PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Gorman E, Lladser ME. 2023 Sparsification of large ultrametric matrices: insights into the microbial Tree of Life. *Proc. R. Soc. A* **479**: 20220847.
https://doi.org/10.1098/rspa.2022.0847

Received: 15 December 2022
Accepted: 22 August 2023

Subject Areas:
applied mathematics, computational biology, computational mathematics

Keywords:
double principal coordinate analysis, Haar-like wavelets, sparsification, phylogenetic covariance matrix, ultrametric matrix, UniFrac

Author for correspondence:
Manuel E. Lladser
e-mail: lladser@colorado.edu

Sparsification of large ultrametric matrices: insights into the microbial Tree of Life

Evan Gorman and Manuel E. Lladser

Department of Applied Mathematics, University of Colorado, PO Box 526 UCB, Boulder, CO 80309, USA

MEL, 0000-0001-6843-6845

Ultrametric matrices appear in many domains of mathematics and science; nevertheless, they can be large and dense, making them difficult to store and manipulate, unlike large but sparse matrices. In this manuscript, we exploit that ultrametric matrices can be represented as binary trees to sparsify them via an orthonormal base change based on Haar-like wavelets. We show that, with overwhelmingly high probability, only an asymptotically negligible fraction of the off-diagonal entries in random but large ultrametric matrices remain non-zero after the base change; and develop an algorithm to sparsify such matrices directly from their tree representation. We also identify the subclass of matrices diagonalized by the Haar-like wavelets and supply a sufficient condition to approximate the spectrum of ultrametric matrices outside this subclass. Our methods give computational access to a covariance matrix model of the microbiologists' Tree of Life, which was previously inaccessible due to its size, and motivate introducing a new wavelet-based (beta-diversity) metric to compare microbial environments. Unlike the established metrics, the new metric may be used to identify internal nodes (i.e. splits) in the Tree that link microbial composition and environmental factors in a statistically significant manner.

1. Introduction

Ultrametric matrices appear across many domains of mathematics and science. They comprise an important class of matrices called inverse-M matrices [1] and are a key object of study in potential theory and Markov

THE ROYAL SOCIETY PUBLISHING

Chains [2]. In scientific applications, ultrametric matrices act as covariance models in phylogenetic comparative analysis [3], network inference [4] and energy models in statistical physics [5]. In the context of modern data science, recent work has shown that the matrix of normalized Euclidean distances between points in some random subsets of $\mathbb{R}^d$ converge in probability to an ultrametric matrix as d tends to infinity [6,7].

In many applications, the underlying ultrametric matrix can be dense and potentially too large to store in computer memory and manipulate. Nonetheless, if a sparse representation of such a matrix can be found, many otherwise impossible tasks may become computationally feasible, such as matrix inversion and eigenvalue decompositions.

This paper tackles the challenge of sparsifying ultrametric matrices, which we define next.

In what remains of this manuscript, $n \geq 1$ is an integer and $[n] := \{1, \ldots, n\}$. Vectors and sometimes functions are represented as column vectors, and the transpose of a vector or matrix A is denoted A'.

Definition 1.1 ([8]). A matrix $S \in \mathbb{R}^{n \times n}$ is ultrametric if it is symmetric with non-negative entries, and $S(i,j) \geq \min\{S(i,k), S(k,j)\}$ for all $i,j,k \in [n]$. In particular, $S(i,i) \geq \max\{S(i,t) : t \neq i\}$ for all $i \in [n]$. Accordingly, S is called strictly ultrametric when it is ultrametric and $S(i,i) > \max\{S(i,t) : t \neq i\}$ for all $i \in [n]$. For $n = 1$, the last inequality is replaced with $S(i,i) > 0$.

Ultrametric matrices have rich properties that are not made evident by their definition [1]. In particular, if S is strictly ultrametric then it is positive definite (hence invertible), S^{-1} is strictly diagonally dominant with non-positive off-diagonal entries, and $S(i,j) = 0$ if and only if $S^{-1}(i,j) = 0$. These properties were initially proved using probabilistic methods [9]. An alternative proof is based on a representation of strictly ultrametric matrices as weighted and rooted binary trees with special characteristics [10, proposition 2.1]. Notably, using a perturbation argument, analogous representations also apply to ultrametric matrices that are not necessarily strictly ultrametric [8]. See also [1, proposition 3.4]. Indeed, if $n > 1$ and S is a symmetric matrix of dimensions $n \times n$ with non-negative entries, then S is ultrametric if and only if there exists a permutation matrix P and ultrametric matrices A and B such that

$$P\big(S - \min(S)\,\mathbf{1}\mathbf{1}'\big)P' = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}, \tag{1.1}$$

where $\min(S)$ is the smallest entry in S, and $\mathbf{1} \in \mathbb{R}^n$ is the column vector of ones. Since A and B are of the same kind as S, this process may be applied recursively and the matrix S encoded as a binary tree (an ordered one). Here, we adopt a slightly different encoding to the one in [8,10], which is more suitable for our purposes. The reader unfamiliar with the jargon and notation of trees may skip ahead to §1b and come back to make better sense of the construction below.

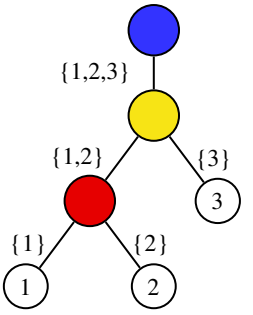
Recall that a bifurcating tree is a tree where each node has degree one or three. We can represent an ultrametric matrix S of dimensions $n \times n$ as a weighted bifurcating tree with $2n$ nodes (of which half are leaves) and hence $(2n-1)$ edges, satisfying the following definition.

Definition 1.2. An out-rooted bifurcating tree (ORB-tree) with n leaves is a weighted rooted tree with the following properties: each vertex has degree 1 or 3; its leaf set is $[n]$ and excludes the root, which has degree 1; each edge is labelled by the subset of leaves that descend from it; and the length $\ell(e)$ of each edge e is non-negative.

ORB-trees are therefore isomorphic to bifurcating trees rooted at a leaf, which is deemed an internal node. This unconventional choice ensures a one-to-one correspondence between their internal nodes and vectors in an associated orthonormal basis (see definition 2.1).

We emphasize that ORB-trees are determined by both their topology and branch lengths.

The representation of an ultrametric matrix S as an ORB-tree may be obtained as follows. The only edge emanating from the root is labelled as $[n]$ and defined to have length $\min(S)$. The only child of the root has two children. One child descends from an edge labelled by the rows (or columns) of S associated with the matrix A before applying the permutation matrix P in (1.1).



e	ℓ(e)	δ_e
{1,2,3}	2	(1,1,1)′
{1,2}	0	(1,1,0)′
{1}	3	(1,0,0)′
{2}	1	(0,1,0)′
{3}	2	(0,0,1)′

Figure 1. ORB-tree and ultrametric matrix correspondence. ORB-tree with interior nodes coloured blue (the root), yellow and red, and branch lengths as given in the table. Accordingly, it encodes a (strictly) ultrametric matrix.

This edge has length min(A). Likewise, the other child descends from an edge labelled by the rows associated with the matrix B and has length min(B). Since A and B are ultrametric, just of smaller dimensions, the tree may be grown recursively from any descendent of the root that is not associated with an ultrametric matrix of dimensions 1×1. The latter represent edges that parent a leaf in the ORB-tree. (When S is strictly ultrametric, these edges must have a strictly positive length because 1×1 strictly ultrametric matrices are strictly positive real numbers.) Conversely, the matrix may be recovered from the ORB-tree as follows. For each edge e in the ORB-tree, let δ_e be the vector of dimension n with entries $\delta_e(i) = 1$ for $i \in e$ and $\delta_e(i) = 0$ for $i \notin e$. It follows that

$$S = \sum_{e \in E} \ell(e)\, \delta_e\, \delta_e'; \text{ in particular, } S(i,j) = \sum_{e \in [i \wedge j, \circ]} \ell(e). \tag{1.2}$$

In particular, an ultrametric matrix may be uniquely recovered from any of its ORB-tree representations. We say 'any' because several trees may represent the same matrix, for example, when the matrix A or B in equation (1.1) is diagonal.

To fix ideas, see figure 1.

We call a matrix with entries such as (1.2) the covariance matrix of the ORB-tree. This terminology is borrowed from the ecology literature where matrices like this are commonly referred to as a tree-structured or phylogenetic covariance matrices [11]. In this setting, the leaves represent organisms, and the matrix entries denote a trait's covariance between pairs of organisms. (The term of cophenetic matrix or cophenetic distance has also been used occasionally in the hierarchical clustering literature [12].) For instance, the matrix encoded by the tree in figure 1 is

$$2\,\delta_{\{1,2,3\}}\delta_{\{1,2,3\}}' + 3\,\delta_{\{1\}}\delta_{\{1\}}' + \delta_{\{2\}}\delta_{\{2\}}' + 2\,\delta_{\{3\}}\delta_{\{3\}}' = \begin{pmatrix} 5 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 4 \end{pmatrix}.$$

Due to the identity in equation (1.2), ultrametric matrices are usually dense; in fact, all the entries in a strictly ultrametric matrix must be non-zero. Nevertheless, precisely because of this identity, their entries contain much redundancy, suggesting they may be amenable to some form of compression. In this manuscript, we apply a change of bases—a discrete wavelet, in fact—with respect to which the covariance matrix of an ORB-tree often becomes sparse.

Wavelets are localized, wave-like functions developed to analyse non-stationary and noisy continuous signals. Traditional wavelets are defined only in Euclidean spaces and have been remarkably successful in identifying multi-scale structures in signals and producing sparse representations of the same [13].

The Haar wavelet is among the oldest and involves averaging a signal locally at different time or space scales [14]. Recently, the authors of [15] extended it past continuous signals introducing the Haar-like wavelet. This new, discrete, wavelet is designed for the multi-scale analysis of discrete datasets equipped with a partition tree—a hierarchical structure that clusters the data

into smaller subsets recursively. Due to the organization of such datasets into different tree levels (i.e. scales) and clusters (i.e. localizations), Haar-like wavelets may identify meaningful patterns in data that would be impossible to detect otherwise—especially in noisy high dimensional datasets.

This paper exploits the representation of ultrametric matrices as ORB-trees to sparsify and sometimes diagonalize the former via a change of basis. This basis is composed of the so-called Haar-like wavelets of the associated ORB-trees. The sparsification achieved by these wavelets can be substantial in large, ultrametric matrices, which would otherwise be inaccessible due to their size. These sparse representations can be valuable in phylogenetic applications [3], network inference [4] and hierarchical clustering problems [12,16] as their models often rely on tree-structured covariance matrices.

(a) Paper organization

In §2, we specialize the Haar-like basis from [15] and give a geometric interpretation of its action on ORB-trees. This results in a closed-form expression for the transformed ultrametric matrix that can be computed efficiently—without having to pre-compute the matrix from the tree. In §3, we present conditions under which the Haar-like basis can be used to sparsify large ultrametric matrices, and show that the basis can substantially sparsify most large random ORB-tree's covariance matrices. Following in §4, we show that the Haar-like basis can be used sometimes to estimate the eigenvalues of an ultrametric matrix. In particular, §4a introduces the concept of trace-balanced trees; these are the ORB-trees whose covariance matrices are diagonalized by their Haar-like basis. Section 4b then details the possible spectrums of trace-balanced trees through a constructive proof that is of consequence for the symmetric non-negative inverse eigenvalue problem [17–19].

Section 5 is devoted to an extensive proof of concept of our methods in metagenomics (i.e. the study of microbial environments based on genetic material extracted directly from them). Specifically, we sparsify the covariance matrix associated with microbiologists' Tree of Life. The significant sparsification achieved by our methods motivates introducing a new but wavelet-based phylogenetic (β-diversity) distance, corresponding to a multi-scale analysis of organism abundances in microbial environments. This new distance gives remarkably similar results to other well-known metrics on a previously studied dataset. However, unlike the established metrics, it can also determine the splits in the Tree responsible for the observed microbial compositions and quantify their respective importance.

Finally, technical proofs can be found in appendix B.

(b) General notation and terminology

For real-vectors $x = (x_i)_{1\le i\le k}$ and $y = (y_i)_{1\le i\le k}$ of dimension k, let $\bar{x} := \frac{1}{k}\sum_{i=1}^{k} x_i$, $\langle x, y\rangle := x'y = \sum_{i=1}^{k} x_i y_i$ and $||x||_2 := \sqrt{\langle x, x\rangle}$. Also, let $[\![\cdot]\!]$ denote the indicator function of the proposition within the parentheses.

In our context, trees are finite undirected connected graphs without cycles.

In what remains of this manuscript, T denotes an ORB-tree with n leaves and branch length function $\ell : E \to [0, \infty)$. We denote the vertex and edge set of T as V and E, respectively. The root of T is denoted as $\circ$. Recall that $\circ$ must have degree one. The set of internal nodes of T is denoted as I, whereas its set of leaves is denoted as L. By definition, $\circ \in I$ and I and L partition V. It follows that $|L| = |I| = n$, hence $|V| = 2n$. Also $|E| = |V| - 1$ because T is a tree. We define $|T| := |V|$ and use this notation when we want to emphasize a direct relationship with the ORB-tree.

For $i, j \in V$, a path of length l between i and j is a sequence $v_0, \ldots, v_l \in V$ such that $v_0 = i$, $v_l = j$ and $\{v_k, v_{k+1}\} \in E$ for $0 \le k < l$. Unless otherwise stated, we write $[i, j]$ to denote the set of edges in the shortest path between i and j in T. This path is unique because trees have no cycles. The depth of i, denoted depth(i), is defined as $|[i, \circ]|$, i.e. the number of edges that connect i with the root. We say that i is an ancestor of j, or alternatively j is a descendent of i, when $i \in [\circ, j]$. In particular,

every node is an ancestor and a descendant of itself. Further, $(i \wedge j)$ denotes the so-called least-common ancestor to i and j. This is the $v \in V$ that maximizes $|[v, \circ]|$, among all the nodes that are ancestors to both i and j.

We define

$$\ell(i,j) := \sum_{e \in [i,j]} \ell(e).$$

In addition, for $J \subset L$ and $i \in V$, define $\ell(J,i)$ as the column vector of dimension $|J|$ with entries $\ell(j,i)$, for $j \in J$. $\ell(i,J)$ is the transpose of $\ell(J,i)$.

For each $i \in V$, $T(i)$ denotes the subtree of T rooted at i. In particular, the vertex set of $T(i)$ is the subset of nodes in T that descend from i, and its edge set is the subset of edges that connect two descendants of i. $L(i)$ denotes the leaf set of $T(i)$. Likewise, for each $e = \{i,j\} \in E$, if i is closer to the root than j, $T(e)$ and $L(e)$ denote $T(j)$ and $L(j)$, respectively.

2. Haar-like basis of ORB-trees

In this section, we specialize the concept of Haar-like basis given in [15] to our setting of ORB-trees. (The authors of [15] consider multi-furcating trees, which encompass bifurcating ones.)

The key result in this section is theorem 2.3, which provides a rather explicit expression for the entries of the covariance matrix of an ORB-tree with respect to its Haar-like basis—see equation (2.2). This expression is helpful because it lets us anticipate entries that vanish when transitioning between bases by directly analysing the tree's topology.

To construct the Haar-like wavelets, it is convenient to represent the nodes in $I \setminus \{o\}$ as binary strings. With this convention, the (only) child of the root is denoted as ε—the so-called empty string. Further, the children of each node $v \in I \setminus \{o\}$ are $v0$ (i.e. the string v with the character zero appended at the end) and $v1$ (i.e. v with the character one appended at the end). In particular, $\varepsilon 0 = 0$ and $\varepsilon 1 = 1$. Keeping this notation in mind, the Haar-like wavelets associated with an ORB-tree are given by the following definition.

Definition 2.1 (Specialization from [15]). The Haar-like basis associated with T is the set of transformations $\{\varphi_v\}_{v \in I}$ defined as follows:

$$\varphi_o(i) := \frac{1}{\sqrt{|L|}}, \quad \text{for all } i \in L;$$

and, for each $v \in I$ with $v \neq o$

$$\varphi_v(i) := \begin{cases} +\sqrt{\dfrac{|L(v1)|}{|L(v0)| \cdot |L(v)|}}, & i \in L(v0); \\ -\sqrt{\dfrac{|L(v0)|}{|L(v1)| \cdot |L(v)|}}, & i \in L(v1); \\ 0, & \text{otherwise.} \end{cases}$$

The Haar-like matrix associated with T is the matrix Φ with columns φ_v, $v \in I$.

That is, for each interior node v, φ_v is supported on $L(v)$, the set of leaves that descend from v. Further, for $v \neq \circ$, there are constants α, β such that $\varphi_v(i) = \alpha$ for leaves on the left-subtree dangling from v, and $\varphi_v(i) = \beta$ for leaves on the right-subtree dangling from v. These constants are such that φ_v has mean zero and unit norm (i.e. $|L(v0)| \cdot \alpha + |L(v1)| \cdot \beta = 0$ and $|L(v0)| \cdot \alpha^2 + |L(v1)| \cdot \beta^2 = 1$). In this manuscript, we adopt the convention that $\alpha > 0 > \beta$. To fix ideas, see figure 2.

We remark that, in the context of compositional data analysis, the Haar-like basis is equivalent to the orthonormal basis used in the isometric log-ratio (ILR) transform [20]. We revisit the implications of this connection in §5.

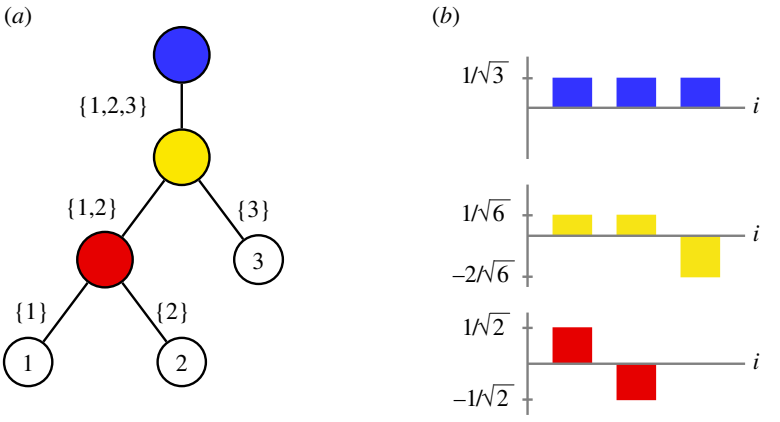


Figure 2. Visualization of the Haar-like wavelet basis associated with an ORB-tree. (*a*) Tree with leaves 1, 2, 3 and coloured interior nodes (blue root). Edges are labelled by the subsets of leaves that descend from them. (*b*) Haar-like wavelets associated with the interior nodes of the ORB-tree. These depend on its topology but not its branch lengths.

The terminology of basis in definition 2.1 is justified by the fact that

$$\text{if } u, v \in I \text{ then } \langle \varphi_u, \varphi_v \rangle = [\![u = v]\!]. \tag{2.1}$$

(Recall that $[\![\cdot]\!]$ denotes the indicator function of the proposition within.) In particular, $\{\varphi_v\}_{v\in I}$ is an orthonormal basis for the vector space of functions from L to $\mathbb{R}$. (See appendix Ba for a self-contained justification of the orthonormality of the wavelets.) Note that the Haar-like matrix Φ has its rows indexed by L and its columns indexed by I. Since $|L| = |I|$, Φ is a square matrix, an orthonormal one. Further, $\Phi' S \Phi$ has its rows and columns indexed by I.

The following definition is useful to understand the relationship between the Haar-like basis of an ORB-tree and its associated covariance matrix.

Definition 2.2. The trace branch length of T is the function $\ell^* : E \to [0, \infty)$ defined as $\ell^*(e) := |L(e)|\, \ell(e)$, for each $e \in E$.

Theorem 2.3. *If $v \in I$ then $S\,\varphi_v = \mathrm{diag}(\ell^*(L, v))\,\varphi_v$.*

It follows from the theorem that for each $u, v \in I$

$$(\Phi' S \Phi)(u, v) = \varphi_u' S \varphi_v = \sum_{i \in L(v) \cap L(u)} \varphi_u(i)\, \varphi_v(i)\, \ell^*(i, v). \tag{2.2}$$

In particular, $(\Phi' S \Phi)(u, v) = 0$ when $u, v \in I$ are such that $L(u) \cap L(v) = \emptyset$. This suggests that the Haar-like matrix can be used to sparsify the covariance matrix of the ORB-tree. The following result is critical to assess how effective this sparsification is in practice.

Lemma 2.4. *For all $u, v \in V$, $L(u) \cap L(v) \neq \emptyset$ if and only if u is an ancestor of v or vice versa.*

(a) Fast sparsification algorithm

A non-trivial challenge to storing and manipulating large ultrametric matrices is that they are very dense in practice. In many applications, however, particularly metagenomics, these matrices are encoded in advance as ORB-trees. This allows us to sparsify them without explicitly computing or storing them in computer memory. It also allows us to anticipate which entries may remain non-zero after sparsification. In fact, due to equation (2.2) and lemma 2.4, all that is required to sparsify an ultrametric matrix from its ORB-tree representation is to precompute the leaves that descend

from each internal node (i.e. the sets $L(v)$, with $v \in I$) and the trace branch length between them (definition 2.2). This can be achieved with two postorder traversals of the ORB-tree. We convey these ideas in the following pseudo-code (algorithm 1), which is fully coded and available on https://github.com/edgor17/Sparsify-Ultrametric.

Algorithm 1. Phylogenetic covariance matrix sparsification.

Input. ORB-tree T with covariance matrix S
Output. Only possibly non-zero entries in $\Phi' S \Phi$
for $v \in I$ in postorder traversal of T **do**
  for $i \in L$ **do**
    if $v = \circ$ **then**
      $\varphi_o(i) \leftarrow \frac{1}{\sqrt{|L|}}$
    else if $i \in L(v0)$ **then**
      $\varphi_v(i) \leftarrow +\sqrt{\frac{|L(v1)|}{|L(v0)| \cdot |L(v)|}}$
      $\ell^*(v, i) \leftarrow \ell^*(i, v0) + |L(v0)| \cdot \ell(v0, v)$
    else if $i \in L(v1)$ **then**
      $\varphi_v(i) \leftarrow -\sqrt{\frac{|L(v0)|}{|L(v1)| \cdot |L(v)|}}$
      $\ell^*(v, i) \leftarrow \ell^*(i, v1) + |L(v1)| \cdot \ell(v1, v)$
    else
      $\varphi_v(i) \leftarrow 0$
    end if
  end for
end for
for $v \in I$ in postorder traversal of T **do**
  while parent$(v) \neq \emptyset$ **do**
    $u \leftarrow$ parent(v)
    $M(u, v) \leftarrow \sum_{i \in L(v) \cap L(u)} \varphi_u(i)\, \varphi_v(i)\, \ell^*(v, i)$ %{equation (**??**)}
  end while
end for
return $M(u, v)$ for $u, v \in I$ such that $L(u) \cap L(v) \neq \emptyset$

3. Sparsification of covariance matrices of ORB-trees

In this section, we quantify how much of the covariance matrix of an ORB-tree can be sparsified by its Haar-like matrix. To state our main result, we require the following definitions.

Definition 3.1. Recall that $|T|$ denotes the total number of nodes in T. The average subtree size of T is the quantity $\mathrm{avg}(T) := \sum_{v \in V} |T(v)|/|T|$.

Definition 3.2 ([21]). The internal and external path lengths of T are the quantities defined as $\mathrm{IPL}(T) := \sum_{v \in I} \mathrm{depth}(v)$ and $\mathrm{EPL}(T) := \sum_{v \in L} \mathrm{depth}(v)$, respectively. The total path length of T is the quantity $\mathrm{TPL}(T) := \mathrm{IPL}(T) + \mathrm{EPL}(T)$.

We note the relationship

$$\mathrm{avg}(T) = 1 + \frac{\mathrm{TPL}(T)}{|T|}, \tag{3.1}$$

because

$$\mathrm{TPL}(T) = \sum_{v \in V} \sum_{u \in V \setminus \{\circ\}} [\![v \in T(u)]\!] = \sum_{u \in V \setminus \{\circ\}} |T(u)| = \left\{ \sum_{u \in V} |T(u)| \right\} - |T|.$$

Definition 3.3. The interior of T is the tree $\mathring{T}$ obtained by trimming the leaves of T.

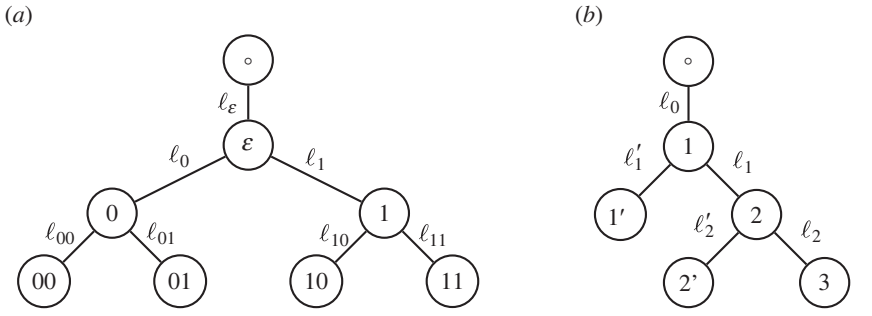


Figure 3. Visualization of (*a*) a perfect binary tree and (*b*) a binary caterpillar tree of heights 3.

Clearly, $\mathrm{IPL}(T) = \mathrm{TPL}(\mathring{T})$.

As mentioned earlier, the identity in (2.2) guarantees that some entries of $\Phi' S \Phi$ vanish. The following result gives a lower-bound for the number of such entries. This bound is independent of the branch lengths and depends—only—on the tree topology.

Theorem 3.4. *Let S be the covariance matrix associated with an ORB-tree T with Haar-like matrix Φ. If ζ denotes the fraction of vanishing entries in $\Phi' S \Phi$ then*

$$\zeta \geq 1 + \frac{1}{|L|} - 2\frac{\mathrm{avg}(\mathring{T})}{|\mathring{T}|} = 1 - \frac{1}{|L|} - 2\frac{\mathrm{TPL}(\mathring{T})}{|\mathring{T}|^2}.$$

It follows from the first lemma in [21, Section 6.4] that for an ORB-tree T, $\mathrm{EPL}(T) - \mathrm{IPL}(T) = 2|I| - 1$, which together with the previous theorem lets us conclude the following asymptotic result.

Corollary 3.5. *If either* $\mathrm{avg}(\mathring{T}) \ll |\mathring{T}|$, $\mathrm{TPL}(\mathring{T}) \ll |\mathring{T}|^2$, $\mathrm{IPL}(T) \ll |I|^2$, *or* $\mathrm{EPL}(T) \ll |L|^2$ *as* $|T| \to \infty$, *then* $\zeta = 1 - o(1)$.

In other words, if T grows so that either of the asymptotic inequalities in the above corollary applies, then an asymptotically negligible fraction of the off-diagonal entries in $\Phi' S \Phi$ will be non-zero.

The last asymptotic condition in the corollary 3.5 (i.e. that $\mathrm{EPL}(T) \ll |L|^2$) is of relevance in phylogenetic studies. In that context, the external path length of a tree is called its Sackin's index [22–24]. This index is used as a measure of the balance or unbalance of phylogenetic trees.

We end this section testing the efficacy of theorem 3.4 in sparsifying the covariance matrices of two types of ORB-trees, each representing an extreme of balance (or imbalance), as given by the following definitions.

For the first definition, recall that ε denotes the empty string.

Definition 3.6. A perfect binary tree of height $h \geq 1$ is the ORB-tree whose nodes are $\circ$ (the root) and all binary strings of length at most $(h-1)$, and whose edges are of the form $\{\circ, \varepsilon\}$ and $\{u, v\}$, with u and v binary strings with lengths that differ by one.

Definition 3.7. A binary caterpillar tree of height $h \geq 1$ is the ORB-tree with nodes $\circ$ (the root), $1, \ldots, h, 1', \ldots, (h-1)'$ and edges of the form $\{\circ, 1\}$, $\{i, i+1\}$, for $i = 1, \ldots, (h-1)$, and $\{j, j'\}$ for $j = 1, \ldots, (h-1)$.

See figure 3 to fix ideas on the above definitions.

Perfect binary trees give the most balanced topology among the ORB-trees. If T is a perfect binary tree then $\mathrm{avg}(\mathring{T}) \ll |\mathring{T}|$ (see appendix Be). In particular, due to theorem 3.5, the associated Haar-like matrix can be used to asymptotically annihilate (via a similarity transformation) the off-diagonal entries of the covariance matrix of a perfect binary tree as its height tends to infinity.

By contrast, binary caterpillar trees give the most unbalanced topology among the ORB-trees. If T is such a tree then $\mathrm{avg}(\mathring{T}) \sim |\mathring{T}|/2$ (see appendix Bf). Hence, the lower-bound provided by theorem 3.4 is trivial, and we cannot guarantee that the Haar-like matrix associated with a sizeable binary caterpillar tree annihilates the off-diagonal entries of the associated ultrametric matrix in any significant way.

(a) Covariance matrices of large random ORB-trees

Perfect binary trees and caterpillar trees are opposite extremes of how balanced (or unbalanced) ORB-trees can be. It is therefore unclear how much sparsification the Haar-like matrix of a large but generic ORB-tree can induce on its covariance matrix. To address this issue, we consider a natural ensemble of random ORB-trees.

In what follows, $\mathbb{T}$ denotes a uniformly at random ORB-tree with n leaves, i.e. with n internal nodes and hence of size $2n$. Such trees may be generated using the Catalan distribution [21, Section 6.7]. This probability model produces full binary trees (i.e. trees in which each node has 0 or 2 children) with a given number of leaves, which we may turn into an ORB-tree by appending their root to a new external one.

Let $\mathbb{S}$ denote the covariance matrix of $\mathbb{T}$, and ζ the fraction of vanishing entries in $\Phi'\mathbb{S}\Phi$, where Φ is the Haar-like matrix associated with $\mathbb{T}$. It turns out that the mean and variance of the internal path length of $\mathbb{T}$ are given by

$$\mathbb{E}\big(\mathrm{IPL}(\mathbb{T})\big) \sim \sqrt{\pi n^3} \tag{3.2}$$

and

$$\mathbb{V}\big(\mathrm{IPL}(\mathbb{T})\big) \sim \left(\frac{10}{3} - \pi\right) n^3. \tag{3.3}$$

The identity in equation (3.2) follows from [25, Proposition VII.3.]. The identity in (3.3) may be regarded a refinement of [25, Note VII.12].

As the following result implies, the Haar-like basis of most large ORB-trees should be highly effective in sparsifying their covariance matrix.

Corollary 3.8. *If $\mathbb{T}$ is a uniformly at random ORB-tree with n leaves, then $\lim_{n\to\infty} \zeta = 1$ in probability. Namely, for each $\delta > 0$, $\zeta > (1-\delta)$ with overwhelmingly high probability as $n \to \infty$.*

4. Spectra of ultrametric matrices

The ORB-tree representation of ultrametric matrices offers new approaches to studying their spectrum. This is of interest for domains such as structural biology [26] and metagenomics [27], where ultrametric matrices emerge as phylogenetic covariance ones. In PCA, a widely used statistical technique to represent high-dimensional data in low dimension (usually two or three), eigenvalue estimates can provide information about the relative importance of different eigenvectors onto which to project data with a correlation structure described by an ultrametric matrix, such as certain trait models in Ecology [27]. Similarly, since ultrametric matrices are symmetric, one can use eigenvalue estimates to establish a threshold for determining the number of singular values required for an accurate low-rank approximation. This section examines how to approximate the spectrum of ultrametric matrices.

In what follows, for a given function $x : L \to \mathbb{R}$ and non-empty $J \subset L$, we define the mean value and variance of x over J naturally as the following quantities:

$$\mathrm{avg}(x; J) := \frac{1}{|J|} \sum_{j \in J} x(j)$$

and

$$\mathrm{var}(x; J) := \frac{1}{|J|} \sum_{j \in J} \big(x(j) - \mathrm{avg}(x; J)\big)^2.$$

For each $v \in V$, let parent(v) denote the parent of node v in T. Define

$$\rho_v := \frac{|L(v)|}{|L(\text{parent}(v))|}.$$

In addition, define for $v \in I$ the quantities

$$\lambda_v := (\Phi' S \Phi)(v, v) = \varphi_v' S \varphi_v = \sum_{i \in L(v)} \varphi_v^2(i)\, \ell^*(i, v), \tag{4.1}$$

where the last identity is based on equation (2.2). Note that, because φ_v has $L(v)$ as its support and $||\varphi_v||_2 = 1$, λ_v is a weighted average of the trace branch length between each leaf in $L(v)$ and v. In particular, since $L(u) \supset L(v)$ when u is an ancestor of v, the closer the internal node v is to the root, the more terms are averaged. (This emulates the averaging at different scales that the standard Haar wavelet transform does to a continuous signal.) Furthermore, since $\ell^*(e) \geq 0$ for all $e \in E$, $\lambda_v \geq 0$.

The next result provides a sufficient condition for λ_v, with $v \in I$, to be a good approximation of an eigenvalue of S. It also quantifies rather explicitly the cosine between $S\varphi_v$ and $\lambda_v \varphi_v$ to assess how close φ_v is to be an eigenvector of S.

In stating and proving the result, we reuse the notation of definition 2.1, and define $\neg 0 := 1$ and $\neg 1 := 0$. Also, for $\lambda \in \mathbb{R}$ and non-empty $A \subset \mathbb{R}$, we define

$$|\lambda - A| := \min_{a \in A} |\lambda - a|.$$

Theorem 4.1. *If $v \in I$ then $\lambda_v = \rho_{v0} \cdot \overline{\ell^*(L(v1), v)} + \rho_{v1} \cdot \overline{\ell^*(L(v0), v)}$, and*

$$|\lambda_v - \sigma(S)| \leq \sqrt{\rho_{v0} \cdot \rho_{v1} \cdot \left\{\overline{\ell^*(L(v0), v)} - \overline{\ell^*(L(v1), v)}\right\}^2 + \sum_{\alpha \in \{0,1\}} \rho_{v\alpha} \cdot \text{var}\left(\ell^*(L, v); L(v\neg\alpha)\right)}.$$

Furthermore, if $\lambda_v = 0$ then $S\varphi_v = 0$. Otherwise, if $\lambda_v \neq 0$ then

$$\cos(S\varphi_v, \lambda_v \varphi_v) = \frac{1}{\sqrt{1 + \{||(S - \lambda_v)\varphi_v||_2 / \lambda_v\}^2}}.$$

(a) Exact spectrum of trace-balanced ultrametric matrices

While theorem 3.4 guarantees that some entries in $\Phi' S \Phi$ vanish—regardless of branch lengths, additional constraints on the latter can lead to further sparsification. In this section, we identify the class of ORB-trees whose Haar-like basis fully sparsifies (i.e. diagonalizes) their covariance matrix.

Due to theorem 4.1, if $\overline{\ell^*(L(v1), v)} = \overline{\ell^*(L(v0), v)}$ and $\text{var}(\ell^*(L, v); L(v\alpha)) = 0$ for $\alpha \in \{0, 1\}$, i.e. $\ell^*(i, v) = \ell^*(j, v)$ for all $i, j \in L(v)$, then $S\varphi_v = \lambda_v \varphi_v$. Conversely, due to theorem 2.3, if $S\varphi_v = \lambda_v \varphi_v$ then $\lambda_v\, \varphi_v(i) = \ell^*(i, v)\, \varphi_v(i)$, for each $i \in L$. That is, $\ell^*(i, v) = \ell^*(j, v)$ for all $i, j \in L(v)$ because $\varphi_v(i) > 0$ when $i \in L(v)$. This motivates the following definition and establishes our next result.

Definition 4.2. An ORB-tree T is called trace-balanced when, for all $v \in I \setminus \{\circ\}$ and all $i, j \in L(v)$, $\ell^*(i, v) = \ell^*(j, v)$.

Corollary 4.3. *The Haar-like basis of T diagonalizes its covariance matrix S if and only if T is trace-balanced. In this case, the spectrum of S is*

$$\sigma(S) = \bigcup_{v \in I} \left\{\ell^*(v, i) \text{ for any } i \in L(v)\right\},$$

and the multiplicity of $\ell^*(v, i)$ is $|\{u \in I : \ell^*(v, i) = \ell^*(u, j), \text{ for some } j \in L(u)\}|$.

Our next corollary generalizes some of the results in [26], which investigated homogeneous (i.e. equal branch lengths) perfect binary trees (definition 3.6) in the phylogenetic setting to predict protein structure features.

Corollary 4.4. *If a perfect binary tree of height h has constant branch lengths at each level, and ℓ_j denotes the common length of the edges that connect a node at depth j with another at depth $(j+1)$, then the spectrum of the associated covariance matrix S is*

$$\sigma(S) = \left\{ \sum_{k=j}^{h-1} 2^{h-1-k} \ell_k, \text{ with } j = 0, \ldots, h-1 \right\}.$$

Furthermore, the multiplicity of an eigenvalue λ is $\sum_{j \in \Lambda} 2^{\max\{0, j-1\}}$, where

$$\Lambda := \left\{ j \in \{0, \ldots, h-1\} \text{ such that } \sum_{k=j}^{h-1} 2^{h-1-k} \ell_k = \lambda \right\}.$$

For example, if the tree in figure 3*a* satisfies that $\ell_\alpha = \ell_\beta$, whenever α and β are binary strings of the same length and at most 2, then it is trace-balanced and the eigenvalues of its associated ultrametric matrix are $\ell_{00}, \ell_{00}, \ell_{00} + 2\ell_0, \ell_{00} + 2\ell_0 + 4\ell_\varepsilon$, repeated according to their multiplicity.

We emphasize that corollary 4.4's hypotheses are sufficient but not necessary for a perfect binary tree to be trace-balanced (see appendix A). In fact, per corollary 4.7 ahead, far more varied spectra are possible when considering covariance matrices of perfect binary trees.

Next, we consider the binary caterpillar tree (definition 3.7). In particular, its internal and leaf set are $I = \{\circ, 1, \ldots, h-1\}$ and $L = \{1', \ldots, (h-1)', h\}$, respectively. Let ℓ_0 denote the branch length of $\{\circ, 1\}$, ℓ_i the length of $\{i, i+1\}$ for $i = 1, \ldots, (h-1)$, and ℓ'_j the branch length of $\{j, j'\}$ for $j = 1, \ldots, (h-1)$. Due to corollary 4.3, we have the following result.

Corollary 4.5. *A binary caterpillar tree of height h is trace-balanced if and only if $\ell'_j = \sum_{k=j}^{h-1} (h-k) \cdot \ell_k$, for $j = 1, \ldots, h-1$. In this case, the eigenvalues of its covariance matrix are $\ell'_0 \geq \ell'_1 \geq \cdots \geq \ell'_{h-1}$, repeated according to their multiplicity, where $\ell'_0 := \sum_{k=0}^{h-1} (h-k) \cdot \ell_k$.*

For instance, the tree in figure 3*b* is trace-balanced if and only if $\ell'_2 = \ell_2$ and $\ell'_1 = \ell_2 + 2\ell_1$. In this case, the associated eigenvalues are ℓ'_2, ℓ'_1, and $\ell'_0 := \ell'_1 + 3\ell_0$, repeated according to their multiplicity. Furthermore, when the branch lengths undergo slight perturbations—as quantified by theorem 4.1—these quantities may be regarded as approximate eigenvalues.

(b) Symmetric non-negative inverse eigenvalue problem

The main result in this section characterizes all the possible spectrums of covariance matrices of trace-balanced ORB-trees. Because its proof is constructive, it can be used to form ultrametric matrices with the desired spectrum and multiplicities. In particular, it is of consequence for the symmetric non-negative inverse eigenvalue problem (SNIEP), which aims to classify the possible spectra of symmetric non-negative matrices [17–19].

Definition 4.6. In a tree T, a function $f : I \to [0, \infty)$ is called decreasing when, for all distinct $u, v \in I$, if u is an ancestor of v then $f(u) \geq f(v)$.

Corollary 4.7. *In a trace-balanced ORB-tree T the function $v \longrightarrow \ell^*(v, i)$, with $v \in I$ and any $i \in L(v)$, is decreasing. Conversely, given any ORB-tree topology T and decreasing function $f : V \to [0, \infty)$, there is a branch length function $\ell : E \to [0, \infty)$ such that $\sigma(S) = f(I)$. Furthermore, the multiplicity of $\lambda \in \sigma(S)$ is $|f^{-1}(\{\lambda\})|$.*

5. New insights into the microbial Tree of Life

In this section, we apply our results to a phylogenetic covariance matrix associated with a standard reference phylogeny. Our main result, motivated by the spectral approximation properties in §4, is the definition of a new wavelet-based metric to compare microbial environments. In a broader context, the significant sparsification of phylogenetic covariance matrices may find use in other phylogenetic comparative methods that require otherwise

computationally impossible tasks related to such models, such as computing the full spectrum or inverse of a covariance matrix.

Many methods in microbiology rely on a phylogenetic tree relating microorganisms. At the microbial level, however, the notions of genus or species are ill-defined because microorganisms do not interbreed. So microbes' taxonomy and phylogeny are often based on the so-called 16S ribosomal RNA (16S rRNA) gene. This gene is present in all known single-cell organisms and can therefore be used as a phylogenetic marker. An operational taxonomic unit (OTU) is a cluster of these markers defined by some least level of DNA sequence similarity among its (highly) conserved regions.

Greengenes is a standarized database based on the 16S rRNA marker. It has been a standard reference in microbial studies, particularly metagenomics, and is the default option in QIITA [28]—a widely used open-source management platform for microbial analyses. Greengenes phylogenetic trees are built using FastTree [29] and their associated taxonomies are assigned using tax2tree [30]. Trees are typically stored in the newick format [31], which encodes their topology and branch lengths, and visualized using software such as FigTree [32] or the ETE Toolkit [33].

Figure 4 displays the Greengenes tree when OTUs are thresholded at a 97% sequence similarity—the average similarity of macro-organisms' DNA in the same species. The tree represents the inferred evolutionary history of modern-day microorganisms from common ancestors. Its root is at the centre of the circular layout, and each OTU is associated with a single leaf in the tree and vice versa. Branch lengths are a proxy of evolutionary time such as the estimated expected number of mutations per nucleotide site [34], and interior nodes (called splits) are inferred speciation events that have led to the present-day microorganisms in the database.

A fundamental problem in microbiology is to link environmental factors (such as acidity, light, nutrients, salinity, temperature, etc.) with microbial composition. A valuable tool for this has been the concept of β-diversity (i.e. a measure of differences between microbial composition across different environments). Early approaches [35,36] ignored the evolutionary relationships between microorganisms when comparing environments. Nonetheless, one would expect microbes with a shared evolutionary history to similarly thrive or struggle in similar environments. Phylogenetic-informed metrics were introduced precisely to convey this idea. These metrics require a phylogenetic tree relating the microorganisms observed in samples from all the environments under study. We emphasize that the construction and selection of these trees are outside the scope of this paper; as is typical in many metagenomics studies, we work with a pre-computed tree. Among other more recent phylogentic trees such as SILVA [37] and WoL [38], Greengenes has been a common choice of representative phylogeny. So, we base our application on the latter—though our methods could be applied to any reference phylogeny.

Double principal coordinate analysis (DPCoA) [27] is a phylogenetically informed β-diversity metric between pairs of microbial environments, which provides similar insights [39] to other more recent though more widely used distances such as unweighted and weighted UniFrac [40].

Let T be the ORB-tree associated with a phylogenetic tree (e.g. the 97% Greengenes tree after adding an external root $\circ$ and connecting it to the original root with a branch of length 0), and S its covariance matrix. Since all edges connected to a leaf have strictly positive length, S is strictly ultrametric [9,10]; in particular, positive definite.

In the context of phylogenetic-informed metrics, environments are represented as probability mass functions over the OTUs (i.e. leaves). We denote those functions with lower-case letters such as a and b, and interpret them as probability models over L. In particular, $a : L \to [0, +\infty)$ satisfies that $\sum_{x\in L} a(x) = 1$ and, for each $e \in E$, $a(e) = \sum_{x\in e} a(x)$. With this convention, the DPCoA distance between two environments a and b is defined as [27,39]

$$d(a, b) := \left\{ \sum_{e\in E} \ell(e) \left(a(e) - b(e)\right)^2 \right\}^{1/2} = \sqrt{(a - b)' S (a - b)}. \tag{5.1}$$

Since S is positive definite, DPCoA corresponds to a Mahalanobis distance [41]; implying that $d(\cdot, \cdot)$ is a metric—in the mathematical sense—in $\mathbb{R}^{|L|}$.

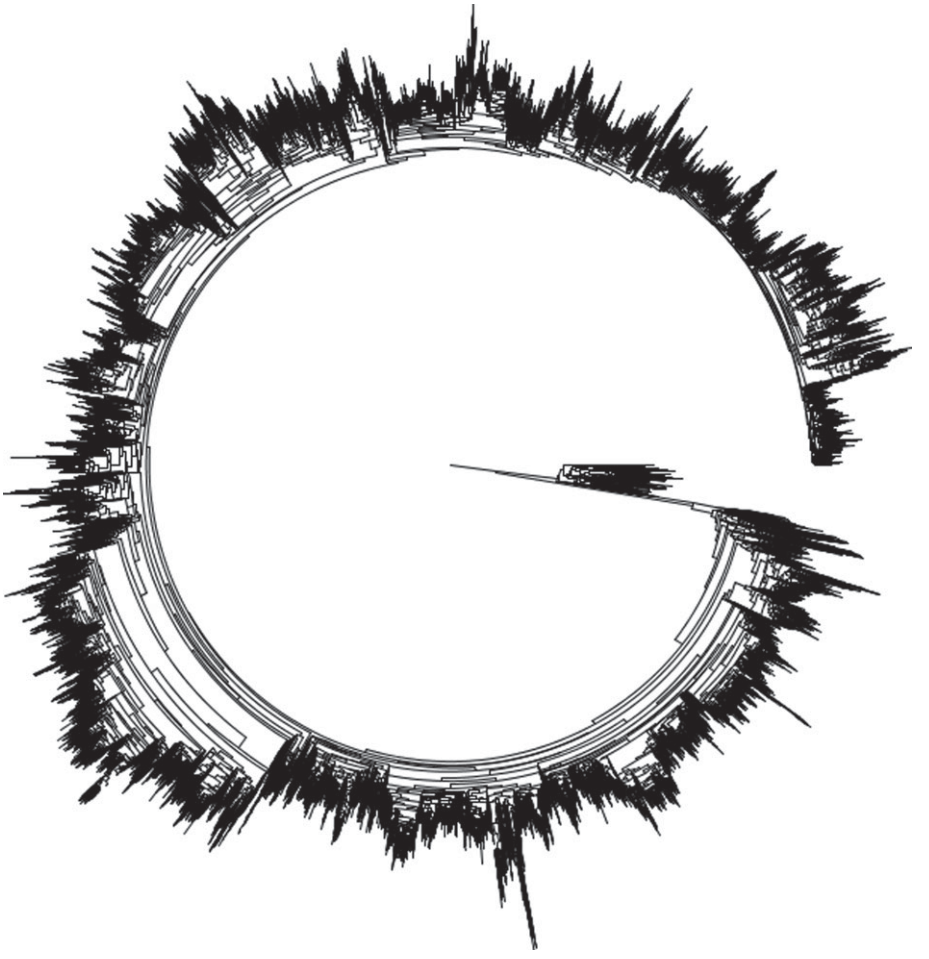


Figure 4. Circular layout of the 97% Greengenes tree. The tree has 99 322 leaves, 198 642 edges and height (i.e. maximal leaf depth) 107. The average branch length is 1.42×10^{-2} units, with lengths varying between 1.5×10^{-4} and 1.0.

The weighted and unweighted UniFrac distances are instead defined as follows [40]:

$$d_w(a,b) := \sum_{e \in E} \ell(e)\,|a(e) - b(e)|$$

and

$$d_u(a,b) := \frac{\sum_{e \in E} \ell(e)\,\left|[\![a(e) > 0]\!] - [\![b(e) > 0]\!]\right|}{\sum_{e \in E} \ell(e)}.$$

Both versions of UniFrac are known to satisfy the triangular inequality [42, Supplementary Methods]. DPCoA is also more robust to unbiased noise but more sensitive to outliers than UniFrac [39].

Regardless of the metric of choice, the standard approach to linking environmental factors with microbial composition goes roughly as follows [43]. First, environmental samples are collected, and each environment is represented by its OTU composition on the leaves of the phylogeny of reference. Then, the pairwise distance matrix between the environments is computed, and the environments are embedded into a low-dimensional Euclidean space using standard techniques such as multidimensional scaling (MDS) [44]. Despite the noisy and high-dimensional nature of microbial datasets [45–47], this approach has been remarkably reliable for the ordination [48] of microbial environments in as little as one to two dimensions, and for correlating environmental

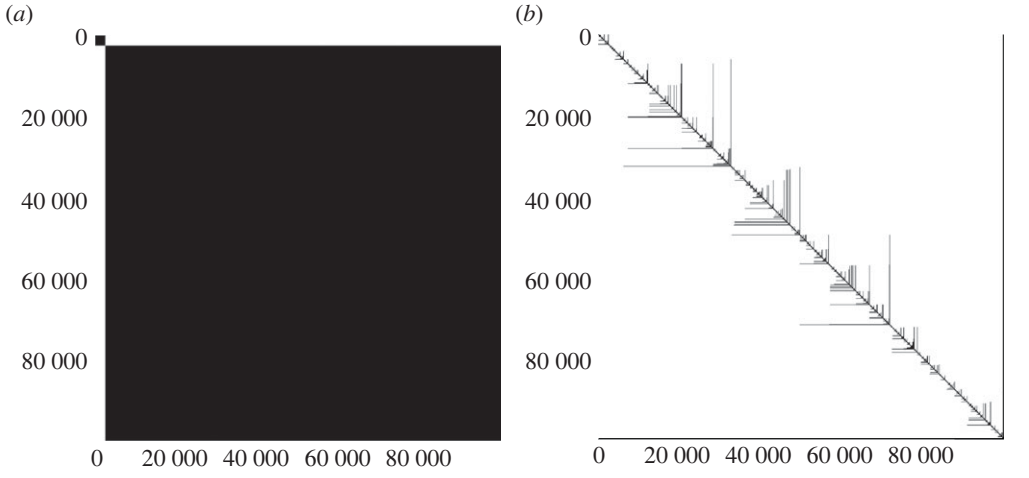


Figure 5. Heatmaps of matrices associated with the 97% Greengenes. Black (white) pixels denote non-zero (vanishing) entries. (*a*) Phylogenetic covariance matrix S of the 97% Greengenes tree. S has dimensions approximately $10^5 \times 10^5$. (*b*) Sparsified matrix $\Phi' S \Phi$.

factors with microorganisms. However, this approach does not usually explain correlations, which need to be justified by other means.

In what remains of this section, we apply our methods to the Greengenes phylogeny. First, we demonstrate significant sparsification of the associated covariance matrix after applying the Haar-like wavelet transform. Then, we motivate a new wavelet-based phylogenetic β-diversity metric corresponding to a multi-scale analysis of the phylogenetic tree. Finally, we show that this wavelet-based metric can give novel insights into the relationship between environmental factors and OTU composition.

(a) Greengenes phylogenetic covariance matrix sparsification

The 97% Greengenes tree has about 100 000 leaves. Let T be the ORB-tree associated with it.

The identity in equation (1.2) implies that the covariance matrix S of T is a 2×2 block diagonal matrix, with each block corresponding to an ORB-subtree. Approximately 94% of the almost 10 billion entries in S are non-zero because one of the ORB-subtrees (corresponding to the Archaea domain) is much smaller than the other—see figure 5*a*. This makes storing the covariance matrix of T challenging. Further, basic computational tasks such as finding the spectrum and inverting S for parameter estimation in phylogenetic comparative methods [49,50] is infeasible because this large matrix is almost fully dense. We may use, however, the Haar-like matrix Φ associated with T to sparsify S. From theorem 3.4, we can guarantee that $\zeta \geq 0.9989$, i.e. at least 99.89% of the entries in the similar matrix $\Phi' S \Phi$ vanish. This significant compression of the matrix S can be appreciated in figure 5*b*.

We implemented algorithm 1 using the sparse matrix packages from SciPy [51] to compute $\Phi' S \Phi$. As proof-of-principle, we used this compressed representation to compute the largest 500 eigenvalues of S to machine precision using SciPy's implementation of the Lanczos algorithm. As seen in figure 6, the eigenvalues of S decay rapidly. In fact, we found that $\lambda_1(S) \sim 1.27 \times 10^5$, $\lambda_2(S) \sim 4.75 \times 10^3$ and $\mathrm{trace}(S) \sim 1.65 \times 10^5$, so the top-two eigenvalues already account for approximately 80% of the trace of S.

As seen in figure 6 also, the sorted diagonal entries in $\Phi' S \Phi$ (i.e. the quantities λ_v, with $v \in I$, as defined in (4.1)) approximate with ample accuracy the spectrum of S. For instance, $\max_{v \in I} \lambda_v$ underestimates $\lambda_1(S)$ with only about a 0.06% relative error. Anticipating this overall accuracy from T alone remains an open problem as neither our mathematical results, particularly theorem

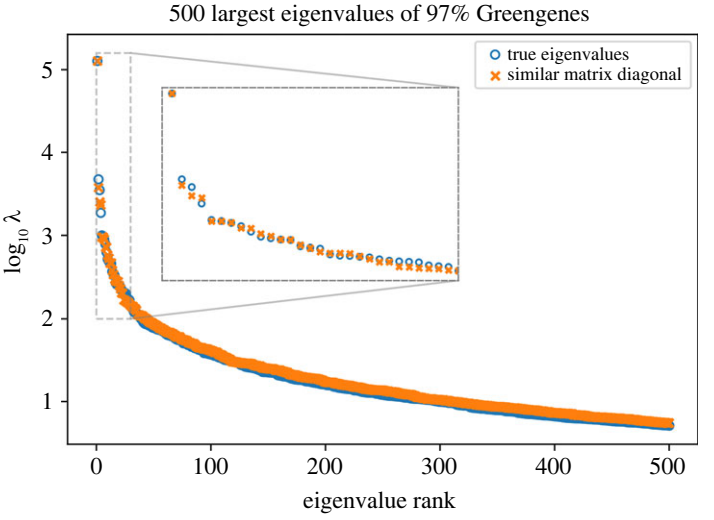


Figure 6. Spectrum decay of 97% Greengenes tree covariance matrix and corresponding approximation using Haar-like wavelets. Only the 500 most dominant eigenvalues of S are plotted as a function of their rank. Logarithms are in base-10.

4.1, nor more general ones such as Gershgorin's circle theorem, Sylvester's determinant theorem, and bounds found in [52–54], have been able to explain it.

(b) A wavelet-based phylogenetic β-diversity metric

Let T be the ORB-tree associated with a phylogenetic tree. In this section, we assume that the covariance matrix S of T is positive definite; in particular, $\lambda_v > 0$, for each $v \in I$.

Recall that φ_v, with $v \in I$, is supported on $L(v)$, and together these functions form an orthonormal base of $\mathbb{R}^{|L|}$. In particular, just as wavelets are traditionally used to localize signals at different scales, we may use the Haar-like basis of T to localize environmental OTU distributions on subsets of leaves defined by splits in the tree. This is particularly appealing from a biological standpoint. Indeed, the opposite signs of φ_v on the leaves of the left and right subtrees dangling from v may be interpreted as a speciation event that conferred more fitness to present-day microorganisms descending from one of the subtrees than the other. We propose the following definition to convey these features into a phylogenetic β-diversity metric.

Note that for a given environment a (i.e. OTU distribution over L), $\Phi' a$ is the projection of a onto the Haar-like basis of the reference tree.

Definition 5.1. The Haar-like distance between two environments a and b is the quantity

$$d_h(a,b) := \sqrt{\sum_{v \in I} \lambda_v \, \Delta_v^2}, \text{ where } \Delta = (\Delta_v)_{v \in I} := \Phi'(a-b).$$

The specifics of this distance can be motivated as follows. On one hand, the terms Δ_v^2, with $v \in I$, convey the idea that d_h regards two environments similar (different) when their OTU compositions project similarly (differently) onto the Haar-like basis of the reference tree. On the other hand, the weights λ_v, with $v \in I$, are motivated by the success of DPCoA in various biological investigations. To explain this, consider the matrices $D := \mathrm{diag}(\lambda_v : v \in I)$ and $E := \Phi' S \Phi - D$. Observe that $d_h(a,b) = \sqrt{\Delta' D \Delta}$; in particular, d_h is a metric in $\mathbb{R}^{|L|}$ because D is positive definite, and $d(a,b) = \sqrt{\Delta' D \Delta + \Delta' E \Delta}$. In large phylogenetic trees, however, we expect E to be mostly filled with zeroes due to corollary 3.8—which suggests considering d_h as an alternative metric to DPCoA.

We have mentioned before that while traditional phylogenetic metrics (in conjunction with embedding techniques) have been remarkably successful at correlating microbial composition with environmental factors, these correlations cannot usually be explained from the metrics alone. The wavelet nature of the Haar-like distance has, however, the potential to explain said correlations. Indeed, the biological interpretation of the Haar-like basis conveyed by their sign flip suggests that if $\lambda_v\,\Delta_v^2$ is comparatively large (small) for some $v \in I$, then the speciation event associated with v has a significant (little) influence differentiating the OTU distributions between two environments a and b. (There may be discrepancies between taxonomy and splits in a tree. In particular, while the aforementioned correlations may be explained by a phylogeny, they are not necessarily explained by a taxonomic classification.)

Previously, we mentioned the equivalence of the Haar-like basis and ILR basis. Accordingly, compositional metrics with similar interpretations [55,56] can be defined by projecting log-ratios of OTU counts onto the Haar-like basis (equivalently ILR basis). However, contrary to the Haar-like distance, these metrics do not account for phylogenetic-induced covariance between OTUs.

(c) Haar-like distances of the Guerrero Negro microbial mat

A microbial mat is a bio-film of layered groups of microorganisms with coupled biochemistries. Their rich biodiversity, combined with the environmental gradients of light, oxygen, etc. offer an ideal setting to test phylogenetic β-diversity metrics.

The Guerrero Negro mat is hypersaline. It is located in Baja California Sur, Mexico. To demonstrate the insights possibly gained from the Haar-like distance, we applied it to a 16S rRNA dataset of 18 samples at different depths of the Guerrero Negro mat [28,57]. We used the 97% Greengenes as the reference phylogeny.

Earlier work [57] based on unweighted UniFrac showed a gradient of microbial composition in the mat with respect to depth—see figure 7*a*. (For a discussion regarding the 'horseshoe' shape in the plot, see [57–59].) As seen on the bottom plot of the same figure, we can practically reproduce this gradient using the Haar-like distance instead. Furthermore, as seen on the bottom two plots, the DPCoA and Haar-like distance produce nearly indistinguishable embeddings.

While the three phylogenetic β-diversity metrics imply that mat depth drives a measurable change in OTU composition, we can go a step further with the Haar-like distance and determine which splits in the 97% Greengenes tree are responsible for this trend and quantify their importance. We demonstrate this by comparing the two extremes in the dataset: let a and b be the OTU compositions of the shallowest and deepest environment, respectively. Define $\Delta = \Phi'(b-a)$. Following the logic described in §b, we computed $v \in I \longrightarrow \lambda_v\,\Delta_v^2$, indexing interior nodes according to a postorder traversal of the 97% Greengenes tree. These values are shown in figure 8*a*. For our analysis, we focus on the largest three values, all of which are statistically significant (via the number of standard deviations they deviate from the mean). These are associated with the Haar-like wavelets φ_{99311} ($\lambda_v\Delta_v^2$-value $\sim 4.84\times10^{-2}$), φ_{6079} ($\lambda_v\Delta_v^2$-value $\sim 4.84\times10^{-3}$) and φ_{67317} ($\lambda_v\Delta_v^2$-value $\sim 4.65\times10^{-3}$). These correspond to splits at depths 10, 18 and 34 of the 97% Greengenes tree, respectively.

Notably, the split associated with φ_{99311} corresponds to the largest $\lambda_v\Delta_v^2$-value. According to the associated taxonomic classification, the (say) left descendants of this split correspond to the phylum level classification of Cyanobacteria. This is consistent with the conclusion in [57], which correlated Cyanobacteria abundance changes with mat depth and explained the correlation by their ability to photosynthesize.

The other two wavelets provide novel insight into other important OTU composition differences driving the observed mat depth gradient in the Guerrero Negro dataset. Indeed, the split associated with φ_{6079} subdivides the Cyanobacteria phylum into further classes, including Oscillatoriophycideae, which is the third most abundant class of the Guerrero Negro dataset. The relevance of this split to differentiate shallow from deep samples may be explained by Oscillatoriophycideae's photoautotrophic capability [61].

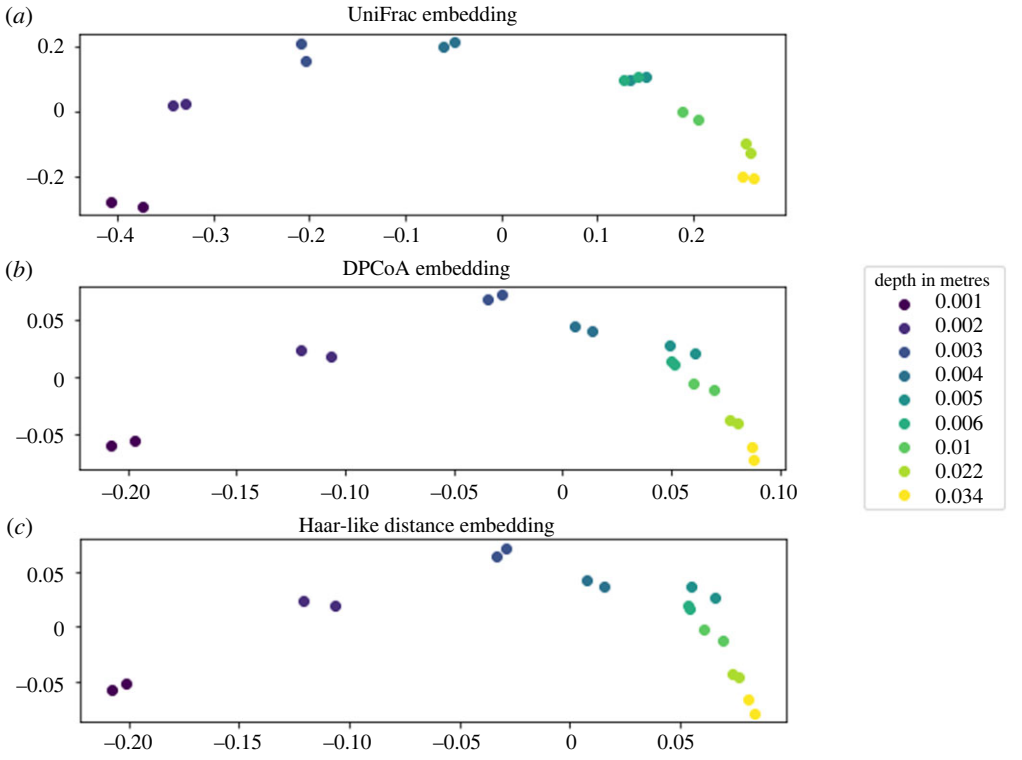


Figure 7. Two-dimensional MDS embeddings of samples from Guerrero Negro w.r.t. different metrics. The embeddings are based on unweighted UniFrac (*a*), DPCoA (*b*) and the Haar-like distance (*c*). Depth varies from 0 to 0.034 m.

Finally, while the descendants of the split associated with φ_{67317} do not exhaust a taxonomic classification, all leaves under that split are classified as Anaerolineae. This differentiation between the shallowest and deepest sample may be due to Anaerolineae's role as an anaerobic digester [62].

Our analysis of the Guerrero Negro mat shows that the Haar-like distance may be a valid alternative to other more common phylogenetic β-diversity metrics, primarily because it provides a systematic method for detecting statistically significant speciation events (and corresponding levels of OTU classification) that can link OTU composition with environmental factor gradients.

6. Conclusion

We have presented an approach to analyse and manipulate ultrametric matrices through their ORB-tree representation. We demonstrated that the Haar-like wavelets associated with an ORB-tree provide an orthonormal basis with respect to which their associated matrix can be sparsified. Subsequently, we detailed a sparsification algorithm and showed that, with overwhelmingly high probability, only an asymptotically negligible fraction of the off-diagonal entries in random but large ultrametric matrices would remain non-zero after its application. The resulting sparse representation may allow otherwise computationally infeasible and standard manipulations of these matrices, such as inverting and factoring them and characterizing their spectrum and eigenvectors. Noteworthily, our results generalize to the class of extended ultrametric covariance matrices and hence may find application in the parameterization of covariance matrices of Gaussian mixture models [16].

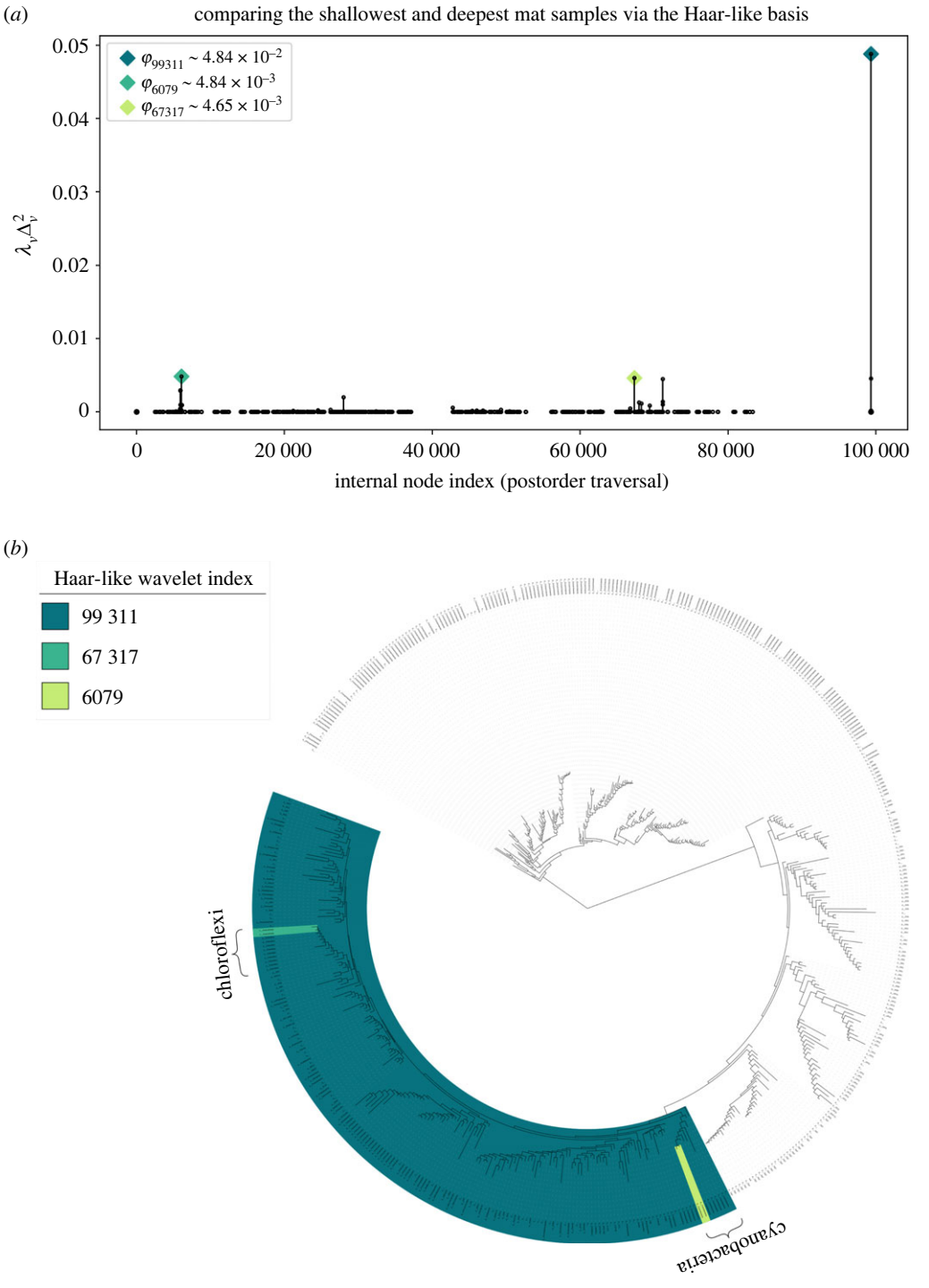


Figure 8. (*a*) Plot of $\boldsymbol{\lambda}_v\ \Delta_v^2$, with $v \in I$, to measure the Haar-like distance between the shallowest and deepest sample in the Guerrero Negro dataset. The three largest projections are shown in the legend and projections with a value of zero are left unmarked. The average non-zero value of $\lambda_v\ \Delta_v^2$ is approximately 2.09×10^{-5}. The standard deviation of these values is approximately 7.55×10^{-4}. (*b*) The support of the three most significant projections visualized on the 97% Greengenes phylogeny using iTOL [60]. Clades have been collapsed for visualization purposes and displayed leaves are labelled by the lowest taxonomic label that applies to all of its (collapsed) children.

Additionally, we provided some exact and approximate spectral results for ultrametric matrices based on the trace branch length 'balancedness' of their ORB-tree. We also characterized the possible spectrums of ultrametric matrices, giving further insight into the symmetric non-negative inverse eigenvalue problem.

Conversely, the ultrametric matrix associated with an ORB-tree corresponds to the phylogenetic covariance matrix of the tree. We applied our methods to the microbiologist's Tree of Life covariance matrix as proof of concept. This covariance model is a standard reference in metagenomic studies, which rely on metrics such as UniFrac and DPCoA. Motivated by the fact that the Tree of Life's covariance matrix is significantly sparsified by the Haar-like wavelets, and that the diagonal of the sparsified matrix approximates with striking accuracy its spectrum, we introduced the Haar-like distance. Like the established metrics, this new metric measures the distance between pairs of microbial environments taking into account the relative abundance of microbes and their evolutionary relatedness. Unlike the established metrics, however, this new distance may be used to identify statistically significant speciation events linking microbial composition with environmental factors.

Data accessibility. All the code and its dependencies, as well as all data files necessary to reproduce the results in this paper, are available at https://github.com/edgor17/Sparsify_Ultrametric.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. E.G.: data curation, investigation, software, validation, visualization, writing—original draft, writing—review and editing; M.E.L.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, writing—original draft, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work has been partially funded by the NSF grant no. 1836914.

Acknowledgements. We thank the reviewers for their time and valuable feedback, which improved our manuscript.

Appendix A. Sparsification and trace-balanced example

Consider a perfect binary tree with four leaves labelled $\{1,2,3,4\}$, and edges of length $\ell(\{1\})=a$, $\ell(\{2\})=b$, $\ell(\{1,2\})=c$, $\ell(\{3\})=d$, $\ell(\{4\})=e$, $\ell(\{3,4\})=f$ and $\ell(\{1,2,3,4\})=g$, all non-negative.

The phylogenetic covariance matrix of this tree is

$$S=\begin{pmatrix} a+c+g & c+g & g & g \\ c+g & b+c+g & g & g \\ g & g & d+f+g & f+g \\ g & g & f+g & e+f+g \end{pmatrix},$$

and the associated Haar-like matrix, when interior nodes are sorted in postorder traversal, is

$$\Phi=\begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{2} & \frac{1}{2} \\ 0 & -\frac{1}{\sqrt{2}} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

As a result

$$\Phi' S \Phi = \begin{pmatrix} \frac{a+b}{2} & 0 & \frac{a-b}{2\sqrt{2}} & \frac{a-b}{2\sqrt{2}} \\ 0 & \frac{d+e}{2} & \frac{e-d}{2\sqrt{2}} & \frac{d-e}{2\sqrt{2}} \\ \frac{a-b}{2\sqrt{2}} & \frac{e-d}{2\sqrt{2}} & \frac{(a+b+4c)+(d+e+4f)}{4} & \frac{(a+b+4c)-(d+e+4f)}{4} \\ \frac{a-b}{2\sqrt{2}} & \frac{d-e}{2\sqrt{2}} & \frac{(a+b+4c)-(d+e+4f)}{4} & \frac{(a+b+4c)+(d+e+4f)+16g}{4} \end{pmatrix}.$$

As expected from lemma 2.4, the entries associated with the first and second interior node in postorder traversal in $\Phi' S \Phi$ must be 0. Also, this matrix is diagonal if and only if $a = b$, $d = e$, and $a + b + 4c = d + e + 4f$. This is equivalent to having $a = b$, $d = e$ and $a + 2c = e + 2f$, i.e. that the tree is traced-balanced, in accordance with corollary 4.3.

Appendix B. Proof of technical results

(a) Orthonormality of Haar-like bases

The statement that the Haar-like basis $\{\varphi_v\}_{v \in I}$ associated with an ORB-tree is orthonormal is based on the concept of multi-resolution analysis of Euclidean spaces in [15]. Here, we justify this fact by first principles.

Let $u, v \in I$. If $u = v$ then

$$\langle \varphi_u, \varphi_v \rangle = \frac{|L(u1)| + |L(u0)|}{|L(u)|} = 1.$$

Instead, there are two possibilities when $u \neq v$. If $L(u) \cap L(v) = \emptyset$ then $\langle \varphi_v, \varphi_u \rangle = 0$ because φ_v and φ_u have disjoint supports. Otherwise, if $L(u) \cap L(v) \neq \emptyset$ then lemma 2.4 lets us assume without any loss of generality that u is an ancestor of v. In particular, $L(v) \subset L(u)$ but also φ_u remains constant over $L(v)$. Therefore, for any given $x \in L(v)$

$$\langle \varphi_u, \varphi_v \rangle = \varphi_u(x) \cdot \sum_{y \in L(v)} \varphi_v(y) = \varphi_u(x) \cdot \left\{ \sqrt{\frac{|L(v1)| \cdot |L(v0)|}{|L(v)|}} - \sqrt{\frac{|L(v0)| \cdot |L(v1)|}{|L(v)|}} \right\} = 0.$$

(b) Proof of theorem 2.3

Consider $v \in I$ and $j \in L$. If $j \notin L(v)$ then $(i \wedge j) = (v \wedge j)$, hence

$$(S\varphi_v)(j) = \sum_{i \in L(v)} \ell(i \wedge j, \circ)\, \varphi_v(i) = \ell(v \wedge j, \circ) \sum_{i \in L(v)} \varphi_v(i).$$

But, if $v = \circ$ then $\ell(v \wedge j, \circ) = 0$. Instead, if $v \neq \circ$ then $\sum_{i \in L(v)} \varphi_v(i) = 0$. In either case: $(S\varphi_v)(j) = 0$. This shows the lemma for $j \notin L(v)$ because the entry associated with j in $\mathrm{diag}(\ell^*(L, v))\, \varphi_v$ is $\ell^*(v, j) \cdot \varphi_v(j)$, and the support of φ_v is $L(v)$.

Next, suppose that $j \in L(v)$. Then

$$\begin{aligned} (S\varphi_v)(j) &= \sum_{i \in L(v)} \sum_{e \in [i \wedge j, o]} \ell(e)\, \varphi_v(i) \\ &= \sum_{i \in L(v)} \sum_{e \in [i \wedge j, v]} \ell(e)\, \varphi_v(i) + \sum_{i \in L(v)} \varphi_v(i) \cdot \sum_{e \in [v, o]} \ell(e) \\ &= \sum_{i \in L(v)} \varphi_v(i) \sum_{e \in [i \wedge j, v]} \ell(e), \end{aligned}$$

where for the last identity we have used that $\sum_{i \in L(v)} \varphi_v(i) = 0$ if $v \neq \circ$, and $\sum_{e \in [v, o]} \ell(e) = 0$ if $v = \circ$. But note that if $i \in L(v)$ is such that $(i \wedge j) = v$ then $\sum_{e \in [i \wedge j, v]} \ell(e) = 0$. Instead, if $(i \wedge j) \neq v$ then

$\varphi_v(i) = \varphi_v(j)$. As a result

$$\begin{aligned}(S\,\varphi_v)(j) &= \varphi_v(j) \sum_{i\in L(v):\, i\wedge j\neq v} \sum_{e\in[i\wedge j, v]} \ell(e)\\ &= \varphi_v(j) \sum_{e\in[j,v]} \sum_{i\in L(v):\, e\in[i\wedge j, v]} \ell(e)\\ &= \varphi_v(j) \sum_{e\in[j,v]} \ell(e)\,|L(e)|\\ &= \varphi_v(j)\,\ell^*(j,v),\end{aligned}$$

which shows the result.

(c) Proof of lemma 2.4

If u is an ancestor of v then $L(v) \subset L(u)$; in particular, $L(u) \cap L(v) = L(v) \neq \emptyset$. The same conclusion applies if v is an ancestor of u. Conversely, suppose that $L(u) \cap L(v) \neq \emptyset$. Without loss of generality assume that $u \neq v$. From the hypothesis, there is $w \in L$ that descends from both u and v. But, since there is a unique path from w to $\circ$, u and v must be both in this path; in particular, either u is an ancestor of v or vice versa.

(d) Proof of theorem 3.4

Recall that $|I| = |L|$ and, for $v \in I$, the support of φ_v is $L(v)$. Hence, from the identity in (2.2), $(\Phi' S \Phi)(u, v) = 0$ when $u, v \in I$ and $L(u) \cap L(v) = \emptyset$. As a result, using that $L(i) \neq \emptyset$ when $i \in I$, and lemma 2.4, we obtain that

$$\begin{aligned}|I|^2\zeta &\geq |I|^2 - |\{(u,v) \in I \times I \text{ such that } L(u) \cap L(v) \neq \emptyset\}|\\ &= |I|^2 - |I| - 2|\{(u,v) \in I \times I \text{ such that } v \neq u \text{ descends from } u \text{ and } L(u) \cap L(v) \neq \emptyset\}|\\ &= |I|^2 - |I| - 2\sum_{u\in I} \left(|\mathring{T}(u)| - 1\right)\\ &= |I|^2 + |I| - 2\sum_{u\in I} |\mathring{T}(u)|,\end{aligned}$$

since $|I| = |L| = |\mathring{T}| = |T|/2$, $|L|^2\zeta \geq |L|^2 + |L| - 2|\mathring{T}| \cdot \mathrm{avg}(\mathring{T})$, which shows the inequality in the theorem. The alternative lower-bound for ζ follows by applying the identity in equation (3.1) to $\mathring{T}$, completing the proof of the theorem.

(e) Perfect binary trees

If T is a perfect binary tree of height $(h + 1)$ (i.e. there are h levels of internal nodes, excluding the root), then $\mathring{T}$ is a perfect binary tree of height h; in particular, $|\mathring{T}| = 2^h$. Further, at depth $1 \leq k \leq h$, $\mathring{T}$ contains 2^{k-1} nodes, each of which is the root of a binary sub-tree of size $(2^{h-k+1} - 1)$. Hence

$$\mathrm{avg}(\mathring{T}) = \frac{2^h + \sum_{k=1}^{h} 2^{k-1} \cdot (2^{h-k+1} - 1)}{2^h} = h + 2^{-h} \ll |\mathring{T}|.$$

(f) Binary caterpillar trees

If T is a binary caterpillar tree of height h then $\mathring{T}$ is a path graph such that $|\mathring{T}| = h$. Further, at depth $0 \leq k \leq (h - 1)$, $\mathring{T}$ contains a single node, which is the root of a path sub-graph of size $(h - k)$. As a result

$$\mathrm{avg}(\mathring{T}) = \frac{\sum_{k=0}^{h-1}(h-k)}{h} = \frac{h+1}{2} \sim \frac{|\mathring{T}|}{2}.$$

(g) Proof of corollary 3.8

Let $t > 0$. Let μ and σ^2 denote the mean and variance of IPL($\mathbb{T}$), respectively. Due to Cantelli's inequality (a one-sided version of the well-known Chebyshev's inequality)

$$\mathbb{P}\big(\mathrm{IPL}(\mathbb{T}) \geq \mu + t\sigma\big) \leq \frac{1}{1+t^2}.$$

But, because of equations (3.2)–(3.3), there is a constant $c > 0$ such that for all n large enough and all $t > 1$, $(\mu + t\sigma) \leq ctn^{3/2}$. In particular, since $\mathrm{TPL}(\mathring{\mathbb{T}}) = \mathrm{IPL}(\mathbb{T})$, we have that

$$\mathbb{P}\left(\frac{\mathrm{TPL}(\mathring{\mathbb{T}})}{n^2} < \frac{ct}{\sqrt{n}}\right) \geq \frac{t^2}{1+t^2}.$$

So, if $t \to \infty$ so that $t = o(\sqrt{n})$ then $ct/\sqrt{n} = o(1)$, and the corollary follows from the second lower-bound for ζ in theorem 3.4.

(h) Proof of theorem 4.1

The proof is based on the following simple result, whose proof is omitted.

Lemma B.1. *If A is a symmetric matrix of dimensions $n \times n$ then, for all $\lambda \in \mathbb{R}$*

$$\mathit{distance}\big(\lambda, \sigma(A)\big) \leq \min_{x \in \mathbb{R}^n:\, ||x||_2 = 1} ||(A - \lambda)x||_2.$$

Fix $v \in I$. To make the λ_v more explicit, observe that if $x : L \to \mathbb{R}$ is a function (or vector) then

$$\sum_{i \in L} \varphi_v^2(i) \cdot x(i) = \rho_{v1} \cdot \mathrm{avg}\,\big(x; L(v0)\big) + \rho_{v0} \cdot \mathrm{avg}\,\big(x; L(v1)\big). \tag{B 1}$$

In particular, due to the definition in equation (4.1)

$$\lambda_v = \varphi_v' S \varphi_v = \sum_{i \in L(v)} \varphi_v^2(i) \cdot \ell^*(v,i) = \rho_{v1} \cdot \overline{\ell^*\big(L(v0), v\big)} + \rho_{v0} \cdot \overline{\ell^*\big(L(v1), v\big)}, \tag{B 2}$$

which shows the first identity in the theorem.

On the other hand, lemma B.1 implies that

$$\mathrm{distance}(\lambda_v, \sigma(S)) \leq ||(S - \lambda_v)\varphi_v||_2.$$

But, from theorem 2.3, we also have for $i \in L$ that $(S\varphi_v - \lambda_v\varphi_v)(i) = \varphi_v(i) \cdot (\ell^*(v,i) - \lambda_v)$. As a result

$$\begin{aligned} ||(S - \lambda_v)\varphi_v||_2^2 &= \sum_{i \in L(v)} \varphi_v^2(i) \cdot \big(\ell^*(v,i) - \lambda_v\big)^2 \\ &= \frac{\rho_{v1}}{|L(v0)|} \sum_{i \in L(v0)} (\ell^*(i,v) - \lambda_v)^2 + \frac{\rho_{v0}}{|L(v1)|} \sum_{i \in L(v1)} (\ell^*(i,v) - \lambda_v)^2, \end{aligned}$$

where for the last identity, we have used equation (B 1). But note that $(\rho_{v0} + \rho_{v1}) = 1$. In particular, from the identity in equation (B 2), we may rewrite

$$\begin{aligned} \sum_{i \in L(v0)} (\ell^*(i,v) - \lambda_v)^2 &= \sum_{i \in L(v0)} \Big(\rho_{v0}\Big\{\overline{\ell^*(L(v0), v)} - \overline{\ell^*(L(v1), v)}\Big\} + \ell^*(i,v) - \overline{\ell^*\big(L(v0), v\big)}\Big)^2 \\ &= |L(v0)|\, \rho_{v0}^2 \Big\{\overline{\ell^*(L(v0), v)} - \overline{\ell^*(L(v1), v)}\Big\}^2 + \sum_{i \in L(v0)} (\ell^*(i,v) - \overline{\ell^*\big(L(v0), v\big)})^2. \end{aligned}$$

Namely,

$$\frac{1}{|L(v0)|} \sum_{i \in L(v0)} (\ell^*(i,v) - \lambda_v)^2 = \rho_{v0}^2 \Big\{\overline{\ell^*(L(v0), v)} - \overline{\ell^*(L(v1), v)}\Big\}^2 + \mathrm{var}\,\big(\ell^*(L,v); L(v0)\big).$$

Similarly,

$$\frac{1}{|L(v1)|}\sum_{i\in L(v1)}(\ell^*(i,v)-\lambda_v)^2=\rho_{v1}^2\left\{\overline{\ell^*(L(v1),v)}-\overline{\ell^*(L(v0),v)}\right\}^2+\operatorname{var}\left(\ell^*(L,v);L(v1)\right).$$

The second identity in the theorem is now a direct consequence of (B 3) and the last two identities.

Finally, if $\lambda_v=0$ then, due to equation (4.1), $\ell^*(i,v)=0$ for all $i\in L(v)$. Hence, due to theorem 2.3 and since φ_v has $L(v)$ as its support, $(S\varphi_v)(i)=\ell^*(i,v)\varphi_v(i)=0$ for all $i\in L$, i.e. $S\varphi_v=0$. Instead, if $\lambda_v\neq 0$ then theorem 2.3 implies that

$$\cos(S\varphi_v,\lambda_v\varphi_v)=\frac{\varphi_v' S\varphi_v}{||S\varphi_v||_2}=\frac{\lambda_v}{\sqrt{\varphi_v'\operatorname{diag}(\ell^*(L,v))^2\varphi_v}}.$$

But, similarly as we argued before

$$\begin{aligned}\varphi_v'\operatorname{diag}(\ell^*(L,v))^2\varphi_v&=\sum_{i\in L}\varphi_v^2(i)\cdot\ell^*(i,v)^2\\&=\lambda_v^2+\sum_{i\in L}\varphi_v^2(i)\cdot(\ell^*(i,v)-\lambda_v)^2\\&=\lambda_v^2+||(S-\lambda_v)\varphi_v||_2^2,\end{aligned}$$

from which the third identity in the theorem follows.

(i) Proof of corollary 4.7

From the definitions of trace-balanced and ORB-tree, it is immediate that the transformation $f: I\to[0,+\infty)$ defined as $f(v):=\ell^*(v,i)$, for any $i\in L(v)$, is well-defined. Also, f is decreasing because if u is an ancestor of v then $L(v)\subset L(u)$, hence for any $i\in L(v)$: $f(u)=\ell^*(u,i)=\ell^*(u,v)+\ell^*(v,i)\geq\ell^*(v,i)=f(v)$. This shows the first statement in the corollary.

For the second statement consider an ORB-tree topology $T=(V,E)$ and function $f: I\to[0,\infty)$ that is decreasing. Due to corollary 4.3, it suffices to show that there is a branch length function $\ell: E\to[0,+\infty)$ such that $f(u)=\ell^*(u,i)$, for all $u\in I$ and any $i\in L(u)$. To do so, let $e=\{u,v\}\in E$ be so that $\operatorname{depth}(u)<\operatorname{depth}(v)$; in particular, $u\in I$. Define

$$\ell(e):=\begin{cases}f(u), & v\in L;\\ \dfrac{f(u)-f(v)}{|L(e)|}, & v\in I.\end{cases}\tag{B 3}$$

Observe that if $v\in L$ then $\ell(e)\geq 0$, and $f(u)=\ell(e)=\ell^*(e)$ because $|L(e)|=1$. Instead, if $v\in I$ then $\ell(e)\geq 0$ because f is decreasing, and $f(u)=f(v)+\ell(e)\cdot|L(e)|=f(v)+\ell^*(e)$. In particular, if we extend the domain of f to all of V defining $f(v):=0$ for $v\in L$ then, for all $e=\{u,v\}\in E$ such that $\operatorname{depth}(u)<\operatorname{depth}(v)$: $f(u)=f(v)+\ell^*(e)$. From this, a simple inductive argument on the difference $d:=\operatorname{depth}(v)-\operatorname{depth}(u)>0$ shows that $f(u)-f(v)=\ell^*(u,v)$; implying that $f(u)=\ell^*(u,i)$, for all $u\in I$ and any $i\in L(u)$, as claimed.

References

1. Dellacherie C, Martinez S, Martín S. 2014 *Inverse M-matrices and ultrametric matrices*, vol. 2118. Lecture Notes in Mathematics. New York, NY: Springer.
2. Dellacherie C, Martínez S, San Martín J. 1996 Ultrametric matrices and induced Markov chains. *Adv. Appl. Math.* **17**, 169–183. (doi:10.1006/aama.1996.0009)
3. Purdom E. 2011 Analysis of a data matrix and a graph: metagenomic data and the phylogenetic tree. *Ann. Appl. Stat.* **5**, 2326–2358. (doi:10.1214/10-aoas402)
4. Ji F, Tang W, Tay WP. 2019 On the properties of Gromov matrices and their applications in network inference. *IEEE Trans. Signal Process.* **67**, 2624–2638. (doi:10.1109/tsp.2019.2908133)

5. Capocaccia D, Cassandro M, Picco P. 1987 On the existence of thermodynamics for the generalized random energy model. *J. Stat. Phys.* **46**, 493–505. (doi:10.1007/bf01013370)
6. Zubarev AP. 2014 On stochastic generation of ultrametrics in high-dimensional Euclidean spaces. *p-Adic Numbers, Ultrametric Anal. Appl.* **6**, 55–165. (doi:10.1134/S2070046614020046)
7. Zubarev AP. 2017 On the ultrametric generated by random distribution of points in Euclidean spaces of large dimensions with correlated coordinates. *J. Classif.* **34**, 366–383. (doi:10.1007/s00357-017-9236-8)
8. Varga RS, Nabben R. 1993 In *On Symmetric Ultrametric Matrices*, pp. 193–200. Berlin, New York: De Gruyter. (doi:10.1515/9783110857658.193)
9. Martinez S, Michon G, San Martín J. 1994 Inverse of strictly ultrametric matrices are of Stieltjes type. *SIAM J. Matrix Anal. Appl.* **15**, 98–106. (doi:10.1137/S0895479891217011)
10. Nabben R, Varga RS. 1994 A linear algebra proof that the inverse of a strictly ultrametric matrix is a strictly diagonally dominant Stieltjes matrix. *SIAM J. Matrix Anal. Appl.* **15**, 107–113. (doi:10.1137/S0895479892228237)
11. Cavalli-Sforza LL, Edwards AW. 1967 Phylogenetic analysis: models and estimation procedures. *Evolution* **21**, 550. (doi:10.2307/2406616)
12. Saraçli S, Doğan N, Doğan I. 2013 Comparison of hierarchical cluster analysis methods by cophenetic correlation. *J. Inequalities Appl.* **2013**, 1–8. (doi:10.1186/1029-242x-2013-203)
13. Mallat SG. 2009 *A wavelet tour of signal processing: the sparse way*. Orlando, FL: Elsevier/Academic Press.
14. Graps A. 1995 An introduction to wavelets. *IEEE Comput. Sci. Eng.* **2**, 50–61. (doi:10.1109/99.388960)
15. Gavish M, Nadler B, Coifman RR. 2010 Multiscale wavelets on trees, graphs and high dimensional data: theory and applications to semi supervised learning. In *Proc. of the 27th Int. Conf. on Machine Learning, Haifa, Israel, 21–25 June*, pp. 367–374.
16. Cavicchia C, Vichi M, Zaccaria G. 2022 Gaussian mixture model with an extended ultrametric covariance structure. *Adv. Data Anal. Classif.* **16**, 399–427. (doi:10.1007/s11634-021-00488-x)
17. Soules GW. 1983 Constructing symmetric nonnegative matrices. *Linear Multilinear Algebra* **13**, 241–251. (doi:10.1080/03081088308817523)
18. Elsner L, Nabben R, Neumann M. 1998 Orthogonal bases that lead to symmetric nonnegative matrices. *Linear Algebra Appl.* **271**, 323–343. (doi:10.1016/s0024-3795(97)00302-9)
19. Devriendt K, Lambiotte R, Van Mieghem P. 2019 Constructing Laplacian matrices with Soules vectors: inverse eigenvalue problem and applications. (http://arxiv.org/abs/10.48550/ARXIV.1909.11282)
20. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. 2003 Isometric logratio transformations for compositional data analysis. *Math. Geol.* **35**, 279–300. (doi:10.1023/A:1023818214614)
21. Sedgewick R, Flajolet P. 2013 *An introduction to the analysis of algorithms*. Boston, MA: Addison-Wesley.
22. Blum MG, François O. 2005 On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited. *Math. Biosci.* **195**, 141–153. (doi:10.1016/j.mbs.2005.03.003)
23. Coronado TM, Mir A, Rosselló F, Rotger L. 2020 On Sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index. *BMC Bioinf.* **21**, 1–17. (doi:10.1186/s12859-020-3405-1)
24. King MC, Rosenberg NA. 2021 A simple derivation of the mean of the Sackin index of tree balance under the uniform model on rooted binary labeled trees. *Math. Biosci.* **342**, 108688. (doi:10.1016/j.mbs.2021.108688)
25. Flajolet P, Sedgewick R. 2013 *Analytic combinatorics*. Cambridge, UK: Cambridge University Press.
26. Qin C, Colwell LJ. 2018 Power law tails in phylogenetic systems. *Proc. Natl Acad. Sci. USA* **115**, 690–695. (doi:10.1073/pnas.1711913115)
27. Pavoine S, Dufour AB, Chessel D. 2004 From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *J. Theor. Biol.* **228**, 523–537. (doi:10.1016/j.jtbi.2004.02.014)
28. Gonzalez A *et al.* 2018 Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798. (doi:10.1038/s41592-018-0141-9)
29. Price MN, Dehal PS, Arkin AP. 2010 FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**(3), 1–10. (doi:10.1371/journal.pone.0009490)

30. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2011 An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618. (doi:10.1038/ismej.2011.139)
31. Felsenstein J, Archie J, Day W, Maddison W, Meacham C, Rohlf F, Swofford D. 1986 The Newick tree format.
32. Rambaut A 2010 FigTree v1.3.1. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh.
33. Huerta-Cepas J, Serra F, Bork P. 2016 ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638. (doi:10.1093/molbev/msw046)
34. Ho SY, Duchêne S. 2014 Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.* **23**, 5947–5965. (doi:10.1111/mec.12953)
35. Bray J, Curtis J. 1957 An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27**, 325–349. (doi:10.2307/1942268)
36. Jaccard P. 1912 The distribution of the flora in the alpine zone. *N. Phytol.* **11**, 37–50. (doi:10.1111/j.1469-8137.1912.tb05611.x)
37. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012 The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**. (doi:10.1093/nar/gks1219)
38. Zhu Q *et al.* 2019 Phylogenomics of 10 575 genomes reveals evolutionary proximity between domains bacteria and Archaea. *Nat. Commun.* **10**, 5477 (doi:10.1038/s41467-019-13443-4)
39. Fukuyama J, McMurdie PJ, Dethlefsen L, Relman DA, Holmes S. 2012 Comparisons of distance methods for combining covariates and abundances in microbiome studies. *Biocomputing* **2012**, 213–224.
40. Lozupone C, Knight R. 2005 UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235. (doi:10.1128/AEM.71.12.8228-8235.2005)
41. Mahalanobis P. 1936 On the generalised distance in statistics. *Proc. Natl Inst. Sci. India* **2**, 49–55.
42. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. 2011 UniFrac: an effective distance metric for microbial community comparison. *ISME J.* **5**, 169–172. (doi:10.1038/ismej.2010.133)
43. Lladser ME, Knight R. 2013 Mathematical approaches for describing microbial populations: practice and theory for extrapolation of rich environments. In *The human microbiota: how microbial communities affect health and disease* (ed. D Fredricks), pp. 85–104. Hoboken, NJ: John Wiley and Sons, Inc.
44. Borg I, Groenen PJ. 2005 *Modern multidimensional scaling: theory and applications*, 2nd edn. New York, NY: Springer.
45. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006 Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl Acad. Sci. USA* **103**, 12 115–12 120. (doi:10.1073/pnas.0605127103)
46. Lladser ME, Gouet R RJ. 2011 Extrapolation of urn models via Poissonization: accurate measurements of the microbial unknown. *PLoS ONE* **6**, e21105. (doi:10.1371/journal.pone.0021105)
47. Hampton J, Lladser ME. 2012 Estimation of distribution overlap of urn models. *PLoS ONE* **7**, e42368. (doi:10.1371/journal.pone.0042368)
48. Jongman RHG. 1995 *Data analysis in community and landscape ecology*. Cambridge, UK: Cambridge University Press.
49. Jhwueng DC, 2022 On the covariance of phylogenetic quantitative trait evolution models and their matrix condition. *Commun. Stat. Simul. Comput.* 1–20. (doi:10.1080/03610918.2022.2037639)
50. Freckleton RP. 2012 Fast likelihood calculations for comparative analyses. *Methods Ecol. Evol.* **3**, 940–947. (doi:10.1111/j.2041-210x.2012.00220.x)
51. Virtanen P *et al.* 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272. (doi:10.1038/s41592-019-0686-2)
52. Thompson R. 1976 The behavior of eigenvalues and singular values under perturbations of restricted rank. *Linear Algebra Appl.* **13**, 69–78. (doi:10.1016/0024-3795(76)90044-6)
53. Bunch JR, Nielsen CP, Sorensen DC. 1978 Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik* **31**, 31–48. (doi:10.1007/bf01396012)
54. Ipsen ICF, Nadler B. 2009 Refined perturbation bounds for eigenvalues of hermitian and non-hermitian matrices. *SIAM J. Matrix Anal. Appl.* **31**, 40–53. (doi:10.1137/070682745)

55. Morton JT *et al.* 2017 Balance trees reveal microbial niche differentiation. *mSystems* **2**, e00162-16 (doi:10.1128/msystems.00162-16)
56. Silverman JD, Washburne AD, Mukherjee S, David LA. 2017 A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* **6**, 1–20. (doi:10.7554/elife.21887)
57. Kirk Harris J *et al.* 2012 Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat. *ISME J.* **7**, 50–60. (doi:10.1038/ismej.2012.79)
58. Morton JT, Toran L, Edlund A, Metcalf JL, Lauber C, Knight R. 2017 Uncovering the horseshoe effect in microbial analyses. *mSystems* **2**, e00166-16 (doi:10.1128/msystems.00166-16)
59. Diaconis P, Goel S, Holmes S. 2008 Horseshoes in multidimensional scaling and local kernel methods. *Ann. Appl. Stat.* **2**, 777–807. (doi:10.1214/08-aoas165)
60. Letunic I, Bork P. 2019 Interactive Tree of Life (iTOL) V4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259. (doi:10.1093/nar/gkz239)
61. Stanier RY, Cohenbazire G. 1977 Phototropic prokaryotes - cyanobacteria. *Annu. Rev. Microbiol.* **31**, 225–274. (doi:10.1146/annurev.mi.31.100177.001301)
62. Xia Y, Wang Y, Wang Y, Chin FY, Zhang T. 2016 Cellular adhesiveness and cellulolytic capacity in anaerolineae revealed by OMICS-based genome interpretation. *Biotechnol. Biofuels* **9**, 1–13. (doi:10.1186/s13068-016-0524-z)