Advances in Applied Mathematics 50 (2013) 327–338

Contents lists available at SciVerse ScienceDirect

Advances in Applied Mathematics

www.elsevier.com/locate/yaama

Polyhedral combinatorics of UPGMA cones

Ruth Davidson, Seth Sullivant *

Department of Mathematics, Box 8205, North Carolina State University, Raleigh, NC 27695-8205, USA

A R T I C L E   I N F O

Article history:
Received 27 July 2012
Accepted 3 September 2012
Available online 11 October 2012

MSC:
92D15
52B05
05C05
06A07

Keywords:
Phylogenetic trees
Polyhedral combinatorics
Partition lattice

A B S T R A C T

Distance-based methods such as UPGMA (Unweighted Pair Group Method with Arithmetic Mean) continue to play a significant role in phylogenetic research. We use polyhedral combinatorics to analyze the natural subdivision of the positive orthant induced by classifying the input vectors according to tree topologies returned by the algorithm. The partition lattice informs the study of UPGMA trees. We give a closed form for the extreme rays of UPGMA cones on n taxa, and compute the spherical volumes of the UPGMA cones for small n.

1. Introduction

The UPGMA algorithm (Unweighted Pair Group Method with Arithmetic Mean) [9] is an agglomerative tree reconstruction method, that takes as input $\binom{n}{2}$ pairwise distances (dissimilarities) between n taxa and returns a rooted, equidistant tree with these n taxa as the leaves. UPGMA is a greedy heuristic that attempts to compute the Euclidean projection onto the space of all equidistant tree metrics [4]. The UPGMA algorithm subdivides the positive orthant $\mathbb{R}_{\geqslant 0}^{n(n-1)/2}$ into regions based on which combinatorial type of tree is returned by the algorithm. The goal of this paper is to study the geometry of these regions in order to understand both how the regions relate to one another as well as the performance of the algorithm.

UPGMA has poor performance if the data is tree-like but does not follow a molecular clock. In spite of this limitation, we find UPGMA an interesting algorithm to study because it is one of the few phylogenetic reconstruction methods that directly returns a rooted tree on a collection of species. One

* Corresponding author.
E-mail addresses: redavids@ncsu.edu (R. Davidson), smsulli2@ncsu.edu (S. Sullivant).

0196-8858/$ – see front matter
http://dx.doi.org/10.1016/j.aam.2012.10.002

motivation for studying the UPGMA algorithm was the work of Aldous [1], where it was observed that rooted trees that have been constructed from data do not typically have the same underlying statistics as familiar speciation models such as the Yule process. This raises the question of whether or not the Yule process is flawed, or the trees that have been constructed are biased because of taxa selection, or there is inherent bias in the reconstruction methods. We believe that analyzing the partition of data space induced by a tree reconstruction method can give some insight into the latter problem: if regions corresponding to some tree shapes are inherently larger than others, this indicates that the algorithm might favor those shapes in the presence of noise or model misspecification of the equidistant assumption.

With these motivating problems in mind, we study the decomposition of space induced by the UPGMA algorithm. For a given binary phylogenetic X-tree T (that is, with leaf labels X but without edge lengths), the region of $\mathcal{P}(T) \subseteq \mathbb{R}_{\geqslant 0}^{n(n-1)/2}$ of dissimilarity maps for which the algorithm returns the phylogenetic X-tree T is a union of finitely many polyhedral cones, one for each ranking function of the interior nodes of T. We give explicit polyhedral descriptions of the cones including facet defining inequalities and extreme rays, for all T and all n. In particular, each cone has $O(n^3)$ facet defining inequalities but exponentially many extreme rays. We compute the spherical volumes of the regions $\mathcal{P}(T)$ for $n \leqslant 7$. These volumes give a measure of the proportion of dissimilarity maps for which UPGMA returns a given combinatorial type of tree. In particular, our computations seem to indicate that highly unbalanced trees have small volume UPGMA cones compared to more balanced trees. Our computation of spherical volumes builds on the Monte Carlo strategy in [3].

2. Ranked phylogenetic trees and the UPGMA algorithm

The UPGMA method is an agglomerative tree reconstruction method that takes as an input $\binom{n}{2}$ pairwise distances between a set of taxa X and returns a rooted equidistant tree metric on X. In this section, we review necessary background on ranked phylogenetic trees and the lattice of set partitions as they pertain to describing the UPGMA algorithm. We refer the reader to [5] and [8] for background on phylogenetics.

Definition 2.1. Let X be a finite set. A *phylogenetic X-tree* is a tree T with leaves bijectively labeled by the set X. A phylogenetic X-tree is rooted if it has a distinguished root node ρ. It is *binary* if every interior vertex that is not a leaf has degree 3 except for the root ρ, which has degree 2.

Throughout this paper, unless stated otherwise, we assume that a *tree T on n taxa* is a rooted binary phylogenetic X-tree where $X = [n]$. In a rooted binary phylogenetic X-tree, ρ is not labeled by an element of X.

A vertex $v \in V(T)$ is a *descendant* of $u \in V(T)$ if the path from ρ to v includes u. This relation induces a partial order on the vertices of T and we can write $u \leqslant_T v$. Let V° denote the set of interior (i.e. nonleaf) vertices of T. A *rank function* on T is a bijection $r : V^\circ \to \{1, 2, \dots, |V^\circ|\}$ satisfying $u \leqslant_T v \to r(u) \leqslant r(v)$. The number of rank functions on T is: $|V^\circ|!/\prod_{v \in V^\circ |\mathrm{de}(v)}|$ where $\mathrm{de}(v)$ denotes the set of descendants of v in the set V° [10]. Note that $v \leqslant_T v$, so that the number of descendants of v will include v itself. A tree with a rank function is called a *ranked phylogenetic tree*.

The lattice of set partitions provides a useful alternate description of ranked phylogenetic trees. See [10] for background and terminology for the theory of partially ordered sets. Let Π_n consist of all partitions of a set with n elements. For simplicity, we identify this underlying set as $[n] = \{1, 2, \dots, n\}$. Partitions are unordered, and consist of unordered elements. The shorthand $A_1|\dots|A_k$ denotes a partition with k parts. For example $12|345$ is shorthand for the partition $\{\{1, 2\}, \{3, 4, 5\}\}$.

Partitions in Π_n are ordered by refinement, so $A_1|\dots|A_k \leqslant B_1|\dots|B_\ell$ if and only if for each $i \in [k]$ there exists a $j \in [\ell]$ satisfying $A_i \subseteq B_j$. Every maximal chain in the lattice of set partitions corresponds to a ranked phylogenetic tree. Indeed, consider a maximal chain

$$C = 1|2|\cdots|n = \pi_n \lessdot \pi_{n-1} \lessdot \cdots \lessdot \pi_2 \lessdot \pi_1 = 12\cdots n$$

in Π_n. We use $\lessdot$ to denote a covering relation in the partial order Π_n, and we use the convention that π_i is always a partition with i parts.

Given $\pi_i \in C$, we write $\pi_i = \lambda_1^i|\lambda_2^i|\cdots|\lambda_i^i$. When $\pi_i \lessdot \pi_{i-1}$, there are exactly two blocks λ_j^i, λ_k^i that are joined in π_{i-1} but distinct in π_i. If $v \in V^\circ$ where $r(v) = n - i$, then π_{i-1} joins the two blocks in π_i that correspond to the subtrees of T induced by the child nodes of v.

The UPGMA algorithm constructs a rooted ranked phylogenetic X-tree from a dissimilarity map d, as well as an equidistant tree metric δ which approximates d. The algorithm works as follows:

Algorithm 2.2 *(UPGMA Algorithm).*

- Input: a dissimilarity map $d \in \mathbb{R}_{\geqslant 0}^{n(n-1)/2}$ on X.
- Output: a maximal chain C in the partition lattice Π_n and an equidistant tree metric δ.
- Initialize $\pi_n = 1|2|\cdots|n$, and set $d^n = d$.
- For $i = n-1, \ldots, 1$ do
  - From partition $\pi_{i+1} = \lambda_1^{i+1}|\cdots|\lambda_{i+1}^{i+1}$ and distance vector $d^{i+1} \in \mathbb{R}_{\geqslant 0}^{(i+1)i/2}$ choose j, k be so that $d^{i+1}(\lambda_j^{i+1}, \lambda_k^{i+1})$ is minimized.
  - Set π_i to be the partition obtained from π_{i+1} by merging λ_j^{i+1} and λ_k^{i+1} and leaving all other parts the same. Let $\lambda_i^i = \lambda_j^{i+1} \cup \lambda_k^{i+1}$.
  - Create new distance $d^i \in \mathbb{R}_{\geqslant 0}^{i(i-1)/2}$ by $d^i(\lambda, \lambda') = d^{i+1}(\lambda, \lambda')$ if λ, λ' are both parts of π_{i+1} and
$$d^i(\lambda, \lambda_i^i) = \frac{|\lambda_j^{i+1}|}{|\lambda_i^i|} d^{i+1}(\lambda, \lambda_j^{i+1}) + \frac{|\lambda_k^{i+1}|}{|\lambda_i^i|} d^{i+1}(\lambda, \lambda_k^{i+1}).$$
  - For each $x \in \lambda_j^{i+1}$ and $y \in \lambda_k^{i+1}$, set $\delta(x, y) = d^{i+1}(\lambda_j^{i+1}, \lambda_k^{i+1})$.
- Return: Chain $C = \pi_n \lessdot \cdots \lessdot \pi_1$ and equidistant metric δ.

Note that that step which recalculates distances, the weighted average

$$d^i(\lambda, \lambda_i^i) = \frac{|\lambda_j^{i+1}|}{|\lambda_i^i|} d^{i+1}(\lambda, \lambda_j^{i+1}) + \frac{|\lambda_k^{i+1}|}{|\lambda_i^i|} d^{i+1}(\lambda, \lambda_k^{i+1})$$

is used to determine the new distance. This is simply a computationally efficient strategy to compute the average of distances

$$d^i(\lambda, \lambda') = \frac{1}{|\lambda| \cdot |\lambda'|} \sum_{x \in \lambda, y \in \lambda'} d(x, y) \tag{1}$$

a formula we will make use of later.

Example 2.3. Let $\mathbf{d} = (1, 2, 1.8, 1.7, 2, 2.6, 3.1, 2.4, 2.6, 1.2) \in \mathbb{R}_{\geqslant 0}^{5(5-1)/2}$, be a dissimilarity map on 5 taxa.

The UPGMA algorithm performs the following steps, where an underline is used to denote the smallest value in the present metric

$$\begin{array}{cccccccccc} 12 & 13 & 14 & 15 & 23 & 24 & 25 & 34 & 35 & 45 \\ (\underline{1}, & 2, & 1.8, & 1.7, & 2, & 2.6, & 3.1, & 2.4, & 2.6, & 1.2) \end{array}$$

$$\begin{array}{cccccc} 12,3 & 12,4 & 12,5 & 34 & 35 & 45 \\ (2, & 2.2, & 2.4, & 2.4, & 2.6, & \underline{1.2}) \end{array}$$

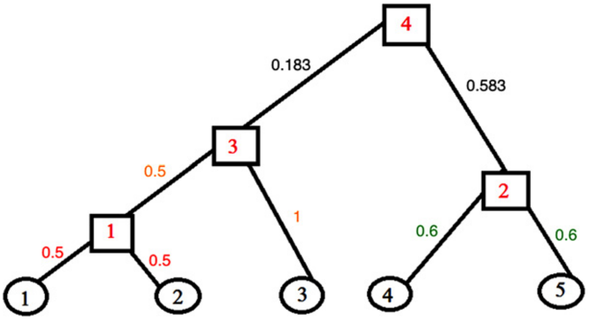

Fig. 1. The tree metric δ.

$$\begin{array}{ccc} 12,3 & 12,45 & 3,45 \\ \underline{2} & 2.3 & 2.5 \end{array}$$

$$\begin{array}{c} 123,45 \\ \underline{2.367} \end{array}$$

where

$$2.367 = \left(\frac{|12|}{|12| + |3|}\right)(2.3) + \left(\frac{|3|}{|12| + |3|}\right)(2.5).$$

The resulting rooted metric tree produced by the UPGMA algorithm is displayed in Fig. 1. The corresponding chain in the lattice of partitions Π_5 is

$$C = 1|2|3|4|5 \lessdot 3|4|5|12 \lessdot 3|12|45 \lessdot 45|123 \lessdot 12345.$$

3. UPGMA regions and UPGMA cones

The UPGMA algorithm takes as input a dissimilarity map $d \in \mathbb{R}^{n(n-1)/2}_{\geqslant 0}$ and returns a rooted equidistant tree metric. If we ignore the resulting metric tree that is output, and only record the rooted tree computed at each step of the algorithm, the UPGMA algorithm produces a rooted tree and a ranking function of the internal nodes corresponding to precisely one maximal chain in the partition lattice. Our goal is to understand the set of dissimilarity maps d, such that the UPGMA returns a rooted tree T, or equivalently a given chain C in the partition lattice Π_n. For a given leaf-labeled rooted tree T let $\mathcal{P}(T) \subseteq \mathbb{R}^{n(n-1)/2}_{\geqslant 0}$ denote the closure of the set of dissimilarity maps such that the UPGMA algorithm returns T. The set $\mathcal{P}(T)$ is called the *UPGMA region* associated to the tree T. Similarly, for a maximal chain C in Π_n, let $\mathcal{P}(C) \subseteq \mathbb{R}^{n(n-1)/2}_{\geqslant 0}$ denote the closure of the set of dissimilarity maps such that the UPGMA algorithm returns the chain C.

Our goal in this section is to describe the sets $\mathcal{P}(T)$ and $\mathcal{P}(C)$. Clearly $\mathcal{P}(T) = \bigcup \mathcal{P}(C)$ where the union is over all maximal chains in Π_n whose associated tree is T.

Theorem 3.1. *For each chain $C \in \Pi_n$ the set $\mathcal{P}(C)$ is a pointed polyhedral cone. The cone has $O(n^3)$ facet defining inequalities, and exponentially many extreme rays. Each covering relation in the chain C determines a collection of facet defining inequalities for $\mathcal{P}(C)$. Each element of the chain C determines a collection of extreme rays of $\mathcal{P}(C)$.*

We refer the reader to [11] for background material on polyhedral geometry. To prove Theorem 3.1, we will provide a more general result for the description of cones associated to partial chains. A partial chain C is a sequence

$$\pi_s \lessdot \pi_{s-1} \lessdot \cdots \lessdot \pi_t$$

for some $n \geqslant s \geqslant t \geqslant 1$. The fact that these are covering relations guarantees that at each step, $\pi_{i+1} \lessdot \pi_i$ we are simply joining a pair of parts together. This means that any partial chain C can be intermediate information that is calculated between steps s and t of the UPGMA algorithm.

For a partial chain C, let $\mathcal{P}(C)$ denote the set of all dissimilarity maps $d \in \mathbb{R}_{\geqslant 0}^{s(s-1)/2}$ which the UPGMA algorithm could produce on steps s through t of the algorithm. The coordinates in the space $\mathbb{R}^{s(s-1)/2}$ are the $s(s-1)/2$ distances $d(\lambda_j^s, \lambda_k^s)$.

Proposition 3.2. *Let C be a partial chain in Π_n. Let $\mathcal{P}(C) \subseteq \mathbb{R}^{s(s-1)/2}$ be the set of dissimilarity maps for which steps s through t of the UPGMA algorithm return the partial chain C. For each covering relation $\pi_i \lessdot \pi_{i-1}$ let $\lambda_{j(i)}^i$ and $\lambda_{k(i)}^i$ be the pair of parts of π_i that are joined in π_{i-1}. Then $\mathcal{P}(C)$ is the solution to the following system of linear inequalities:*

$$d\left(\lambda_j^s, \lambda_k^s\right) \geqslant 0 \text{ for all } j, k$$

$$\text{for } i = s, \ldots, t-1, \text{ and for all pairs } j, k \neq j(i), k(i)$$

$$\frac{1}{|\lambda_{j(i)}^i||\lambda_{k(i)}^i|} \sum_{\lambda_j^s \subseteq \lambda_{j(i)}^i, \lambda_k^s \subseteq \lambda_{k(i)}^i} \left|\lambda_j^s\right|\left|\lambda_k^s\right| d\left(\lambda_j^s, \lambda_k^s\right) \leqslant \frac{1}{|\lambda_j^i||\lambda_k^i|} \sum_{\lambda_j^s \subseteq \lambda_j^i, \lambda_k^s \subseteq \lambda_k^i} \left|\lambda_j^s\right|\left|\lambda_k^s\right| d\left(\lambda_j^s, \lambda_k^s\right).$$

Note that if $s > t$ we only need the nonnegativity constraint $d(\lambda_{j(s)}^s, \lambda_{k(s)}^s) \geqslant 0$, as the other inequalities $d(\lambda_j^s, \lambda_k^s) \geqslant 0$ follow from $d(\lambda_{j(s)}^s, \lambda_{k(s)}^s) \leqslant d(\lambda_j^s, \lambda_k^s)$.

Proof. At step i of the UPGMA algorithm, we choose the pair of $\lambda_{j(i)}^i$ and $\lambda_{k(i)}^i$ to merge such that $d^i(\lambda_{j(i)}^i, \lambda_{k(i)}^i)$ is minimized. Using the formula

$$d^i\left(\lambda_j^i, \lambda_k^i\right) = \frac{1}{|\lambda_j^i||\lambda_k^i|} \sum_{x \in \lambda_j^i, y \in \lambda_k^i} d(x, y)$$

twice shows that

$$d^i\left(\lambda_j^i, \lambda_k^i\right) = \frac{1}{|\lambda_j^i||\lambda_k^i|} \sum_{\lambda_j^s \subseteq \lambda_j^i, \lambda_k^s \subseteq \lambda_k^i} \left|\lambda_j^s\right|\left|\lambda_k^s\right| d\left(\lambda_j^s, \lambda_k^s\right).$$

This yields precisely the inequalities in the statement of the proposition at step i. $\square$

Proposition 3.3. *Given a maximal chain $C \in \Pi_n$, there are $O(n^3)$ facet defining inequalities for $\mathcal{P}(C)$.*

Proof. At step t, there are $\binom{t}{2}$ ways to merge two blocks of π_t, and the pair of parts $d(\lambda_{j(t)}^t, \lambda_{k(t)}^t)$ merged at step t can be paired with $\binom{t}{2} - 1$ other pairs of parts. So $\binom{t}{2} - 1$ new inequalities are introduced at step t. An elementary identity for binomial coefficients tells us that for $a, b \geqslant 0$, $\sum_{r=b}^{a} \binom{r}{b} = \binom{a+1}{b+1}$. Thus there are

$$\sum_{t=2}^{n}\left(\binom{t}{2}-1\right)=\binom{n+1}{3}-n+1$$

facet defining inequalities. □

Now we provide a description of the extremal rays of the cones of partial chains $\mathcal{P}(C)$, for partial chains starting with the bottom element $\pi_n = 1|2|\cdots|n$. The polyhedral description of the cones $\mathcal{P}(C)$ for more general partial chains is used in the proof of the main cases of interest.

Definition 3.4. Given a partition $\pi_k = \lambda_1|\lambda_2|\cdots|\lambda_k \in \Pi_n$ a *traversal* of π_k is a subset $F \subset \binom{[n]}{2}$ of size $\binom{k}{2}$, where each element of F is a pair $\{p, p'\} \in \pi$ satisfying $p \in \lambda$, $p' \in \lambda'$. There is precisely one such pair p, p' for every pair of parts λ, λ' of π_k.

For example, the partition 12|3|45 has $2^2 \cdot (2 \cdot 1) \cdot (2 \cdot 1) = 16$ traversals.

Definition 3.5. Let $\pi_k = \lambda_1|\lambda_2|\cdots|\lambda_k \in \Pi_n$. Let F be a traversal of π_k. The *induced vector* of F, denoted $v(F)$, is the vector in $\mathbb{R}^{\binom{n}{2}}$ such that

(1) $v(F)_{ij} = 0$ if the pair i, j is not in the traversal F.
(2) $i, j \in F$, $v(F)_{ij} = |\lambda_{k(i)}||\lambda_{k(j)}|$ where $i \in \lambda_{k(i)}$ and $j \in \lambda_{k(j)}$.

Consider the traversal $\{\{1,3\},\{1,4\},\{3,5\}\}$ of the partition 12|3|45. This traversal induces the vector $(0,2,4,0,0,0,0,0,2,0)$.

Theorem 3.6. *Let $C = \pi_n \lessdot \pi_{n-1} \lessdot \cdots \lessdot \pi_t$ be a grounded partial chain in Π_n. Then $\mathcal{P}(C)$ is a cone with extreme rays given by the set of vectors*

$$\{e(k,l)\colon\ k, l \text{ are not in the same part of the partition } \pi_t\} \cup \bigcup_{i=t+1}^{n}\{v(F)\colon\ F \text{ is a traversal of } \pi_i\}.$$

Note that $e(k,l)$ denotes the standard unit vector in $\mathbb{R}^{n(n-1)/2}$ with a 1 in the k,l position and a 0 elsewhere. Note that if $t = 1$, the first set in the union is empty.

The remainder of this section consists of the proof of Theorem 3.6 and completes our description of the cones $\mathcal{P}(C)$. The proof will be broken into a number of pieces, and will work by induction on both t and n.

Let $\mathbf{1}^t$ denote the vector in $\mathbb{R}^{t(t-1)/2}$ all of whose coordinates are equal to one. Note that $\mathbf{1}^n$ is the induced vector of the single traversal associated to the partition $1|2|\cdots|n$, which appears in every partial chain.

Lemma 3.7. *Let $C = \pi_s \lessdot \cdots \lessdot \pi_t$ be a partial chain in Π_n with $s > t$. Then*

(1) *$\mathbf{1}^s$ is an extreme ray of $\mathcal{P}(C)$, and*
(2) *$\mathbf{1}^s$ is the only extreme ray of $\mathcal{P}(C)$ that has a nonzero $(\lambda^s_{j(s)}, \lambda^s_{k(s)})$ coordinate where $(\lambda^s_{j(s)}, \lambda^s_{k(s)})$ is the pair of parts joined together in the partition π_{s-1}.*

Proof. First of all, all the inequalities of Proposition 3.2 are satisfied with equality by $\mathbf{1}^s$ so that $\mathbf{1}^s \in \mathcal{P}(C)$, except for the single inequality $d(\lambda_{j(s)}, \lambda_{k(s)}) \geqslant 0$, which is satisfied strictly. Hence the extreme ray $\mathbf{1}^s$ is in the intersection of all the facet defining inequalities except for one. Since $\mathcal{P}(C)$ is a pointed cone because it is contained in the positive orthant, this implies that $\mathbf{1}^s$ is an extreme ray.

This proves part (1). Furthermore, since every extreme ray of a cone is the intersection of some of its facet defining inequalities, every other extreme ray must have the inequality $d(\lambda_{j(s)}, \lambda_{k(s)}) \geqslant 0$ as an active inequality. This proves part (2). □

Note that Lemma 3.7 implies that if $s > t$, the vertex figure of $\mathcal{P}(C)$ is a pyramid with apex $\mathbf{1}^s$.

Let $C = \pi_s \lessdot \cdots \lessdot \pi_t$ be a partial chain, and C' a partial chain obtained as a final segment of C, that is, there is a $s < u \leqslant t$, such that $C' = \pi_u \lessdot \cdots \lessdot \pi_t$. The UPGMA algorithm induces a natural linear map $A(C, C') : \mathbb{R}^{s(s-1)/2} \to \mathbb{R}^{u(u-1)/2}$. In particular, it is defined by

$$\big(A\big(C, C'\big)d\big)\big(\lambda, \lambda'\big) = \frac{1}{|\lambda||\lambda'|} \sum_{\substack{\mu, \mu' \in \pi_s \\ \mu \subseteq \lambda, \mu' \subseteq \lambda'}} |\mu|\big|\mu'\big| d\big(\mu, \mu'\big)$$

where λ, λ' are parts of π_u. Note, in particular, the quantity $d(\mu, \mu')$ only appears in the formula for $(A(C, C')d)(\lambda, \lambda')$, so that $A(C, C')$ is a coordinate substitution map (Definition 3.9) when restricted to the coordinates $d(\mu, \mu')$ where μ, μ' are in different parts of π_s.

With the preceding paragraph in mind, we let $\tilde{\mathcal{P}}(C)$ denote the intersection of $P(C)$ with the hyperplane $\{d\colon\ d(\lambda_{j(s)}, \lambda_{k(s)}) = 0\}$.

Proposition 3.8. *Let $C = \pi_s \lessdot \cdots \lessdot \pi_t$ be a partial chain and with final segment $C' = \pi_{s-1} \lessdot \cdots \lessdot \pi_t$. Then $A(C, C') : \tilde{\mathcal{P}}(C) \to \mathcal{P}(C')$ is surjective, and $\tilde{\mathcal{P}}(C) = A(C, C')^{-1}(\mathcal{P}(C')) \cap \mathbb{R}_{\geqslant 0}^{s(s-1)/2-1}$.*

Proof. Note that by definition of the UPGMA algorithm, the map $A(C, C') : \mathcal{P}(C) \to \mathcal{P}(C')$ is surjective. If a vector $d^s \in \mathcal{P}(C)$, then so is the vector

$$d' = d^s - d^s\big(\lambda^s_{j(s)}, \lambda^s_{k(s)}\big) e\big(\lambda^s_{j(s)}, \lambda^s_{k(s)}\big),$$

obtained by zeroing out the $(\lambda^s_{j(s)}, \lambda^s_{k(s)})$ coordinate. However, $A(C, C')d^s = A(C, C')d'$, which implies that $A(C, C') : \tilde{\mathcal{P}}(C) \to \mathcal{P}(C')$ is surjective.

To see that $\tilde{\mathcal{P}}(C) = A(C, C')^{-1}(\mathcal{P}(C')) \cap \mathbb{R}_{\geqslant 0}^{s(s-1)/2-1}$, note that the inequalities that describe $\tilde{\mathcal{P}}(C)$ are precisely the pullbacks of the inequalities that describe $\mathcal{P}(C')$, plus nonnegativity constraints, since none of the inequalities on $\mathcal{P}(C)$ coming from the covering relation $\pi_s \lessdot \pi_{s-1}$ are needed. □

Definition 3.9. A linear transformation $\phi : \mathbb{R}^n \to \mathbb{R}^m$ is a *coordinate substitution* if for each of the coordinate vectors e_i, $\phi(e_i) = c_i e_{\alpha(i)}$ with $c_i > 0$, where $\alpha : [n] \to [m]$. That is, each coordinate maps to a scaled version of another coordinate.

Lemma 3.10. *Let $D \subseteq \mathbb{R}^m$ be a polyhedral cone, $\phi : \mathbb{R}^n \to \mathbb{R}^m$ be a coordinate substitution with associated map α, and $C \subseteq \mathbb{R}^n$ a polyhedral cone such that $\phi(C) = D$. Suppose that $C = \mathbb{R}^n_{\geqslant 0} \cap \phi^{-1}(D)$. Let V be the set of extreme rays of C. Then extreme rays of D consist of all vectors obtained by the following procedure*:

For each extreme ray $\sum_j a_j e_j \in V$, consider all vectors of the form $\sum_j a_j / c_{\beta(j)} e_{\beta(j)}$ ranging over all functions $\beta : [m] \to [n]$ such that $\alpha(\beta(j)) = j$ for all j.

Proof. It suffices to show that under the hypotheses of the lemma, every extreme ray of C maps onto an extreme ray of D. Indeed, if that is the case, the extreme rays of C are precisely the vertices of the polytopes $\phi^{-1}(v) \cap \mathbb{R}^n_{\geqslant 0}$ as v ranges over the extreme rays of V. Note that since ϕ is a coordinate substitution $\phi^{-1}(v)$ is isomorphic to a product of simplices, the simplices being defined over coordinate subsets over the form $\alpha^{-1}(j)$. The vertices of these products of simplices have the form of the statement of the lemma.

Hence, it suffices to show the claim that every extreme ray of C maps onto an extreme ray of D. So suppose that v' is an extreme ray of C such that $\phi(v') = v$ is not an extreme ray of D. Then there exists $w, u \in D$, not equal to v such that $v = w + u$. Using these vectors, we construct $w', u' \in C$ not equal to v' such that $v' = w' + u'$. For each i such that $\alpha(i) = j$ define

$$w'_i = \frac{w_j}{v_j} v'_i \quad \text{and} \quad u'_i = \frac{u_j}{v_j} v'_i.$$

Clearly with this choice, we have $v' = w' + u'$ since $v_j = w_j + u_j$, and both w' and u' consist of nonnegative vectors. Also, since w, u not equal v, neither are w', u' equal to v'. So we must show that $\phi(w') = w$ and $\phi(u') = u$. But

$$\phi(w')_j = \sum_{i:\alpha(i)=j} \frac{w_j}{v_j} c_i = \frac{w_j}{v_j} \sum_{i:\alpha(i)=j} c_i = \frac{w_j}{v_j} v_j = w_j.$$

Similarly for u', which completes the proof. $\square$

We now have all the ingredients to prove Theorem 3.6.

Proof of Theorem 3.6. Let $C = \pi_s \lessdot \cdots \lessdot \pi_t$. First of all, note that if $s = t$, then $\mathcal{P}(C)$ is the positive orthant in $\mathbb{R}^{s(s-1)/2}$, whose extreme rays are the standard unit vectors.

Now assume that $s > t$. According to Lemma 3.7, the vector $\mathbf{1}^s$ is an extreme ray of $\mathcal{P}(C)$. Letting $C' = \pi_{s-1} \lessdot \cdots \lessdot \pi_t$, Proposition 3.8 we see that all other extreme rays of $\mathcal{P}(C)$ can be obtained by applying Lemma 3.10 to the extreme rays of $\mathcal{P}(C')$. Repeating this procedure for the extreme rays of $\mathcal{P}(C)$ that do not map to $\mathbf{1}^{s-1} \in \mathcal{P}(C')$, we see that every extreme ray of $\mathcal{P}(C)$ besides $\mathbf{1}^s$ can be obtained as a vertex of $A(C, C_u)^{-1}(\mathbf{1}^u)$ where $C_u = \pi_u \lessdot \cdots \lessdot \pi_t$, plus the vertices of $A(C, C_t)^{-1}(e(\lambda_k, \lambda_l))$.

To complete the proof of the theorem we must analyze the vertices of $A(C, C_t)^{-1}(e(\lambda_k, \lambda_l))$ and show that the vertices of $A(C, C_u)^{-1}(\mathbf{1}^u)$ are precisely the induced vectors from the traversals of π_u. For both of these statements, we can use Lemma 3.10.

Indeed, $A(C, C_u)$ is the map such that

$$\big(A(C, C_u)d\big)\big(\lambda, \lambda'\big) = \frac{1}{|\lambda| \cdot |\lambda'|} \sum_{\substack{x \in \lambda \\ y \in \lambda'}} d(x, y).$$

This implies, by Lemma 3.10 that the vertices of

$$A(C, C_t)^{-1}\big(e\big(\lambda, \lambda'\big)\big)$$

are $|\lambda| \cdot |\lambda'| e(k, l)$ such that $k \in \lambda$ and $l \in \lambda'$. Since we can ignore the scaling factor $|\lambda| \cdot |\lambda'|$ when describing extreme rays, taking the union over all pairs $\lambda, \lambda' \in \pi_t$, yields the set of rays $\{e(k, l){:}\ k, l$ are not in the same part of the partition $\pi_t\}$ from Theorem 3.6.

Similarly, applying Lemma 3.10 to the map $A(C, C_u)$ and the vector $\mathbf{1}^u$ yields the set of induced vectors $v(F)$ associated to the partition π_u. Indeed, the coordinate 1 in the (λ, λ') position of $\mathbf{1}^u$ produces an entry of $|\lambda| \cdot |\lambda'|$ in exactly one of the positions $d(x, y)$ such that $x \in \lambda$, $y \in \lambda'$. This completes the proof of Theorem 3.6. $\square$

We now show that Theorem 3.6 implies that the UPGMA cones have exponentially many extreme rays.

Proposition 3.11. *The cones $\mathcal{P}(C)$ have exponentially many extreme rays.*

Proof. Given $\pi_s = \lambda_1^s | \cdots | \lambda_s^s$, the number of traversals is the product of the pairwise products of the cardinalities of the blocks of π_s. So the number of extreme rays induced by π_s is

$$\prod_{\{i,j\} \subset \binom{[s]}{2}} |\lambda_i^s||\lambda_j^s| = \prod_{i=1}^{s} |\lambda_i^s|^{s-1}.$$

Given a maximal chain $C \in \Pi_n$, the total number of extreme rays will be

$$\sum_{s=2}^{n} \prod_{i=1}^{s} |\lambda_i^s|^{s-1}$$

which is exponential in n. $\square$

Note that Propositions 3.2, 3.3, 3.11 and Theorem 3.6 yield Theorem 3.1.

4. Applications of Theorem 3.6

We use the characterization of the extreme rays of the cones $\mathcal{P}(C)$ to provide easy geometric applications. First, the set $\mathcal{P}(T)$ of all dissimilarity maps for which UPGMA returns a given tree is not a convex set in general. Second, the partition of the positive orthant into the cones $\mathcal{P}(C)$ does not have the structure of a polyhedral fan, which means cones do not intersect in their boundary in an especially nice way. Third, we show the comb tree topology minimizes the number of rays in a UPGMA cone.

Corollary 4.1. *The UPGMA regions $\mathcal{P}(C)$ are not convex in general.*

Proof. We give an example for $n = 4$. Let $T = ((12)(34))$. Then $\mathcal{P}(T) = \mathcal{P}(C_1) \cup \mathcal{P}(C_2)$ where

$$C_1 = 1|2|3|4 \lessdot 3|4|12 \lessdot 12|34 \lessdot 1234$$

$$C_2 = 1|2|3|4 \lessdot 1|2|34 \lessdot 34|12 \lessdot 1234.$$

Now $v_1 = (0, 0, 2, 2, 0, 1)$ is an extreme ray of $P(C_1)$ induced by a traversal of 3|4|12 and $v_2 = (1, 0, 2, 2, 0, 0)$ is an extreme ray of $\mathcal{P}(C_2)$ induced by a traversal of 1|2|34. Let d be the convex combination

$$d = \frac{1}{2} v_1 + \frac{1}{2} v_2 = \left(\frac{1}{2}, 0, 2, 2, 0, \frac{1}{2}\right).$$

If d is input into UPGMA, the algorithm will return a tree with either $(1, 3)$ or $(2, 4)$ as a cherry, so d is not in $\mathcal{P}(T)$. So, in general, UPGMA regions are not convex unless $\mathcal{P}(T) = \mathcal{P}(C)$ for a single chain C in Π_n. $\square$

A *fan* is a family $\mathcal{F}$ of cones in $\mathbb{R}^n$ such that

(1) if $P \in \mathcal{F}$ then every nonempty face of P is in $\mathcal{F}$;
(2) if $P_1, P_2 \in \mathcal{F}$ then $P_1 \cap P_2 \in \mathcal{F}$.

Corollary 4.2. *The UPGMA cones do not partition $\mathbb{R}^{\binom{n}{2}}$ into a fan.*

Proof. Consider the two chains in Π_4

$$C_1 = 1|2|3|4 \lessdot 3|4|12 \lessdot 4|123 \lessdot 1234$$
$$C_2 = 1|2|3|4 \lessdot 2|4|13 \lessdot 4|123 \lessdot 1234.$$

The vector $(0,0,0,1,1,1)$ generates an extreme ray of $P(C_1) \cap P(C_2)$ which we verified using polymake [6]. If $P(C_1) \cap P(C_2)$ was a face of $P(C_1)$ and $P(C_2)$, then $(0,0,0,1,1,1)$ would generate a ray of $P(C_1)$ and $P(C_2)$. However by Theorem 3.6, extreme rays of $P(C_1)$ and $P(C_2)$ must correspond to partitions in Π_4. Only partitions with 3 blocks induce vectors with 3 nonzero coordinates, and no partition of the set [4] has 3 blocks of equal cardinality. So, no traversal of a partition in Π_4 induces a multiple of $(0,0,0,1,1,1)$. Therefore the UPGMA cones are not a fan. □

Corollary 4.3. *For each n, the comb tree topology minimizes the number of extreme rays over all UPGMA cones in* $\mathbb{R}^{\binom{n}{2}}$.

Proof. Fix n. We will show that for each $1 \leqslant s \leqslant n$, the partitions whose parts have cardinalities $1, 1, \dots, 1, n-s+1$ minimize the number of traversals for all partitions with s parts. For all integers $x, y > 0$, we have $xy \geqslant (x+y-1)(1)$. So for $\pi_s = \lambda_1^s | \cdots | \lambda_s^s$, the number of extreme rays induced by π_s satisfies

$$\prod_{\{i,j\} \subset \binom{[s]}{2}} |\lambda_i^s||\lambda_j^s| \geqslant \prod_{\{i,j\} \subset \binom{[s]}{2}} (1)\left(|\lambda_i^s| + |\lambda_j^s| - 1\right).$$

The only type of partition in Π_n with s parts such that all pairs $\{i,j\} \subset \binom{[s]}{2}$ satisfy either $|\lambda_s^i| = 1$ or $|\lambda_s^j| = 1$ is the type with $s-1$ singleton parts and one part of size $n-s+1$. Therefore partitions of this type minimize the number of associated induced vectors.

If C is a maximal chain in Π_n such that every π_s in C is of this type, then the tree returned by $d \in \mathcal{P}(C)$ has the comb tree topology. Therefore this tree topology minimizes the number of extreme rays for the cone $\mathcal{P}(C)$. □

5. Spherical volumes of UPGMA regions

A natural way to measure the region $\mathcal{P}(T)$ is to calculate the $\binom{n}{2} - 1$ dimensional measure of the surface arising as the intersection of the cones $\mathcal{P}(C) \subset \mathcal{P}(T)$ with the unit sphere S in $\mathbb{R}^{\binom{n}{2}}$. We refer to this measure as *spherical volume*.

We estimated the spherical volume of UPGMA cones in two ways using Mathematica, polymake [6], and the software [7]. For the first method, we sampled points from the positive orthant using a spherical distribution and input the samples into UPGMA, recording which ranked tree the algorithm returned on the input point. The volume of $\mathcal{P}(T)$ is then the fraction of the total sample points returning T. We calculate volumes for $n = 4, 5, 6, 7$ using this method.

For the second method, we used a Monte Carlo strategy to estimate the surface area of the cones. For $n = 4, 5, 6$, we used the software [7] for $n = 4, 5, 6$. This software requires as input triangulations of point configurations that we computed using polymake [6]. For $n = 7$, some triangulations for maximal chains in Π_7 were too large to compute and use. We used Mathematica to implement a modification of the sampling strategy employed in [3] along with the UPGMA algorithm.

The basic strategy using Monte Carlo integration to compute spherical volumes can be described as follows. Given a simplicial cone cone(V) spanned by vectors $V = v_1, \dots, v_n$, it is easy to generate uniform samples from the simplex conv(V). The map that takes a point $x \in \text{conv}(V)$ onto cone$(V) \cap S$ is simply $x \to x/\|x\|_2$. The spherical volume is then the average value of the Jacobian of this map. To calculate the spherical volume of a cone $\mathcal{P}(C)$ of a full chain in situations where we could only compute a triangulation of a cone from a partial chain $\mathcal{P}(C')$, we generate random points from the partial cone $\mathcal{P}(C')$ and compute the average of the product of Jacobian and the indicator function of lying in the cone $\mathcal{P}(C)$.

We summarize the results here of those computations for $n = 4, 5, 6, 7$ leaf trees, only displaying results for the regions $\mathcal{P}(T)$. In the tables below, we give estimates of the spherical volumes of the regions $\mathcal{P}(T)$. The column "Tree" gives the tree in Newick format. The column "# Chains" refers to the number of cones producing the given tree. The column "Volume" gives the total volume of all of the cones associated to the given tree, and the column "Fraction of orthant" gives the portion of the positive orthant in $\mathbb{R}^{\binom{n}{2}}$ that returns the given tree *topology* under UPGMA.

Recall that $\mathcal{P}(T) = \bigcup \mathcal{P}(C)$ where C ranges over the chains in Π_n corresponding to T. So, the number of cones associated to a tree T depends on the number of rank functions that T admits. For example, in the table for $n = 5$, the tree $T_2 = (((12)3)(45))$ has $4!/(4 \cdot 2 \cdot 1 \cdot 1) = 3$ rank functions, and there are 3 cones in $\mathcal{P}(T_2)$.

A more detailed explanation of the volume computations, as well as software and input files, is available at [2].

	Tree	# Chains	Volume	Fraction of orthant
1	(((12)3)4)	1	0.0238	0.5895
2	((12)(34))	2	0.0662	0.4099

	Tree	# Chains	Volume	Fraction of orthant
1	((((12)3)4)5)	1	8.57×10^{-5}	0.206
2	(((12)3)(45))	3	5.01×10^{-4}	0.604
3	(((12)(34))5)	2	3.14×10^{-4}	0.189

	Tree	# Chains	Volume	Fraction of orthant
1	(((((12)3)4)5)6)	1	2.05×10^{-8}	0.042
2	((((12)3)4)(56))	4	2.10×10^{-7}	0.216
3	((((12)3)(45))6)	3	2.16×10^{-7}	0.223
4	(((12)3)((45)6))	6	4.50×10^{-7}	0.229
5	((((12)(34))5)6)	2	1.05×10^{-7}	0.054
6	(((12)(34))(56))	8	9.06×10^{-7}	0.231

	Tree	# Chains	Volume	Fraction of orthant
1	((((((12)3)4)5)6)7)	1	2.75×10^{-13}	0.0050
2	(((((12)3)4)5)(67))	5	4.82×10^{-12}	0.0435
3	(((((12)3)4)(56))7)	4	6.32×10^{-12}	0.0570
4	((((12)3)4)((56)7))	10	1.95×10^{-11}	0.1762
5	(((((12)3)(45))6)7)	3	4.45×10^{-12}	0.0402
6	((((12)3)(45))(67))	15	5.72×10^{-11}	0.2581
7	((((12)3)((45)6))7)	6	1.66×10^{-11}	0.0747
8	(((12)3)((45)(67)))	20	9.00×10^{-11}	0.2030
9	(((((12)(34))5)6)7)	2	1.73×10^{-12}	0.0078
10	((((12)(34))5)(67))	10	2.63×10^{-11}	0.0593
11	((((12)(34))(56))7)	8	3.33×10^{-11}	0.0753

The computations suggest some observations which might hold true for large n. As Corollary 4.3 shows, the cone associated to the single rank function on the comb tree yields the cone $\mathcal{P}(C)$ with the fewest number of extreme rays. Our computations up to $n = 7$ suggest that this is also the cone with the smallest spherical volume. See [2] for those values. The size of the region $\mathcal{P}(T)$ appears be roughly proportional to the number of chains C that yield the tree T and appears to be smallest for the comb tree. Furthermore, the relative proportion of the positive orthant taken up by the comb tree topology appears to be the smallest. We predict that these patterns hold for larger number of taxa as well.

Acknowledgments

Ruth Davidson was partially supported by the US National Science Foundation (DMS 0954865). Seth Sullivant was partially supported by the David and Lucille Packard Foundation and the US National Science Foundation (DMS 0954865).

References

[1] David Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, Statist. Sci. 16 (2001) 23–34.
[2] R. Davidson, S. Sullivant, Supplementary materials for "Polyhedral combinatorics of UPGMA cones", http://www4.ncsu.edu/~smsulli2/Pubs/UPGMACones/UPGMACones.html.
[3] Kord Eickmeyer, Peter Huggins, Lior Pachter, Ruriko Yoshida, On the optimality of the neighbor-joining algorithm, Algorithms Mol. Biol. 3 (2008).
[4] C. Fahey, S. Hosten, N. Krieger, L. Timpe, Least squares methods for equidistant tree reconstruction, arXiv:0808.3979, 2008.
[5] J. Felsenstein, Inferring Phylogenies, 2nd edition, Sinauer Associates, 2003.
[6] Ewgenij Gawrilow, Michael Joswig, Polymake: a framework for analyzing convex polytopes, in: Polytopes – Combinatorics and Computation, Oberwolfach, 1997, in: DMV Sem., vol. 29, Birkhäuser, Basel, 2000, pp. 43–73.
[7] P. Huggins, NJBMEVolume: Software for computing volumes, http://bio.math.berkeley.edu/NJBME, 2008.
[8] C. Semple, M. Steel, Phylogenetics, Oxford University Press, Oxford, 2003.
[9] R.R. Sokal, P.H.A. Sneath, Numerical Taxonomy, W.H. Freeman, San Francisco, 1963.
[10] R. Stanley, Enumerative Combinatorics, vol. I, Cambridge University Press, 1997.
[11] G. Ziegler, Lectures on Polytopes, Springer, 2006.