Ecological Monographs, 89(2), 2019, e01353
© 2019 by the Ecological Society of America

Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data

Alex D. Washburne,[1,7] Justin D. Silverman,[2,3] James T. Morton,[4,5] Daniel J. Becker,[1] Daniel Crowley,[1] Sayan Mukherjee,[3,6] Lawrence A. David,[3] and Raina K. Plowright[1]

[1]*Department of Microbiology and Immunology, Montana State University, Bozeman, Montana 59717 USA*
[2]*Program for Computational Biology and Bioinformatics, Duke University, Durham, North Carolina 27708 USA*
[3]*Center for Genomic and Computational Biology, Duke University, Durham, North Carolina 27708 USA*
[4]*Department of Computer Science, University of California San Diego, La Jolla, California 92037 USA*
[5]*Department of Pediatrics, University of California San Diego, La Jolla, California 92037 USA*
[6]*Department of Statistical Science, Mathematics, and Computer Science, Duke University, Durham, North Carolina 27708 USA*

Citation: A. D. Washburne, J. D. Silverman, J. T. Morton, D. J. Becker, D. Crowley, S. Mukherjee, L. A. David, and R. K. Plowright. 2019. Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data. Ecological Monographs 89(2):e01353. 10.1002/ecm.1353

Abstract. The problem of pattern and scale is a central challenge in ecology. In community ecology, an important scale is that at which we aggregate species to define our units of study, such as aggregation of "nitrogen fixing trees" to understand patterns in carbon sequestration. With the emergence of massive community ecological data sets, there is a need to objectively identify the scales for aggregating species to capture well-defined patterns in community ecological data. The phylogeny is a scaffold for identifying scales of species-aggregation associated with macroscopic patterns. Phylofactorization was developed to identify phylogenetic scales underlying patterns in relative abundance data, but many ecological data, such as presence-absences and counts, are not relative abundances yet may still have phylogenetic scales capturing patterns of interest. Here, we broaden phylofactorization to a graph-partitioning algorithm identifying phylogenetic scales in community ecological data. As a graph-partitioning algorithm, phylofactorization connects many tools from data analysis to phylogenetically informed analyses of community ecological data. Two-sample tests identify five phylogenetic factors of mammalian body mass which arose during the K-Pg extinction event, consistent with other analyses of mammalian body mass evolution. Projection of data onto coordinates connecting the phylogeny and graph-partitioning algorithm yield a phylogenetic principal components analysis which refines our understanding of the major sources of variation in the human gut microbiome. These same coordinates allow generalized additive modeling of microbes in Central Park soils, confirming that a large clade of Acidobacteria thrive in neutral soils. The graph-partitioning algorithm extends to generalized linear and additive modeling of exponential family random variables by phylogenetically constrained reduced-rank regression or stepwise factor contrasts. All of these tools can be implemented with the R package phylofactor.

Key words: community ecology; dimensionality reduction; graph partitioning; microbiome; phylofactorization; phylogeny.

Introduction

The problem of pattern and scale is a central problem in ecology (Levin 1992). Ecological patterns of observable features across communities, such as regular differences in carbon sequestration, species abundance distributions, epidemics, ecosystem services, and more, are often the result of processes that operate at multiple scales. The common "scales" of interest in ecology are space, time, and levels of ecological organization ranging from individuals to populations to ecosystems.

Predicting patterns in spatial variation over different scales–millimeters, meters, or kilometers–requires incorporating different processes driving the patterns. The relevant processes determining patterns in abundance over the scale of meters may not be the most relevant processes determining patterns in abundance over the scale of kilometers. Similarly for time, predicting climatic and weather patterns over days, years, or millennia requires different data, processes, and models. Similarly for levels of ecological organization, predicting the collective behavior of a school of fish requires interfacing individual behavior with interaction networks of those individuals (Katz et al. 2011) and predicting the ability of a forest

Manuscript received 31 December 2017; revised 25 June 2018; accepted 10 October 2018. Corresponding Editor: Nathan J. B. Kraft.

[7] E-mail: alex.d.washburne@gmail.com

to act as a carbon sink requires interfacing abiotic features and competition between trees with different traits, such as nitrogen fixation (Farrior et al. 2013). Understanding emergent infectious diseases requires interfacing processes over scales ranging from animal population dynamics, reservoir epizootiology, and human epidemiology (Plowright et al. 2017). Ecological theory requires interfacing phenomena across scales believed to be important, and continually updating our beliefs about which scales are important to interface.

A scale of particular interest in community ecology is the scale at which we group organisms into units: species, functional ecological groups, guilds, and more. For a novel or unfamiliar pattern, such as a change in microbial community composition along environmental gradients, how can one objectively identify the appropriate scales for grouping species into units? In macroscopic systems, a researcher will typically use intuition derived from natural history knowledge to determine scales of interest, selecting functional ecological groups based on processes or traits previously demonstrated to be important. Models of how the natural history traits affect the pattern will be constructed, and the goodness of fit to the pattern of interest will be used as a metric for the successful identification of relevant ecological scales. However, for some patterns and communities, such as inflammation or fatty acid production associated with the human gut microbiome, there is limited natural history knowledge to draw on to assist the decision of the appropriate scales of interest. Even familiar communities can be more objectively analyzed and compared with the help of rules, algorithms, and laws to identify the dominant scales of community ecological units.

All communities exist as a hierarchical assemblage of entities, many of whose relationships and evolutionary history can be estimated and organized into a phylogeny. The estimated phylogeny contains edges along which mutations occur and new traits arise. When the phylogeny correctly captures the evolution of discrete, functional ecological traits underlying a pattern of interest, the phylogeny is a natural scaffold for simplification, aggregation, and scaling in ecological systems (Washburne et al. 2018). Patterns whose functional ecological traits are laterally transferred can still be simplified by constructing a phylogeny of the laterally transferred genes, such as using a phylogeny for beta-lactamases (Hall and Barlow 2004) to understand microbial responses to antibiotics.

Graham et al. (2018) develop the term “phylogenetic scale” to refer to the depth of the tree over which we aggregate information from a clade, but functional ecological traits often arise at different depths of the tree and thus many ecological phenomena are driven by traits not properly aggregated by mowing the phylogeny along a constant depth. Instead, there may be multiple phylogenetic scales, or grains, underlying an ecological pattern of interest, and such scales need to be partitioned from one another while avoiding the obvious nested dependence caused by clades within clades. For example, the patterns of vertebrate abundances on land and water are simplified by nested clades—Tetrapods, Cetaceans, Pinnipeds, etc.—and ancestors immediately before an affected clade, say the ancestors before Tetrapods, are prone to misclassification due to the nestedness of a clade with a strong effect. For more complicated community ecological data, such as breeding bird surveys or microbiome data sets, there is a need for general statistical methods to partition the phylogeny into the grains with significantly different associations with or contributions to ecological patterns of interest. Such a method can objectively identify the phylogenetic scales underlying an ecological pattern of interest and assist community ecological theory in both familiar and unfamiliar systems.

Phylofactorization (Washburne et al. 2017) was developed to identify the phylogenetic scales in compositional (relative abundance) data by iteratively constructing variables corresponding to edges in the phylogeny separating species with different patterns of abundance. The variables used to identify phylogenetic scales were a common transform from compositional data analysis (Aitchison 1982), referred to as the isometric log-ratio transform (Egozcue et al. 2003, Egozcue and Pawlowsky-Glahn 2005), which contrast the relative abundances of species separated by an edge in the phylogeny. A coordinate in an isometric log-ratio transform aggregates relative abundances within clades by a geometric mean and contrasts clades through log-ratios of the clades’ geometric mean relative abundances. The isometric log-ratio transform also allows the construction of non-overlapping contrasts, thereby reducing an obvious source of nested dependence in phylogenetic variables. The isometric log-ratio transform is used to identify phylogenetic scales, capture large blocks of variation in relative-abundance data and construct coordinates that correspond to edges along which hypothesized functional ecological traits arose.

However, many ecological data are not appropriately analyzed as compositions. For example, the presence/absence of bird species across continents are best modeled as Bernoulli random variables, not compositions. There is a need to generalize phylofactorization to identify phylogenetic scales in any data type. In this paper, we extend phylofactorization to broader classes of data types by generalizing the logic of phylofactorization to three operations: aggregation, contrast, and an objective function defined by the pattern of interest. The nested dependence of clades within clades is avoided by defining phylofactorization as a graph-partitioning algorithm that contrasts species separated by edges and iteratively partition the phylogeny along edges that best differentiate species by maximizing the objective function. After defining phylofactorization as a graph-partitioning algorithm, we illustrate the generality of the algorithm through several examples.

First, we show that two-sample tests, such as *t* tests and Fisher’s exact test, provide natural operations for phylofactorization. Two-sample tests aggregate data from two groups through means or proportions, contrast the aggregates via a difference of means or proportions, and have natural objective functions defined by their test statistics. We illustrate the use of two-sample tests by performing phylofactorization of a data set of mammalian body mass.

Then, we show how the phylogeny serves as a scaffold for changing variables in biological data through a contrast basis. The same basis used in the isometric log-ratio transform can be used to identify the phylogenetic scales providing low-rank, phylogenetically-interpretable factorizations of matrices. The contrast basis allows us to introduce a phylogenetic analog of principal components analysis, phylogenetic components analysis, which identifies the dominant, phylogenetic scales capturing variance in a data set. Phylogenetic components analysis of the American gut microbiome data set (McDonald et al. 2018) reveals that some of the dominant clades explaining variation in the American gut correspond to clades within Bacteroides and Firmicutes, thereby providing finer phylogenetic resolution of the taxonomic-based Bacteroides/Firmicutes ratios found to be associated with obesity (Turnbaugh et al. 2006), age (Mariat et al. 2009), and more. Another phylogenetic factor of variance in the American gut is a clade of Gammaproteobacteria strongly associated with inflammatory bowel disease (IBD), corroborating a recent study's use of phylofactorization to diagnose patients with IBD (Vázquez-Baeza et al. 2017).

The contrast basis can also be used for regression-based analyses if the data are assumed to be approximately normal or related to the normal distribution through a monotonic transformation such as a logarithm. We illustrate regression-phylofactorization through a generalized additive model analysis of how microbial abundances change across a range of pH, nitrogen, and carbon concentrations in soils. The resulting contrast basis and its fitted values from generalized additive modeling yield a low-rank representation of biological big data and translates to clear biological hypotheses aiming to identify the traits driving observed non-linear patterns of abundance across environmental gradients (Ramirez et al. 2014).

Data sets comprised of non-Gaussian, exponential family random variables can also be formally analyzed through regression-phylofactorization. We present and compare four algorithms using reduced-rank and shared-coefficient models for generalized regression-phylofactorization of exponential family data. We discuss the relation of the presented algorithms to the contrast basis and graph partitioning algorithm and we finish with a discussion of the challenges and opportunities for future development of phylofactorization.

All analyses and the R package phylofactor are available online; see Data Availability.

Conceptual Overview

We first motivate the need for phylofactorization and introduce the graph-partitioning algorithm built on contrasting species separated by edges. In the context of the graph-partitioning algorithm, we consider two examples. The first, simple example of phylofactorization is the use of two-sample tests as a measure of contrast. The second example is the use of the contrast basis, a linear change of variables which facilitates phylogenetically interpretable, low-rank approximations of data matrices, connecting phylofactorization to everything from principal components analysis to regression-based approximations of data matrices. We extend regression-based phylofactorizatzion to exponential family random variables via generalized linear models. Four algorithms that can embed phylofactorization in generalized linear models are presented and compared. Finally, we discuss how the regression-phylofactorization methods introduced above can be incorporated into spatially and temporally explicit data analyses. In an effort to promote honest development of phylofactorization as an inferential tool, we examine several statistical challenges of phylofactorization that we are aware of.

Why phylofactorization?

Which vertebrates live on land and which vertebrates live in the sea (Fig. 1A)? Most children have enough natural history knowledge to say "fish live in the sea," thus correctly identifying one of the most important phylogenetic factors of land/sea associations in vertebrates. The statement "fish live in the sea" can be mathematically formalized by noting that one edge in the vertebrate phylogeny separates sea-dwelling "fish" from predominantly land-dwelling "non-fish" (Fig. 1B). Partitioning the phylogeny along the edge basal to tetrapods separates vertebrates fairly well into groups with different land/sea associations. An algorithm identifying the edge basal to tetrapods using only land/sea associations would correctly identify the edge along which important, functional ecological traits arose: comparisons of fish/non-fish would reveal clear morphological and physiological adaptations to sea/land. There are a few more phylogenetic factors of land/sea associations in vertebrates. Controlling for the previously identified edge, one might be able to later identify the edges basal to Cetaceans, Pinnipeds, and other tetrapods that live in the sea (Fig. 1B). Using such an algorithm, a few edges can capture most of the variation in land/sea associations across thousands of vertebrate species.

Ancestral state reconstruction of habitat association is a well-known means of making inferences about trait differences arising along edges. However, some traits and ecological patterns of interest are more complicated and their ancestral state reconstruction dubious. For instance, how can we identify the phylogenetic scales of microbial community composition changes along a pH gradient, allowing possible nonlinear associations that could be detected through generalized additive modeling (Fig. 1C)? Answering such a question through ancestral state reconstruction requires conceiving and analyzing

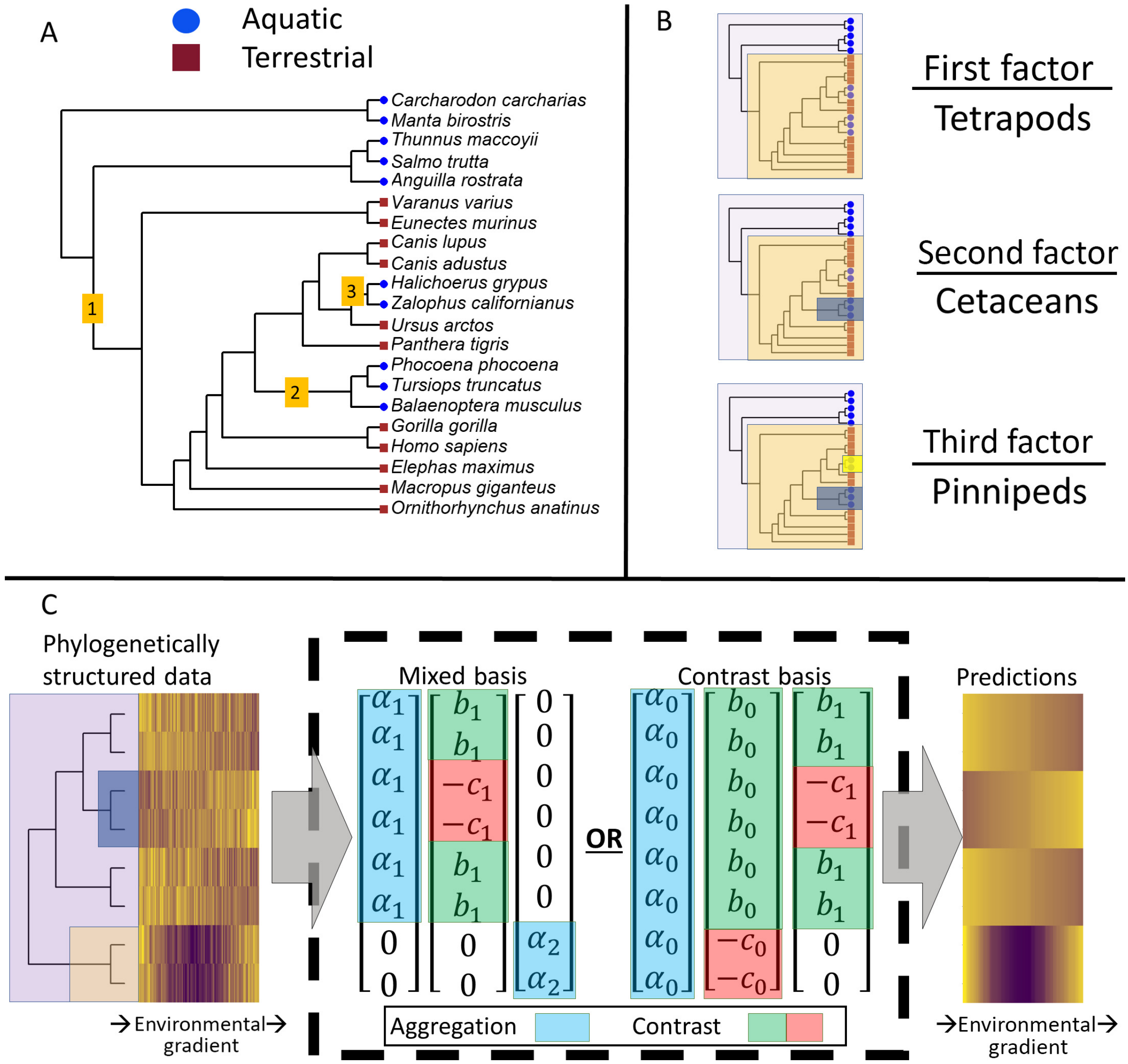


FIG. 1. Phylofactorization generalizes the logic of how to simplify phylogenetically structured data sets. (A) Vertebrate land/water associations can be simplified by partitioning the tree into the edges along which major traits arose. (B) The first phylogenetic factor of vertebrate land/water associations is the edge along which tetrapods arose, an edge along which lungs and limbs evolved that allowed colonization of land. Downstream factors can refine the original partitioning to identify the Cetaceans, Pinnipeds, and other aquatic tetrapods. (C) Phylogenetic factorization uses the operations of aggregation and contrast to generalize this same logic for phylogenetically structured data in which traits might not be known or their evolution easily modeled, including traits like a nonlinear relationship between abundance and an environmental gradient. Pure aggregations (blue) sum data within a clade, whereas contrasts (green/red) are differences between two clades. Low-rank, phylogenetically interpretable predictions of our data can be obtained through a mixed basis containing a series of aggregations and contrasts, or a "contrast basis" containing a global aggregate partitioned with subsequent contrasts.

an evolutionary model of how generalized additive models evolve along a tree.

Phylofactorization is a graph-partitioning algorithm, generalizing the phylogenetic logic used above to simplify land/sea associations by iteratively identifying edges in the phylogeny along which meaningful differences arise. With data-driven definitions of "meaningful differences" between groups of species, phylofactorization can identify phylogenetic scales underlying more complicated ecological patterns, patterns for which ancestral state reconstruction would be dubious.

Graph-Partitioning Algorithm

Phylofactorization requires a phylogeny spanning the set of species considered in the data. All phylogenies are rooted or unrooted graphs with no cycles, containing and connecting the units of interest in our data (the units

Box 1. Table of mathematical notation.

Terms	Description
$A(.)$	Aggregation operator
$C(.,\ .)$	Contrast operator
$\mathcal{F}(\theta)$	Distribution parameterized by θ
F_e	F statistic for edge e
K_t	Number of edges considered in iteration t of phylofactorization
N	Size of a binomial random variable
Q	A group $Q = R \cup S$ aggregated at a current or previous iteration
$R,\ S$	Two groups contrasted containing r and s species, respectively
P	Partitioning variables for phylofactorization
$\mathcal{T}$	Phylogenetic tree
$\mathbf{B}$	$m \times p$ coefficient matrix
$\mathbf{W}$	Matrix of component scores corresponding to $\mathbf{V}$
$\mathbf{V}$	m matrix of contrast basis elements
$\mathbf{X}$	$m \times n$ data matrix used for phylofactorization
$\mathbf{Y}$	$K \times n$ matrix of component scores, one for each edge considered
$\mathbf{Z}$	$n \times p$ matrix of meta-data used in regression-phylofactorization
a	Coefficient in aggregation vector
$b,\ c$	Coefficients in a contrast vector
e_k	Edge k
$e*$	Winning edge
e_t^*	Winning edge at iteration t
$f(.)$	Transformation in generalized f mean
$i,\ j,\ k,\ l$	Indexes. Often, i is the index for species and j for samples
m	Number of species
n	Number of samples
p	Number of meta data types for each sample
`phylo`	Categorical variable indicating which side of an edge a species is found
q	Number of pure aggregates in a basis for $\mathbb{R}^{>}$
$r,\ s$	Numbers of species in groups $R,\ S$ respectively
$s(.)$	Smoothing spline notation for term in generalized additive model
t	Iteration of phylofactorization
$x_{i,j}$	The $i,\ j$th element of data matrix $\mathbf{X}$
$x_{R,j}$	Aggregate, $A(\mathbf{x}_j)$ of group R for sample j, if j is missing then sample is arbitrary
$x_{S,j}$	See $x_{R,j}$
x_i	A random variable (assumed to be a single species i for arbitrary sample)
$[x]_{i,j}$	$i,\ j$th entry of data matrix, $\mathbf{X}$
z_i	Column of meta data matrix, $\mathbf{Z}$
$\mathbf{v}_{Q,i}$	ith element of aggregation basis element for set Q
$\mathbf{v}_{C_{R\|S}}$	Contrast vector splitting groups R and S
$\mathbf{v}_{C_e}$	Contrast vector for edge e (which splits sub-tree into two disjoint groups)
$\mathbf{x}_{R,j}$	r vector containing only the species in group R for sample j
$\mathbf{x}_{S,j}$	See $\mathbf{x}_{R,j}$
$\mathbf{x}$	m vector of species' data for an arbitrary sample
$\bar{\mathbf{x}}$	Sample mean of vector $\mathbf{x}$
$\mathbf{y}_e$	n vector of component scores for edge e
$\mathbf{z}_k$	Vector of meta data of type k
β_i	Coefficients for linear model
η	Natural parameter for exponential-family random variable
κ	Scale parameter for Gamma distribution
π	Number of failures parameter for Negative Binomial distribution
ρ	Probability of success for Bernoulli, Binomial, Negative Binomial distributions
σ	Standard deviation for Gaussian random variable
θ	Arbitrary parameters for probability distribution
–	

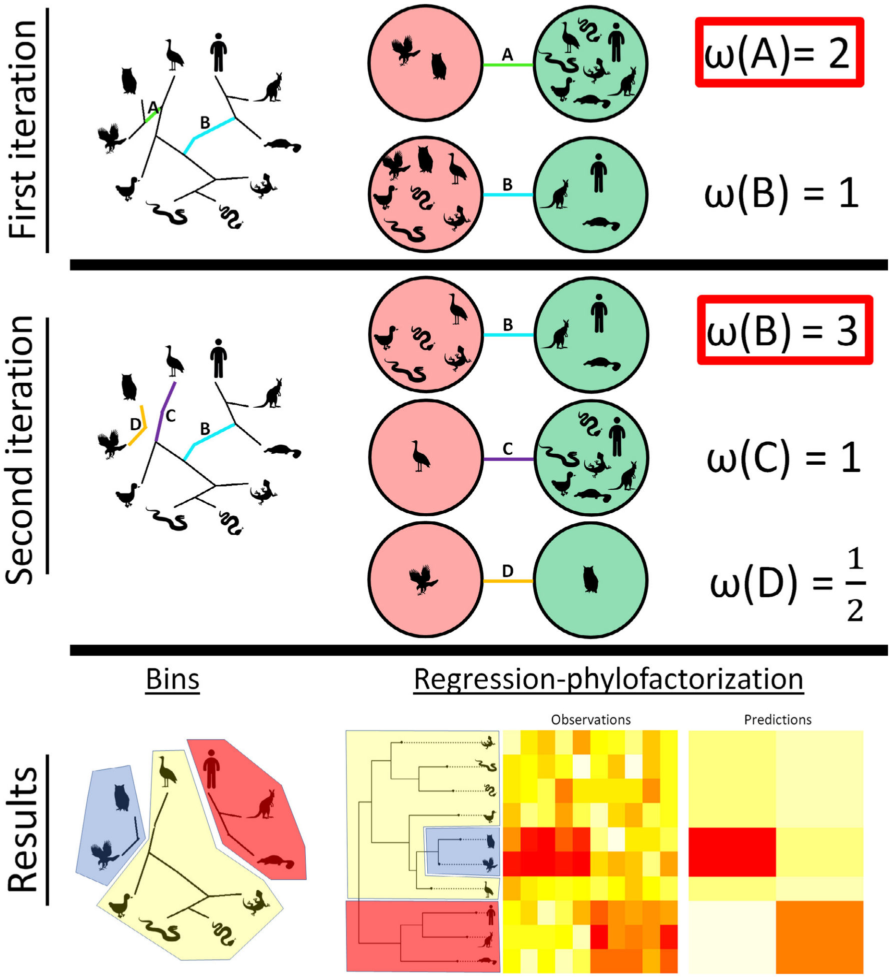


FIG. 2. Phylofactorization is a graph-partitioning algorithm. An objective function, ω, of a contrast of species separated by an edge allows one to iteratively partition the phylogeny along edges maximizing the objective function (first iteration). After partitioning the phylogeny, the objective functions are recomputed to contrast species in the same sub-tree separated by an edge. Edge B in the first iteration contrasted mammals from non-mammals, but in the second iteration, it contrasts mammals from non-mammals, excluding raptors (partitioned in the first iteration). The result of k iterations of phylofactorization is a set of $k + 1$ bins of species. Regression-phylofactorization defines an objective function through regression. Regression-phylofactorization can identify clades with similar patterns of association with environmental meta data and obtain low-rank, phylogenetically interpretable representations of a data matrix.

can be species, genes, or other evolving units of interest; we use "species" from here on). Phylofactorization can be implemented with disjoint phylogenies, such as viral phylogenies for which there are not clear common ancestors, and the sub-phylogenies can either be kept separate or joined at a polytomous root. The phylogeny may have an arbitrary number and degree of polytomies. Definitions of mathematical terms can be found in Box 1.

Let $\mathbf{X}$ be the data matrix of interest for phylofactorization whose rows are species and columns are samples, with $x_{i,j}$ being the data for species $i = 1, \ldots, m$ in sample $j = 1, \ldots, n$. Let $\mathbf{X}_R$ be the sub-matrix of $\mathbf{X}$ containing only a subset of species, R, and let $\mathbf{x}_{R,j}$ be the jth column of $\mathbf{X}_R$. Let $\mathbf{Z}$ be the $n \times p$ matrix containing p additional meta data variables for each sample. Let $\mathcal{T}$ be the phylogenetic tree, $\{\mathcal{T}_s\}$ a set of sub-trees whose tips span all species, and let edge e in the phylogeny separate the disjoint groups R and S. Phylofactorization requires (1) an aggregation function, $A(\mathbf{X}_R, \mathcal{T}, e) \in \mathbb{R}$ which aggregates any subset, R, of species within samples, possibly using information from the tree, $\mathcal{T}$ and species' proximity to the edge, e; (2) a contrast function, $C(A(\mathbf{X}_R, \mathcal{T}, e), A(\mathbf{X}_S, \mathcal{T}, e), \mathbf{Z}, \mathcal{T}, e) \in \mathbb{R}$ which contrasts the aggregates of two disjoint subsets of species, R and S, spanning the species in $\mathcal{T}$, possibly using meta data, $\mathbf{Z}$, and edge, e; and (3) an objective function, $\omega(C)$.

With these operations, phylofactorization is defined iteratively as a special case of a graph partitioning algorithm (Fig. 2). The steps of phylofactorization are as follows:

1) For each edge, e, in $\{\mathcal{T}_s\}$ separating disjoint groups of species R_e and S_e within the sub-tree $\mathcal{T}_e$

containing e, compute $C_e = C(A(\mathbf{X}_{R_e}, \mathcal{T}_e, e), A(\mathbf{X}_{S_e}, \mathcal{T}_e, e), \mathbf{Z}, \mathcal{T}_e, e)$

2) Compute edge objective $\omega_e = \omega(C_e)$ for each edge, e

3) Select winning edge $e^* = \text{argmax}_e(\omega_e)$

4) Update $\{\mathcal{T}_s\}$ by removing $\mathcal{T}_e$ and adding the two sub-trees formed by partitioning $\mathcal{T}_e$ along e^*.

5) Repeat 1–5 until a stopping criterion is met.

Unlike more general graph-partitioning algorithms, phylofactorization does not impose a balance constraint that would require the partitions have a similar size or weight. Furthermore, phylofactorization is the particular application of graph partitioning in which the graph is a phylogeny capturing the evolutionary relationships between organisms, thereby allowing an evolutionary and ecological interpretation of the partitions.

Aggregation and contrast operations used are the principle operations for defining scales and units of ecological organization, and by working with phylogenies the units aggregated will have many shared traits and the units contrasted will have traits or evolutionary histories that separate them. Phylofactorization is limited to contrasts of non-overlapping groups. The constraint of contrasting aggregates forces researchers to define a priori the method of aggregating data from groups of species partitioned by phylofactorization, thereby ensuring data from groups of species are subsequently summarized with the same method by which they were discovered to be different from data from other groups of species. The incorporation of the tree, $\mathcal{T}$, in the contrast function encompasses a class of ancestral state reconstruction reconstruction methods. Ancestral state reconstruction with non-overlapping contrasts can be done with time-reversible models of evolution; in this case, phylofactorization contrasts the root ancestral states obtained in which the two nodes adjacent an edge are considered roots of the subtrees separated by that edge. Finally, as we discuss in detail in the section *The Contrast Basis*, the use of aggregation and contrast as the central operations in phylofactorization connect the graph partitioning algorithm with a method for constructing a basis that can be used for matrix factorization and low-rank approximations of data sets.

We use the term "phylogenetic factor" to refer to the results from a particular iteration of the algorithm. "Factors" have two groups, R_e and S_e, separated by an edge or link of edges, e, and thus the term "factor", as opposed to "iteration", is chosen to allude to latent variables (traits, evolutionary regimes, etc.) sensu factor analysis and the basis elements used for matrix factorization (Washburne et al. 2017). It's possible to define objective functions through pure aggregation, such as $A(\mathbf{X}_R, \mathcal{T}, e)$, but we limit our focus to contrast-based phylofactorizations which identify edges along which meaningful differences arose due to the non-orthogonality of nested aggregates and the orthogonality of contrasts, discussed in greater detail in *The Contrast Basis* section.

The result of phylofactorization after t iterations is a set of t inferences on edges or links of edges. Links of edges occur following a previous partition, when two adjoining edges separate the same two groups in the resultant sub-tree. Partitioning the phylogeny along t edges results in $t+1$ bins of species, referred to as "binned phylogenetic units" (BPUs). In general, the problem of maximizing some global objective function, $\omega(e_1^*, \ldots, e_t^*)$, for a set of t edges, $\{e_1^*, \ldots, e_t^*\}$, is NP hard (Buluç et al. 2016). However, stochastic searches of the space of possible partitions, via a stochastic computation of ω_e in step 2 or a weighted draw of e^* in step 3, may yield better approximations of a global maximum (Metropolis et al. 1953, Hastings 1970, Jerrum and Sorkin 1998).

Aggregation, contrast, and objective functions are decision points to define and interpret meaningful quantities and outcomes from data analysis. Explicit decisions about aggregation formalize how a researcher would summarize data from an arbitrary set of species. Explicit decisions about contrasts formalize how a researcher differentiates two arbitrary, disjoint groups of species. The operations of aggregation and contrast operationalize the concept of phylogenetic scales. Many mathematical operations can be aggregations, including but not limited to addition, multiplication, generalized means, and maximum likelihood estimation of ancestral states under models of trait diffusion away from the focal node. Likewise, contrasts can be differences, ratios, two-sample tests, and more complicated metrics of dissimilarity such as the deviance of a factor contrast in a generalized additive model. Researchers must decide how best to aggregate information in groups of species, contrast two groups, and decide which group maximizes the objective for a research goal pertaining to a particular ecological pattern. Doing so allows objective, a priori definitions of what makes an informative phylogenetic scale.

Below, we show examples of the algorithm along with results from phylofactorization of real data. These examples were run using the R package phylofactor, using relevant functions for analyzing and visualizing phylogenies from the R packages ape (Paradis et al. 2004), phangorn (Schliep 2011), phytools (Revell 2012), and ggtree (Yu et al. 2017).

Two-sample tests

If the data are a single vector of observations, $\mathbf{x}$, such as average body mass estimated for a set of m species, phylofactorization can be implemented through standard tests for differences of means or rate parameters in the two sets of species, R and S.

To illustrate, we phylofactorize a data set of mammalian body mass from PanTHERIA (Jones et al. 2009) and the open tree of life using the R package rotl (Michonneau et al. 2016). A single vector of data assumed to be log-normal can be factored based on a two-sample t test (Fig. 3). In this case, our aggregation function $A(\mathbf{x}_R) = \overline{\log(\mathbf{x}_R)}$ is the arithmetic mean of the log body mass; our contrast function

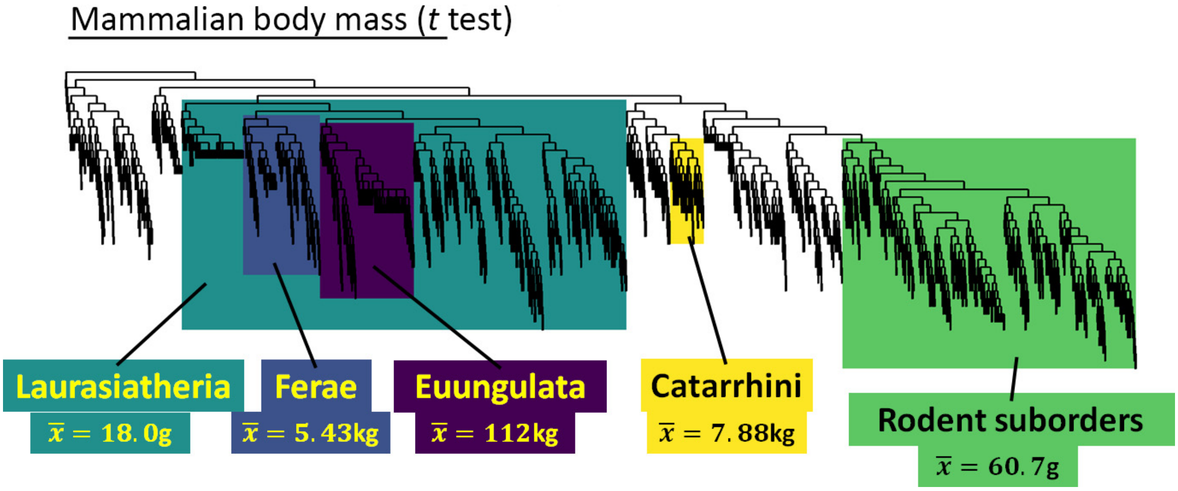


FIG. 3. Two-sample tests, such as *t* tests, can be used as objective functions for phylofactorization of vectors of data. Using a *t* test of equal variance, five iterations of phylofactorization on a data set of mammalian log body mass yields five clades with very different body masses. These *t* statistics are identical to projections of the data onto contrast basis elements discussed in the section *The Contrast Basis.*

$$C = \frac{A(\mathbf{x}_R) - A(\mathbf{x}_S)}{\sqrt{\frac{1}{r} + \frac{1}{s}}} \tag{1}$$

is a standardized difference of means, and the objective function $\omega_e = |C_e|$. With these operations, our objective function is the test-statistic for a two-sample *t* test with the assumption of constant variance. Maximization of the objective function yields edges separating mammals with the most significant difference in body mass.

The first five phylogenetic factors of mammalian body mass in these data are Euungulata, Ferae, Laurasiatheria (excluding Euungulata and Ferae), a clade of rodent suborders Myodonta, Anomaluromorpha, and Castorimorpha, and the simian parvorder Catarrhini. Five factors produce six binned phylogenetic units of species with different average body mass (Fig. 3). The most significant phylogenetic partition of mammalian body mass occurs along the edge basal to Euungulata, identifying a clade of 296 species with significantly larger body mass than other mammals. The second partition corresponds to Ferae, containing 242 species which have body masses larger than other mammals, excluding Euungulata. The third partition corresponds to 864 remaining species in Laurasiatheria, excluding Euungulata and Ferae, which contains Chiroptera, Erinaceomorpha, and Soricomorpha. These mammals have lower body mass than non-Laurasiatherian mammals. The fourth partition identifies three rodent sub-orders comprising 926 species with lower body mass than non-Laurasiatherian mammals. Finally, 106 species comprising the Simian parvorder Catarrhini are factored as having higher body mass than the remaining mammals. These factors are fairly robust: 3,000 replicates of stochastic Metropolis-Hasting phylofactorization, drawing edges in proportion to C^{λ} with $\lambda = 6$ (producing a 1/4 probability of drawing the most dominant edge at the first iteration) could not improve upon these five factors.

The first two phylogenetic factors of mammalian body size partition the mammalian tree at deep edges with ancestors near the K-Pg extinction event, corroborating evidence of ecological release (Alroy 1998, 1999) and the exponential growth of maximum body sizes following the K-Pg extinction event (Smith and Lyons 2011) for these two dominant clades. The crown group of modern Euungulata are thought to have originated in the late Cretaceous (Zhou et al. 2011) and its representatives may have expanded into previously dinosaur-occupied niches during the rapid evolution of body size in mammals immediately after the K-Pg extinction event at the Cretaceous/Paleogene boundary (Smith et al. 2010). Cope's rule posits that lineages tend to increase in body size over time, and a recent study (Baker et al. 2015) confirms Cope's rule and found that mammals have, along all branch lengths in their phylogeny, tended to increase in size. The phylogenetic factors of mammalian body size discovered here illustrate an important feature of phylofactorization: correlated evolution within a clade, such as a consistent directional evolution among lineages in a clade, can cause the edge basal to a clade to be an important partition for capturing variance in a trait. A more robust phylofactorization may be done through iterative ancestral-state reconstruction of the roots of subtrees partitioned by each edge (where the subtrees are re-rooted at the nodes adjacent the edge), but this unsupervised phylogenetic factorization body masses in 3,374 mammals takes 15 s on a laptop and yields partitions which simplify the story of mammalian body-mass variation to a set of five edges forming six binned phylogenetic units.

Two-sample tests can be used for phylogenetic factorization of any vector of trait data. For another example, Bernoulli trait data, such as presence/absence of a trait, can be factored using Fisher's exact test that there is the same proportion of presences in two groups, R and S. In this case, the aggregation operation $A(\mathbf{x}_R) = \sum_{i \in R} x_i$ counts the number of successes in group R, the contrast function, C, is the P value of Fisher's exact test with the contingency table shown in Table 1.

An objective function can be defined as the inverse of the P value from Fisher's exact test, $\omega_e = C_e^{-1}$. The phylofactorization of vertebrates by land/water association in Fig. 1, using an ad-hoc selection of vertebrates for illustration, was performed using Fisher's exact test, and the factors obtained correspond to Tetrapods, Cetaceans, and Pinnipeds. Unlike the phylofactorization of mammalian body mass, all three factors obtained from phylofactorization of vertebrate land/water association correspond to a set of traits. Tetrapods evolved lungs and limbs which allowed them to live on land. Cetaceans evolved fins and blowholes, and Pinnipeds evolved fins, all traits adaptive to life in the water.

Two-sample tests are used when partitioning a vector of traits and not controlling for additional meta data such as sampling effort or other confounding effects. Phylofactorization of body mass and land/water associations illustrate two potential evolutionary models under which edges are important: correlated evolution of members of a clade caused by different evolutionary regimes (e.g., ecological release, niche partitioning or geographic separation) and punctuated equilibria in which functional traits of large importance arise infrequently. More complicated methods of phylofactorization will yield similar evolutionary interpretations: factors may correspond to traits or evolutionary regimes shared among extant members of a clade and/or their ancestors.

The Contrast Basis

How can we identify the phylogenetic scales in an arbitrary matrix of data, $\mathbf{X}$, such as the data obtained when measuring abundances or traits of species across a range of environments? Low-rank approximations of matrices are useful tools for simplifying big data, and often rely on choosing a small set of vectors $\{\mathbf{v}_i\}_{i=1}^{K}$ and their coordinates $\{\mathbf{w}_i\}_{i=1}^{K}$ to minimize the distance between the matrix, $\mathbf{X}$, and some low-rank matrix $\mathbf{VW}$.

In this section, we introduce the contrast basis, a set of vectors that connect phylofactorization's graph-partitioning algorithm to various methods for low-rank approximations of data matrices. We then use the contrast basis for a phylogenetic analog of principal components analysis to analyze gut microbiomes across hundreds of patients, and for regression-based dimensionality reduction to identify the phylogenetic scales of community compositional changes in central park soils.

Table 1. Fisher's Exact test for two-sample phylofactorization of Bernoulli trait data.

Successes	Failures	Total
$A(\mathbf{x}_R)$	$r - A(\mathbf{x}_r)$	r
$A(\mathbf{x}_S)$	$s - A(\mathbf{x}_S)$	s
$A(\mathbf{x}_R) + A(\mathbf{x}_S)$	$r + s - (A(\mathbf{x}_r) + A(\mathbf{x}_S))$	$r + s$

Note: Variables are defined in Box 1.

The phylogeny provides a natural scaffold for low-rank, phylogenetically interpretable approximations of the data. As a sphere defines a natural set of coordinates for GPS data, the phylogeny defines a natural set of coordinates for community ecological data (Washburne et al. 2018). One example of a natural coordinate in the phylogeny is an aggregation; the total abundance of species within a clade is obtained by projecting the data onto a vector containing 1 for all elements corresponding to species in that clade and 0 for all other elements. Another example of a natural coordinate in the phylogeny is a contrast; the difference of total abundance between two clades is obtained by projecting the data onto a vector containing 1 for all elements in one clade and −1 for all elements in the other clade. These operations allow one to construct natural coordinates for more sophisticated analyses of phylogenetically structured ecological data.

Phylogenetically interpretable, low-rank approximations of data can be obtained by constructing basis elements through aggregation and contrast vectors (Fig. 1C). If two groups, R and S, are separated by an edge of interest for phylofactorization, an aggregation basis element for the group $Q = R \cup S$ can be constructed through a vector, $\mathbf{v}_{A_Q}$, whose ith element is

$$\mathbf{v}_{A_Q,i} = \begin{cases} a & i \in Q \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

If, for example, there are 10 species in Q, projecting the data onto $\mathbf{v}_{A_Q}$ with $a = 1/10$ is equivalent to taking the mean of those 10 species, whereas if $a = 1$ then projection onto $\mathbf{v}_{A_Q,i}$ is equivalent to summing the data of those 10 species. A natural complement to an aggregation vector is a vector contrasting the groups R and S whose ith element is

$$\mathbf{v}_{C_{R|S},i} = \begin{cases} b & i \in R \\ -c & i \in S \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $b > 0$ and $c > 0$. By meeting the criteria

$$\mathrm{rb} - \mathrm{sc} = 0 \tag{4}$$

$$\mathrm{rb}^2 + \mathrm{sc}^2 = 1 \tag{5}$$

one can ensure that the aggregation vector, $\mathbf{v}_{A_Q}$, and the contrast vector of the two disjoint sets comprising Q, $\mathbf{v}_{C_{R|S}}$, are orthogonal to one another (Eq. 4) and have unit norm (Eq. 5). Solving Eqs. 4 and 5 yields a choice of b and c for contrast basis elements:

$$b = \sqrt{\frac{s}{r(r+s)}} \tag{6}$$

$$c = \sqrt{\frac{r}{s(r+s)}}. \tag{7}$$

Phylogenetic scales of interest in data matrices can be identified through analysis of data projected onto aggregation and contrast vectors. In the language of phylofactorization's graph partitioning algorithm, projecting data from sample j, $\mathbf{x}_j$, onto the contrast vector $\mathbf{v}_{C_{R|S}}$ is equivalent to defining the sample-wise aggregation and contrast operations as

$$\begin{aligned} A(\mathbf{x}_{R,j}) &= \bar{\mathbf{x}}_{R,j} \\ C\big(A(\mathbf{x}_{R,j}), A(\mathbf{x}_{S,j})\big) &= \sqrt{\frac{\mathrm{rs}}{r+s}}\left(\bar{\mathbf{x}}_{R,j} - \bar{\mathbf{x}}_{S,j}\right) \end{aligned} \tag{8}$$

where $\bar{\mathbf{x}}_{R,j}$ is the sample mean of species in group R and sample j. Projecting an entire data set, $\mathbf{X}$, onto $\mathbf{v}_{C_{R|S}}$ yields coordinates, one for each sample, which are a standardized difference of means identical to Eq. 1. Since the contrast vector is comprised of two sub-aggregations of opposite sign, one for group R and the other for group S, it will be orthogonal to a subsequent a contrast vector partitioning either R or S into two disjoint groups. Thus, the non-overlapping contrasts produced by phylofactorization's graph-partitioning coupled with the criterion in Eq. 4 allow one to construct an orthogonal contrast basis during phylofactorization. The orthonormal contrast basis can be used to make low rank approximations of $\mathbf{X} = \mathbf{VW} + \boldsymbol{\epsilon}$ where the low-rank matrix $\mathbf{VW}$ corresponds to important phylogenetic scales in the data.

One can construct a complete basis using only aggregation and their contrast vectors. By disallowing overlapping aggregations (e.g., aggregations of nested clades) while maintaining the criteria in Eqs. 4 and 5 for contrast basis elements, one can ensure the basis is orthonormal. With m species, first define a set of $q \leq m$ orthogonal aggregation vectors aggregating disjoint sets of species Q_l such that the entire set of aggregations, $\bigcup_{l=1}^{l=q} Q_l = \{1, \ldots, m\}$, covers the entire set of m species. Then, $m - q$ contrast vectors partitioning the aggregations and all multi-species sub-aggregations within contrast vectors can complete the basis (Fig. 1C). It's worth noting that the span of any aggregate and its contrast is equal to the span of the contrasts' sub-aggregates, i.e., for $R \cup S = Q$

$$\mathrm{span}\left(\mathbf{v}_{A_Q}, \mathbf{v}_{C_{R|S}}\right) = \mathrm{span}(\mathbf{v}_{A_R}, \mathbf{v}_{A_S}) \tag{9}$$

(Fig. 1C). Thus, these two natural pairs of basis elements, an aggregate of species and its orthogonal contrast (grouping species and partitioning the group) or two orthogonal aggregates (two disjoint groups of species), are rotations of one another.

Aggregation vectors as defined in Eq. 2 can be defined a priori based on non-overlapping traits or clades of species thought to be important for the question at hand (e.g., aggregate "terrestrial" and "aquatic" animals), or they can be learned through clustering algorithms or even phylofactorization based purely on aggregation by converting steps 1 and 2 in the phylofactorization algorithm into a single step: maximizing an objective function of the aggregate of a clade. A special case occurs when data are compositional (Aitchison 1982), in which case the sum of the data for all species in the community will equal 1 and thus the data are constrained by an aggregation element: the aggregate of all species. Consequently, changes in compositional data are always orthogonal to the **1** vector, and, for compositional data, variation is best described through contrast basis elements. For this reason, phylofactorization via contrasts of log-relative abundance data allows one to construct an isometric log-ratio transform, a commonly used and well-behaved transform for the analysis of compositional data (Egozcue et al. 2003, Egozcue and Pawlowsky-Glahn 2005, Silverman et al. 2017). For non-compositional data, since the span of an aggregate and its contrast is equal to the span of the contrasts' two aggregates (Eq. 9), we simplify the identification of phylogenetic scales and the construction of a phylogenetic basis by considering, from here on out, only the "contrast basis" similar to that used in compositional data whereby an initial aggregate of all species is partitioned with a series of contrasts.

Phylogenetic components analysis

Principal components analysis obtains a set of orthogonal directions, called loadings, which sequentially maximize the variance of the data projected onto the loadings. Similarly, orthogonal contrast vectors allow researchers to partition the variance in a community ecological data set along each of a set of orthogonal directions corresponding to discrete, interpretable features in the phylogeny.

An edge, e, separating groups of species R and S has a corresponding candidate basis element, $\mathbf{v}_{C_{R|S}}$, that we will refer to as $\mathbf{v}_{C_e}$. Projecting the data matrix onto the contrast basis element yield what we'll call component scores $\mathbf{y}_e = \mathbf{v}_{C_e}^T \mathbf{X}$. The component scores can be used to identify phylogenetically interpretable directions capturing variance in the data through the objective function

$$\omega_e = \mathrm{Var}[\mathbf{y}_e]. \tag{10}$$

Phylofactorization via the objective function in Eq. 10 yields a phylogenetic decomposition of variance we call "phylogenetic components analysis" or PhyCA. PhyCA is a constrained version of principal components analysis, allowing researchers to identify the dominant axes of variation corresponding to contrasts of species separated by an edge.

The component score for sample j, $\mathbf{y}_{e,j}$, can be written as

$$\mathbf{y}_{e,j} = \sqrt{\frac{rs}{r+s}}(\bar{\mathbf{x}}_{R,j} - \bar{\mathbf{x}}_{S,j}) \tag{11}$$

where $\bar{\mathbf{x}}_{R,j}$ is the sample mean of $x_{i,j}$ for $i \in R$ and $\bar{\mathbf{x}}_{S,j}$ is the sample mean of $x_{i,j}$ for $i \in S$. Consequently, the variance of the component score is

$$\mathrm{Var}[\mathbf{y}_e] = \frac{rs}{r+s}(\mathrm{Var}[\bar{\mathbf{x}}_R] + \mathrm{Var}[\bar{\mathbf{x}}_S] - 2\mathrm{Cov}[\bar{\mathbf{x}}_R, \bar{\mathbf{x}}_S]). \tag{12}$$

The variance of $\mathbf{y}_e$ increases through a combination of variances of the aggregations of groups R and S across samples ($\bar{\mathbf{x}}_R$ and $\bar{\mathbf{x}}_S$, respectively) and a high negative covariance between aggregations for groups R and S across samples. Negative covariance may be caused by competitive exclusion, different habitat associations across the samples, and more - such ecological phenomena can be identified through PhyCA.

We use PhyCA to identify 10 factors from a sub-sample of the American gut data set (McDonald et al. 2018) and the greengenes phylogeny (DeSantis et al. 2006) containing m = 1,991 species and n = 788 samples from human feces (Fig. 4). The American gut data set was filtered to fecal samples with over 50,000 sequence counts and, among those samples, operational taxonomic units (OTUs) with an average sequence per sample greater than 1. After performing PhyCA, possible ecological explanations of variance were explored via least squares regression predicting the winning component score, $\mathbf{y}_{e^*}$, using seven explanatory variables: types_of_plants (a question asking participants how many types of plants they've eaten in the past week), age, bmi, alcohol consumption frequency, sex, antibiotic use (ABX), and inflammatory bowel disease (subset_ibd) (Fig. 4). The raw P values from t tests of the coefficients are presented below; the P value threshold for a 5% family-wise error rate, given the 70 tests run, is 7.1×10^{-4}.

The first factor splits 1,229 Firmicutes OTUs from the remaining 782 OTUs. The component score for the first factor, $y_{e^*_1}$, is strongly associated with antibiotic use ($P = 3.6 \times 10^{-4}$), showing dramatic decreases in relative abundance in patients who have taken antibiotics in the past week or month. The second factor identifies 217 species of several genera of Lachnospiraceae, a clade contained within the Firmicutes of factor 1. These Lachnospiraceae are contrasted from the remaining Firmicutes by a strong association with age ($P = 1.2 \times 10^{-15}$), bmi ($P = 3.2 \times 10^{-6}$), and alcohol ($P = 6.4 \times 10^{-3}$). The third

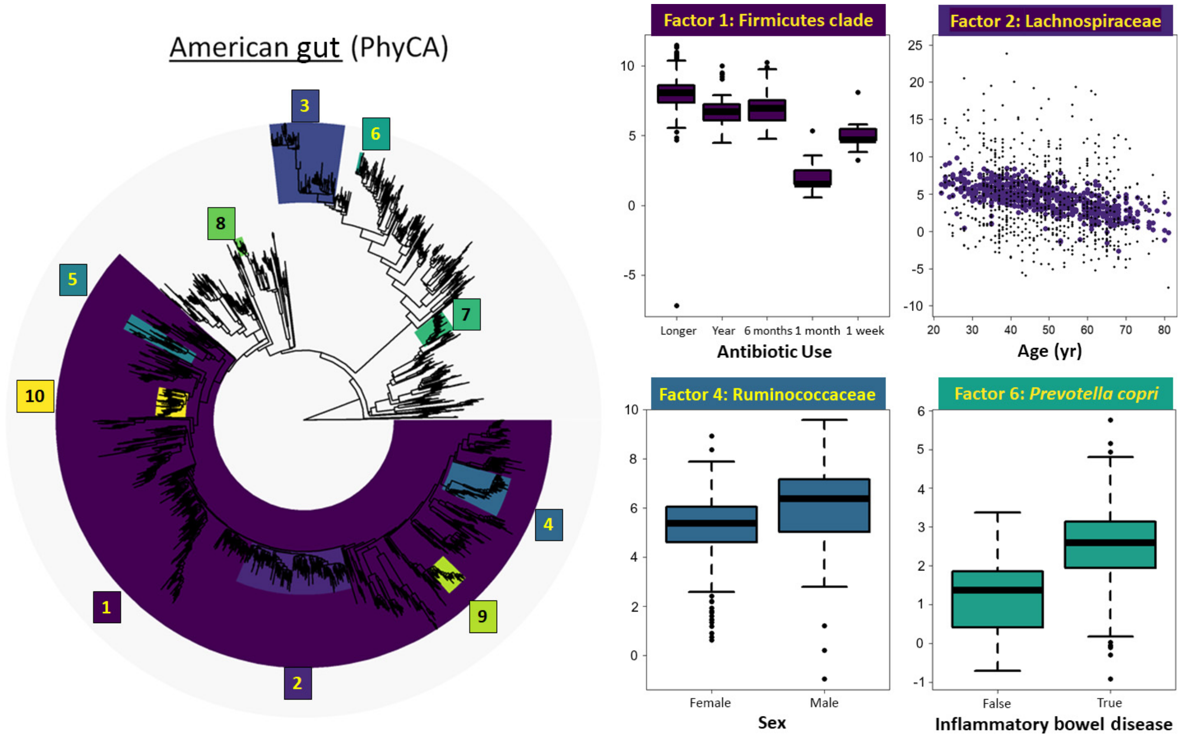


FIG. 4. Maximizing the variance of component scores, y_e, of log-relative abundance data produces a "phylogenetic components analysis" (PhyCA) of the American gut data set. The most variable clades cover a range of phylogenetic scales. Downstream analysis of component scores are tested for associations with biologically relevant meta data. Plotted are linear predictors against relevant meta data for one exception: the plot of Lachnospiraceae includes the raw data as black dots.

factor is a clade of 81 Bacteroides most strongly associated with types_of_plants ($P = 2 \times 10^{-9}$). By identifying clades of Firmicutes and Bacteroides as major axes of variation, factors 1 and 3 refine the Firmicutes to Bacteroidetes ratio commonly used to describe variation in the gut microbiome and found associated with obesity and other disease states (Ley et al. 2006, Clemente et al. 2012). It's been found that the Firmicutes/Bacteroidetes ratio changes with age (Mariat et al. 2009), but the picture from phylofactorization is more nuanced: the large clade of Firmicutes in the first factor does not change with age relative to the complement set of all species, but the Lachnospiraceae within that clade decrease strongly with age relative to the remaining Firmicutes, while the Bacteroides show only a moderate decrease with age. The strong decrease with age in Lachnospiraceae is found in a few other clades within the Firmicutes: the fourth factor identified a clade of Firmicutes of the family Ruminococcaceae strongly associated with types of plants ($P = 3.6 \times 10^{-5}$), sex ($P = 5.9 \times 10^{-4}$) and decreasing with age ($P = 9.2 \times 10^{-4}$), and the fifth factor identified a group of Firmicutes of the family Tissierellaceae that decrease strongly with age ($P = 1.9 \times 10^{-5}$).

The sixth factor partitions small group of five OTUs of *Prevotella copri* associated with types_of_plants ($P = 2.8 \times 10^{-4}$) and weakly associated with inflammatory bowel disease ($P = 2.5 \times 10^{-3}$). Previous studies have found that *Prevotella copri* abundances are correlated with rheumatoid arthritis in humans and inoculation of *Prevotella copri* exacerbates colitis in mice. Consequently, *Prevotella copri* is hypothesized to increase inflammation in the mammalian gut (Scher et al. 2013), and the discovery of *Prevotella copri* as one of the dominant phylogenetic factors of the American gut, as well as the discovery of its association with IBD, corroborates the hypothesized relationship between *Prevotella copri* and inflammation. Likewise, the seventh factor is a clade of 41 Gammaproteobacteria of the order Enterobacteriales also associated with types_of_plants ($P = 6.7 \times 10^{-8}$) and weakly associated with inflammatory bowel disease ($P = 0.022$). Gammaproteobacteria were used as biomarkers of Crohn's disease in a recent study (Vázquez-Baeza et al. 2017) and their associations with IBD in the American gut project corroborates the use of Gammaproteobacterial abundances for diagnosis of IBD from stool samples.

Gaussian-based regression-phylofactorization

When the data are assumed to be Gaussian or easily mapped to Gaussian with a monotonic function, such as logistic-normal compositional data or log-normal data, objective functions can be defined directly from regression on component scores. While $\mathbf{y}_e$ can be used as either an independent or dependent variable, the transformed-Gaussian assumption of the data is particularly important when $\mathbf{y}_e$ are used as dependent variables.

Maximizing the explained variance from regression identifies clades through the product of a high contrast variance from Eq. 10 and a high percentage of explained-variance from regression – such clades can capture large blocks of explained variance in the data set. Maximizing the deviance or F statistic from regression identifies clades with more predictable responses: such clades can be seen as bioindicators or particularly sensitive clades, even if they are not particularly large or variable clades. Regression-phylofactorization uses the component scores as a response or explanatory variable, the latter being used in the phylofactorization-based classification of Crohn's disease (Vázquez-Baeza et al. 2017). For multiple regression, one can define objective functions based on the explanatory power of the entire model or the explanatory power of a subset of the model. More complicated regression models can be considered, including generalized additive models, regularized regression, and more.

To identify phylogenetic scales corresponding to non-linear patterns of abundance-habitat associations, we perform a generalized additive model analysis of the Central Park soils data set (Ramirez et al. 2014) previously analyzed with least squares. To identify non-linear associations between clades and pH, carbon, and nitrogen, we perform a generalized additive model of the form

$$\mathbf{y}_e \sim s(\mathrm{pH}) + s(\mathrm{carbon}) + s(\mathrm{nitrogen}) \tag{13}$$

where $s()$ indicates a smoothing spline. Our objective function was the explained variance of the entire model. The resultant phylofactorization (Fig. 5) identified the same four factors as the least squares model and nonlinear patterns of community compositional changes along environmental gradients. The four factors partition over 3,000 species into five binned phylogenetic units; aggregating abundances within BPUs while sorting the data along pH (the dominant explanatory variable for all four factors) allows clear, phylogenetically interpretable, low-rank visualization of otherwise complex behavior of how a community of several thousand microbes changes across several hundred soil samples. Phylofactorization through generalized additive modeling identifies a clade of Acidobacteria, the Chloracidobacteria, which have their highest relative abundances in neutral pH soils.

Generalized Phylofactorization

Many ecological data are not Gaussian. Presence–absence data or count data with many zeros cannot be easily transformed to yield approximately Gaussian random variables. Data assumed to be exponential family random variables can be analyzed with regression-phylofactorization by adapting concepts used in generalized linear models for aggregation & contrast of species separated by edges.

For an example of why this is important, consider a data matrix, $\mathbf{X}$, whose entries are either 0 or 1 (i.e. presence–absence data). Projecting these data onto $\mathbf{v}_{C_e}$ will yield component scores, $\mathbf{y}_e$, which are discrete and for

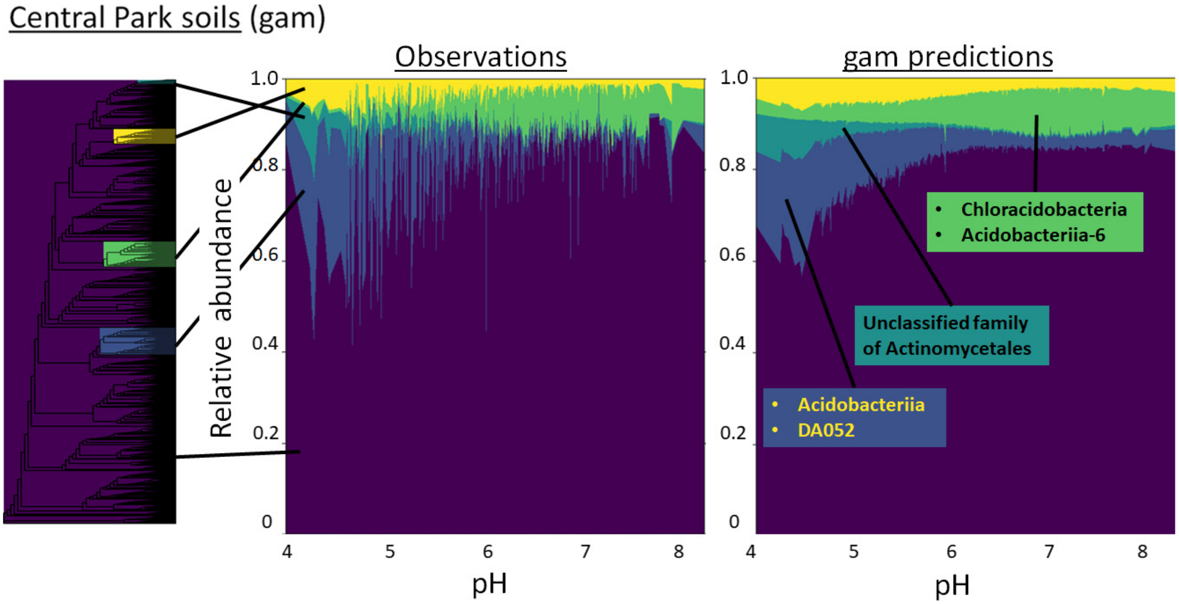


FIG. 5. Projecting the data onto contrast basis elements permits a broad range of analyses. Here, the component scores y_e from projections of log-relative abundances are analyzed to find phylogenetic factors of changing community composition in Central Park soils. The model $y_e \sim s(\log(\text{carbon})) + s(\log(\text{nitrogen})) + s(\text{pH})$, where $s()$ indicates a smoothing spline, was combined with the objective of maximizing the explained variance. The relative importance of pH in the generalized additive models, and the exact clades with a high amount of variance explained by pH, allows a projection of 3,000 species into 5 binned phylogenetic units (BPUs) for clear visualization and prediction of nonlinear community compositional changes along a key environmental gradient.

which least squares regression is not appropriate; among other things, the regression model could predict values of $\mathbf{y}_e$ beyond what is possible given the number of species in R and S. As one would use logistic regression in generalized linear models to analyze patterns in the presence/absence of a single species, we present algorithms to use generalized linear models for regression-phylofactorization of such non-Gaussian, exponential family random variables.

We present four algorithms to identify edges separating groups of species with high within-group similarity and high between-group differences in regression coefficients estimated through generalized linear modeling. The algorithms either explicitly use the contrast basis to approximate the regression coefficient matrix or implicitly use an analog of the contrast basis in the likelihood function via categorical factor contrasts in a shared coefficients model. The algorithms we propose are (1) coefficient contrast, which uses the contrast basis to identify sets of species with significantly different regression coefficients; (2) `phylo` factor contrasts, which uses surrogate categorical variables, `phylo`, to contrast regression coefficients; (3) marginally stable aggregation, which aggregates data to marginally stable distributions, then use of `phylo` factor contrasts; and (4) mixed, which uses algorithm 1 (the fastest) to identify a subset of edges as candidates for algorithm 2 (the slowest but most accurate).

At the end of this section, we compare the computational costs and scaling of these algorithms. The broader use of phylofactorization through generalized linear modelling is referred to as "generalized phylofactorization".

Algorithm 1: Coefficient contrast

Matrix factorization, $\mathbf{X} = \mathbf{V}\mathbf{W} + \boldsymbol{\epsilon}$, can be used for low-rank approximations of the coefficient matrix. The first algorithm, related to reduced rank regression for vector generalized linear models (Yee and Hastie 2003), uses the contrast basis to provide a reduced-rank approximation of the coefficient matrix from multivariate generalized linear models. Multivariate (vector) generalized linear models assume the data $\mathbf{X}$ are drawn from an exponential family distribution with canonical parameters for each species, $\boldsymbol{\eta} \in \mathbb{R}^m$, related to the meta data $\mathbf{Z}$ through a linear model

$$\boldsymbol{\eta} \sim \mathbf{B}\mathbf{Z} \tag{14}$$

where $\mathbf{B} \in \mathbb{R}^{m\times p}$ is the coefficient matrix and $\mathbf{Z} \in \mathbb{R}^{p\times n}$ is the matrix of meta data. Instead of using $m \times p$ coefficients, one can approximate the coefficient matrix $\mathbf{B}$ through contrast basis elements and their component scores

$$\mathbf{B} = \mathbf{1}\mathbf{w}_0^T + \mathbf{V}\mathbf{W} + \boldsymbol{\epsilon} \tag{15}$$

where $\mathbf{1} \in \mathbb{R}^m$ is the one vector, $\mathbf{w}_0 \in \mathbb{R}^p$ contains the mean of the regression coefficients for each of the p predictors, $\mathbf{V} \in \mathbb{R}^{m\times t}$ is a matrix whose columns are contrast basis elements obtained from t iterations of phylofactorization and $\mathbf{W} \in \mathbb{R}^{t \times p}$ is a matrix whose rows are the component scores for each contrast basis element and whose columns are the set of component scores for each of the p predictors. If one is interested in partitioning species based on a subset, P, of the explanatory variables, one can implement Eq. 15 for the matrix $\mathbf{B}_P$ containing only the partitioning variables for phylofactorization.

The approximation of $\mathbf{B}$ in Eq. 15 is best done by directly testing differences in the standardized regression coefficients obtained by dividing regression coefficients, $\beta_{i,j}$, by their standard error. We refer to the matrix of such standard coefficients for partitioning variables as the "standardized coefficient matrix," $\tilde{\mathbf{B}}_P$.

The Euclidean norm of the projection of the standardized coefficient matrix onto contrast basis elements can serve as an objective function

$$\omega_e = \|\mathbf{v}_{C_e}^T \tilde{\mathbf{B}}_P\| \tag{16}$$

capturing the extent to which coefficients in $\tilde{\mathbf{B}}_P$ differ between the sets of species partitioned by the edge e. Combined with the asymptotic normality of regression coefficients in generalized linear models and assuming independence of $\beta_{i,j}$ across meta data j, a given objective function ω_e is a chi-squared statistic with p degrees of freedom. Coefficient contrasts are fast and easy to compute, but the algorithm described here minimizes the distance between $\mathbf{VW}$ and $\tilde{\mathbf{B}}_P$ and do not necessarily maximize the likelihood of the reduced-rank regression. Other algorithms described below construct reduced-rank approximations which maximize the likelihood of the data under an explicit model of within-group shared coefficients.

Algorithm 2: Stepwise `phylo` *factor contrasts*

A surrogate variable `phylo` $\in \{R, S\}$, indicating which group a species is in, can be used to explicitly model shared coefficients within-groups andcontrast the coefficients between-groups, all while finding the maximum-likelihood estimates of the shared coefficients. Stepwise, maximum-likelihood selection of `phylo` factor contrasts are a more accurate yet computationally intensive algorithm for generalized phylofactorization.

To see how `phylo` factors are constructed, a data frame contrasting how the counts of "birds" and "non-birds" are associated with meta data z_2 while controlling for z_1 can be constructed as shown in Table 2. The monophyletic group of birds is always takes the value R for the variable `phylo`, whereas non-birds always take on the value S. Phylofactorization can be implemented through a generalized linear model for a count family (e.g., Poisson, binomial, or negative binomial) using the formula

$$\text{Abundance} \sim z_1 + \text{phylo} \times z_2. \tag{17}$$

TABLE 2. Constructing `phylo` factor corresponding to edge separating birds from non-birds.

Site	Species	Abundance	z_1	z_2	`phylo`
1	Pigeon	10	1	0.5	*R*
1	Dove	8	1	0.5	*R*
1	Lizard	1	1	0.5	*S*
1	Mouse	3	1	0.5	*S*
1	Cat	1	1	0.5	*S*
2	Pigeon	2	0	−2	*R*
2	Dove	1	0	−2	*R*
2	Lizard	10	0	−2	*S*
2	Mouse	4	0	−2	*S*
2	Cat	3	0	−2	*S*
…	…	…	…	…	…

Note: Variables are defined in Box 1.

The `phylo` factor contrasts groups separated by an edge; using its deviance as the objective function will find the edge e^* whose `phylo` factor maximizes the likelihood of the data under a model of shared coefficients.

In `phylo` factor contrasts, aggregation occurs within the likelihood function. The likelihood $\mathcal{L}(\mathbf{x}_j; \boldsymbol{\eta})$ for a vector of binomial random variables $\mathbf{x}_j \in \mathbb{R}^m$ can be written in exponential family form

$$\mathcal{L}(\mathbf{x}_j; \boldsymbol{\eta}) = h(\mathbf{x}_j) \exp\{\boldsymbol{\eta}'\mathbf{x} - \mathcal{A}(\boldsymbol{\eta})\}\,. \tag{18}$$

A two-factor model, such as $\mathbf{x} \sim \text{phylo}$, will reduce the likelihood function from m parameters in $\boldsymbol{\eta}$ to two parameters, $\boldsymbol{\eta}_i \in (\eta_R, \eta_S)$, yielding

$$\mathcal{L}(\mathbf{x}_j; \text{phylo}) = h(\mathbf{x}_j) \exp\left\{ \eta_R \sum_{i \in R} x_{i,j} + \eta_S \sum_{i \in S} x_{i,j} - \mathcal{A}(\boldsymbol{\eta}) \right\}.$$

Aggregation within the likelihood function above is summation of data within-groups; more generally, aggregation is given by the sufficient statistic, $T(\mathbf{x})$, in the exponential family random variable's likelihood function (e.g., $T(\mathbf{x}) = \sum_i \log(x_i)$ for the Pareto and chi-squared distributions). With the maximum likelihood estimates, $\hat{\eta}_R$ and $\hat{\eta}_S$, a contrast function can be defined as a difference of η_R and η_S, or the test-statistic from a hypothesis test that $\eta_R = \eta_S$.

Stepwise selection of maximum-likelihood `phylo` factor contrasts, constructed for non-overlapping sets of species via the graph partitioning in phylofactorization, is an accurate yet extremely computationally intensive method for regression-phylofactorization of exponential family random variables. A faster yet less accurate algorithm, which still performs maximum-likelihood estimation of `phylo` factor contrasts, is the use of marginally stable aggregation (Fig. 6).

Algorithm 3: Marginally stable (mStable) aggregation

Another option, aimed to reduce the computational costs of explicit maximum-likelihood estimation of `phylo` factors, is to aggregate the raw data $\mathbf{X}$ prior to

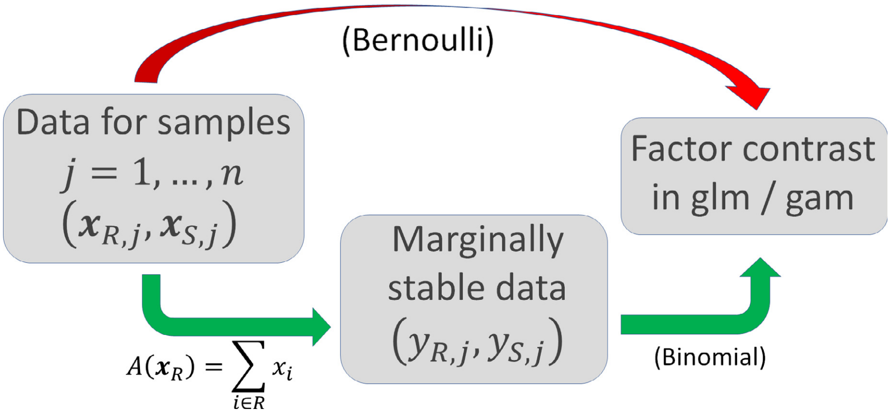


FIG. 6. Exponential family random variables can be explicitly analyzed in regression-phylofactorization either directly through `phylo` factor contrasts or through marginally stable aggregation. Aggregating data to a marginally stable distribution, such as addition of Bernoulli random variables with the same probability of success to a binomial random variable, can dramatically reduce computational costs while allowing maximum-likelihood estimation of regression coefficients under assumptions of within-group homogeneity. A comparison of the two methods' accuracy is provided in the Appendix (see Appendix S1).

evaluating the generalized linear model. The method we present is to assume within-group homogeneity and aggregate exponential family random variables to a "marginally stable" exponential family random variable that can be used for downstream analysis. Marginal stability, to the best of our knowledge, has not been explicitly defined elsewhere, and thus we introduce the term here by loosening the definition of stable distributions (Sato 1999). Given a non-trivial aggregation operator $A{:}\Omega^m \mapsto \Omega$ defined for any natural number m, a distribution with parameters $\{\theta_1, \theta_2\}$, $\mathcal{F}(\theta_1, \theta_2)$, defined over Ω, is said to be marginally stable on θ_1 with respect to A if for all $\mathbf{x} \in \Omega^m$ with independent elements $x_i \sim \mathcal{F}(\theta_1, \theta_{2,i})$ for $i = 1, \dots, m$, $A(\mathbf{x}) \sim \mathcal{F}(\theta_1, \theta_{2,m+1})$ conditioned on θ_1 being fixed.

The Gaussian distribution is stable: the sum of two Gaussian random variables is also Gaussian. Meanwhile, binomial random variables are marginally stable on the probability of success; random variables $x_i \sim \mathrm{Binom}(\rho, \mathrm{N}_i)$ can be summed to yield $A(\mathbf{x}) \sim \mathrm{Binom}(\rho, \sum \mathrm{N}_i)$. Marginal stability opens up more distributions to stable aggregation. Presence absence data, for instance, can be assumed to be Bernoulli random variables. The assumption of within-group homogeneity for the probability of presence, ρ, allows addition of Bernoulli random variables within each group, R and S, to yield a respective binomial random variable, x_R and x_S. Likewise, the addition of a set of binomial random variables with the same probability of success, ρ, yields an aggregate binomial random variable. A set of exponential random variables with the same rate parameter, λ, can be added to form a gamma random variable. Gamma random variables, $x_i \sim \mathrm{Gamma}(\kappa_i, \theta)$, parameterized by their shape, κ_i, and scale, θ, are marginally stable on θ. Addition of geometric random variables with the same rate parameter forms a negative binomial, and the addition of a set of negative binomial random variables, $x_i \sim \mathrm{NB}(\pi_i, \rho)$, with the same probability of success ρ but different numbers of failures, π_i, can be aggregated into $x_R = \sum_{i \in R} x_i$ where $x_R \sim \mathrm{NB}\left(\sum_{i \in R} \pi_i, \rho\right)$. All of these distributions are not stable, but they are marginally stable. Marginal stability, for the purposes of phylofactorization, must be on the parameter of interest in generalized linear modeling.

Marginal stability can also be used with transformations connecting the assumed distribution of the data to a marginally stable distribution. Log-normal random variables can be converted to Gaussians through exponentiation; chi random variables can be converted to chi-squared through squaring; random variables from many distributions may be analyzed by transformation to a stable or marginally stable family of distributions. Such transformation-based analyses implicitly define aggregation through a generalized f-mean

$$A_f(\mathbf{x}_R) = f^{-1}\left(\sum_{i \in R} f(x_i)\right) \qquad (19)$$

where $f(x) = \log(x)$ for log-normal random variables, $f(x) = x^2$ for chi random variables, etc. The goal of such aggregation, whether through exploiting marginal stability or generalized f-means or other algebraic-group operations in the exponential family, is to produce summary statistics for each group of species, R and S, in a manner that permits generalized linear modeling of the summary statistics. By ensuring summary statistics are also exponential-family random variables, one can perform a factor-contrast style analysis as described above using only two summary statistics and not all $r + s$ species. Doing so can greatly reduce the computational load of phylofactorizing large data sets and can increase the power of edge identification even when the within-group homogeneity assumption does not hold (see Appendix S1).

Marginally stable aggregation can be made efficient by matrix multiplication onto one-vectors $\mathbf{1}_R$ and $\mathbf{1}_S$ whose ith entries are 1 for all $i \in R, S$, respectively, and 0 otherwise. Assuming a Poisson or negative binomial count model for the bird/non-bird data frame above, the data frame is reduced to Table 3 and the same equation (Eq. 17) can be used for phylofactorization through `phylo` factor-contrasts. Thus, marginally stable aggregation and `phylo` factor contrasts present two options for generalizing regression-phylofactorization to data from the exponential family (Fig. 6).

TABLE 3. Marginally stable aggregation and `phylo` factor construction for edge separating birds from non-birds.

Site	Species	Abundance	z_1	z_2	`phylo`
1	Bird	18	1	0.5	R
1	Non-Bird	5	1	0.5	S
2	Bird	3	0	−2	R
2	Non-Bird	17	0	−2	S
…	…	…	…	…	…

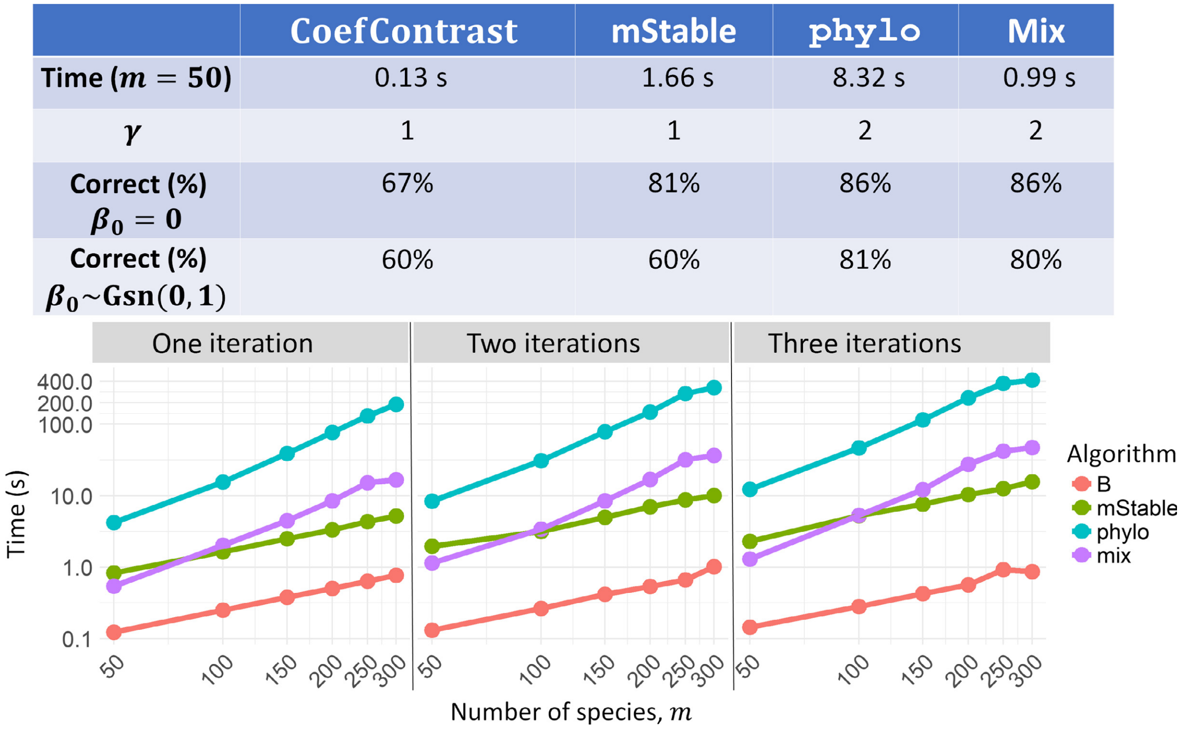
Note: Variables are defined in Box 1.

Algorithm 4: Mixed algorithm

Coefficient contrasts are computationally easy yet less accurate for edge identification, whereas stepwise `phylo` factor selection (without marginally stable aggregation) is accurate yet computationally demanding (Fig. 7). It's possible to develop mixed algorithms with accuracy similar to stepwise `phylo` factor selection and reduced computational costs more similar to coefficient contrasts or marginally stable aggregation. For each iteration, coefficient contrasts (Eq. 16) can be used to narrow down the set of possible edges, $\{e\}_{\text{top}}$, to a set of edges with high objective functions from standardized coefficient contrasts. We use the top 20% of edges based on ω_e in Eq. 16, resulting in an approximately 80% speed-up compared to the brute-force `phylo` factor contrast algorithm. For only these edges, `phylo` factors are considered and the winning edge is the top-quantile edge which maximizes the deviance of its `phylo` factor contrast.

Algorithm comparison

We compare the performance of the four algorithms listed above. The algorithms are compared on how well



	CoefContrast	mStable	phylo	Mix
Time ($m = 50$)	0.13 s	1.66 s	8.32 s	0.99 s
γ	1	1	2	2
Correct (%) $\beta_0 = 0$	67%	81%	86%	86%
Correct (%) $\beta_0 \sim \text{Gsn}(0,1)$	60%	60%	81%	80%

FIG. 7. Accuracy, computation time, and scaling of four algorithms for generalized phylofactorization. Algorithms are compared by the baseline time for two factors with m = 50 species, the scaling coefficient γ in time $\propto m^{\gamma}$, and percent of correctly identified edges in simulated data with m = 50 species and two affected clades. Stepwise `phylo` factor contrasts have high accuracy but are computationally costly and scale quadratically with the number of species. Marginally stable (mStable) aggregation scales linearly with m but only performs well when β_0 = 0. Computation time can be reduced and accuracy preserved if coefficient contrasts in Eq. 16 are used to narrow the set of edges considered for rigorous `phylo` factor contrasts.

they can correctly identify two edges with a known effect, $\{e_1^*, e_2^*\}$, and how long they take to extract a variable number of factors across a range of species, m, in the data set.

To compare edge-identification accuracy, presence/absence data were simulated for a set of $m = 50$ species and $n = 40$ samples. The logit probabilities of null species i being present were

$$\eta_i \sim \beta_{i,0} + 0.1z_1 + 0.1z_2 \tag{20}$$

where either $\beta_0 = 0$ for all species or $\beta_{i,0} \overset{i.i.d.}{\sim} \mathrm{N}(0, 1)$ to violate the within-group homogeneity in mean probability of presence/absence. The other two explanatory variables, z_1 and z_2, were the partitioning variables differentiating species separated by edges. Two non-nested clades, one containing 21 species and the other containing five species, had a different association with the meta-data:

$$\eta_i \sim z_{0,i} - 0.2z_1 + 0.6z_2$$

for species i in either of the two affected clades. To add an additional level of complexity, the two meta-data variables were given multicolinearity by simulating $z_1 \sim \mathrm{Gsn}(0, 1)$ and $z_2 \sim \mathrm{Gsn}(z_1, 1)$. The algorithms were run for two factors and the number of correctly identified edges (out of two) was tallied across 1,000 replicates (e.g., an algorithm that was 80% correct identified 1,600 correct edges over 1,000 replicates).

The time it took for each of these algorithms to compute two factors above was also recorded. To compare the scaling of the algorithms with increasing number of species, null data were simulated across a range of species richness $m \in \{50, 100, 150, 200, 250, 300\}$ and across a range of factors $t \in \{1, 2, 3\}$.

Deviance-maximization in the stepwise `phylo` factor contrasts had the greatest accuracy but also the slowest computation time (Fig. 7). The time required to compute `phylo` factor contrasts scale quadratically with the number species whereas coefficient contrasts and marginally stable (mStable) aggregation scale linearly. Marginally stable aggregation only performs well when $\beta_{i,0} = 0$ for all species, i, and when the within-group heterogeneity is small. The accuracy of `phylo` factor contrasts can be preserved and the computation time reduced by selecting the top 20% of edges based on coefficient contrasts.

Summary of generalized phylofactorization

We have presented algorithms to perform regression-phylofactorization for non-Gaussian data. The stepwise selection of `phylo` factor contrasts is best able to correctly identify edges but is computationally costly for large data sets. The computation time of stepwise `phylo` factor contrasts can be reduced by narrowing the set of considered edges to those with high coefficient contrasts. Marginally stable aggregation may be a promising alternative for faster algorithms as it scales linearly with the number of species, but marginally stable aggregation only performs well when there is little difference in the mean, $\beta_{i,0}$, across species, i.

These algorithms are intimately related to reduced rank regression and generalized linear modeling with shared coefficients. Reduced-rank regression uses gradient ascent over a compact set of possible basis vectors to find maximum-likelihood estimates. The constrained, countable set of contrasts defined by the phylogeny precludes gradient ascent and produces problems directly analogous to those in phylogenetic components analysis. Consequently, we have focused on explicit testing of all possible allowable contrasts in the phylogeny or, in the case of the mixed algorithm, testing a subset of contrasts believed to contain the winning edge, e^*. These methods can extend to generalized additive models and, as we discuss below, spatial and time-series data as well.

Phylogenetic Factors of Space and Time

Phylofactorization can be used in explicit analyses of spatial and temporal patterns. For Gaussian data, or for data used as an explanatory variable, samples of a community over space and time can be projected onto contrast basis elements or other contrast functions and the resulting component scores analyzed directly using standard spatial or temporal methods. Similarly, `phylo` factor contrasts can be used in spatially explicit analyses. Multivariate Autoregressive Integrated Moving Average (ARIMA) models can be constructed either as ARIMA models of the component scores, y_e, or as multivariate ARIMA models with `phylo` factor contrasts, to identify phylogenetic partitions based on differences in drift, volatility, and other time-series features of interest. Coefficient matrices, including spatial and temporal autocorrelation matrices or coefficients of association with extrinsic meta-data $\mathbf{Z}$, can be approximated with phylogenetic contrast-bases as in Eq. 15.

Marginally stable aggregation in spatial and temporal data requires a brief consideration of the marginal stability of spatially explicit random variables and stochastic processes. "Stability," for spatially and temporally explicit random variables, must preserve the underlying model for the spatial or temporal process assumed to produce the data. An example of a less obvious marginally stable aggregation of time-series data is the stability of neutral drift (sensu Hubbell 2001) to grouping.

Neutral communities fluctuate, and those fluctuations have a drift and volatility unique to neutral drift. Neutral drift can also be defined either by discrete, finite-community-size urn processes or stochastic differential equations serving as continuous approximations of large communities' neutral drift. Washburne et al. (2016) articulated the importance of a mathematical property of neutral drift which enables time-series neutrality tests: its invariance to grouping of species. If a stochastic process of relative abundances, $\mathbf{X}_t$, obeys the probability law defined by neutral drift, then any disjoint groupings of all species in $\mathbf{X}_t$ also obeys the probability law for a lower-dimensional neutral drift. Thus, neutral processes are stable to aggregation by summation of relative abundances. Collapsing all species into two disjoint groups, R and S, yields a two-dimensional neutral drift with a well-defined neutrality test for time-series data. Specifically, if $\mathbf{X}_t$ is a Wright Fisher process and R and S are disjoint groups whose union covers the entire community, the quantity

$$\nu_t = \arcsin\left(\left(\sum_{i\in R} X_{i,t}\right) - \left(\sum_{j\in S} X_{j,t}\right)\right) \qquad (21)$$

has a constant volatility whose constancy can be tested in order to test neutrality.

Phylofactorization can use these process-specific operations for marginally stable aggregation and contrast of neutral drift to partition edges across which the dynamics appear to be the least neutral. For the test developed by Washburne et al., the aggregation operation is the L_1 norm and the contrast operation is the arcsine of the differences of groups:

$$\begin{aligned} A(\mathbf{x}_R) &= |\mathbf{x}_R| \\ C(A(\mathbf{x}_R), A(\mathbf{x}_S)) &= \arcsin(A(\mathbf{x}_R) - A(\mathbf{x}_S)) \end{aligned}. \qquad (22)$$

An objective function, ω, for edge e can be the test statistic of a homoskedasticity test of C_e. Neutrality is a relative measure, biological units are neutral relative to one-another, and thus the use of aggregation of species into a unit and a contrast of two units is a natural connection between the theory and operations of phylofactorization and the biologically important null model of neutrality.

Statistical Challenges

There are many statistical connections and challenges which illuminate phylofactorization as a statistical tool. Phylofactorization is formally defined as a graph-partitioning algorithm, but maximizing the variance of the data projected onto contrast basis elements is a constrained principal components analysis. The use of regression-based objective functions and the iterative construction of a low-rank approximation of a data matrix is similar to factor analysis. The selection of a sequence of orthogonal factor contrasts in generalized linear models is a form of stepwise/hierarchical regression, and the factorization of a coefficient matrix $\mathbf{B}$ is a method for reduced-rank regression. The maximization of the objective function at each iteration is a greedy algorithm. Each connection between phylofactorization and other classes of methods produces a body of related literature which could inform phylofactorization and facilitate development of exploratory phylofactorization into a robust, inferential tool.

In this section, we enumerate some of the statistical challenges and discuss work that has been done so far. First, as with any method using the phylogeny as a scaffold for creating variables or making inferences, the uncertainty of the phylogeny and the common use of multiple equally likely phylogenies warrant consideration and further method development (Washburne et al. 2018). Other challenges discussed here are: the propagation of error; the use of Metropolis algorithms to better arrive at global maxima; the appropriateness and error rates of phylofactorization under various evolutionary models underlying the data; the graph-topological biases and confidence regions; cross-validation of partitions and inferences from phylofactorization across communities with different species; the appropriate number of factors and stopping criteria to stop a running phylofactorization algorithm; and the null distribution of test statistics when objective functions being maximized are themselves test-statistics from a well characterized distribution. Any exploratory data analysis tool can be made into an inferential tool with appropriate understanding of its behavior under a null hypothesis, and the connections of phylofactorization to related methods can accelerate the development of well calibrated statistical tests for phylogenetic factors.

Phylogenetic inference.—So far we have assumed that the phylogeny is known and error free, but the true evolutionary history is not known, it is estimated. Consequently, phylofactorization makes inferences on an uncertain scaffold; all else being equal, the more certain the scaffold, the more certain our inferences about a clade. Two challenges remain for dealing with phylofactorization on an uncertain phylogeny.

For a consensus tree, there is the question of what statistics of the consensus tree can yield precise statements of uncertainty in phylofactorization inferences. Bootstrapped confidence limits for monophyly (Felsenstein 1985*a*) are the most common metric of certainty for a consensus tree, but there may be others as well. Since phylofactorization can still be performed on a tree with polytomies and reducing the number of edges considered at each iteration can focus statistical effort (and chances of false discovery) on clades about which the researcher is more certain, trees containing clades with low bootstrap monophyly can be collapsed to improve the certainty of phylofactorization inferences. Different organisms will have different leverages in regression or two-sample test phylofactorization, and thus monophyly is only part of the picture: leverage and other statistics will also determine the stability of an inference to changing tree topology. Last, for a set of equally likely bootstrapped trees, there is a need to integrate phylofactorization across trees. Phylofactorization of sets of equally likely phylogenies has not yet been done, but is a fruitful avenue for future research.

Propagation of error.—Phylofactorization is a greedy algorithm. Like any greedy algorithm, its deterministic application is non-recoverable. Choosing the incorrect edge at one iteration can cause errors to propagate, potentially leading to decreased reliability of downstream edges. Little research has been done toward managing the propagation of error in phylofactorization, but recognizing the method as a greedy algorithm suggests options for improving performance. Stochastic-optimization schemes, such as replicate phylofactorizations using Metropolis algorithms and stochastic sampling as implemented in the mammalian tree phylofactorization (sampling of edges with probabilities increasing monotonically with ω_e and picking the phylofactor object which maximizes a global objective function), may reduce the risk of error cascades in the final, resulting phylofactorization (Hastings 1970).

Behavior under various evolutionary models.—Phylofactorization is hypothesized to work well under a punctuated-equilibrium model of evolution or jump-diffusion processes (Gould 1972, Landis et al. 2012) in which jumps are infrequent and large, such as the evolution of vertebrates to land or water. Phylofactorization may also work well when infrequent life-history traits arise or evolutionary events occur along edges which cause correlated, directional evolution among descendants. Phylofactorization of mammalian body sizes yielded an example of the second category of evolutionary scenarios under which phylofactorization works well. Both aggregation and contrast functions can incorporate phylogenetic structure and edge lengths to partition the tree based on likelihoods of such evolutionary models. The sensitivity of phylofactorization to alternative models, such as continuous Brownian motion and Ornstein-Uhlenbeck models commonly used in phylogenetic comparative methods (Felsenstein 1985*b*, Hansen 1997), remains to be tested and will likely vary depending on the particular method used.

Basal/distal biases.—Researchers may be interested in the graph topological distribution of factored edges in the tree. If a microbial community is exposed to antibiotics and regression-phylofactorization results in many tips being selected, a researcher suspecting lateral transfer of antibiotic resistance may be interested in quantifying the probability of drawing a certain number of tips given t iterations of phylofactorization. Alternatively, if several edges are drawn in close proximity, researchers may wonder the probability of drawing such clustered edges under a null model of phylofactorization. For another example, researchers may ask if an unusually high/low number of factors appear in a particular historical time window due to some hypothesis of important evolutionary event or environmental change. All of these tests require an accurate understanding of the probability of drawing edges in different locations of the tree.

All methods described here, save the Fisher exact test, have a bias for tips in the phylogeny (Fig. 8). Graph-topological biases affect the calibration of statistical tests of the location of phylogenetic factors, such as a

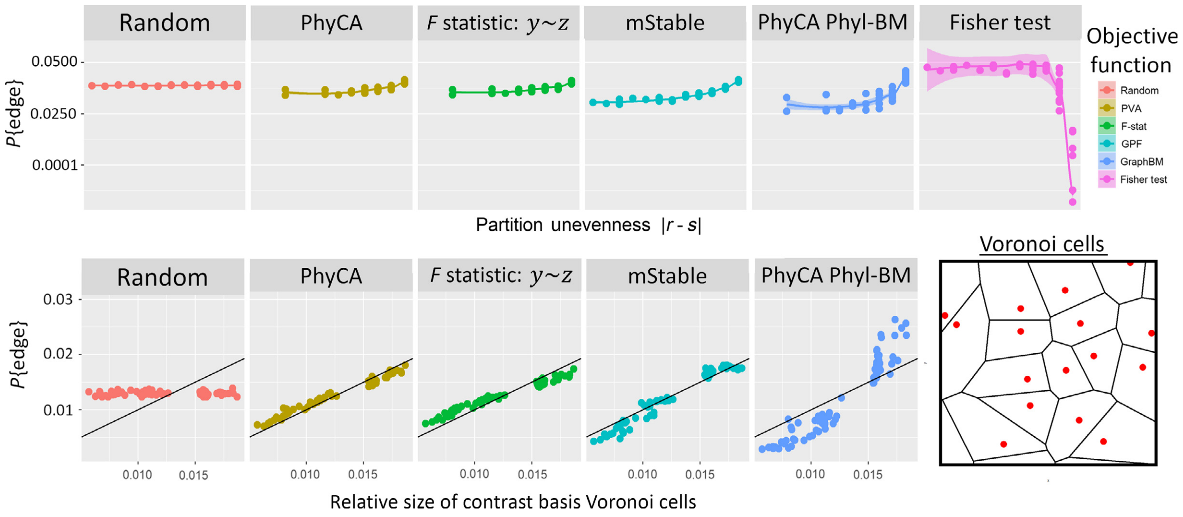


FIG. 8. Graph topological bias in null data and the relative size of Voronoi cells of contrast basis elements. The method and the null distribution of the data determine graph-topological bias of phylofactorization. A random draw of edges does not discriminate against edges based on the relative sizes of two groups contrasted by the edge, but 16,000 replicate phylofactorizations of null data reveal that contrast-basis methods are slightly biased toward uneven splits (e.g., tips of the phylogeny). Standard Gaussian null data were used for PhyCA, F statistics from regression on contrast basis elements ($y_e \sim z$), and binomial null data was used for generalized phylofactorization (GPF) through marginally stable aggregation. Other methods, such as Fisher's exact test of a vector of Bernoulli random variables, have opposite biases. The tip-bias of contrast-basis analysis is amplified for marginally-stable aggregation, and amplified even more if the null data have residual structure from a Brownian motion diffusion along the phylogeny (Phyl-BM). The common bias when using contrast bases across a range of objective functions is related to the uneven relative sizes of Voronoi cells produced by the bases, simulated here by Eq. 24.

test of whether/not there is an unusually large number of differentiating edges in mammalian body mass during or after the K-Pg extinction event.

Phylofactorization using the contrast basis is biased towards the tips of the tree. Some progress can be made towards understanding the source of basal/distal biases in phylofactorization by examining the contrast basis. The biases from analyses of contrast basis coordinates, $\mathbf{y}_e$, stem from a common feature of the set of K_t candidate basis elements $\{\mathbf{v}_{C_e}\}_{e=1}^{K_t}$ considered at iteration t of phylofactorization. In the t test phylofactorization of a vector of data, $\mathbf{x}$, the winning edge e^* is

$$e^* = \underset{e}{\mathrm{argmax}}\, |\mathbf{v}_{C_e}^T \mathbf{x}| \tag{23}$$

and thus the objective function is monotonically related to the angular distance between the vector of data and the contrast basis elements.

Since the basis elements have unit norm, each basis element corresponds to a point on an m-dimensional unit hypersphere. If the data, $\mathbf{x}$, are drawn at random, such that no direction is favored over another, the probability that a particular edge e is the winning edge is proportional to the relative size of its Voronoi cell on the surface of the unit m hypersphere. Thus, the basal/distal biases for contrast-basis analyses with null data assumed to be drawn from a random direction can be boiled down to calculating the relative sizes of Voronoi cells. For our simulations reported in Fig. 8, the size of Voronoi cells was estimated through matrix multiplication

$$\mathbf{Y}_{\mathrm{null}} = \mathbf{V}^T \mathbf{X}_{\mathrm{null}} \tag{24}$$

were $\mathbf{V}$ is a matrix whose columns j is the contrast basis elements for edge e_j being considered and $\mathbf{X}_{\mathrm{null}}$ is a null dataset whose entries are standard Gaussian random variables. Each column of $\mathbf{Y}_{\mathrm{null}}$ contains the projections of a single random vector and the element of each column with the largest absolute value is the edge closest to that random vector.

Graph-topology and confidence regions.—As a graph-partitioning algorithm, phylofactorization invites a novel description of confidence regions over the phylogeny. The graph-topology of our inferred, edges, and their proximity to other edges, both on the phylogeny and in the m-dimensional hypersphere discussed above, can be used to refine our statements of uncertainty. 95% Confidence intervals for an estimate of a real-valued quantity give bounds within which the true value is likely to fall 95% of the time in random draws of the estimate. Confidence regions are multi-dimensional extensions of confidence intervals. Conceptually, it's possible to make similar statements about phylogenetic factors, confidence regions on a graph indicating the regions in which the true, differentiating edge is likely to be.

Extending the concept of confidence regions to the graph-topological inferences from phylofactorization requires useful notions of distance and "regions" in graphs. One example of such a distance between two edges is a walking distance: the number of nodes one crosses along the geodesic path between two edges. Alternatively, one could define regions in terms of years or branch lengths. For phylofactorization using the contrast basis, confidence regions may be well-characterized by angular distances to nearby contrast basis elements and their Voronoi cells.

Defining confidence regions in any phylofactorization must combine the uneven Voronoi cell sizes and the proximity of contrast basis elements to one another. For low effect sizes, graph-topological confidence regions extend to distant edges on the graph whose contrast basis element have a large relative Voronoi cell size (e.g., the tips). As the effect sizes increase, confidence regions over the graph are better described in terms of angular distances between the contrast basis elements and that of the winning edge, e^* (Fig. 9).

Cross-validation.—How do we compare phylofactorization across data sets to cross-validate our results? If a researcher observes a pattern in the ratio of squamates to mammalian abundances in North America, say a decrease in the ratio of lizard and snake to mammal abundance with increasing altitude, they may wish to cross-validate their findings in other regions, including regions with few or none of the same species found in the original study. Researchers replicating the study in Australia and New Zealand would have to grapple with whether or not to include monotremes in their grouping of "mammals" and whether or not to include the tuatara, a close relative of squamates, in their grouping of "squamates": such branches were basal to the squamate and mammalian clades contrasted in the hypothetical North American study.

Phylofactorization formalizes the issues arising with such phylogenetic cross-validation (Fig. 10). If all species in the training/testing data sets can be located on a universal phylogeny, phylofactorization of a training data set identifies edges or links of edges in the training phylogeny which are guaranteed to correspond to edges or links of edges in the universal phylogeny. New species in the testing data set may introduce new edges to the phylogeny which interrupt the links of edges in the universal phylogeny along which factors were found in the training data. In the example above, the tuatara and monotremes all interrupt the link of edges separating North American mammals from North American reptiles on the universal phylogeny.

Cross-validating phylofactorization requires addressing the issues arising from the interruptions of edges produced by novel species. Interruptions may be ignored or used to refine the location of a factor on the universal tree by placing the interrupting clade into one of the two groups contrasted at that factor. Returning to the previous example, one can use the presence of monotremes and tuatara to refine the definition of North American mammals to

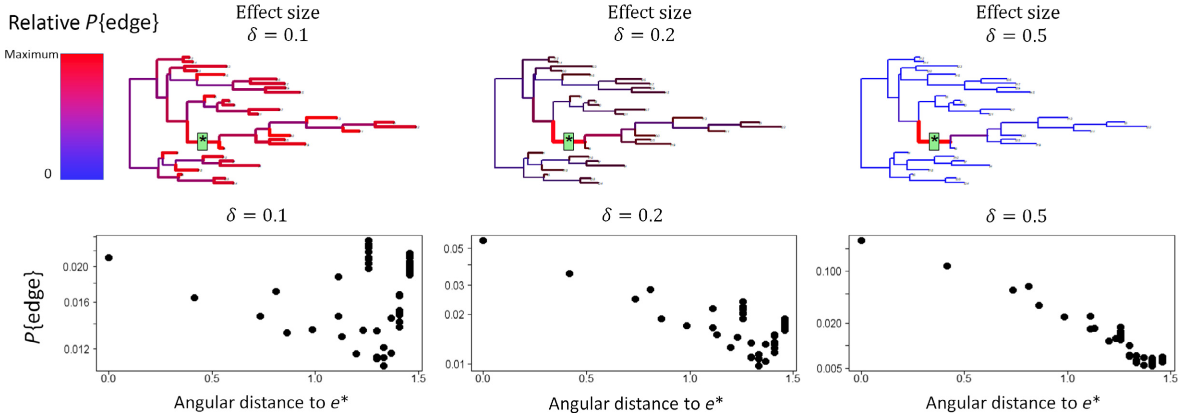


FIG. 9. Graph-topological confidence regions for phylofactorization. Confidence regions around inferred edges must use distances relevant to the method and graph topology. A tree with 30 species and 10 samples was given a fixed effect about edge e*. The effects were an association with meta data, z, modeled as $x_{i,j} = (\pm\delta/2)z_j + \epsilon_{i,j}$ where $\varepsilon_{i,j}$ and z_j are i.i.d. standard Gaussian random variables. A total of 7 × 10^5 iterations of regression phylofactorization on $\mathbf{y}_e$ were run and the relative probability of drawing each edge was visualized through both the color and width of the edge. The relationship between the angular distance of an edge's contrast basis element to that of e* and the probability of drawing the edge indicate that for low effects, confidence regions must incorporate a mix of tip-bias and angular distance, but for larger effect sizes, in which the edge drawn is reliably in the neighborhood of e*, the angular distance of contrast basis elements capture confidence regions around the location of inferred phylogenetic factors.

mean "all mammals" or "all placental and marsupial mammals", and likewise one can refine the definition of "squamates" to the broader "Lepidosauria" clade.

Stopping criteria.—For computational and conceptual purposes, it's desirable to obtain a minimal set of partitions to prioritize findings, simplify high-dimensional data, and focus downstream effort on more certain inferences. Doing so requires a method for stopping phylofactorization as the algorithm is running. There are two broad options for stopping phylofactorization: null simulations and stopping criteria for a running algorithm. Null simulations may allow statistical statements stemming from a clear null model, but stopping criteria can be far more computationally efficient.

Washburne et al. (2017) proposed a stopping criterion for regression phylofactorization which extends to all methods of phylofactorization using an objective function whose null-distribution for a single edge is known. The original stopping criterion is based on the fact that, if the null hypothesis is true, the distribution of P values from multiple hypothesis tests is uniform. Phylofactorization performs multiple hypothesis tests at each iteration. At each iteration, one can perform a one-tailed Kolmogorov-Smirnov (KS) test on the uniformity of the distribution of the P values from the test statistics on each edge; if the KS test is nonsignificant, stop phylofactorization. KS test stopping criteria can conservatively stop simulations at the appropriate number of factors when there is a discrete subset of edges with effects. Such a method performs similarly to Horn's stopping criterion for factor analysis (Horn 1965), whereby one stops factorization when the scree plot from the data crosses that expected from null data (Fig. 11). One can also use a stopping criterion and subsequently run null simulations to understand the likelihood of observed results under a null model of the researcher's choice (Fig. 11). Other stopping criteria may outperform the KS test, such as using bonferroni cutoffs or sequentially-rejective cutoffs, stopping the algorithm when the lowest P value falls above the cutoff for a desired family-wise error rate or false-discovery rate.

Calibrating statistical tests for ω_{e^*}.—Often, the objective function for phylofactorization is a well understood test statistic. Applying a standard test for the winning test statistic, however, will lead to a high false-positive rate and an overestimation of the significance of an effect, because the winning statistic was drawn as the best of many. Even when using a test statistic not equal to the objective function, researchers should be cautious of dependence between their test statistic and the objective function as a possible source of high false-positive rates. Two methods for calibrating statistical tests of $\omega_{e^*}$are multiple-comparisons corrections to control a family-wise error rate or false-discovery rate, and conservative bounds on the distribution of the maximum of many independent, identically distributed statistics. For example, if each edge of one of the K_t edges considered at iteration t resulted in an independent F statistic, F_e, then the distribution of the maximum F statistics, F_{e^*}, is

$$\begin{aligned} P\{F_{e^*} > F\} &= P\{F_{e_1} > F \cap F_{e_2} > F \cap \ldots \cap F_{e_K}\} \\ &= P\{F_e > F\}^{K_t}. \end{aligned} \tag{25}$$

Such an approximation may be used to yield conservative estimates, but the F statistics are not independent

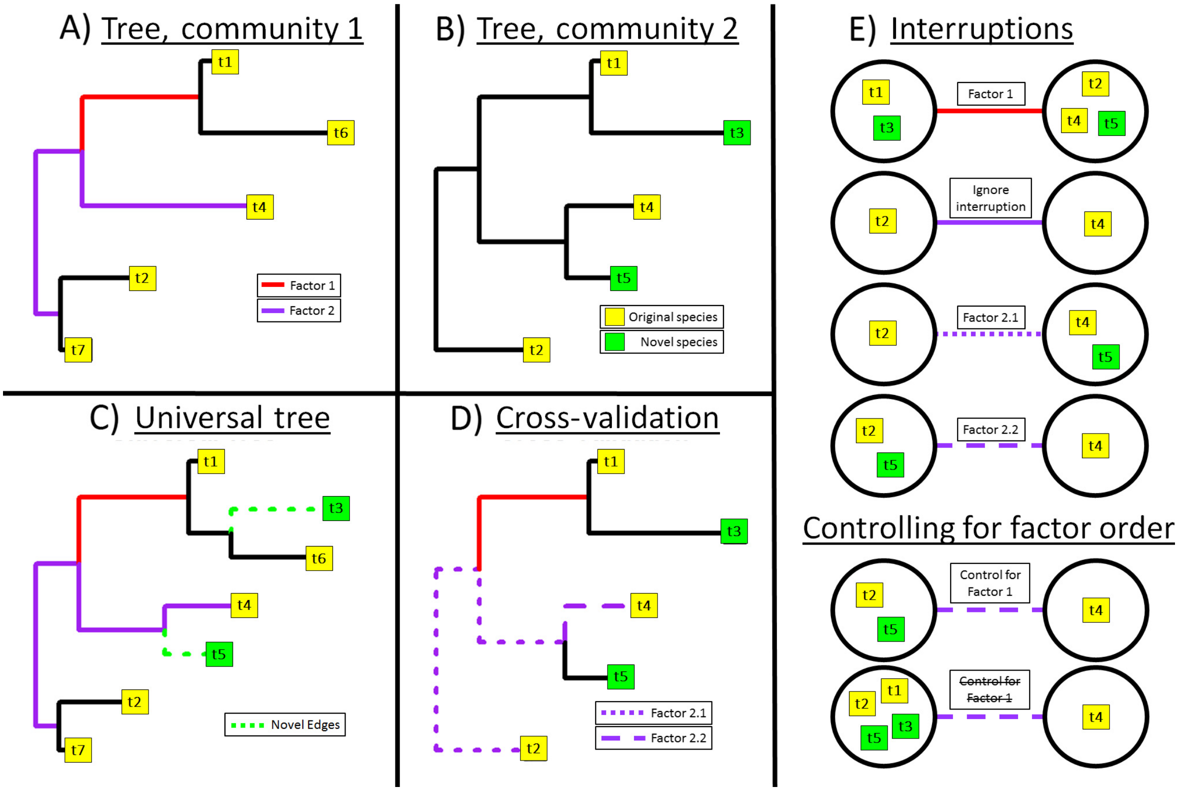


FIG. 10. Graph-topological considerations with cross-validation. (A) The training community has five species (yellow boxes) split into two factors. The first factor partitions {t1,t6} from {t2,t4,t7}. The second factor partitions t4 from {t2,t7}. The second factor does not correspond to a single edge, but instead a chain of two edges. (B) A second, testing community is missing species t6 and t7 and contains novel species t3 and t5 (green boxes). (C) All factors can be mapped to chains of edges on a universal phylogeny. Novel species "interrupt" edges in the original tree; cross-validation requires deciding what to do with novel species and interrupted edges. Species t3 does not interrupt a factored edge, and so t3 can be reliably grouped with t1 in factor 1. However, species t5 interrupts one of the edges in the edge-path of factor 2. (D,E) Interruptions can be ignored, or they can be used to refine the location of important edges (illustrated in Factor 2.1 and Factor 2.2). Another topological and statistical question is whether/not to control for factor order. For instance, controlling for factor order with Factor 2.2 would partition t4 from {t2,t5}. Not controlling for factor order would partition t4 from {t1,t2,t3,t5}.

and thus more nuanced analyses are needed for well calibrated statistical tests. Unpublished simulations suggest that the order statistics of Eq. 25 break down for downstream factors. More research is needed to obtain conservative bounds on test-statistics in phylofactorization.

Summary of limitations.—Phylofactorization can be a reliable statistical tool with a careful understanding of the statistical challenges inherent in the method and shared with related methods such as graph partitioning, greedy algorithms, factor analysis, and the use of a constrained basis for matrix factorization. Phylofactorization is an exploratory tool, but all exploratory tools can be made inferential with suitable understanding of their behavior under an appropriate null model. For example, principal components analysis was and still is primarily an exploratory tool, but the discovery of the Marcenko-Pastur distribution (Marčenko and Pastur 1967) has improved the calibration of statistical tests on principal components for standardized, mean-centered data. Improved understanding of how uncertainties in phylogenetic inference translate to uncertainties in phylofactorization, conservative stopping criteria, null distributions of test statistics for winning edges, propagation of error, graph-topological biases, and confidence regions on a graph can all improve the reliability of phylofactorization as an inferential tool.

While phylofactorization was built with an evolutionary model of punctuated equilibria in mind, it may also work well under other evolutionary models such as correlated evolution among descendants of an edge. There are also many evolutionary models under which phylofactorization does not perform well. For instance the graph-topological biases of PhyCA are increased under a Brownian motion model of evolution. All statistical tools operate well under appropriate assumptions, and understanding the assumptions, as well as the known limitations, are necessary for responsible and academically fruitful use of statistical tools like phylofactorization. Diagnostic tools to visualize and analyze the appropriateness of a phylofactorization, such as those used to test

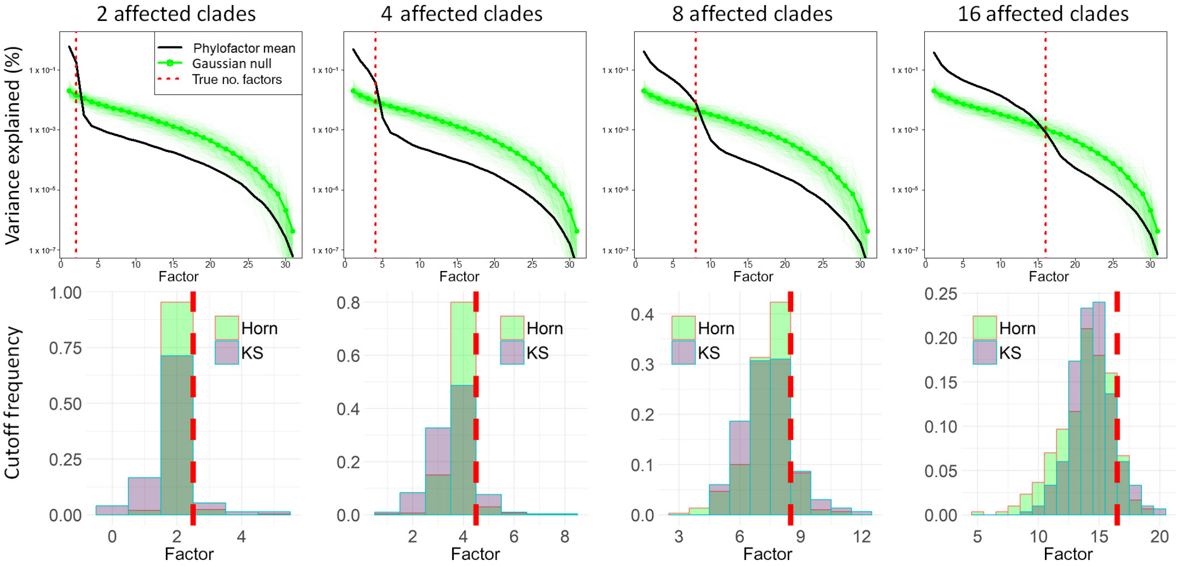


Fig. 11. Null simulations and stopping criteria. A challenge of phylofactorization is determining the number of factors to include in an analysis. Null simulations allow quantile-based cutoffs such as those in Horn's parallel analysis from factor analysis. Stopping criteria use features available during phylofactorization to stop a running algorithm. Abundances of m = 32 species across n = 10 samples were simulated as i.i.d. standard Gaussian random variables. A set of u clades were associated with environmental meta data, $\mathbf{z}$, where $z_j \overset{i.i.d.}{\sim} \mathrm{Gsn}(0,1)$. Regression-phylofactorization on the contrast-basis scores y_e was performed on 300 data sets for each $u \in \{2, 4, 8, 16\}$ and on data with and without effects. The objective function was the total variance explained by regression $y_e \sim z$. The top row shows the percentage of the variance in the data set explained at each factor (EV) decreases with factor, t, and the mean EV curve for data with u affected clades intersects the mean EV curve for null data near $t = u$, motivating a stopping criterion (Horn) based on phylofactorization of null data sets. The bottom rows shows that the two algorithms did not have extremely different rates of over-factorization. Both criteria can be modified to be made more conservative. The KS stopping criterion is far less computationally intensive for large data sets as it requires running phylofactorization only once. Null simulations, however, can allow inferential statistical statements regarding the null distribution of test statistics in phylofactorization.

heteroskedasticity and leverage in generalized linear models, can greatly improve the robustness of analyses.

Discussion

Early physicists studying patterns of projectile motion were not plagued with challenges of how to group matter to define a bowling ball. Community ecologists, however, must grapple with the challenge of how to group organisms into units for experiments, modeling, analysis, and management. As a starting point, it's often used to group organisms based on some measure of high within-group similarity and between-group differences. Ecological patterns are determined by organisms' interactions with the biotic and abiotic conditions of their environment; such interactions are determined by traits. Functional ecological traits thus underlie many observed patterns in ecology and, where an ecological pattern of interest is associated with heritable traits, the phylogeny provides a scaffold for functional groupings of organisms with a common role in the ecological pattern of interest.

Traits arise along edges in the phylogeny. Contrasting taxa on opposing sides of an edge allows one to uncover sets of species that are meaningfully different and whose differences may be due to heritable traits. By noting that each edge partitions the phylogeny into two disjoint sets of species, by generalizing the operations of aggregation and contrast, and by defining the objective function of interest, we have developed a universal method for identifying the relevant phylogenetic scales underlying ecological patterns in community ecological data sets.

Phylofactorization is a graph-partitioning algorithm which can use community ecological data to separate the phylogeny into binned phylogenetic units with high within-group similarity and high between-group differences. For a vector of data, two-sample tests are a natural method for making such partitions. The quantities used in two-sample tests can be extended to larger, real-valued data sets by analyzing a contrast basis. Objective functions for choosing the appropriate contrast basis include maximizing variance, a phylogenetic analog of principal components analysis, maximizing explained variance from regression, maximizing F statistics from regression, and more. For regression on community ecological data assumed to be exponential family random variables, phylofactorization can be extended to generalized linear models, generalized additive models, and analyses of spatial and temporal patterns in ecological data by use of `phylo` factor contrasts and marginally stable aggregation within the exponential family. All algorithms discussed here can be extended to analysis of spatially and temporally explicit ecological patterns.

We've illustrated that two-sample tests can partition a data set of mammalian body mass into groups with very different average body masses. Maximizing the variance of data projected onto contrast basis elements can identify major clades of bacteria in human feces known at a coarser resolution to be highly variable. Additionally, one of the top phylogenetic factors in the American gut data set is a clade of Gammaproteobacteria associated with inflammatory bowel disease (IBD) used recently in an effort to diagnose patients with Crohn's disease. We've shown that analyses of contrast bases can use nonlinear regression, sorting 3,000 species into five binned phylogenetic units producing a simplified story of nonlinear community compositional changes in Central Park soils across a gradient of pH, carbon concentrations, and nitrogen concentrations.

One can also perform phylofactorization when doing maximum-likelihood regression of exponential family random variables. The coefficient matrix can be approximated using the contrast basis, resulting in a phylogenetically interpretable reduced-rank regression. Alternatively, `phylo` factor contrasts specify a shared-coefficients model and which can be used to select edges based explicitly on likelihood maximization of a shared coefficients model. One can perform the factor contrasts on the raw data, or, for many exponential family random variables, aggregate the data within each group to a marginally stable distribution for more computationally efficient factor contrasts. All methods discussed here can be implemented with the R package phylofactor, and scripts for running all analyses in this paper are available on Zenodo (see Data Availability).

As with any method, there are limitations to be aware of. First, the general problem of separating species into k bins that maximize a global objective function is NP hard. Second, like any greedy algorithm, phylofactorization may fall into ruts and errors in one step that might propagate into downstream inferences. Third, the null distribution of test-statistics resulting from phylofactorization is not known; the resultant test statistics are biased towards extreme values. Null simulations, conservative stopping functions, and/or extremely stringent multiple comparisons corrections can be used to make inferences through phylofactorization while maintaining conservative bounds in family-wise error rates or false-discovery rates. When the objective function being maximized has a well-characterized null distribution for a single edge, one-sided KS-tests of the P values of the test statistic can serve as a computationally efficient and conservative stopping function. Fourth, common objective functions using the contrast basis will be biased due to the unequal relative sizes of the Voronoi cells of the contrast basis elements in the unit hypersphere in which they lie; contrast basis elements corresponding to tips of the phylogeny tend to have larger relative Voronoi cell size than contrast basis elements corresponding to interior edges. Understanding the graph-topology of errors can assist the description of graph-topological confidence regions for each inference. Finally, phylofactorization formalizes the logic and challenges of cross-validating ecological comparisons even when the training and testing sets of species are completely disjoint. Many of these limitations may be resolved with future work, allowing the exploratory algorithm to become a fast, well-calibrated inferential tool.

Phylofactorization can objectively identify phylogenetic scales for ecological data and produce avenues for future natural history research. By iteratively identifying clades, phylofactorization provides a sequence of low-rank approximations of a data set that correspond to groups of species with a shared evolutionary history. What traits characterize the Chloracidobacteria which don't like acidic soils? What traits characterize the monophyletic clade of Gammaproteobacteria associated with IBD? What ecological or immunological mechanisms underlie the *Prevotella* species' variability in the American gut? The low-rank approximations of ecological data obtained by phylofactorization motivate subsequent questions best answered by life history comparisons, comparative genomics, physiological studies, and other avenues of future research contrasting the species partitioned.

Relation to other phylogenetic methods.—Phylofactorization is proposed during an explosion of literature in phylogenetic comparative methods and various other phylogenetic methods for analyzing ecological data sets (Lozupone and Knight 2005, Purdom 2011, Garamszegi 2014), and some careful thinking is beneficial to clarify the distinctions between phylofactorization and other methods.

Phylogenetic generalized least squares (Grafen 1989) aims to control for residual structure in the response variable expected under a model of trait evolution, and is thus used when performing regression on a trait, whereas phylofactorization aims to partition observed trait values or abundances into groups, separated by edges, with different means or associations with meta data. Thus, while methods of phylogenetic signal, such as Pagel's λ (Pagel 1999) or Blomberg's κ (Blomberg et al. 2003), summarize global patterns of phylogenetic signal by parameterizing the extent to which a particular model of evolution can be assumed to underlie the residual structure of observed traits (often for downstream use in PGLS), phylofactorization iteratively identifies precise locations of putative changes and precise locations partitioning phylogenetic signal or structure.

Phylofactorization can be implemented by a contrast of ancestral state reconstructions of nodes separated by edges, for example by looking for edges with nodes whose reconstructed ancestral states are most different, but is limited by disallowing the descendant clade of an edge to impact the ancestral state of the edge's basal node; a proper non-overlapping contrast would separate the groups of species being used to reconstruct each node, and thus phylofactorization can be implemented with ancestral state reconstruction under the assumption of time-reversible evolutionary models.

Phylogenetically independent contrasts (PIC; Felsenstein 1985*b*) produces variables corresponding to contrasts of descendants from each node, whereas phylofactorization uses contrasts of species separated by an edge, picks out the best edge, splits the tree, and repeats. The contrasts used in PIC for comparison of sister clades (standardized differences of means) can be used as the contrast function for phylofactorization to identify edges with standardized differences of means that maximize some objective function. While the contrast basis proposed here is fixed regardless the observed data across samples, PIC divides the difference of group means by empirically observed standard deviations for each sample. Consequently, the contrasts from PIC can be used as a contrast function but can't be interpreted as a projection of the data onto a fixed basis.

Phylofactorization develops a set of variables and an orthonormal basis to describe ecological data, but limits itself to bases interpretable as non-overlapping contrasts along edges; eigenvectors of phylogenetic distances matrices or covariance matrices under diffusion models of traits (Pagel 1999), are not encompassed in phylofactorization as they do not construct non-overlapping contrasts along edges. Such eigenvector methods construct quantities whose evolutionary and functional ecological interpretation is less clear. Unlike many modern methods for redefining distances, such as UniFrac distances (Lozupone and Knight 2005) or phylogenetically defined inner products (Purdom 2011), phylofactorization is principally about discovering phylogenetically interpretable directions: contrast basis vectors that characterize primary axes of variation in the community or a basis made of aggregations of the binned phylogenetic units.

R package: phylofactor.—An R package is in development and publicly available (see Data Availability). The R package contains detailed help functions and supports flexible definition of two-sample tests (the function `twoSampleFactor`), contrast-basis analyses with the function `PhyloFactor`, and generalized phylofactorization with the function `gpf`. Phylofactorization is highly parallelizable, and the R package functions have built-in parallelization. The R package also works with phylogenies containing polytomies, allowing researchers to collapse clades with low bootstrap support to make more robust inferences. The output from phylofactorization is a "phylofactor" object containing the contrast basis and other useful features, allowing one to input the object into various functions which summarize, plot, cross-validate and do other tricks to parse out the information from phylofactorization. Researchers are invited to beta-test the package and contact the first author Alex Washburne for assistance with the package, including how to produce their own customized phylofactorizations. Such feedback will be invaluable for a user-friendly stable release to CRAN.

Until then, the Supporting Information contains the data and scripts used for all analyses done in this manuscript along with a tutorial for the R package in an effort to accelerate method development in this field.

Everything makes sense in light of evolution.—Phylogenetic factorization is a new paradigm for analyzing a large class of biological data. Ecological data, as Thomas Dhobzansky noted about biology in general, makes sense "in light of evolution." Phylofactorization connects a broad set of data analyses—two sample tests, generalized linear modeling, factor analysis and PCA, and analysis of spatial and temporal patterns—to a natural set of variables and operations defined by the phylogeny. Phylofactorization localizes inferences to particular edges or chains of edges on the phylogeny and, in so doing, accelerates our understanding of the phylogenetic scales underlying ecological patterns of interest. The problem of pattern and scale is central to biology, and phylofactorization uses flexible definitions of patterns to objectively uncover the relevant phylogenetic scales in ecological data sets.

Acknowledgments

This work is published in loving memory of Diana Nemergut. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA; D16AP00113).

Literature Cited

Aitchison, J. 1982. The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological) 44:139–177.

Alroy, J. 1998. Cope's rule and the dynamics of body mass evolution in North American fossil mammals. Science 280:731–734.

Alroy, J. 1999. The fossil record of North American mammals: evidence for a Paleocene evolutionary radiation. Systematic Biology 48:107–118.

Baker, J., A. Meade, M. Pagel, and C. Venditti. 2015. Adaptive evolution toward larger size in mammals. Proceedings of the National Academy of Sciences USA 112:5093–5098.

Blomberg, S. P., T. Garland Jr., A. R. Ives, and B. Crespi. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745.

Buluç, A., H. Meyerhenke, I. Safro, P. Sanders, and C. Schulz. 2016. Recent advances in graph partitioning. Pages 117–158 *in* L. Kliemann, and P. Sanders, editors. Algorithm engineering: Lecture Notes in Computer Science. Springer, Cham, Switzerland.

Clemente, J. C., L. K. Ursell, L. W. Parfrey, and R. Knight. 2012. The impact of the gut microbiota on human health: an integrative view. Cell 148:1258–1270.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology 72:5069–5072.

Egozcue, J. J., and V. Pawlowsky-Glahn. 2005. Groups of parts and their balances in compositional data analysis. Mathematical Geology 37:795–828.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal. 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35:279–300.

Farrior, C. E., R. Dybzinski, S. A. Levin, and S. W. Pacala. 2013. Competition for water and light in closed-canopy forests: a tractable model of carbon allocation with implications for carbon sinks. American Naturalist 181:314–330.

Felsenstein, J. 1985*a*. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

Felsenstein, J. 1985*b*. Phylogenies and the comparative method. American Naturalist 125:1–15.

Garamszegi, L. Z. 2014. Modern phylogenetic comparative methods and their application in evolutionary biology. *In* Concepts and practice. Springer, London, UK. https://link.springer.com/book/10.1007%2F978-3-662-43550-2

Gould, N. E.-S. J. 1972. Punctuated equilibria: an alternative to phyletic gradualism. Pages 82–115 *in* F. J. Ayala, and J. C. Avise, editors. Essential readings in evolutionary biology. JHU Press, Baltimore, Maryland, USA.

Grafen, A. 1989. The phylogenetic regression. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 326:119–157.

Graham, C. H., D. Storch, and A. Machac. 2018. Phylogenetic scale in ecology and evolution. Global Ecology and Biogeography 27:175–187.

Hall, B. G., and M. Barlow. 2004. Evolution of the serine β-lactamases: past, present and future. Drug Resistance Updates 7:111–123.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Horn, J. L. 1965. A rationale and test for the number of factors in factor analysis. Psychometrika 30:179–185.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography (MPB-32). Princeton University Press, Princeton, New Jersey, USA.

Jerrum, M., and G. B. Sorkin. 1998. The metropolis algorithm for graph bisection. Discrete Applied Mathematics 82:155–175.

Jones, K. E., et al. 2009. Pantheria: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology 90:2648.

Katz, Y., K. Tunstrøm, C. C. Ioannou, C. Huepe, and I. D. Couzin. 2011. Inferring the structure and dynamics of interactions in schooling fish. Proceedings of the National Academy of Sciences USA 108:18720–18725.

Landis, M. J., J. G. Schraiber, and M. Liang. 2012. Phylogenetic analysis using lévy processes: finding jumps in the evolution of continuous traits. Systematic Biology 62:193–204.

Levin, S. A. 1992. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. Ecology 73:1943–1967.

Ley, R. E., P. J. Turnbaugh, S. Klein, and J. I. Gordon. 2006. Microbial ecology: human gut microbes associated with obesity. Nature 444:1022–1023.

Lozupone, C., and R. Knight. 2005. Unifrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71:8228–8235.

Marčenko, V. A., and L. A. Pastur. 1967. Distribution of eigenvalues for some sets of random matrices. Mathematics of the USSR-Sbornik 1:457.

Mariat, D., O. Firmesse, F. Levenez, V. Guimarães, H. Sokol, J. Doré, G. Corthier, and J. Furet. 2009. The firmicutes/bacteroidetes ratio of the human microbiota changes with age. BMC Microbiology 9:123.

McDonald, D., et al. 2018. American gut: an open platform for citizen science microbiome research. mSystems 3:e00031-18.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21:1087–1092.

Michonneau, F., J. W. Brown, and D. J. Winter. 2016. rotl: an R package to interact with the open tree of life data. Methods in Ecology and Evolution 7:1476–1481.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Paradis, E., J. Claude, and K. Strimmer. 2004. Ape: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Plowright, R. K., C. R. Parrish, H. McCallum, P. J. Hudson, A. I. Ko, A. L. Graham, and J. O. Lloyd-Smith. 2017. Pathways to zoonotic spillover. Nature Reviews Microbiology 15:502–510.

Purdom, E. 2011. Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. Annals of Applied Statistics 5:2326–2358.

Ramirez, K. S., et al. 2014. Biogeographic patterns in below-ground diversity in New York City's central park are similar to those observed globally. Proceedings of the Royal Society B: Biological Sciences 281:20141988.

Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3:217–223.

Sato, K.-i. 1999. Lévy processes and infinitely divisible distributions. Cambridge University Press, Cambridge, UK.

Scher, J. U., et al. 2013. Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. Elife 2:e01202.

Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

Silverman, J. D., A. D. Washburne, S. Mukherjee, and L. A. David. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. Elife 6:e21887.

Smith, F. A., and S. K. Lyons. 2011. How big should a mammal be? A macroecological look at mammalian body size over space and time. Philosophical Transactions of the Royal Society of London B: Biological Sciences 366:2364–2378.

Smith, F. A., et al. 2010. The evolution of maximum body size of terrestrial mammals. Science 330:1216–1219.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1131.

Vázquez-Baeza, Y., A. Gonzalez, Z. Z. Xu, A. Washburne, H. H. Herfarth, R. B. Sartor, and R. Knight. 2017. Guiding longitudinal sampling in IBD cohorts. Gut 67:1743–1745.

Washburne, A. D., J. W. Burby, and D. Lacker. 2016. Novel covariance-based neutrality test of time-series data reveals asymmetries in ecological and economic systems. PLoS Computational Biology 12:e1005124.

Washburne, A. D., J. D. Silverman, J. W. Leff, D. J. Bennett, J. L. Darcy, S. Mukherjee, N. Fierer, and L. A. David. 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ 5:e2969.

Washburne, A. D., J. T. Morton, J. Sanders, D. McDonald, Q. Zhu, A. M. Oliverio, and R. Knight. 2018. Methods for phylogenetic analysis of microbiome data. Nature Microbiology 3:652.

Yee, T. W., and T. J. Hastie. 2003. Reduced-rank vector generalized linear models. Statistical Modelling 3:15–41.

Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8:28–36.

Zhou, X., S. Xu, J. Xu, B. Chen, K. Zhou, and G. Yang. 2011. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the laurasiatherian mammals. Systematic Biology 61:150–164.

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecm.1353/full

Data Availability

Additional data are available on Zenodo: https://doi.org/10.5281/zenodo.1490224. The R package phylofactor is available at https://github.com/reptalex/phylofactor.