Syst. Biol. 65(3):495–507, 2016

DOI:10.1093/sysbio/syv116
Advance Access publication December 12, 2015

Characterizing and Comparing Phylogenies from their Laplacian Spectrum

ERIC LEWITUS* AND HELENE MORLON

Institut de Biologie (IBENS), École Normale Supérieure, Paris, France;
**Correspondence to be sent to: Institut de Biologie (IBENS), École Normale Supérieure, Paris, France; E-mail: lewitus@biologie.ens.fr.*

Received 23 March 2015; reviews returned 14 July 2015; accepted 4 December 2015
Associate Editor: Olivier Gascuel

Abstract.—Phylogenetic trees are central to many areas of biology, ranging from population genetics and epidemiology to microbiology, ecology, and macroevolution. The ability to summarize properties of trees, compare different trees, and identify distinct modes of division within trees is essential to all these research areas. But despite wide-ranging applications, there currently exists no common, comprehensive framework for such analyses. Here we present a graph-theoretical approach that provides such a framework. We show how to construct the spectral density profile of a phylogenetic tree from its Laplacian graph. Using ultrametric simulated trees as well as non-ultrametric empirical trees, we demonstrate that the spectral density successfully identifies various properties of the trees and clusters them into meaningful groups. Finally, we illustrate how the eigengap can identify modes of division within a given tree. As phylogenetic data continue to accumulate and to be integrated into various areas of the life sciences, we expect that this spectral graph-theoretical framework to phylogenetics will have powerful and long-lasting applications. [biodiversity; diversification; graph theory; influenza; Laplacian; macroevolution; microbiology; phylodynamics; phylogenetics]

Phylogenies are essential to many areas of the life sciences. In population genetics and phylogeography, they are used to infer past demography and historical migration events (Avise 2000). In epidemiology, they are key to understanding how best to control the spread of infectious disease (Popinga et al. 2015). In microbiology, they provide one of the most natural and powerful measures of diversity (Lozupone and Knight 2008). Phylogenies are also increasingly effective in ecology, where they can inform our understanding of community assembly (Webb et al. 2002), interspecific interactions (Rezende et al. 2007), and species responses to environmental change (Condamine et al. 2013), as well as guide conservation efforts (Faith 1992; Purvis et al. 2005). Finally, phylogenies are essential to comparative phylogenetics (Pennell and Harmon 2013) and comparative genomics (Burki 2014) and therefore to our understanding of diversification (Morlon 2014), trait evolution (Harmon et al. 2010), and the genetic underpinnings of both (e.g., Lewitus and Huttner 2015).

Despite the importance of phylogenetics in the life sciences, the current techniques aimed at extracting information from tree shapes (i.e., unlabeled phylogenies) are limited. One of these techniques is built on summary statistics. In microbiology, ecology, and conservation biology, summary statistics based on measures of phylogenetic diversity, such as total phylogenetic branch length (Faith 1992; Cadotte et al. 2010), are often used. In diversification analyses, traditional summary statistics quantify either the stem-to-tip (e.g., γ Pybus and Harvey 2000 and Lineage-Through-Time plots Nee et al. 1992) or lineage-to-lineage (e.g., β and the Colless index Blum and François 2006) distribution of branching events across trees. These summary statistics disregard much of the data — and therefore the biological information — encoded in trees: they are simply too crude to precisely capture the complexity of events recorded in empirical trees. Recent computational and conceptual advances based on maximum-likelihood techniques have been able to take better advantage of the full sweep of information provided by empirical trees. Accordingly, they have become the yardstick for determining how clades and traits behave over evolutionary time (Pennell and Harmon 2013; Garamszegi 2014; Morlon 2014), the selection pressures acting on different genes (Kosakovsky Pond et al. 2011), and changes in rates of infection as a function of time (Vijaykrishna et al. 2014). However, all such model-based approaches rely on the *a priori* formulation of a model, which can be problematic, because we cannot exhaustively model the many dynamics potentially generating all empirical trees. Finally, both the summary-statistics and the model-based approaches mentioned above are limited to the analysis of ultrametric trees (i.e., trees in which the distances from the root to every tip are equal), therefore limiting their domain of applicability. In this article, we introduce an approach to phylogenetics that does not require any *a priori* assumption about how the phylogeny behaves and can be applied to ultrametric as well as non-ultrametric trees.

We develop an approach based on spectral graph theory that allows a systematic characterization and comparison of the entirety of information encoded in tree shapes. In various configurations, graph theory has been successful in understanding the organizing principles behind biological phenomena at every scale, including the regulation of gene expression (Shen-Orr et al. 2002), protein–protein interactions (Szklarczyk et al. 2015), metabolic networks (Ravasz et al. 2002), and ecological food webs (Dunne et al. 2002). Graph theory and associated spectral analyses have also been useful in phylogenetics, particularly in developing approaches for tree inference (Chen et al. 2007) or for comparing the phylogenetic composition of microbial samples (Matsen IV and Evans 2013). Metrics like the

Robinson–Foulds distance (Robinson and Foulds 1981) and nearest neighbor interchange (Moore et al. 1973), too, for example, are used to compare different trees representing the same set of organisms by counting the number of steps needed to transform one into the other (or both into a third); while others take a geometric approach to define polytopic contours around a reconstructed tree in order to define "confidence regions" in the tree (Billera et al. 2001). Typically, such distance metrics have been used to identify outliers among or discordance between gene trees, in order to derive a consensus tree or define the "space" that a set of gene trees occupies (Hillis et al. 2005; Matsen 2006). They are not, however, built (or adapted) to function as comparative metrics between species trees representing different sets of organisms. Hence, despite the utility of characterizing and comparing tree shapes sampled from different species trees for understanding general principles in the evolution of biological systems, there exists no graph-theoretical approach designed to do so.

The approach we develop here also provides a way to identify distinct modes of division within a tree, which may, for example, reflect distinct modes and/or rates of diversification. Previous attempts in this direction have focused on identifying shifts in diversification rates under a presumed model of diversification. These can, among other things, examine distributions in species richness across the tips of a tree or use other types of imbalance measures (Agapow and Purvis 2002; Chan and Moore 2002). More recently, methods such as MEDUSA (Alfaro et al. 2009) and BAMM (Rabosky 2014) have been developed to detect the location(s) of rate shifts on phylogenies in a likelihood or Bayesian framework, whereas other methods, based on non-parametric comparisons of branch-length distributions between subclades, identify shifts in rates as well as modes of diversification (Shah et al. 2013). The latter approach, however, has been implemented only for pairwise comparisons and is therefore not suited for exploring multiple possible modes of division in trees. Furthermore, all above-mentioned approaches are limited to the analysis of ultrametric trees.

In the current work, we describe how to construct the spectral density of phylogenetic trees and demonstrate how to interpret this density in terms of specific properties of the trees. We show how to compute the distance between trees based on their spectral densities and how to identify distinct modes of division within individual trees. We use simulations to demonstrate that spectral densities cluster phylogenetic trees into meaningful classes and can identify meaningful modes of division within trees. We illustrate the unique utility of this approach for testing hypotheses on non-ultrametric trees by analyzing different influenza strains as well as an archaeal tree. Finally, we discuss potential extensions of the approach with implications for the study of community ecology, macroevolution, microbiology, and epidemiology.

Materials and Methods

Implementation

Below, we describe how to construct the spectral density profile of a phylogenetic tree, how to compute the spectral distance between trees, and how to cluster trees based on this distance. We also describe how to identify modalities within a given phylogenetic tree and to compute associated support values. We implemented these functionalities in the R package *RPANDA* freely available on CRAN (Morlon et al. 2015).

Construction of the Spectral Density

Our goal is to provide a common, comprehensive framework for characterizing phylogenetic trees, comparing them, and identifying distinct branching patterns within them. Given a (potentially labeled) phylogenetic tree, we discard its labeling and consider the resulting tree shape as a particular kind of graph, G = (N,E,w), composed of nodes (N) representing extant and ancestral species, edges (E) delineating the relationships between nodes, and a weight function (w) defining the phylogenetic distances between nodes. We consider fully resolved (i.e., bifurcating) trees throughout for illustrative purposes, but our framework is equally applicable to unresolved trees (i.e., displaying polytomies). We consider trees with explicit branch lengths, but trees with only topological information could be analyzed using a weight function of 1 for each edge. The framework is equally applicable to ultrametric and non-ultrametric trees, as illustrated below in our empirical applications.

We begin by constructing the modified graph Laplacian (MGL) of a phylogenetic tree, defined as the difference between its degree matrix (the diagonal matrix where diagonal element i is the sum of the branch lengths from node i to all the other nodes in the phylogeny) and its distance matrix (where element (i,j) is the branch-length between nodes i and j) (Fig. 1, Supplemental Fig. 1). Each row and column therefore sums to zero. For a rooted tree with n tips, there are $N=2n-1$ nodes, and the MGL is a N x N matrix. The graph Laplacian is said to be "modified" insofar as it takes a distance matrix, rather than an adjacency matrix, as its subtrahend. We also consider a normalized version of the MGL (nMGL), defined as the MGL divided by its degree matrix. The nMGL emphasizes phylogeny shape at the expense of size, which can be useful for comparing phylogenies on considerably different time scales. There is a wealth of knowledge on the MGL that we may draw from the physical sciences (Mohar 1997). In particular, the MGL is a positive semidefinite matrix, meaning that it has N non-negative eigenvalues, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{N-1} \geq \lambda_N \geq 0$. Each of these λ reflects the connectivity of the tree — in terms of both density of nodes and weights — in a particular neighborhood of the tree (Noh and Rieger 2004). Large λ are characteristic of sparse neighborhoods (few nodes) typical of deep branching events, whereas

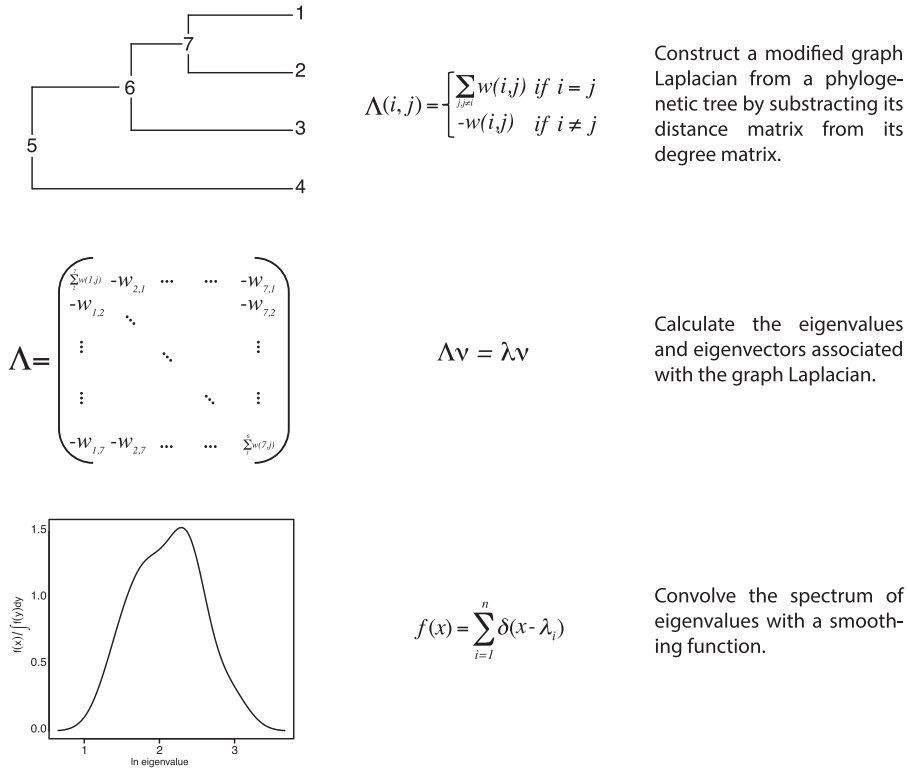

FIGURE 1. Pipeline for constructing the spectral density of a phylogenetic tree. Graphical depictions (column 1), equations (column 2), and brief descriptions (column 3) for each step in constructing the spectral density are shown. Given a phylogenetic tree with numbered tips and nodes (top left), the Modified Graph Laplacian Λ is computed as the difference between its diagonal degree matrix (where diagonal element i is the sum of the branch lengths $w_{i,j}$ from node i to all the other nodes j in the phylogeny) and its distance matrix (where element (i, j) is defined as the branch length (i.e., the "weight") between nodes i and j). Next, the eigenvalues λ and eigenvectors v of Λ are computed (middle row). Finally, the spectral density is obtained by convolving the eigenvalues with a smoothing function (bottom row). See Supplemental Fig. 1 for a toy example.

small λ are characteristic of dense neighborhoods (many nodes) typical of shallow branching events (Martins and Housworth 2002).

The entire organization of the tree is best represented as a density profile of the spectrum of eigenvalues λ (Fig. 1, Supplemental Figure 1), the so-called spectral density profile (Banerjee and Jost 2008), obtained by convolving λ with a smoothing function. Here, we use a Gaussian kernel,

$$f(x) = \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|x-\lambda_i|^2}{2\sigma^2}\right), \qquad (1)$$

where N is the number of λ and $\sigma = 0.1$. The choice of kernel does not considerably change the distribution (Banerjee 2012) and the value of σ is selected for the degree of desired resolution (i.e., smaller values will highlight finer details at the expense of global ones). The spectral density of a tree is then plotted as a function of $\ln(\lambda)$ as $f^*(x) = \frac{f(x)}{\int f(y)dy}$. Throughout, spectral densities constructed from the MGL and nMGL are referred to as standard and normalized spectral densities, respectively.

Importantly, there are heuristic arguments and evidence (although not a formal proof) that it is possible to reconstruct a graph from its spectral density (Ipsen and Mikhailov 2002). Although this does not guarantee that each spectral density uniquely represents a tree (i.e., there may be multiple trees with the same spectral density), this is ostensibly the case for nearly all trees of intermediate size (Matsen and Evans 2012). Therefore, although the spectral density loses information on the labeling of the tree, no (or minimal) information is lost on the shape of the tree. In the physical sciences, spectral density analyses have been successful in differentiating graphs from different domains (Banerjee and Jost 2009), uncovering network modularity (Arenas et al. 2006), and characterizing synchronization dynamics (McGraw and Menzinger 2008). We therefore hypothesized that the spectral densities of phylogenetic trees would provide powerful tools for characterizing and comparing tree shapes as well as identifying modules within them.

The spectral density can in principle be constructed for trees of any size. However, spectral density profiles of trees with fewer than ~20 tips can be erratic and difficult to compare to larger ones. We therefore discard any trees with fewer than 20 tips.

Interpreting Spectral Density Profiles

The global distribution of λ from a MGL is indicative of the total structure of the tree. Each eigenvector v_i describes a branching event in the tree, and the eigenvalue λ_i associated with v_i describes the inverse diffusion time of the branching event between two nodes (Noh and Rieger 2004). Because the diffusion time between two nodes operates principally as a function of the number of branching events between them (Shen and Cheng 2010), large λ represent short diffusion times characteristic of branching events in speciation-poor regions of the tree (sparse nodes separated by long branches) and small λ represent long diffusion times characteristic of speciation-rich regions (dense nodes separated by short branches).

In order to confirm algebraic interpretations of various characteristics of spectral density profiles, we analyzed different spectral properties directly on simulated trees. We simulated rooted birth-death trees with constant speciation (0.2) and extinction (0.05) rates, with 20 time units each, using the R package *TESS* (Höhna 2013). Trees were pruned of extinct lineages and discarded if fewer than 20 lineages survived to the present. A total of 530 trees remained. We constructed the MGL and nMGL and corresponding spectral densities of each tree. The corresponding skewness and kurtosis were computed as $\frac{\mu_4}{\mu_2^2}$ and $\frac{\mu_3}{\mu_2^{3/2}}$, respectively, where μ_i is the ordinary i-th moment of the distribution. Negative and positive skewness reflect a relative abundance of large and small λ, respectively. Lower and higher kurtosis reflect an even and uneven distribution of λ values, respectively. We compared the principal λ, skewness, and kurtosis of spectral density profiles of each simulated tree to four classical phylogenetic metrics: species richness, phylogenetic diversity, the γ statistic (Pybus and Harvey 2000), and the Colless index (Agapow and Purvis 2002). Phylogenetic diversity was measured as the sum of phylogenetic branch length (Faith 1992) using the R package *picante* (Kembel et al. 2010). γ is a popular summary statistic reflecting the stem-to-tip structure of a tree: negative γ values characterize stemmy trees, whereas positive values characterize tippy trees (Pybus and Harvey 2000). The Colless index is a measure of the lineage-to-lineage structure of a tree: smaller Colless indices characterize balanced trees, whereas larger indices characterize imbalanced trees (Agapow and Purvis 2002). The γ statistic and the Colless index were calculated using the R packages *ape* and *apTreeshape*, respectively.

Measuring the Distance between Spectral Densities

To measure the distance between two phylogenies Λ_1 and Λ_2, we begin by computing their spectral densities f_1^* and f_2^*, and then use a probability distribution distance, the Jensen–Shannon distance, defined as:

$$D(\Lambda_1,\Lambda_2)=\sqrt{\frac{1}{2}KL(f_1^*,f^*)+\frac{1}{2}KL(f_2^*,f^*)}, \qquad (2)$$

where $D(\Lambda_1,\Lambda_2)=D(\Lambda_2,\Lambda_1)$, $f^*=\frac{1}{2}(f_1^*+f_2^*)$, and KL is the Kullback–Leibler divergence measure for the probability distribution, where

$$KL(f_2^*,f_1^*)=\int f_1^*(x)\ln\frac{f_1^*(x)}{f_2^*(x)}dx. \qquad (3)$$

Notably, because each tree does not necessarily have a unique spectral density profile, it is possible for $D(\Lambda_1,\Lambda_2)=0$ and $\Lambda_1\neq\Lambda_2$. Although this is statistically unlikely for trees of intermediate size (Matsen and Evans 2012), it means that D is not strictly speaking a metric on tree shape.

Clustering Phylogenies from their Spectral Density Profiles

To cluster a given a set of phylogenies, we begin by constructing their respective spectral densities. Next, we compute the Jensen–Shannon distance for each pair. Finally, we cluster the results with energy-based hierarchical and k-medoids clustering, defining the optimal number of clusters by both an expectation-maximization based on the Bayesian Information Criterion (*BIC*) and medoid partitioning. Energy-based hierarchical clustering is a particularly powerful tool for maximizing among-cluster means and minimizing within-cluster means (Székely and Rizzo 2005) and can show partitioning at different levels of resolution. K-medoids clustering, on the other hand, makes no soft assignments, so each spectral density profile is assigned to a single cluster, and each assignment is given a support estimate based on silhouette width (Reynolds et al. 2006).

In order to check the performance of clustering phylogenies based on their spectral density profiles, we implemented the method on a set of trees simulated under different diversification processes using the R package *TESS*. We simulated trees according to six models of diversification, simulating 100 trees under each model, for a total of 600 trees, during 50 time units each. All models had a constant background extinction rate held at μ = 0.05. The models had either a (i) constant speciation rate, (ii) decreasing speciation rate, (iii) decreasing speciation rate dipping below μ, (iv) increasing speciation rate, or constant speciation rate with an (v) ancient or (vi) recent mass extinction (Supplemental Fig. 2). Trees were pruned of extinct lineages and any tree with fewer than 20 tips surviving to the present was discarded. We then tested the efficacy of clustering in three ways: using the spectral density profile, using summary statistics of the spectral density profile (principal λ, skewness, and kurtosis), and using traditional phylogenetic summary statistics (species richness, phylogenetic diversity, the Colless index, γ, mean branch length, and branch length standard deviation). Spectral density profiles were clustered as described above; summary statistics were normalized and then clustered using hierarchical and k-medoids clustering on principal components.

Assessing the Sensitivity of Spectral Density Profiles to Undersampling

To assess the effect of undersampling on spectral density profiles, we picked 3 trees from the simulations detailed above (a constant speciation rate tree, an increasing speciation rate tree, and an ancient mass-extinction tree) and jackknifed (i.e., sampled without replacement) each of them at 90%, 80%, 70%, 60%, 50%, and 40%. One hundred replicate trees were used for each sampling value. We then compared the spectral densities of the complete and undersampled trees, using the Jensen–Shannon distance and spectral properties.

Identifying Modalities within a Phylogenetic Tree

To identify modes of division (or modalities) within a phylogenetic tree, we first compute the λ from the MGL of the tree and rank them in descending order of magnitude. In graph theory, the ranked λ reflect the connectivity of the graph. If there are i ideal clusters in the graph (i.e., high between-cluster and low within-cluster variation), then each of the i largest λ represents a separation between clustered points in the graph. Furthermore, there will be a single largest gap between λ_i and λ_{i+1}, where $\lambda_{>i} << \lambda_{\leq i}$ (von Luxburg 2007). For this reason, the eigengap, identified as the largest difference between two consecutive λ, is an indicator of the number of clusters in the graph (Shen and Cheng 2010). Transposing this heuristic to a phylogenetic tree, if the eigengap is between λ_i and λ_{i+1}, then there are i clusters, or modes of division, in the tree, and these modalities can be identified using k-medoids clustering on the graph by setting $k = i$ (Newman 2006). These clusters need not represent monophyletic regions of the tree, because the $\lambda_{\leq i}$ describe branching events distributed anywhere in the tree. A cluster could, in principle, be composed of non-adjacent branches sampled from across the tree.

Once an indication of the number of modalities, i, in a tree of interest has been obtained from the eigengap, it is possible to get a confidence measure for i. This can be done by comparing *BIC* values for detecting i modalities in the distance matrix of the tree of interest (BIC_{test}) and in randomly bifurcating trees parameterized on that tree (BIC_{random}) (Pelleg and Moore 2000). Here, $BIC = D + \log(N) * m * q$, where D is the total within-cluster sum of squares based on posterior probability estimates from k-medoids clustering of the nodes of the matrix, N is the total number of nodes in the matrix, and m and q are the dimensions of the clusters of nodes in the x,y-plane, respectively. The random trees can be ultrametric or not. In the former case, tips are randomly coalesced with the same branch length distribution and number of tips as the tree of interest; similarly in the latter case, except the branches are randomly split from the stem. The number of modalities is then considered significant if $BIC_{random} / BIC_{test} \geq 4$ for at least 95% of the random trees, which provides a conservative test of the significance of the modalities (Kass and Raftery 1995). Random trees were constructed using the R package *ape*.

To test this approach, we investigated the ability of the eigengap heuristic to recover simulated shifts in ultrametric trees. We started by simulating a 100-tip pure-birth tree with speciation rate 0.15. A given shift on that tree was simulated by randomly choosing a node located in the middle of the tree (i.e., excluding the first and last quartile of tree length), pruning all branches descending from that node, and grafting onto that node a new tree with proper length simulated under either a pure-birth model with a different speciation rate (ranging from 0.05 to 0.5) or a different diversification model (randomly chosen among the set of models represented in Supplemental Fig. 2). In a single tree, we iteratively simulated up to 10 shifts all shifts being composed of either shifts in speciation rate or diversification pattern (never both). Trees were pruned and grafted using our own code with functions from the R package *phytools* (Revell 2012). In total, 200 trees with 0–10 shifts in speciation rate and 200 trees with 0–10 shifts in diversification pattern were simulated. The recovery reliability of the eigengap heuristic was compared with MEDUSA (Alfaro et al. 2009) and BAMM (Rabosky 2014), the most commonly used methods for identifying rate shifts. We ran MEDUSA using the *medusa* function from the R package *geiger* (v2.0.3) with the initial speciation rate set to 0.15 and extinction rate constrained to 0. We ran BAMM after setting priors for each simulated tree with the R package *BAMMtools* (Rabosky 2014). We did not compare our results to the non-parametric rate comparison (PRC) of (Shah et al. 2013), because the approach is implemented only for pairwise comparisons and becomes prohibitively computationally expensive when iterated.

Empirical Applications

To illustrate our approach, we used two empirical data sets: the first, representing influenza A strains spanning two animal hosts and 25 countries, was used to illustrate the usefulness of clustering on the spectral density profile; the second, composed of 350 16s rRNA archaeal sequences collected from the sediment of Lake Dagow (Barberan et al. 2011), was used to illustrate the eigengap heuristic. We purposely chose applications on non-ultrametric trees, as their analysis is not typically available to current techniques.

For the viral data set, we collected trees for a range of influenza A virus strains from the NCBI Flu Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html). Protein-coding sequences for haemagglutinin (HA), matrix protein 1 (M1), neuraminidase (NA), nucleoprotein (NP), nonstructural protein 1 (NS1), polymerase acid protein (PA), and polymerase basic protein 2 (PB2) were obtained for strains from avian and human hosts originating from 25 different countries. Neighbor-joining trees were constructed for each protein separately from multiple sequence alignments in MUSCLE (Edgar 2004). Identical sequences were collapsed. Trees with fewer than 25 or more than 1000 tips were not considered (in a few

instances, to meet this criterion, a particular subtype of the strain was selected). In total, 324 trees were constructed with an average of 200, minimum of 25, and maximum of 915 tips. We clustered the standard and normalized spectral density profiles of these trees and compared them with spectral density summary statistics, using peak height as a measure of evenness. (Peak height, defined as the largest *y*-axis value of the spectral density profile, is a measure of evenness that we found to be better behaved than kurtosis on non-ultrametric trees.) We tested the effect of country of origin on clustering by randomly assigning strains to clusters and comparing the actual versus randomized distribution of strains for 500 randomizations.

The archaeal phylogenetic tree was taken from (Barberan et al. 2011). We removed the four out-groups from the original tree, applied the eigengap heuristic described above to the resulting tree, and characterized each of the identified clusters with their respective spectral density profiles.

Results

Interpreting the Spectral Density of a Phylogenetic Tree

Different aspects of the shape of spectral density profiles may be interpreted in terms of the underlying shape of the phylogeny. In particular, the shift (right bound), asymmetry (skewness), peakedness (kurtosis or peak height), and number of peaks (modalities) of the spectral density profile are illustrative of specific interpretable patterns in the phylogenetic tree (Fig. 2). These interpretations are validated using trees simulated under a birth–death model with constant speciation and extinction rates (Supplemental Fig. 3). It also demonstrates that spectral density profile summary statistics and traditional phylogenetic summary statistics are not perfectly correlated, meaning that spectral density profile summary statistics potentially capture different aspects of tree shape.

The *shift* corresponds to the principal (or largest) λ, which is related to the largest phylogenetic distance between tip species and may be an indicator of species richness and phylogenetic diversity (Fig. 2a). For the normalized spectral density profile, the principal λ is not significantly correlated to species richness and is negatively correlated with phylogenetic diversity (Supplemental Fig. 3a,b), demonstrating that the nMGL essentially removes the effect of tree size.

The *asymmetry* of the density profile, which can be quantified by its skewness (a measure based on the 3rd and 2nd moments), is primarily indicative of the stem-to-tip structure of the tree (Fig. 2b; Supplemental Fig. 3c). Intuitively, a positive skewness indicates a relative abundance of small λ corresponding to shallow branching events, and therefore characterizes tippy phylogenies, whereas a negative skewness indicates a relative abundance of large λ corresponding to deep branching events, and therefore characterizes stemmy phylogenies.

The *peakedness* of a spectral density profile, which can be quantified by its kurtosis (a measure based on the 4th and 2nd moments), is primarily indicative of the lineage-to-lineage structure of the tree (Fig. 2c; Supplemental Fig. 3d). Intuitively, a flat peak indicates that there is an even distribution of λ values, meaning that branch lengths are homogeneously distributed in the tree and so the tree is balanced — whereas a steep peak means the tree is imbalanced. Another way to measure this peakedness is by directly measuring peak height. We found that peak height is better behaved than kurtosis on non-ultrametric trees.

The *number of peaks* in the density plot is indicative of the different number of modalities within the tree (Fig. 2d). For example, if clade A is composed of tippy branching events and its sister clade B is composed of stemmy branching events, then the spectral density profile of the tree encompassing clades A and B has two peaks, one at small λ representing clade A and one at large λ representing clade B.

Comparing and Clustering Phylogenies Using Spectral Density Profiles

We aim to measure the distance between phylogenetic trees (an inverse measure of their similarity) and to cluster them according to their similarity. Once phylogenies have been transformed into their spectral density profiles, the distance between them can easily be computed using a probability distribution distance metric and clustered using a traditional clustering algorithm. We used the popular Jensen–Shannon distance metric (Endres and Schindelin 2003) as our distance metric. This metric quantifies the square root of the total divergence to the average probability distribution; it has the advantage of being symmetric and finite. We used this metric to cluster phylogenies with energy-based divisive clustering, which is a bottom-up hierarchical approach that provides resolution within clusters, and k-medoids clustering, which does not, although it is possible to get support values for cluster assignment using the silhouette width of each data point. Both clustering methods deal well with high-dimensional data.

In order to investigate the power to identify different types of tree shapes from their spectral density profiles, and to compare this performance to that obtained from traditional summary statistics, we simulated trees under six different models of diversification and checked whether their spectral density profiles clustered into distinct classes of trees. The six models mimicked constant speciation and extinction rates, decreasing or increasing speciation rates (with speciation remaining above or decreasing below extinction rates), and ancient or recent mass-extinction events (Supplemental Fig. 2). We found that the spectral density profiles were optimally clustered into six groups according to diversification model using both hierarchical clustering (bootstrap probability

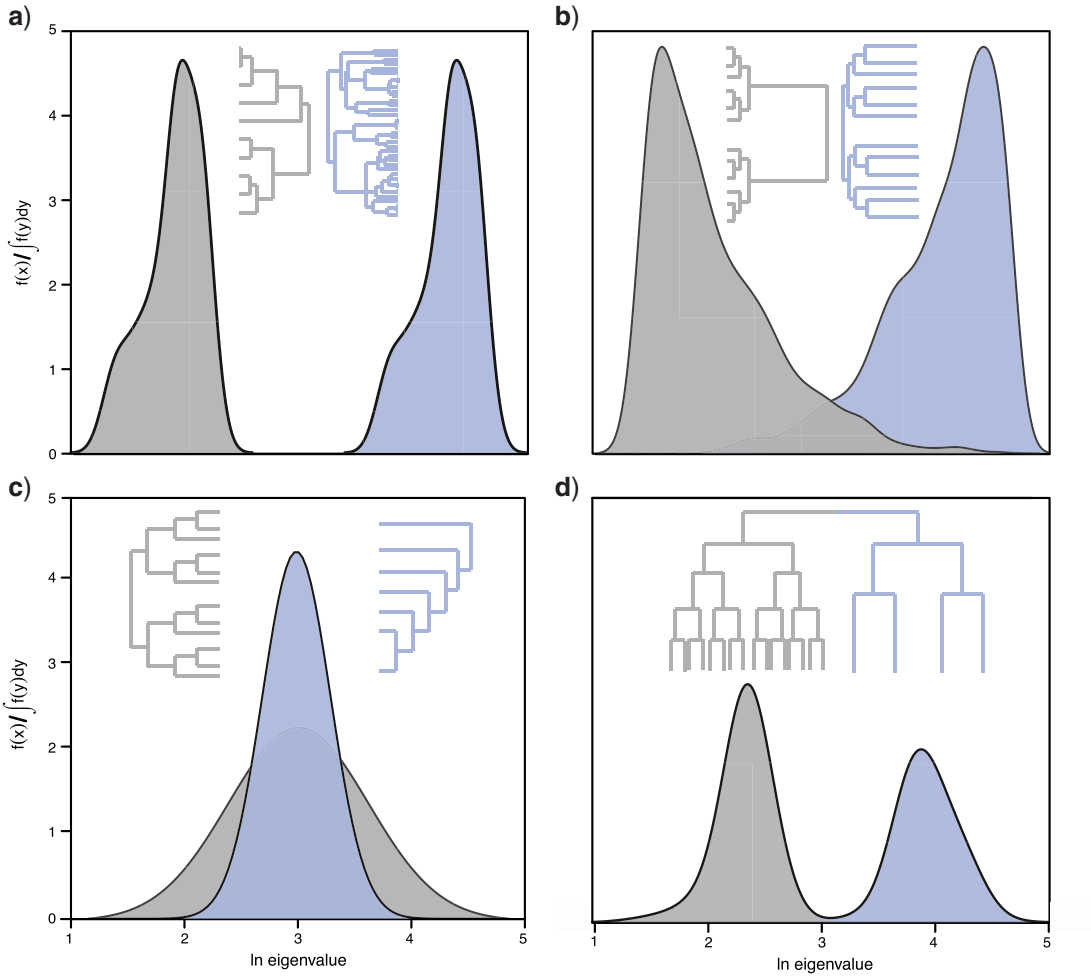

FIGURE 2. Global properties of the spectral density are indicative of specific patterns in the phylogenetic tree. a) Trees with high (blue) and low (gray) species richness are characterized by large (blue) and small (gray) principal λ, respectively. b) Stemmy (blue) and tippy (gray) trees are characterized by negative (blue) and positive (gray) skewness. c) Imbalanced (blue) and balanced (gray) trees are characterized by high (blue) and low (gray) kurtosis. d) Different modalities within a tree, such as one with stemmy (blue) and one with tippy (gray) branching events, appear as peaks in the spectral density (here at large and small λ, respectively).

>0.95) and k-medoids clustering ($P < 0.05$), suggesting that spectral density profiles provide an efficient way to distinguish and cluster different types of tree shapes. The least and most within-cluster variation was observed in the constant speciation-rate model and the ancient mass-extinction model, respectively (Fig. 3a). A follow-up principal component analysis on summary statistics for the spectral density profile showed comparable influence from principal λ (38% of total explained variance), skewness (34%), and kurtosis (28%). Specifically, each acts on a different dimension, with skewness acting orthogonally to principal λ (Fig. 3a). Inspection of spectral density profiles representative of each cluster reveals local and global differences between clusters in their distributions of λ (Fig. 3b). By comparison, clustering on traditional phylogenetic summary statistics retrieved only three modes of diversification (Supplemental Fig. 4a), which explained 79% of the variance among trees, compared to 93% for spectral density summary statistics. The principal components derived from traditional phylogenetic summary statistics were unable to distinguish between the two decreasing speciation-rate models or between the constant speciation-rate and recent and ancient mass-extinction models (Supplemental Fig. 4b).

In order to further test whether phylogeny size was primarily responsible for clustering together the different models, we clustered the same trees using spectral density profiles computed from their nMGLs. We found that these spectral density profiles also clustered by model (Supplemental Fig. 5a), suggesting that trees under a magnitude size difference are not clustered on size alone. K-medoids clustering on principal components derived from summary statistics calculated on the nMGL, however, retrieved only four clusters, showing an inability to distinguish between the two decreasing speciation-rate models or the constant speciation-rate and recent mass-extinction models (Supplemental Fig. 5b).

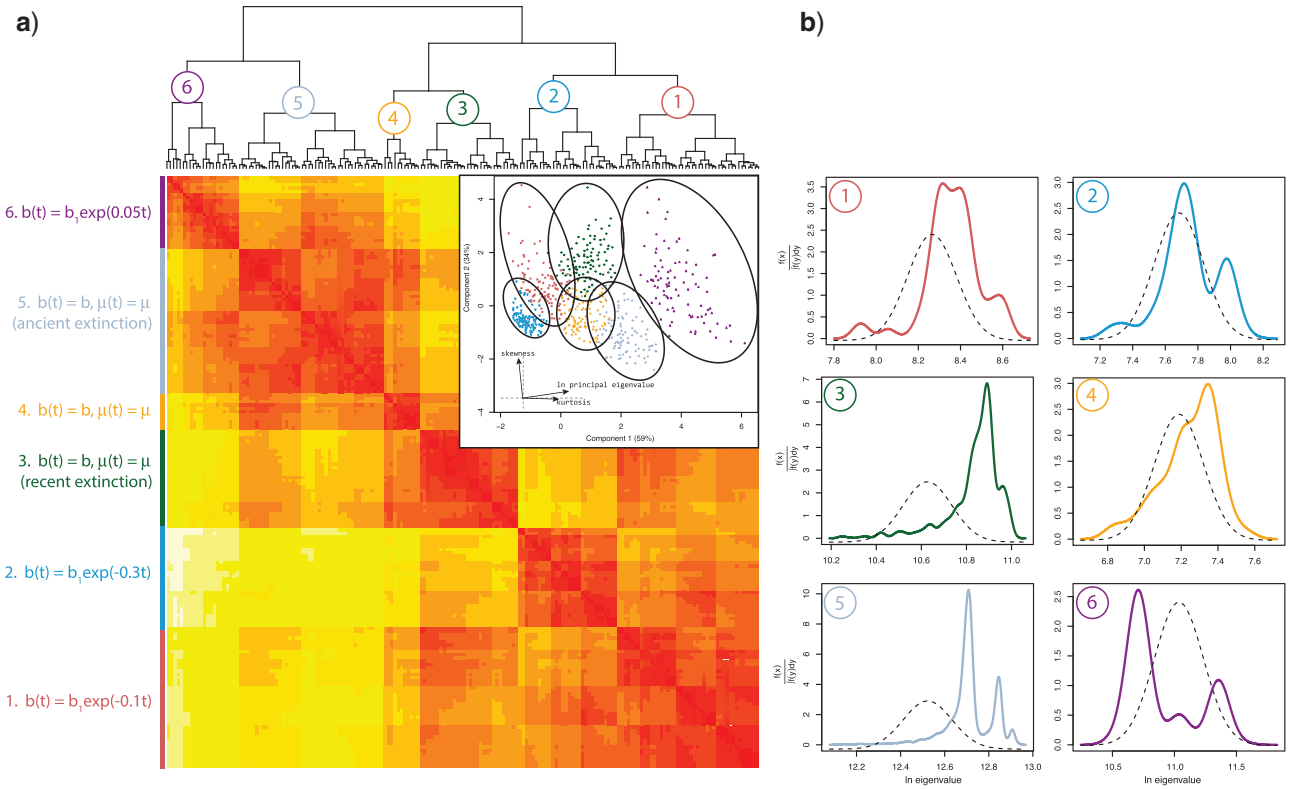

FIGURE 3. Clustering of spectral density profiles identifies distinct modes of diversification in simulated trees. a) Hierarchical clustering on the MGL based on the Jensen–Shannon distances between spectral density profiles of trees simulated under different diversification models. Both hierarchical and k-medoids clustering techniques identify 6 clusters of trees from 600 trees ($P < 0.05$), each corresponding to a distinct underlying diversification model, whose property in terms of speciation-extinction rate variation is summarized in the left column. Partitions in the hierarchical cluster are collapsed below a threshold height of 2, so that less variation between individual trees is represented as fewer partitions in the hierarchical cluster and fewer cells in the heatmap. a), (inset) K-medoids clustering on principal components derived from spectral density profile summary statistics: ln principal λ, skewness, and kurtosis. Shapes correspond to the cluster assignment of trees based on highest silhouette width; colors correspond to diversification type. Ellipsoids represent confidence intervals for each cluster, such that each tree could, based on silhouette width support, be assigned to any cluster whose ellipsoid encompasses it. Each tree is assigned to the cluster for which it has the most support. The inset shows the relative contribution of each statistic in the dimensionality of the principal component analysis. b) A representative spectral density profile for each cluster, defined as the median spectral density profile according to the Jensen–Shannon distance, against a normal distribution (dashed lines) with the same mean and variance, but a fixed height (2.5) to emphasize differences between spectral density profiles, is shown for the 6 groups. Note the different *x*- and *y*-axis ranges.

Testing the Effect of Undersampling on the Spectral Density Profile

Because trees are often incomplete, undersampling is a common issue to consider in phylogenetic analyses. We tested the extent to which (and how) undersampling modifies the shape of spectral density profiles by jackknifing simulated trees. As expected, the spectral density of a tree is sensitive to undersampling and begins to become visually misrepresentative of the complete tree at ~80% complete, although many features of the plot may persist until ~40% (Supplemental Fig. 6a-c). The spectral distance between original and jackknifed trees increased linearly with the level of undersampling. As the trees became less complete, the skewness decreased in constant speciation rate ($1.11 \rightarrow 0.52 \pm 0.10$), increasing speciation rate ($0.84 \rightarrow 0.46 \pm 0.09$), and recent mass extinction ($1.87 \rightarrow 1.31 \pm 0.14$) trees; as did kurtosis, for constant speciation rate ($-0.04 \rightarrow -1.28 \pm 0.12$) and recent mass-extinction ($2.03 \rightarrow 0.94 \pm 0.14$), but not increasing speciation rate ($-0.70 \rightarrow -0.78 \pm 0.21$) trees, which showed a sharp increase in kurtosis in some samples at ≤50% complete. So, according to their spectral density profiles, undersampled trees are increasingly stemmy, as expected, and, in general, increasingly balanced.

Identifying Modalities within Phylogenies

To assess the ability of the eigengap to recover shifts in diversification, we generated trees with simulated shifts in modes and rates of diversification. We then applied the eigengap heuristic (which includes the post-hoc *BIC* analysis), MEDUSA, and BAMM to those trees and compared the number of recovered *versus* simulated shifts. The eigengap heuristic did not artificially detect shifts in their absence. The eigengap heuristic and MEDUSA performed comparably well on trees with shifts in only speciation rate (Supplemental

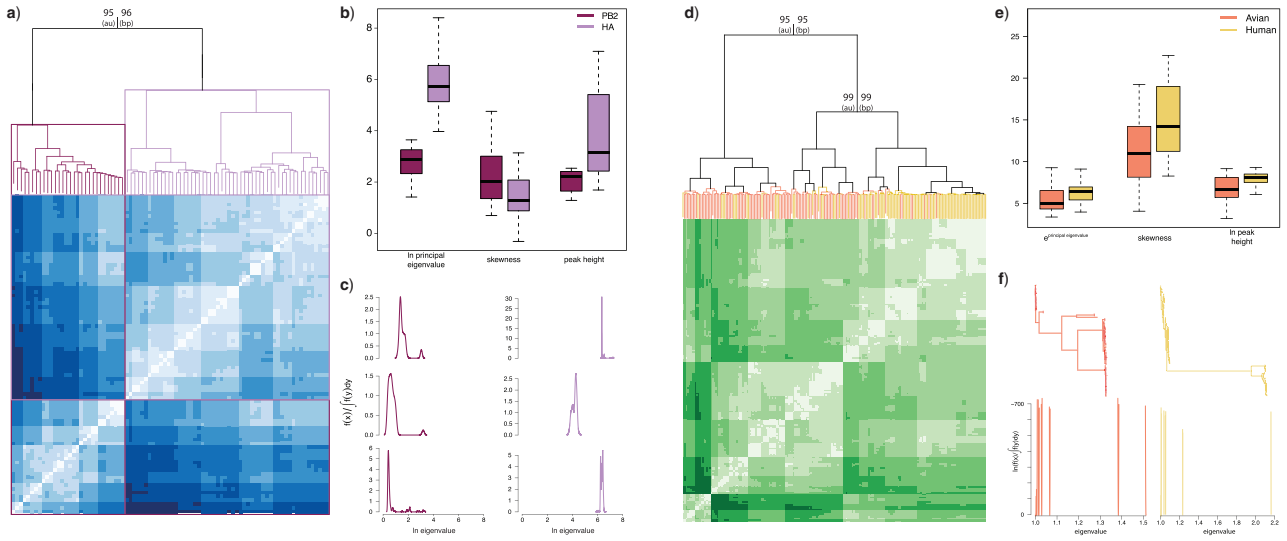

FIGURE 4. Clustering of influenza A viral strains from 25 countries and 2 animal hosts on standard and normalized graph Laplacians. a) Heatmap and hierarchical cluster of standard spectral density profiles for strains constructed from HA and PB2 sampled across 25 countries in avian and human hosts. Approximately unbiased (au) and bootstrap probability (bp) support are shown at branching events. PB2 phylogenies are overrepresented in cluster 1 (94% of all strains in cluster) and underrepresented in cluster 2 (18%). b) Boxplot of spectral density profile summary statistics for clusters 1 (light purple) and 2 (dark purple) in (a). All mean differences are significant at $P < 0.01$. c) Sample of spectral density profiles from cluster 1 (left) and cluster 2 (right). Note the different y-axis ranges. d) Heatmap and hierarchical cluster of normalized spectral density profiles for trees constructed from HA, M1, NA, NP, NS2, PA, and PB2 across 2 hosts and 25 countries. Columns calculated from strains sampled from avian (red) and human (gold) hosts are shown. Phylogenies across all proteins and countries of origin are largely distinguished by host (bootstrap probability > 0.95) e) Boxplot of mean values for spectral density profile summary statistics. All mean differences are signfiicant at $P < 0.05$. f) Trees and spectral density profiles for strains sampled from avian and human hosts.

Fig. 7a), with both methods routinely underestimating the number of shifts, something previously reported for MEDUSA (Rabosky 2014). For trees with shifts in diversification patterns, however, the eigengap heuristic consistently outperformed MEDUSA (Supplemental Fig. 7b). MEDUSA commonly underestimated the number of shifts, whereas the eigengap heuristic was on average within ± 1 the number of shifts. Using the priors estimated from *BAMMtools*, BAMM was not sensitive enough to detect more than three shifts in any tree.

Empirical Applications

Traditional phylogenetic approaches are typically incapable of dealing with non-ultrametric trees. The ability of our approach to deal with such trees opens up the possibility to analyze the diversification of groups that are rarely studied in macroevolutionary terms. For example, little is known about the diversification patterns of viruses, despite the significance of their evolution for epidemiology (but see (Poon et al. 2013)). We compared the spectral density profiles of 324 influenza A trees constructed independently for 6 protein segments, 25 countries, and 2 animal hosts. Results from profiles constructed using the MGL and nMGL showed consistent differences (and similarities) in diversification dynamics across phylogenies derived from different protein segments, originating in different countries, and hosted in different animals. Although qualitatively consistent, results from the nMGL were quantitatively more emphatic than those from the MGL, suggesting that phylogenetic shape (not size) was the main effector of differences in diversification. Both showed considerably different profiles for HA and NA compared with the other proteins (Fig. 4a–c). Specifically, diversification patterns in HA and NA, in both avian and human hosts, were more expansionary, tippy, and imbalanced than those in the five other protein segments (Supplemental Fig. 8). We furthermore found, using k-medoids clustering on profiles computed from the MGL and nMGL, 6 and 4 ($P < 0.05$) clusters, respectively, across countries and hosts. Strains from the same country of origin were more likely to fall into the same cluster than expected by chance for avian ($D \geq 0.584$, $P < 0.001$) and human ($D \geq 0.399$, $P < 0.001$) hosts, although this effect decreased when analyzed across hosts ($D \leq 0.185$, $P \geq 0.044$) (Supplemental Fig. 9). Finally, we found significant differences between hosts for individual strains and across all strains (Fig. 4d–f, Supplemental Fig. 8). We have demonstrated, therefore, how our approach based on the graph Laplacian makes it possible to test macroevolutionary hypotheses on life forms heretofore largely unavailable to diversification analyses; and additionally exemplified how the MGL and nMGL may be used to corroborate different aspects of those tests.

To further illustrate the empirical applicability of our approach, we examined the spectral density of an archaeal phylogenetic tree of microbial species. Using the described framework for finding modes of division within trees, we identified the eigengap to be between λ_6 and λ_7, indicative of six modalities (Fig. 5), which was

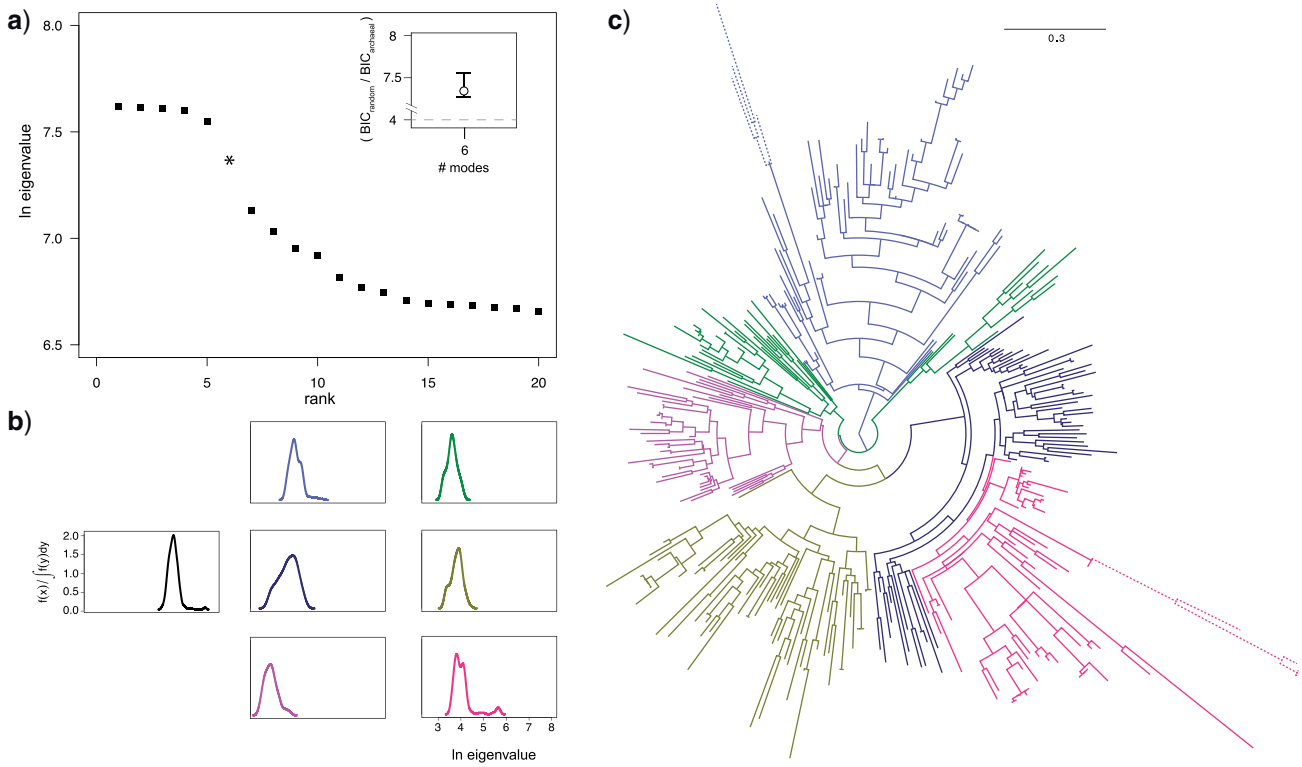

FIGURE 5. Identifying modes of diversification within a single tree. a) λ calculated from the MGL for the archaeal tree are ranked in descending order and the eigengap is identified between λ_6 and λ_7, suggestive of 6 modes of division. a), (inset) The ratio of *BIC* values for finding 6 modes in the distance matrices of 100 randomly bifurcating trees and the distance matrix of the archaeal tree. The *BIC* ratio significance cutoff is indicated (gray dashed line). Error bars are drawn from *BIC* ratios against 100 random trees. b) Spectral density profiles for the original species tree (black) and for each of the modality trees. c) Lineages in the archaeal tree showing different modes of division as identified by the eigengap heuristic, with each mode of diversification shown by a different color corresponding to (b). The dashed branches have been shortened for presentation.

supported by post-hoc analysis ($BIC_{random}/BIC_{archaeal} \geq 7.39$). These results suggest that the archaeal community from Lake Dagow is made of six groups of sequences with distinctive evolutionary histories. The spectral density profile is therefore useful, not only in finding clusters of nodes within trees, but also for assessing what makes these clusters distinct.

DISCUSSION

We have introduced an approach based on the spectrum of the graph Laplacian for reducing unlabeled phylogenetic trees to their constituent properties. We have shown how to compute the spectral density profiles of phylogenies, and how to use these profiles to characterize, compare, and cluster trees, as well as to find distinct modes of division within them. This provides a comprehensive framework for (i) summarizing the information contained in phylogenies, (ii) identifying similarities and dissimilarities between trees, and (iii) picking out distinctive branching patterns within individual trees, without making any *a priori* assumptions about underlying behavior. The ability of this approach to analyze non-ultrametric trees, in particular, fills a largely empty gap in the field of diversification dynamics.

Approaches for summarizing phylogenetic information are required across multiple domains of the life sciences. They are necessary for studying phylogenetic diversity in both the macro- and microbial worlds (Faith 1992; Lozupone and Knight 2008), for measuring how closely related species are within community assemblies (Webb et al. 2002), for understanding how diversification varies in time and across lineages (Morlon 2014), and for tracking genealogical diversity of infectious diseases through time (Vijaykrishna et al. 2014). Such approaches are also particularly useful in phylogenetic modeling where they allow us to evaluate how closely a specific ecological, epidemiological, or macroevolutionary model reproduces empirical trees. The ability provided by approaches summarizing phylogenetic information to quantify the distance between trees allows us to measure the distance between trees simulated under a specific model and empirical trees, which is crucial to fitting approaches such as Approximate Bayesian Computation (Janzen et al. 2015) or posterior predictive simulations (Lewis et al. 2014).

Given the importance of summarizing the information contained in phylogenetic trees, our study is not the first attempt at doing so. However, our approach is unprecedented insofar as spectral densities account for

almost the entirety of phylogenetic structure: for trees of intermediate size, no (or minimal) information is lost on tree shape when expressing a phylogeny as its spectral density. It is therefore superior to previously proposed summary statistics that limit themselves to certain properties of the tree summarized by a single statistic. When reduced to its constituent properties (i.e., principal λ, skewness, and kurtosis), the spectral density profile still manages to better identify diversification types among trees than a combination of the most widely used traditional summary statistics. An additional advantage of spectral density profiles compared with many traditional summary statistics is that they can be computed irrespective of whether the tree is dated, ultrametric, or fully resolved.

There are many potential applications of our approach. For example, assuming that coevolution and codiversification lead to similarities in branching patterns, clades undergoing codiversification could be identified based on similarities in their spectral density profiles without any *a priori* information about their interaction. This could be particularly useful in the case of microbes and viruses, for which interactions and coevolution cannot directly be observed in nature. In viruses, especially, similarities in spectral density profiles can be used to identify convergence across lineages, where diversification may be driven by, for example, an ecological parameter, trait adaptation, or even site-specific substitution. In this respect, our analyses for the various diversification patterns in influenza A strains — although they are meant here only for illustrative purposes and should be taken with caution — are of some interest.

We find differential effects of protein segment, host, and country of origin on the diversification of influenza A. For most segments of the virus, diversification patterns are similar, although there are marked differences between both HA and NA and other segments. These two segments show significantly higher mean values for principal λ and peak height, indicative of highly expanded, imbalanced diversification, which corroborates previous observations of especially high substitution rates in these proteins (Bhatt et al. 2011). Contrary to previous work, however, we do not find similarities in the spectral density profiles of HA and M1, which have been suggested to have comparable phylogenetic histories due to their interaction during viral assembly (Rambaut et al. 2008). Although these segments may be mechanically interdependent, the considerable variation between their diversification patterns suggests that their strategies of coevolution, while compatible, are not equivalent. Finally, the exceptional differences between HA and PB2, in particular, with the former exemplifying disproportionately more expansive, imbalanced, and stemmy trees than those constructed with the latter, evince distinctive evolutionary trajectories for two proteins in a single virus, as well as strong constraints on those trajectories across distant phylogenetic hosts. We furthermore see a significant influence of country of origin on patterns of diversification within each host, where strains from the same country diversify more similarly than expected by chance. However, for both the standard and the normalized spectral density profiles, the single strongest impact on the shape of virus diversification is the animal host. These results illustrate the utility of our approach to deal with non-ultrametric trees and to explore the diversification behavior of many organisms previously unavailable to macroevolutionary hypotheses.

Finding shifts in diversification processes is a major interest in macroevolution. Methods for identifying rate shifts in trees (e.g., Alfaro et al. 2009; Shah et al. 2013; Rabosky 2014) have been invaluable in establishing, for example, adaptive radiations in large clades (Alfaro et al. 2009; Shi and Rabosky 2015). We introduce the eigengap heuristic, an approach for finding different modes of diversification within a single tree. Our approach shows considerable — albeit imperfect — success in recovering rate shifts in simulated trees, comparable (or superior) to the most widely used methods. But it is important to emphasize that the analytic difference in this approach bespeaks a conceptual difference as well: the eigengap heuristic does not strictly identify rate shifts in the tree, but identifies branches of similar diversification processes. So it is not surprising that it underperforms, if only slightly, against an existing method in identifying shifts in diversification rate, but outperforms the same method in identifying shifts in diversification pattern. The eigengap heuristic, therefore, distinguishes itself by its power to recognize modes of diversification patterns present in a tree. Our illustration of this approach with an archaeal tree demonstrates how the eigengap heuristic may be used to pinpoint disparately evolving populations of microbial species in a single environment (in this case, Lake Dagow). Specifically, it reveals subtrees with considerably different diversification patterns, which do not vary by phylogenetic relatedness.

Most previous graph-theoretical work in phylogenetics has focused on developing methods to estimate the "tree space" that different hypotheses for the same phylogenetic tree occupy (Hillis et al. 2005; Huang and Li 2013; Whidden and Matsen 2015). These methods have been very successful and we think that, by assessing the congruence of spectral densities for different gene-based trees for the same species tree, our approach may also be useful for estimating confidence intervals for trees. Similarly, it may be possible to investigate the coevolution of traits (and genes) based on the (dis)similarities between the spectral density profiles of trait-trees (and gene-trees) sampled from the same species. Generally, comparing spectral density profiles for many phylogenies, whether or not they are sampled from the same species tree, is useful for identifying characteristic patterns of diversification as well as natural limits to those patterns.

There are also many potential variations on our approach. We illustrated the approach on bifurcating trees, yet the degree matrix can take any form, such that reticulated trees (i.e., phylogenetic networks) can also

be analysed. Reticulated trees have so far been largely unnavigable by conventional phylogenetic techniques and, as a result, studies of microbial phylogenies have typically assumed the trees to be bifurcating (Martin et al. 2004), which is often not accurate given the level of lateral gene transfer in the microbial world. Given that microbes constitute the majority of biodiversity on the planet, it is crucial to develop such approaches.

Finally, there are many potential extensions of our approach. For example, graph Laplacians are used in synchronization dynamics (McGraw and Menzinger 2008) to analyze if and how a given part of a network affects the dynamics of other parts of that network. Applied to phylogenies, this could allow for analyzing the interaction effects of some clades on others. There are also techniques from differential geometry, based on the so-called trace formula (Horton et al. 2006), that could be used to analyze the behavior of suites of spectral densities, such as the spectral densities measured for a tree at different times from its origin. Such analyses could inform us about the evolution of a clade. A third potential extension would be to use signed graphs, where a signed matrix maps data onto the edges of the graph Laplacian (Shames et al. 2014) to analyze how certain information not encoded in the molecular phylogeny (e.g., geographic or phenotypic distance) affects local structures in the tree.

We have developed an approach, implemented in user-friendly software, which gives researchers access to questions underserved by current phylogenetic techniques.

Funding

Funding was provided by the CNRS and grants ECOEVOBIO-CHEX2011 from the French National Research Agency (ANR) and PANDA from the European Research Council (ERC) attributed to H.M.

Acknowledgments

We would like to thank Mike Steel, two anonymous reviewers, and Olivier Gascuel for their counsel, as well Julien Clavel, Jonathan Drury, Nancy Irwin, Marc Manceau, and Olivier Missa for helpful comments on the article. EL would like to thank Evan Charles for helpful discussion.

References

Agapow P., Purvis A. 2002. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. Syst. Biol. 51:866–872.

Alfaro M.E., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky D.L., Carnevale G., Harmon L.J. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl Acad. Sci. 106:13410–13414.

Arenas A., Diaz-Guilera A., Perez-Vicente C.J.. 2006. Synchronization reveals topological scales in complex networks. Phys. Rev. Lett. 96:114102.

Avise J. 2000. Phylogeography: the history and formation of species. Cambridge (MA): Harvard University Press.

Banerjee A. 2012. Structural distance and evolutionary relationship of networks. Biosystems 107:186–196.

Banerjee A., Jost J. 2008. On the spectrum of the normalized graph laplacian. Linear Algebra Appl. 428:3015–3022.

Banerjee A., Jost J. 2009. Graph spectra as a systematic tool in computational biology. Networks Comput. Biol. 157:2425–2431.

Barberan A., Fernandez-Guerra A., Auguet J.-C., Galand P.E., Casamayor E.O. 2011. Phylogenetic ecology of widespread uncultured clades of the kingdom euryarchaeota. Mol. Ecol. 20:1988–1996.

Bhatt S., Holmes E.C., Pybus O.G. 2011. The genomic rate of molecular adaptation of the human influenza A virus. Mol. Biol. Evol. 28: 2443–2451.

Billera L.J., Holmes S.P., Vogtmann K. 2001. Geometry of the space of phylogenetic trees. Adv. Appl. Math. 27:733–767.

Blum M.G.B., François O. 2006. Which random processes describe the tree of life? a large-scale study of phylogenetic tree imbalance. Syst. Biol. 55:685–691.

Burki F. 2014. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harbor Perspect. Biol. 6:a016147.

Cadotte M.W., Jonathan Davies T., Regetz J., Kembel S.W., Cleland E., Oakley T.H. 2010. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. Ecol. Lett. 13:96–105.

Chan K.M.A., Moore B.R. 2002. Whole-tree methods for detecting differential diversification rates. Syst. Biol. 51:855–865.

Chen D., Burleigh G.J., Fernández-Baca D. 2007. Spectral partitioning of phylogenetic data sets based on compatibility. Syst. Biol. 56:623–632.

Condamine F.L., Rolland J., Morlon H. 2013. Macroevolutionary perspectives to environmental change. Ecol. Lett. 16(Suppl 1):72–85.

Dunne J.A., Williams R.J., Martinez N.D. 2002. Food-web structure and network theory: the role of connectance and size. Proc. Natl Acad. Sci. USA 99:12917–12922.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Endres D., Schindelin J. 2003. A new metric for probability distributions. IEEE Trans. Informat. Theory 49:1858–1860.

Faith D.P. 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61:1–10.

Garamszegi L. 2014. Modern phylogenetic comparative methods and their application in evolutionary biology. Concepts and Practice. London, UK: Springer.

Harmon L.J. Losos J.B., Jonathan Davies T., Gillespie R.G., Gittleman J.L., Bryan Jennings W., Kozak K.H., McPeek M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte Ii J.A., Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., Mooers A. Ø. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evol. Int. J. Organic Evol. 64:2385–2396.

Hillis D.M., Heath T.A., St John K. 2005. Analysis and visualization of tree space. Syst. Biol. 54:471–482.

Höhna S. 2013. Fast simulation of reconstructed phylogenies under global time-dependent birth-death processes. Bioinformatics 29:1367–1374.

Horton M.D., Stark H., Terras A.A. 2006. What are zeta functions of graphs and what are they good for? Contemp. Math. 415:173–190.

Huang H., Li Y. 2013. MASTtreedist: visualization of tree space based on maximum agreement subtree. J. Comput. Biol. J. Comput. Mol. Cell Biol. 20:42–49.

Ipsen M., Mikhailov A.S. 2002. Evolutionary reconstruction of networks. Phys. Rev. E 66:046109.

Janzen T., Höhna S., Etienne R.S. 2015. Approximate bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT. Methods Ecol. Evol. 6:566–575.

Kass R. E., Raftery A.E. 1995. Bayes factors. J. Am. Stat. Assoc. 90: 773–795.

Kembel S.W., Cowan P.D., Helmus M.R., Cornwell W.K., Morlon H., Ackerly D.D., Blomberg S.P., Webb C.O. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464.

Kosakovsky Pond S.L., Murrell B., Fourment M., Frost S.D.W., Delport W., Scheffler K. 2011. A random effects branch-site model for

detecting episodic diversifying selection. Mol. Biol. Evol. 28: 3033–3043.
Lewis P.O., Xie W., Chen M.-H., Fan Y., Kuo L.. 2014. Posterior predictive Bayesian phylogenetic model selection. Syst. Biol. 63:309–321.
Lewitus E., Huttner W.B. 2015. Neurodevelopmental LincRNA microsyteny conservation and mammalian brain size evolution. PloS One 10:e0131818.
Lozupone C.A., Knight R. 2008. Species divergence and the measurement of microbial diversity. FEMS Microbiol. Rev. 32: 557–578.
Martin A.P., Costello E.K., Meyer A.F., Nemergut D.R., Schmidt S.K. 2004. The rate and pattern of cladogenesis in microbes. Evol. Int. J. Organic Evol. 58:946–955.
Martins E.P., Housworth E.A. 2002. Phylogeny shape and the phylogenetic comparative method. Syst. Biol. 51:873–880.
Matsen F. 2006. A geometric approach to tree shape statistics. Syst. Biol. 55:652–661.
Matsen F.A., Evans S.N. 2012. Ubiquity of synonymity: almost all large binary trees are not uniquely identified by their spectra or their immanantal polynomials. Algorithms Mol. Biol. 7:14.
Matsen IV F.A., Evans S.N. 2013. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. PLoS One 8:e56859.
McGraw P.N., Menzinger M. 2008. Laplacian spectra as a diagnostic tool for network structure and dynamics. Phys. Rev. E 77:031102.
Mohar B. 1997. Some applications of laplace eigenvalues of graphs. *in* Graph symmetry: algebraic methods and applications. Netherlands: Springer.
Moore G.W., Goodman M., Barnabas J. 1973. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. J. Theoret. Biol. 38:423–457.
Morlon H. 2014. Phylogenetic approaches for studying diversification. Ecol. Lett. 17:508–525.
Morlon H., Lewitus E., Condamine F.L., Manceau M., Clavel J., Drury J. 2015. RPANDA: Phylogenetic ANalyses of Diversification in R. R package version 1.0. Methods Ecol. Evol.
Nee S., Mooers A.O., Harvey P.H. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proc. Natl Acad. Sci. USA 89:8322–8326.
Newman M.E.J. 2006. Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74:036104.
Noh J.D., Rieger H. 2004. Random walks on complex networks. Phys. Rev. Lett. 92:118701.
Pelleg D., Moore A. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In: Proceedings of the 17th International Conference on Machine Learning. Morgan Kaufmann. P. 727–734.
Pennell M.W., Harmon L.J. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Annals N. Y. Acad. Sci. 1289:90–105.
Poon A.F.Y., Walker L.W., Murray H., McCloskey R.M., Harrigan P.R., Liang R.H. 2013. Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. PloS One 8:e78122.
Popinga A., Vaughan T., Stadler T., Drummond A.J. 2015. Inferring epidemiological dynamics with bayesian coalescent inference: the merits of deterministic and stochastic models. Genetics 199:595–607.
Purvis A., Gittleman J.L., Brooks T. 2005. Phylogeny and conservation. No. 8 in Conservation Biology. Cambridge (MA): Cambridge University Press.
Pybus O.G., Harvey P.H. 2000. Testing macro–evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. B Biol. Sci. 267:2267–2272.
Rabosky D.L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PloS One 9:e89543.
Rambaut A., Pybus O.G., Nelson M.I., Viboud C., Taubenberger J.K., Holmes E.C. 2008. The genomic and epidemiological dynamics of human influenza A virus. Nature 453:615–619.
Ravasz E., Somera A.L., Mongru D.A., Oltvai Z.N., Barabási A. 2002. Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555.
Revell L.J. 2012. phytools: An r package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3:217–223.
Reynolds A., Richards G., de la Iglesia B., Rayward-Smith V. 2006. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. J. Math. Modell. Algorithms 5:475–504.
Rezende E.L., Lavabre J.E., Guimarães P.R., Jordano P., Bascompte J.. 2007. Non-random coextinctions in phylogenetically structured mutualistic networks. Nature 448:925–928.
Robinson D., Foulds L. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Shah P., Fitzpatrick B.M., Fordyce J.A. 2013. A parametric method for assessing diversification-rate variation in phylogenetic trees. Evolution 67:368–377.
Shames I., Summers T., Cantoni M. 2014. Manipulating factions evolved in signed networks. Proceedings of the 21st International Symposium on Mathematical Theory of Networks and Systems (MTNS), Groningen, The Netherlands.
Shen H.-W., Cheng X.-Q. 2010. Spectral methods for the detection of network community structure: a comparative analysis. J. Stat. Mechan. Theory Exp. 2010:P10020.
Shen-Orr S.S., Milo R., Mangan S., Alon U. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. Nat. Genetics 31:64–68.
Shi J.J., Rabosky D.L. 2015. Speciation dynamics during the global radiation of extant bats: BAT SPECIATION DYNAMICS. Evolution 69:1528–1545.
Székely, G. and M. Rizzo. 2005. Hierarchical clustering via joint between-within distances: extending ward's minimum variance method. J. Classification 22:151–183.
Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A., Tsafou K.P., Kuhn M., Bork P., Jensen L.J., von Mering C. 2015. v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43:D447–452.
Vijaykrishna D., Holmes E.C., Joseph U., Fourment M., Su Y.C., Halpin R., Lee R.T., Deng Y.-M., Gunalan V., Lin X., Stockwell T.B., Fedorova N.B., Zhou B., Spirason N., Kühnert D., Bošková V., Stadler T., Costa A.-M., Dwyer D.E., Huang Q.S., Jennings L.C., Rawlinson W., Sullivan S.G., Hurt A.C., Maurer-Stroh S., Wentworth D.E., Smith G.J., Barr I.G. 2014. The contrasting phylodynamics of human influenza b viruses. eLife 4:e05055.
von Luxburg U. 2007. A tutorial on spectral clustering. Stat. Comput. 17:395–416.
Webb C.O., Ackerly D.D., McPeek M.A., Donoghue M.J. 2002. Phylogenies and community ecology. Ann. Rev. Ecol. Syst. 33: 475–505.
Whidden C., Matsen T., Frederick A. 2015. Quantifying MCMC exploration of phylogenetic tree space. Syst. Biol. 64:472–491.