possible character states. Although previous work based upon simulated data has suggested that Mk-based approaches outperform parsimony (Wright and Hillis 2014, Puttick et al. 2017), the extent and conditions under which this is the case in empirical datasets is unclear (Goloboff et al. 2017). Emprical datasets are also likely to depart signficantly from the assumptions of the Mk model. It is unclear how sensitive analyses are to these violations.

For all of these reasons, continuous traits have been suggested as a feasible alternative (Felsenstein 1988, MacLeod 2001, Parins-Fukuchi 2017). Tools that quantify morphological size and shape have the capacity to alleviate many of the concerns relating to bias and subjectivity that occur with discrete characters. Approaches such as geometric morphometrics offer the potential to wholistically incorporate all dimensions of shape to inform phylogeny. The continuous state space of morphometric data might also increase the amount of information that can be extracted from morphological datasets, which may be beneficial when analyzing poorly-sampled fossil data.

Traditional linear morphometric measurements have long been employed in morphological phylogenetics, but are typically discretized to more easily analyze them alongside present-absence data. However, these transformations may decrease the amount of information in continuous datasets by binning fine-scaled variation into shared discrete categories, and are susceptible to the difficulties in modelling under the Mk model described above. Geometric morphometric data have shown utility in several previous phylogenetic studies using parsimony-based methods (González-José et al. 2008, Catalano and Goloboff 2010, Smith and Hendricks 2013), however, they have not gained substantial traction. This may be in part due to the lack of available tools to analyze continuous trait data in a probabilistic framework.

The earliest studies investigating probabilistic methods of phylogenetic inference were developed using continuous characters modelled under Brownian Motion (BM) (Cavalli-Sforza and Edwards 1967, Felsenstein 1973). Due in part to the abundant discrete character data that became available with the emergence of DNA sequencing, these approaches were quickly overshadowed in popularity by discrete trait approaches based upon Markov nucleotide substitution models. Continuous trait models have since gained significant popularity in phylogenetic comparative methods, but still are rarely used for phylogenetic inference. As a result, few implementations exist, with only ContML in the PHYLIP package and RevBayes providing such functionality (cite phylip and revbayes). The approaches used in these packages are also fairly minimalistic, with no real tailoring to the challenges that might be presented by empirical datasets.

In this paper, I describe a new set of approaches that place fossils on molecular trees using quantitative characters modelled under BM. These methods seek to tackle some of the most pressing obstacles associated with the use of traditional and geometric morphometric data in phylogenetic inference. Using simulated data, I validate and explore the behavior of the implementation. I also analyze empirical datasets representing the Vitaceae family of flowering plants and carnivoran mammals (Jones et al. 2015) comprised of traditional and geometric morphometric measurements, respectively. These methods use Markov Chain Monte Carlo (MCMC) to infer the evolutionary placements of fossils and branch lengths. Although MCMC is generally associted with Bayesian inference, my implementation can perform inference both with and without the use of priors. It is thus philosophically agnostic, and offers exploration of a diverse range of options to best accommodate diverse morphometric datasets. These approaches are implemeted in the *cophymaru* package.

Methods and Materials:

Brownian motion model

The approaches that I describe in this paper all rely upon the familiar BM model of evolution. Under BM, traits are assumed to be multivariate distributed, with variances between taxa defined by the product of their evolutionary distance measured in absolute time and the instantaneous rate parameter (σ^2):

$$dX(t) = \sigma dB(t);$$

(Eqn. 1)

where *dX(t)* is the time derivative of the change in trait *X* and *dB(t)* corresponding to normally distributed random variables with mean 0 and variance *dt*. This leads to the expectation that over time *t*,

$Expect(X<sub>t</sub>) = X<sub>0</sub> $, Eqn. 2

with

$Var(X<sub>t</sub>) = $sigma;*t*. $ Eqn. 3

The methods that I describe use a slightly different parameterization and likelihood calculation than most conventional implementations used in modern phylogenetic comparative methods (PCMs). These generally construct a variance-covariance (VCV) matrix from a dated, ultrametric phylogeny to calculate the likelihood of the data, assuming a multivariate normal distribution (see Felsenstein 1973 or O’Meara 2004 for a detailed explanation). Since these methods treat the topology and branching times as known, the goal is typically to obtain the maximum likelihood estimate (MLE) of the rate parameter (σ^2) to examine evolutionary rate across clades.

One drawback to the use of this version of the phylogenetic BM model in the reconstruction of topology is its requirement that phylogenies be scaled to absolute time. Although it is possible to simultaneously estimate divergence times and topology while analyzing continuous traits, this can cause additional error and requires the specification of a tree prior that can accommodate non-ultrametric trees that include fossils. This requirement would also cause circularity in cases where researchers are interested in obtaining estimates and error in fossil placements in order to more rigorously inform molecular clock calibrations. To overcome the need for simultaneously estimating divergence times and fossil placements, I estimate the product $\sigma^2 t$ together. As a result, rate and absolute time become confounded. Branch lengths, which reflect the morphjological disparity between taxa, are thus measured in units of morphological standard deviations per site. This interpretation could be thought roughly of as a continuous analogue to the branch lengths obtained from discrete substitution models. Similarly to the discrete case, long branch lengths could reflect either a rapid rate of evolution or a long period of divergence (in absolute time) along that lineage.

Calculation of the likelihood:

Rather than use the computationally expensive VCV likelihood calculation, I use the reduced maximum likelihood (REML) calculation described by Felsenstein (1973). This calculates the likelihood on the phylogenetic independent contrasts (PICs) using a ‘pruning’ algorithm.

Markov chain Monte Carlo:

This method uses a Metropolis-Hastings (MH) MCMC algorithm to simulate the posterior or confidence distribution of fossil insertion points along a fixed reference tree and branch lengths. Rearrangements of the topological positions of fossil taxa are performed by randomly pruning and reinserting a fossil taxon to generate a proposal. This is a specific case of the standard subtree pruning and regrafting (SPR) move for unrooted tees that yields the MH proposal ratio:

EQN here

Branch lengths are updated both individually and by randomly applying a multiplier to subclades of the tree. This uses a proposal that constrains branch lengths > 0, and uses the MH proposal ratio:

EQN here.

Branch length priors:

Since the estimation of branch lengths from continuous traits is relatively uncharted territory in phylogenetics, I implemented and tested three different branch length priors derived from the molecular canon: 1) flat (uniform), 2) exponential, and 3) a compound dirichlet prior after Rannala et al. (2011). The compound dirichlet prior also offers an emprical Bayes option that uses an initial ML estimate of the branch lengths to specify the parameter corresponding to mean tree length.

Generating a rough ML starting tree:

To generate an initial estimate of fossil placements and branch lengths, I estimate an ML starting tree. Initial placements are achieved using stepwise addition. Each fossil is individually inserted along all existing branches of the tree, with the insertion point that yields the highest likelihood retained. At each step, MLEs of the branch lengths are computed using the iterative procedure introduced by Felsensten (1981). In this procedure, the tree is rerooted along each node. PICs are calculated to each of the three subtending edges, and then the MLE of each edge is computed analytically using the expressions:

vhat1 =

vhat2 =

vhat3 =

This process is iterated until the branch lengths and likelihoods converge. One the optimal placement of all of the fossils has been identified, the branch lengths are recalulated and can be used to inform the empirical Dirichlet parior described above. One problem with this approach is that it has a strong propensity to becoming trapped in local optima. As a result, it should be interpreted and deployed cautiously in especially messy datasets.

Filtering for concordant sites:

One major hurdle involved in the use of morphological data is its frequent tendency toward displaying noisy and discordant signal. This problem might be expected to manifest even more intrusively in morphometric datasets than in discrete datasets, since traits are much less likely to be excluded *a priori* on the basis of perceived unreliability. As a result, there is a need to filter through noisy signal to favor more reliable sites. I developed a procedure adapted from Berger and Stamatakis (2010) for this purpose. This computes a set of weights based upon the concordance of each site with the reference tree. In this procedure, the likelihood of each site is calculated on the reference tree (excluding fossil taxa). Next, the likelihood of each site is calculated along 100 randomly generated phylogenies. If the likelihood of the site is higher along the reference tree than the current random tree, the weight of the site is incremented by one. This yields a weight vector that is the same length as the character matrix, with each site possessing a weight between 0 and 100. The sites are then weighted using one of three schemes: 1) whole integer values, where the weight is retained as a whole number between 0 and 100, 2) a floating point value between 0 and 1, where the value generated from the random comparison is divided by 100, and 3) a binary value where the weight is equal to 1 if the site displayed a higher likelihood in the reference tree than 95 or more of the random trees, and 0 if less than 95.

In application, I found that the integer weighting scheme caused poor MCMC mixing, and so the floating and binary schemes are probably most practical in most cases. Since it filters out discordant sites completely, the binary scheme is enforces a harsher penalty than the floating and integer schemes, and so might be of greatest use in particularly noisy datasets.

Simulations:

To explore the behavior of these approaches under different settings and to validate the implementation, I performed a set of simulations to ascertain accuracy and reliability. From a single simulated tree, I pruned five “fossil” taxa and estimated their positions along the tree using 100 datasets of 50 characters simulated under BM. To compare the floating and binary weighting schemes, I also generated alignments consisting of 50 “clean” traits simulated along the true tree, and combined with partitions of “dirty” traits in intervals of 10, 25, and 50 generated from random trees.

References:

Catalano, S. A., P. A. Goloboff, and N. P. Giannini. 2010. Phylogenetic morphometrics (I): the use of landmark data in a phylogenetic framework. Cladistics 26:539–549. Blackwell Publishing Ltd.

Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic Analysis: Models and Estimation Procedures. Evolution 21:550.

González-José, R., I. Escapa, W. A. Neves, R. Cúneo, and H. M. Pucciarelli. 2008. Cladistic analysis of continuous modularized traits provides phylogenetic signals in Homo evolution. Nature 453:775–778. Nature Publishing Group.

MacLeod, N. 2015. Use of landmark and outline morphometrics to investigate thecal form variation in crushed gogiid echinoderms. Palaeoworld 24:408–429.

Smith, U. E., and J. R. Hendricks. 2013. Geometric morphometric character suites as phylogenetic data: extracting phylogenetic signal from gastropod shells. Syst. Biol. 62:366–385.