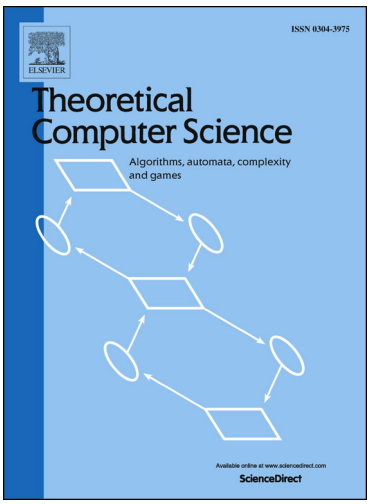
Accepted Manuscript



UPGMA and the normalized equidistant minimum evolution problem

Vincent Moulton, Andreas Spillner, Taoyang Wu

PII: S0304-3975(18)30069-0
DOI: https://doi.org/10.1016/j.tcs.2018.01.022
Reference: TCS 11464

To appear in: *Theoretical Computer Science*

Received date: 12 April 2017
Revised date: 22 January 2018
Accepted date: 23 January 2018

Please cite this article in press as: V. Moulton et al., UPGMA and the normalized equidistant minimum evolution problem, *Theoret. Comput. Sci.* (2018), https://doi.org/10.1016/j.tcs.2018.01.022

UPGMA and the normalized equidistant minimum evolution problem

Vincent Moulton[a], Andreas Spillner[b], Taoyang Wu[a,*]

[a] *School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK*
[b] *Department of Engineering and Natural Sciences, Merseburg University of Applied Sciences, Germany*

Abstract

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a widely used clustering method. Here we show that UPGMA is a greedy heuristic for the normalized equidistant minimum evolution (NEME) problem, that is, finding a rooted tree that minimizes the minimum evolution score relative to the dissimilarity matrix among all rooted trees with the same leaf-set in which all leaves have the same distance to the root. We prove that the NEME problem is NP-hard. In addition, we present some heuristic and approximation algorithms for solving the NEME problem, including a polynomial time algorithm that yields a binary, rooted tree whose NEME score is within $O(\log^2 n)$ of the optimum.

Keywords: UPGMA, minimum evolution, balanced minimum evolution, hierarchical clustering

2000 MSC: 68Q17, 05C05, 05C85, 92B05

1. Introduction

Clustering (i.e. subdividing a dataset into smaller subgroups or clusters) is a fundamental task in data analysis, and has a wide range of applications (see, e.g. [1]). An important family of clustering methods aim to produce a clustering of a dataset in which the clusters form a hierarchy where the clusters nest within one another. Such hierarchies are typically represented by leaf-labeled tree structures known as dendrograms or rooted phylogenetic trees. Introduced in 1958 [2], average linkage analysis, usually referred to as UPGMA (Unweighted Pair Group Method with Arithmetic Mean), is arguably the most

*Corresponding author
Email addresses: vincent.moulton@cmp.uea.ac.uk (Vincent Moulton), mail@andreas-spillner.de (Andreas Spillner), taoyang.wu@gmail.com (Taoyang Wu)

popular hierarchical clustering algorithm in use to date, and remains widely cited[1] and extremely popular (see, e.g. [3]). This is probably because UPGMA is conceptually easy to understand and fast in practice, an important consideration as big data sets are becoming the norm in many areas. UPGMA is commonly used in phylogenetics and taxonomy to build evolutionary trees [4, Chapter 11] as well as in related areas such as ecology [5] and metagenomics [6]. In addition, it is used as a general hierarchical clustering tool in bioinformatics and other areas including data mining and pattern recognition [7, Chapter 2].

UPGMA is a text-book algorithm that belongs to the family of agglomerative clustering methods that share the following common bottom-up scheme (cf. e.g. [4, p.162]). They take as input a dissimilarity D on a set X, i.e. a real-valued, symmetric map on $X \times X$ which vanishes on the diagonal, and build a collection of clusters or subsets of X which correspond to a rooted tree with leaf-set X. To do this, at each step two clusters with the minimum inter-cluster dissimilarity are combined to create a new cluster, starting with the collection of clusters consisting of singleton subsets of X, and finishing when the cluster X is obtained. Different formulations of the inter-cluster dissimilarity, which specifies the dissimilarity of sets as a function of the dissimilarities observed on the members of the sets, lead to different heuristic criteria of the agglomerative methods. UPGMA, as the name average linkage analysis suggests, uses the mean dissimilarity across all pairs of elements that are contained within the two clusters. Formally, two clusters $A, B \subseteq X$ are selected for merging at each iteration step of UPGMA if the average

$$\frac{1}{|A||B|} \sum_{a \in A,\, b \in B} D(a, b)$$

is minimized over all possible pairs of clusters. Since the arithmetic mean is used, UPGMA is often more stable than linkage methods in which only a subset of the elements within the clusters are used (e.g. the single-linkage method).

UPGMA is commonly thought of as a method that greedily constructs a rooted phylogenetic tree that is closest to the input dissimilarity matrix in the least squares sense [8]. However, it is not guaranteed to do so, although it often does quite well in practice [4,

[1] According to Google Scholar, the method has been cited over 17,200 times during the period between 2011 and 2015.

p.162]. In [9] it was shown that the related Neighbor-Joining [10] method for constructing unrooted phylogenetic trees from dissimilarity matrices can be thought of as a greedy heuristic that minimizes the so-called *balanced minimum evolution* score. Here we shall observe that (see Section 3), in a similar way, UPGMA is a greedy heuristic for computing a rooted phylogenetic tree that minimizes the so-called *minimum evolution* score [11] over all rooted phylogenetic trees on the same fixed leaf-set in which all leaves have the same distance in the tree to the root. We refer to this optimization problem as the *normalized equidistant minimum evolution* (NEME) problem, and expect that a better understanding of this problem will provide further insights into the behavior of the UPGMA algorithm.

Theoretical properties of discrete optimization problems arising in the construction of evolutionary trees have been studied for many years (for some earlier work see, e.g. [12, 13, 14]). Among these, the problems falling under the name of minimum evolution alone form a quite diverse family (see, e.g. [15]), in which the so-called balanced minimum evolution problem [16] is a particularly well-studied member. For this problem it was recently shown in [17] that for general $n \times n$-input dissimilarity matrices there exists a constant $c > 1$ such that no polynomial time algorithm can achieve an approximation factor of c^n unless P equals NP. We note that this hardness result does not rely on the often imposed restriction (see, e.g. [18, 13]) that the edge lengths of the constructed tree must be integers. Moreover, in contrast to general input dissimilarity matrices, for inputs that are metrics (i.e. matrices that also satisfy the triangle inequality) a polynomial time algorithm with an approximation factor of 2 is presented in [17]. Interestingly, the proof of this approximation factor uses the fact that the balanced minimum evolution score of an unrooted tree can be interpreted as being the average length of a spanning cycle compatible with the structure of the tree [19].

Another recent, related direction of work considers the algebraic structure of the space of rooted phylogenetic trees induced by the UPGMA method (see, e.g. [8, 20]). This algebraic structure is tightly linked with the property of consistency of a tree construction method, that is, those conditions under which the method is able to reconstruct a tree that has been used to generate the input dissimilarity matrix (see, e.g. [21]). In the context of our work, we are particularly interested in the consistency of methods that perform a local search of the space of all rooted phylogenetic trees on a fixed set of leaves (see, e.g. [16]).

Again, balanced minimum evolution is the variant of minimum evolution for which some consistency results of this type are known [22, 23].

After presenting some preliminaries in the next section, in Section 3 we begin by giving an explicit formula of the minimum evolution score of a rooted tree T as a linear combination of the input dissimilarities. This formula allows us to interpret the minimum evolution score of T in terms of the average length of a minimum spanning tree compatible with the set of clusters induced by T.

Using this observation, we explain how UPGMA can be regarded as a greedy heuristic for the NEME problem for binary rooted trees. In addition, we show that there are rooted phylogenetic trees with n leaves on which some input dissimilarity matrix has an optimal least squares fit while the NEME score of that tree for the same dissimilarity matrix is worse than the minimum NEME score by a factor in $\Omega(n^2)$. This highlights the fact that the NEME problem and searching for trees with minimum least squares fit are quite distinct problems.

Next, in Section 4, we explore solving the NEME problem by performing a local search of the space of binary rooted phylogenetic trees using so-called rooted nearest neighbor interchanges as the moves in the local search. We show that this approach is consistent. More specifically, for any input dissimilarity matrix that can be perfectly represented by a unique binary rooted phylogenetic tree with all leaves having the same distance from the root, we prove that the local search will arrive at this tree after a finite number of moves.

In Section 5 we show that the NEME problem is NP-hard even for $n \times n$ input distance matrices that satisfy the triangle inequality and only take on $O(\log n)$ different values. In light of this fact, in Section 6 we consider some approximation algorithms for solving the NEME problem. More specifically, we first show that the tree produced by UPGMA can have a score that is worse than the minimum score by a factor in $\Omega(n)$. Then, for dissimilarity matrices that satisfy the triangle inequality, we present a polynomial time algorithm that yields a binary rooted phylogenetic tree whose NEME score is within $O(\log^2 n)$ of the optimum. We conclude in Section 7 by mentioning two possible directions for future work.

2. Preliminaries

In this section we give a formal definition of the NEME problem and introduce some of the notation and terminology that will be used throughout this paper.

Let X be a finite non-empty set. A *dissimilarity* on X is a symmetric map $D : X \times X \to \mathbb{R}$ with $D(x, x) = 0$ for all $x \in X$. In this paper, a *rooted phylogenetic tree* on X is a rooted tree $T = (V, E, \rho)$ with (i) root ρ of degree 1, (ii) leaf set X and (iii) every vertex not in $X \cup \{\rho\}$ having degree at least 3. Note that even though we require the root to have degree 1, we do not consider it a leaf of the tree. A *normalized equidistant edge weighting* (NEEW) of a rooted phylogenetic tree $T = (V, E, \rho)$ on X is a map $\omega : E \to \mathbb{R}$ such that the total weight of the edges on the path from ρ to x is 0 for all $x \in X$. More generally, for all $u, v \in V$, we denote the total weight of the edges on the path between vertices u and v by $\ell_{(T,\omega)}(u, v)$. The *height* $h_{(T,\omega)}(v)$ of any vertex $v \in V$ is defined as $\ell_{(T,\omega)}(v, x)$ for any leaf x in the subtree of T rooted at v. The *length* $\ell_\omega(T)$ of T under the edge weighting ω is $\sum_{e \in E} \omega(e)$, that is, the total weight of all edges of T.

Note that the length of any rooted phylogenetic tree $T = (V, E, \rho)$ on X with a normalized equidistant edge weighting ω can also be expressed as follows:

$$\ell_\omega(T) = \sum_{v \in (V - (X \cup \{\rho\}))} (deg(v) - 2) \cdot h_{(T,\omega)}(v), \tag{1}$$

where $deg(v)$ denotes the degree of vertex v in T. Note that the restriction of $\ell_{(T,\omega)}$ to $X \times X$ yields a dissimilarity $D_{(T,\omega)}$ on X. Moreover, this dissimilarity is an *ultrametric*, that is, $D_{(T,\omega)}(x, z) \leq \max\{D_{(T,\omega)}(x, y), D_{(T,\omega)}(y, z)\}$ holds for all $x, y, z \in X$, if and only if the edge weighting ω assigns a non-negative real number to each edge not adjacent to a vertex in $X \cup \{\rho\}$. We call any such edge weighting *interior positive*.

Let D be a dissimilarity on X and $T = (V, E, \rho)$ a rooted phylogenetic tree on X. For any vertex $v \in V$ let $ch(v)$ denote the set of *children* of v, that is, the set of vertices u that are adjacent to v and for which v lies on the path from u to ρ. Moreover, we refer to v as the *parent* of the vertices $u \in ch(v)$ and we denote by $C(v)$ the *cluster* of elements in X induced by v, that is, the set of those leaves x of T for which the path from x to ρ contains v. In [24] it is shown that, for any dissimilarity D and any rooted phylogenetic tree T on

X, there is a unique normalized equidistant edge weighting $\omega = \omega_{(D,T)}$ with

$$\Delta(D, D_{(T,\omega)}) = \sum_{\{x,y\}\in\binom{X}{2}} \big(D(x,y) - D_{(T,\omega)}(x,y)\big)^2$$

minimum, where $\binom{X}{2}$ denotes the set of 2-element subsets of X. More precisely, using the notation $\ell_{(T,\omega)}(.,.)$ introduced above to represent the relevant sums of edge weights, this edge weighting ω is the unique solution of the system of linear equations

$$\ell_{(T,\omega)}(v,y) = \frac{1}{2\sum_{\{u,u'\}\in\binom{ch(v)}{2}} |C(u)|\cdot|C(u')|} \sum_{\{u,u'\}\in\binom{ch(v)}{2}} \sum_{\substack{x\in C(u)\\ x'\in C(u')}} D(x,x'), \tag{2}$$

for all $v \in V - (X \cup \{\rho\})$, and $\ell_{(T,\omega)}(v,y) = 0$, for all $v \in X \cup \{\rho\}$, $y \in C(v)$. Note that this is the analogue to Vach's theorem for unrooted trees [25]. For later use, we put $\Delta(D, D_{(T,\omega_{(D,T)})}) = \Delta(D,T)$.

Now, the normalized equidistant minimum evolution score of a rooted phylogenetic tree T on X with respect to a dissimilarity D on X is formally defined as

$$\sigma_D(T) = \ell_{\omega_{(D,T)}}(T),$$

that is, the length of T under the edge weighting $\omega_{(D,T)}$. The NEME problem is to compute, for an input dissimilarity D on X, a rooted phylogenetic tree on X with minimum NEME score. Formally, it can be stated as below.

Problem NEME Problem

Instance: A distance matrix D on a finite set X and a number p.

Question: Does there exist a rooted phylogenetic tree T on X such that $\sigma_D(T) \leq p$ holds.

3. UPGMA and the NEME problem

We begin this section by explaining how the UPGMA algorithm can be reinterpreted as a greedy approach to solving the NEME problem for binary rooted trees. First note that it follows directly from Equations (1) and (2) that the NEME score of a rooted phylogenetic tree $T = (V, E, \rho)$ can be written as the following linear combination of the

given dissimilarity values:

$$\sigma_D(T) = \sum_{\{x,y\}\in\binom{X}{2}} \alpha^T_{\{x,y\}} \cdot D(x,y) \quad \text{with} \tag{3}$$
$$\alpha^T_{\{x,y\}} = \frac{deg(lca(x,y)) - 2}{2\sum_{\{u,u'\}\in\binom{ch(lca(x,y))}{2}} |C(u)| \cdot |C(u')|},$$

where $lca(u,v)$ is the lowest common ancestor in T for any two vertices $u, v \in V$. In particular, in case T is a *binary* tree, that is, every vertex not in $X \cup \{\rho\}$ has degree precisely 3, we obtain, for any $\{x,y\} \in \binom{X}{2}$ the coefficient

$$\alpha^T_{\{x,y\}} = \frac{1}{2\prod_{u\in ch(lca(x,y))} |C(u)|}. \tag{4}$$

As an immediate consequence of (3) we obtain that the score $\sigma_D(T)$ is linear in D, that is, when D can be written as $D = \lambda_1 \cdot D_1 + \lambda_2 \cdot D_2$ for non-negative real numbers λ_i and dissimilarities D_i, $i \in \{1,2\}$, then

$$\sigma_D(T) = \lambda_1 \cdot \sigma_{D_1}(T) + \lambda_2 \cdot \sigma_{D_2}(T) \tag{5}$$

To link the Formula (3) with the UPGMA algorithm, recall that this algorithm constructs from a given dissimilarity D on X a binary rooted phylogenetic tree $T = (V, E, \rho)$ on X by generating the list of clusters associated to the vertices of T. It starts with the singleton clusters associated to the leaves of T. Then, in each iteration of the algorithm it considers the partition $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ of X into $m \geq 2$ clusters associated to those vertices of T generated by the algorithm so far whose parents in T have not been generated yet. UPGMA then selects a pair of two distinct clusters $A, B \in \mathcal{C}$ that minimizes

$$\frac{1}{|A||B|} \sum_{a\in A,\, b\in B} D(a,b),$$

generates the vertex v of T with $C(v) = A \cup B$ and then considers in the next iteration the partition $(\mathcal{C} - \{A,B\}) \cup \{A \cup B\}$.

Next note that fixing, for every $v \in (V - (X \cup \{\rho\}))$, an arbitrary ordering of the two children u_1 and u_2 of v and putting $A(v) = C(u_1)$ and $B(v) = C(u_2)$, we can rewrite Formula (3) for the binary rooted phylogenetic tree T generated by UPGMA as

$$\sigma_D(T) = \frac{1}{2} \sum_{v\in(V-(X\cup\{\rho\}))} \left[\frac{1}{|A(v)| \cdot |B(v)|} \sum_{a\in A(v),\, b\in B(v)} D(a,b) \right]. \tag{6}$$

This suggests the following interpretion: In each iteration UPGMA greedily pairs two clusters so as to minimize the contribution to the score $\sigma_D(T)$ by the vertex v generated during that iteration.

Interestingly, the coefficients in (4) suggest the following alternative interpretation of the NEME score of a binary rooted phylogenetic tree $T = (V, E, \rho)$: Consider the complete graph G with vertex set X. Each edge $\{x, y\}$ of G is weighted with the value $D(x, y)$. We construct a random subgraph of G as follows. For each vertex $v \in V - (X \cup \{\rho\})$ select a random edge that has precisely one end point in each cluster associated with the two children of v. Let H denote the resulting subgraph of G. It is easy to see that H is always a spanning tree of G. Thus, in case T is binary, $\sigma_D(T)$ can be interpreted as half the average length of a random spanning tree H of G that is compatible with the clusters of T. Based on (3), this interpretation can be extended to the non-binary case where, instead of a spanning tree, a random spanning forest in G with $|X| - 1$ edges, some of which selected more than once, arises.

Next we present a technical lemma summarizing some simple observations about the NEME score that will be used later. Let $\mathbf{R}_X$ denote the set of all rooted phylogenetic trees on X. In addition let $\mathbf{BR}_X$ denote the subset of those trees in $\mathbf{R}_X$ that are binary.

Lemma 1. *Let D be a non-negative dissimilarity on a finite set X with $|X| = n \geq 2$. Then, for all $T \in \mathbf{R}_X$, we have:*

(i)
$$\sum_{\{x,y\} \in \binom{X}{2}} \alpha^T_{\{x,y\}} = \frac{1}{2}(n-1),$$

(ii)
$$\frac{2}{n^2} \leq \min\{\alpha^T_{\{x,y\}} : \{x, y\} \in \binom{X}{2}\} \leq \max\{\alpha^T_{\{x,y\}} : \{x, y\} \in \binom{X}{2}\} \leq \frac{1}{2},$$

(iii)
$$\sigma_D(T) \leq \frac{n^2}{4} \min\{\sigma_D(T') : T' \in \mathbf{R}_X\}.$$

Proof. (i): We use induction on n. For $n = 2$ the equality clearly holds. Next assume $n \geq 3$ and consider any $T \in \mathbf{R}_X$. Let u be the single child of ρ. Put $k = deg(u) - 1$ and

let $v_1, v_2, \ldots, v_k$ denote the children of u. Then we have, by induction,

$$\sum_{\{x,y\}\in\binom{X}{2}} \alpha^T_{\{x,y\}} = \left[\sum_{i=1}^{k} \sum_{\{x,y\}\in\binom{C(v_i)}{2}} \alpha^T_{\{x,y\}}\right] + \left[\sum_{\substack{\{x,y\}\in\binom{X}{2}\\ u=lca(x,y)}} \alpha^T_{\{x,y\}}\right]$$
$$= \left[\sum_{i=1}^{k} \frac{1}{2}(|C(v_i)|-1)\right] + \frac{k-1}{2} = \frac{1}{2}(n-1),$$

as required.

(ii): Consider any $T \in \mathbf{R}_X$. The inequality $\max\{\alpha^T_{\{x,y\}} : \{x,y\} \in \binom{X}{2}\} \leq \frac{1}{2}$ follows immediately from the definition of the coefficients $\alpha^T_{\{x,y\}}$. And the inequality $\frac{2}{n^2} \leq \min\{\alpha^T_{\{x,y\}} : \{x,y\} \in \binom{X}{2}\}$ follows from the fact that, for any integer $k \geq 2$, the function

$$f : \mathbb{R}^k \to \mathbb{R} : (z_1, z_2, \ldots, z_k) \mapsto \sum_{1\leq i<j\leq k} z_i z_j$$

attains its maximum among all non-negative $(z_1, z_2, \ldots, z_k) \in \mathbb{R}^k$ with $\sum_{i=1}^k z_i = 1$ at $z_1 = z_2 = \cdots = z_k = \frac{1}{k}$. Hence, for any $\{x,y\} \in \binom{X}{2}$ with $deg(lca(x,y)) - 1 = k$ and $|C(lca(x,y))| = m$, we have $\alpha^T_{\{x,y\}} \geq \frac{k}{m^2}$.

(iii): This is an immediate consequence of (ii):

$$\sigma_D(T) \leq \sum_{\{x,y\}\in\binom{X}{2}} \frac{1}{2} D(x,y)$$
$$= \frac{n^2}{4}\left[\sum_{\{x,y\}\in\binom{X}{2}} \frac{2}{n^2} D(x,y)\right] \leq \frac{n^2}{4}\min\{\sigma_D(T') : T' \in \mathbf{R}_X\}.$$

■

We end this section presenting a family of dissimilarities D for which the closest rooted equidistant tree T in the least squares sense (i.e. the tree with $\Delta(D,T)$ minimum) has an NEME score that is worse than the minimum NEME score by a quadratic factor. This illustrates, as mentioned in the introduction, that the NEME problem is quite different from the problem of finding a closest rooted tree.

Lemma 2. *There exist dissimilarities D on a set X with n elements for which there exists a rooted phylogenetic tree T on X together with a normalized equidistant edge weighting ω with $D(x,y) = D_{(T,\omega)}$ for all $x, y \in X$ but*

$$\sigma_D(T) \geq \frac{n^2}{4}\min\{\sigma_D(T') : T' \in \mathbf{R}_X\}$$

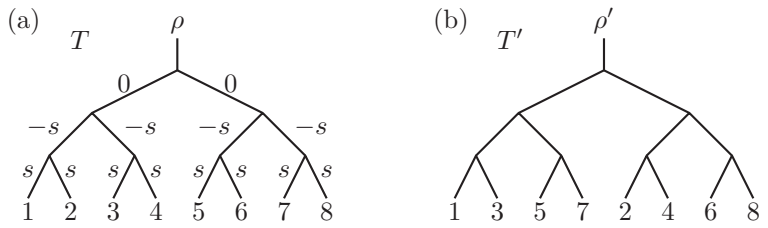


Figure 1: Examples of rooted phylogenetic trees considered in the proof of Lemma 2. For the dissimilarity $D = D_{(T,\omega)}$ induced by the rooted phylogenetic tree in (a) we have $\sigma_D(T) = 4s$. For the tree T' in (b) we obtain $\sigma_D(T') = \frac{s}{4}$.

Proof. Assume $X = \{1, 2, \ldots, n\}$ with $n = 2^k$ for some integer $k \geq 1$. Define the dissimilarity D on X by putting $D(i, i+1) = 2s$ for all odd $i \in X$, where $s > 0$ is a real number, and $D(x, y) = 0$ for all other $x, y \in X$. Note that in any rooted phylogenetic tree T on X for which $D = D_{(T,\omega)}$ holds with $\omega = \omega_{(D,T)}$ each pair $\{i, i+1\}$, i odd, must form a cherry (cf. Figure 1(a)). Moreover, for any such tree T we have $\sigma_D(T) = \frac{n}{2}s$. In contrast, in any tree T' with minimum NEME score for D the vertex $lca(i, i+1)$ must be the single child of the root for all odd $i \in X$ (cf. Figure 1(b)). This implies $\sigma_D(T') = \frac{2}{n}s$. ■

4. Searching tree space for an optimal NEME tree

In this section we shall first establish that performing a local search on the space $\mathbf{BR}_X$ for trees with minimum NEME score is a *consistent* approach, that is, if the input dissimilarity D can be represented by a binary rooted phylogenetic tree with an interior positive normalized equidistant edge weighting then this tree has minimum NEME score and, under some mild technical conditions, the local search will arrive at precisely this tree after a finite number of steps.

Note that consistency is an important property and there are general conditions known that imply consistency for approaches that construct unrooted phylogenetic trees (see e.g. [21, 26]). We first show that for any *generic* ultrametric, that is, a dissimilarity $D = D_{(T,\omega)}$ where $T = (V, E, \rho) \in \mathbf{BR}_X$ and ω is a normalized equidistant edge weighting for T with $\omega(e) > 0$ for all edges e not incident to a vertex in $X \cup \{\rho\}$, a local search in $\mathbf{BR}_X$ starting from any $T' \in \mathbf{BR}_X$ using *rooted nearest neighbor interchanges* (rNNI) will terminate in T. For unrooted trees an analogous result is established in [22] employing, however, a different, less local, tree rearrangement operation than we employ here. Recall that an

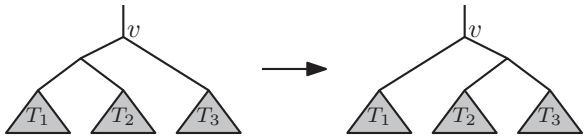


Figure 2: A rooted nearest neighbor interchange prunes a subtree T_2 from one child of v, suppresses the resulting vertex of degree 2 and then grafts T_2 onto the edge incident to the other child of v.

rNNI modifies a rooted phylogenetic tree locally around a vertex v as depicted in Figure 2. In the following, for any vertex $v \neq \rho$ of a rooted phylogenetic tree $T = (V, E, \rho)$ on X, the *subtree* of T induced by v consists of the parent p of v together with all the vertices u of T for which the path from ρ to u contains v. Note that such a subtree can be viewed as a phylogenetic tree with root p on the cluster $C(v)$ of elements in X induced by v.

Note that for a generic ultrametric we require that the edge weighting is on a binary tree and strictly positive on the edges not incident to a vertex in $X \cup \{\rho\}$. This is more restrictive than an interior positive edge weighting which might also assign weight 0 to some edges incident to a vertex in $X \cup \{\rho\}$. In the following we will use the well known fact that, for any generic ultrametric D on X, the binary rooted phylogenetic tree $T \in \mathbf{BR}_X$ with $D = D_{(T,\omega)}$ for the edge weighting $\omega = \omega_{(D,T)}$ is unique [27, Theorem 7.2.8]. We will say that T *represents* D, for short.

Lemma 3. *Let D be a generic ultrametric on X and $T \in \mathbf{BR}_X$ the unique binary rooted phylogenetic tree on X that represents D. Then, for any $T' \in (\mathbf{BR}_X - \{T\})$, there exists an rNNI that changes T' to $T'' \in \mathbf{BR}_X$ with $\sigma_D(T') > \sigma_D(T'')$.*

Proof. We use induction on $n = |X|$. The statement in the lemma clearly holds for $n \in \{1, 2\}$ in view of the fact that $|\mathbf{BR}_X| = 1$. So assume $n \geq 3$ and consider any $T' \in (\mathbf{BR}_X - \{T\})$. The situation is depicted in Figure 3(a). Let A and B, respectively, denote the set of leaves in the rooted subtrees T_1 and T_2 of T. Similarly, let A' and B', respectively, denote the set of leaves in the rooted subtrees T'_1 and T'_2 of T'. Note that the restriction $D_{|Y}$ of D to any non-empty subset $Y \subseteq X$ is again a generic ultrametric.

First consider the case that T'_1 does not represent $D_{|A'}$. Then, by induction, there exists an rNNI in T'_1 that results in a rooted phylogenetic tree T''_1 on A' with strictly smaller NEME score. Hence, applying the same rNNI to T' yields a tree T'' with $\sigma_D(T') > \sigma_D(T'')$. The case that T'_2 does not represent $D_{|B'}$ is completely analogous.

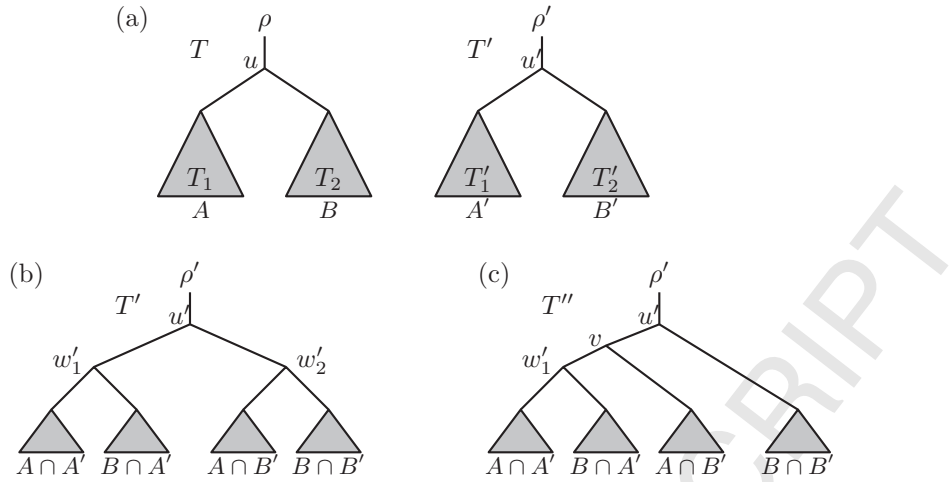


Figure 3: (a) The two trees T and T' considered in the proof of Lemma 3. (b) The detailed structure of T' in one of the cases considered in the proof. (c) The tree T'' resulting from a suitable rNNI applied to T'.

It remains to consider the case that T'_1 and T'_2 represent $D_{|A'}$ and $D_{|B'}$, respectively. Then, $A = A'$ and $B = B'$ immediately implies $T_1 = T'_1$ and $T_2 = T'_2$ and, thus, $T = T'$. Otherwise, there exists at least one $\{x, y\} \in \binom{X}{2}$ with $\{x, y\} \subseteq A$ or $\{x, y\} \subseteq B$ but $x \in A'$ and $y \in B'$. Thus, swapping the roles of either A and B or A' and B', we can assume without loss of generality that the sets $A \cap A'$, $A \cap B'$ and $B \cap B'$ are non-empty. Moreover, the structure of the tree T' must be as depicted in Figure 3(b) in view of the fact that T'_1 and T'_2 represent $D_{|A'}$ and $D_{|B'}$, respectively, and, therefore, the structure of T'_1 and T'_2 must be isomorphic to the smallest subtree of T containing the vertices in $A' \cup \{\rho\}$ and $B' \cup \{\rho\}$ (after suppresing all vertices of degree 2), respectively. Put $\omega' = \omega_{(D,T')}$.

First assume that $B \cap A' \neq \emptyset$. Put $n_1 = |A \cap A'|$, $n_2 = |B \cap A'|$, $n_3 = |A \cap B'|$ and $n_4 = |B \cap B'|$. Without loss of generality we assume $n_1 + n_2 > n_4$. We perform an rNNI pruning and regrafting the subtree with leaf set $A \cap B'$ to obtain the tree T'' depicted in Figure 3(c). Put $\omega'' = \omega_{(D,T'')}$. To show $\sigma_D(T') > \sigma_D(T'')$ it suffices to show $h_{(T',\omega')}(u') + h_{(T',\omega')}(w'_2) > h_{(T'',\omega'')}(u') + h_{(T'',\omega'')}(v)$. To establish the latter inequality, recall that T represents D and, therefore, we can assume that D has been scaled so that $D(a, b) = 1$ for all $a \in A$, $b \in B$. We put

$$\delta_A = \sum_{\substack{a' \in (A \cap A') \\ b' \in (A \cap B')}} D(a', b') \quad \text{and} \quad \delta_B = \sum_{\substack{a' \in (B \cap A') \\ b' \in (B \cap B')}} D(a', b').$$

Note that all distances that contribute to δ_A and δ_B are strictly less than 1, implying

$\delta_A < n_1n_3$ and $\delta_B < n_2n_4$. Using this notation, we obtain

$$\begin{aligned}
&h_{(T',\omega')}(u') + h_{(T',\omega')}(w_2') - h_{(T'',\omega'')}(u') - h_{(T'',\omega'')}(v)\\
&= \frac{1}{2} + \frac{n_1n_4 + n_2n_3}{2(n_1+n_2)(n_3+n_4)} - \frac{n_2}{2(n_1+n_2)} - \frac{n_1+n_3}{2(n_1+n_2+n_3)}\\
&\quad + \frac{1}{2}\left[\frac{1}{(n_1+n_2)(n_3+n_4)} - \frac{1}{(n_1+n_2)n_3}\right]\cdot\delta_A\\
&\quad + \frac{1}{2}\left[\frac{1}{(n_1+n_2)(n_3+n_4)} - \frac{1}{(n_1+n_2+n_3)n_4}\right]\cdot\delta_B\\
&= g(\delta_A,\delta_B).
\end{aligned}$$

Note that $g(\delta_A, \delta_B)$ is a linear function in δ_A and δ_B and that the coefficient of δ_A is negative. Moreover, the assumption $n_1 + n_2 > n_4$ implies that the coefficient of δ_B is negative too. Thus, using the fact that $\delta_A < n_1n_3$ and $\delta_B < n_2n_4$, we have

$$g(\delta_A,\delta_B) > g(n_1n_3, n_2n_4) = 0,$$

from which $\sigma_D(T') > \sigma_D(T'')$ follows, as required.

It remains to consider the case that $B \cap A' = \emptyset$, that is, $n_2 = 0$. We apply the same rNNI to T' as in the previous case and, using similar calculations, we obtain

$$\sigma_D(T') - \sigma_D(T'') = \frac{n_4}{2(n_3+n_4)} + \frac{1}{2}\left[\frac{1}{n_1(n_3+n_4)} - \frac{1}{n_1n_3}\right]\cdot\delta_A > 0,$$

using again $\delta_A < n_1n_3$. ■

In the following main result of this section we note that even for the non-generic case a weak form of consistency holds.

Theorem 4. *Let $T = (V, E, \rho) \in \mathbf{BR}_X$ and ω an interior-positive normalized equidistant edge weighting for T. Put $D = D_{(T,\omega)}$. Then we have*

$$\sigma_D(T) = \min\{\sigma_D(T') : T' \in \mathbf{BR}_X\}.$$

If D is generic, then T is the unique tree in $\mathbf{BR}_X$ minimizing the NEME score for D and a local search using rooted nearest neighbor interchanges starting from any tree in $\mathbf{BR}_X$ will arrive at T after a finite number of steps.

Proof. For generic D, the theorem is an immediate consequence of Lemma 3. So, assume that D is not generic and, for a contradiction, that there exists some $T' \in \mathbf{BR}_X$ with $\sigma_D(T') < \sigma_D(T)$. For any real number $\varepsilon > 0$, define the NEEW ω_ε of T by putting $\omega_\varepsilon(e) = \omega(e) + \varepsilon$ for all edges e of T not adjacent to a vertex in $X \cup \{\rho\}$ and $\omega_\varepsilon(e) = \omega(e)$ for all other edges e of T. Put $D_\varepsilon = D_{(T,\omega_\varepsilon)}$ and note that, by construction, D_ε is a generic ultrametric that is represented by T. As a consequence, $\sigma_{D_\varepsilon}(T) < \sigma_{D_\varepsilon}(T')$ must hold. But this contradicts $\sigma_D(T') < \sigma_D(T)$ in view of the fact that, as ε tends to 0, $\sigma_{D_\varepsilon}(T)$ tends to $\sigma_D(T)$ while $\sigma_{D_\varepsilon}(T')$ tends to $\sigma_D(T')$. ■

The result in Theorem 4 immediately raises the question whether the NEME score for any input distance matrix is minimized by some binary rooted phylogenetic tree. It is known [28] that for unrooted phylogenetic trees the balanced minimum evolution score is indeed always minimized for some unrooted binary phylogenetic tree. We end this section establishing that, in contrast, the answer to the above question for the NEME score is no. Therefore, requiring that the rooted phylogenetic tree generated by an algorithm is binary (such as, for example, in the UPGMA algorithm) is, in general, a non-redundant constraint in the NEME problem. The next lemma gives a family of examples that illustrate this.

Lemma 5. *There exist dissimilarities D with*

$$\min\{\sigma_D(T) : T \in \mathbf{R}_X\} < \min\{\sigma_D(T) : T \in \mathbf{BR}_X\}.$$

Proof. Consider a binary rooted phylogenetic tree T whose structure is as depicted in Figure 4(a). It consists of two rooted binary subtrees T_1 and T_2, each having $m \geq 3$ leaves. In addition, there is a single leaf adjacent to vertex u. Let ω be an interior positive normalized equidistant edge weighting for T such that $h_{(T,\omega)}(v) = 1$ and $h_{(T,\omega)}(u) = s$ for some $s > 1$. Put $D = D_{(T,\omega)}$.

Next, consider the non-binary rooted phylogenetic tree T' depicted in Figure 4(b). It is constructed from T by contracting the edge between u and v into the vertex w. Put $\omega' = \omega_{(D,T')}$. To show that $\sigma_D(T) > \sigma_D(T')$, it suffices, by Equation (1), to show that

$$h_{(T,\omega)}(u) + h_{(T,\omega)}(v) = s + 1 > 2h_{(T',\omega')}(w) = \frac{2m^2 + 4ms}{m^2 + 2m},$$

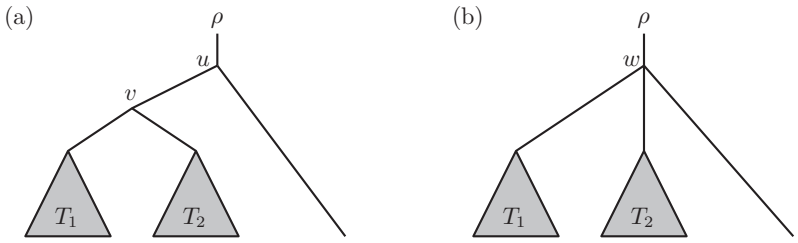


Figure 4: The construction referred to in the proof of Lemma 5.

which can easily be checked to be the case, in view of $m \geq 3$, for any $s > 1$. Hence, by Theorem 4, we have $\min\{\sigma_D(T') : T' \in \mathbf{R}_X\} < \sigma_D(T) = \min\{\sigma_D(T') : T' \in \mathbf{BR}_X\}$. ■

5. The NEME problem is NP-hard

To establish NP-hardness of the NEME problem, we use a reduction from the well-known NP-hard graph coloring problem (see e.g. [29]). More specifically, we consider the following variant of this problem:

INPUT: A graph $G = (V, E)$ with $|V| = 4n$.
QUESTION: Can V be partitioned into 4 subsets V_1, V_2, V_3, V_4 with $|V_1| = |V_2| = |V_3| = |V_4| = n$ such that no edge $e \in E$ has both endpoints in the same set V_i for some $i \in \{1, 2, 3, 4\}$? We call any such partition a *4-coloring* of G.

Note that the additional constraint that the sets in a 4-coloring are all of the same size is merely added to simplify the description of the reduction. It preserves the NP-hardness of the graph coloring problem in view of the fact that adding isolated vertices to any graph G does not change the minimum number of colors that suffice to color G. We first present a technical lemma that will be used in the construction below.

Lemma 6. *Let X be a set with $2(m+k)$ elements, $m \geq 1$, $k \geq 1$, that is partitioned into the sets A, B and C with $|A| = |B| = m$ and $|C| = 2k$. In addition, let $s > 0$ be a real number and D a dissimilarity on X with $D(x, y) = s$ for all $x \in A$, $y \in B$ and $D(x, y) \leq \frac{s}{3(m+k)^5}$ for all other $x, y \in X$. Then any binary rooted phylogenetic tree $T = (V, E, \rho)$ on X with $\sigma_D(T) = \min\{\sigma_D(T) : T \in \mathbf{BR}_X\}$ must contain two distinct vertices $v, w \in V$ with (i) $C(v) \cap C(w) = \emptyset$, (ii) $|C(v)| = |C(w)| = (m + k)$, (iii) $A \subseteq C(v)$ and $B \subseteq C(w)$.*

Proof. First consider a binary rooted phylogenetic tree T on X that contains vertices v and w with properties (i)-(iii). Note that this implies that v and w have the same parent u and that u is the single child of the root ρ. Moreover, for all $a \in A$, $b \in B$, we have $\alpha^T_{\{a,b\}} = \frac{1}{2(m+k)^2}$ and, by Lemma 1(ii), this is the smallest possible value for a rooted phylogenetic tree with $2(m+k)$ leaves.

Next consider any binary rooted phylogenetic tree $T' = (V', E', \rho')$ on X that does not contain two vertices v and w satisfying properties (i)-(iii). Let u' denote the single child of ρ' and consider the two children v' and w' of u'. In particular, v' and w' must violate at least one of the properties (i)-(iii). By construction we have $C(v') \cap C(w') = \emptyset$. Hence, one of the properties (ii) or (iii) must be violated.

First consider the case that (iii) is violated. This implies, without loss of generality, that there exist $a \in A$ and $b \in B$ with $\{a, b\} \subseteq C(v')$. In view of $|C(v')| \leq 2(m+k)-1$ and Lemma 1(ii) we have

$$\alpha^{T'}_{\{a,b\}} \geq \frac{2}{(2(m+k)-1)^2} = \frac{1}{2(m+k)^2 - 2(m+k) + 1/2}$$

Next consider the case that property (iii) is satisfied but (ii) is violated for v' and w'. Then, for any $a \in A$ and any $b \in B$, we have

$$\alpha^{T'}_{\{a,b\}} \geq \frac{1}{2(m+k+1)(m+k-1)} = \frac{1}{2(m+k)^2 - 2}$$

Noting that $2(m+k)^2 - 2 \geq 2(m+k)^2 - 2(m+k) + 1/2$, we calculate a lower bound on the difference between the coefficients in T and T':

$$\frac{1}{2(m+k)^2 - 2} - \frac{1}{2(m+k)^2} \geq \frac{1}{2(m+k)^4}.$$

This implies, using Lemma 1(i) to obtain the upper bound in the second line below:

$$\begin{aligned}
\sigma_D(T) &= \sum_{\{x,y\} \in \binom{X}{2}} \alpha^T_{\{x,y\}} D(x,y) \\
&\leq \left[\sum_{a \in A,\, b \in B} \alpha^T_{\{a,b\}} D(a,b) \right] + \frac{s(m+k)}{3(m+k)^5} \\
&< \left[\sum_{a \in A,\, b \in B} \alpha^T_{\{a,b\}} s \right] + \frac{s}{2(m+k)^4} \\
&\leq \sum_{a \in A,\, b \in B} \alpha^{T'}_{\{a,b\}} s \leq \sum_{\{x,y\} \in \binom{X}{2}} \alpha^{T'}_{\{x,y\}} D(x,y) = \sigma_D(T')
\end{aligned}$$

Hence, T' cannot be an optimal tree for D. ■

Next we describe how to construct for a given graph $G = (V(G), E(G))$ with $|V(G)| = 4n$ a suitable dissimilarity $D = D(G)$. First construct a set X that is the disjoint union of $V(G)$, Y and W with $|Y| = 2^k$ and $|W| = (2^k - 4)n$ where $k = \lfloor \log_2(n+1) + 4 \rfloor$. Note that

$$|X| = 2^k(n+1) \leq 2^{\log_2(n+1)+4}(n+1) = 16(n+1)^2.$$

Put $m_i = 2^{k-i}(n+1)$, $i \in \{0, 1, \ldots, k\}$. In addition, put $s_1 = 1$ and, for $i \in \{1, 2, \ldots, k\}$, $s_{i+1} = \frac{s_i}{3m_i^5}$. Moreover, put $s^* = \frac{s_{k+1}}{2^{k+2}(n+1)^3}$. The values $s_1 > s_2 > \cdots > s_{k+1} > s^* > 0$ will be the possible distances between elements in X.

Now, recursively partition the set Y so as to force a fully balanced binary tree as a backbone structure. More precisely, put $Y_{0,0} = Y$ and define, for all $i \in \{0, 1, \ldots, k-1\}$ and all $j \in \{0, 1, \ldots, 2^i - 1\}$, sets $Y_{i+1,2j}$ and $Y_{i+1,2j+1}$ so that $Y_{i+1,2j} \cap Y_{i+1,2j+1} = \emptyset$, $Y_{i+1,2j} \cup Y_{i+1,2j+1} = Y_{i,j}$ and $|Y_{i+1,2j}| = |Y_{i+1,2j+1}|$ hold. Select an element $y_l^* \in Y_{2,l}$ for each $l \in \{0, 1, 2, 3\}$.

Next we construct the dissimilarity $D = D(G)$ on X:

(a) For all $w \in W$ and all $x \in X$ we put $D(w, x) = 0$. The elements in W are just used to fill subtrees so that we get a fully balanced backbone tree.

(b) For all $y, y' \in Y$, $y \neq y'$, we put $D(y, y') = s_{i+1}$ where i is the largest index in $\{0, 1, \ldots, k-1\}$ with $\{y, y'\} \subseteq Y_{i,j}$ for some $j \in \{0, 1, \ldots, 2^i - 1\}$. The distances between the elements in Y force a fully balanced backbone tree by Lemma 6 (cf. Figure 5).

(c) For all $v \in V(G)$ and all $y \in (Y - \{y_1^*, y_2^*, y_3^*, y_4^*\})$ we put $D(v, y) = s_{k+1}$. And for all $v \in V(G)$ and all $y \in \{y_1^*, y_2^*, y_3^*, y_4^*\}$ we put $D(v, y) = 0$. This will force that a subset of n vertices of G is grouped together with each y_l^*, $1 \leq l \leq 4$, in the same subtree.

(d) For all $v, v' \in V(G)$, $v \neq v'$, we put $D(v, v') = s^*$ if $\{v, v'\} \in E(G)$ and $D(v, v') = 0$ otherwise. These distances capture the structure of G and are so small that they do not interfere with forming the fully balanced backbone tree.

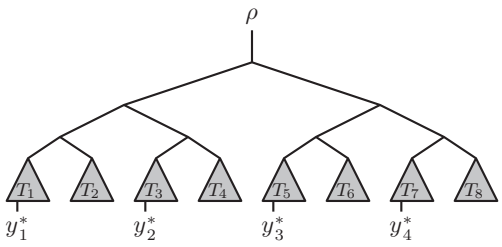


Figure 5: The structure of the backbone tree for $k = 3$.

Lemma 7. *Let G be a graph with $4n$ vertices and let $D = D(G)$ be the dissimilarity on X constructed above. Then G has a 4-coloring if and only if there exists a binary rooted phylogenetic tree T on X with $\sigma_D(T)$ not larger than*

$$\left[2^{k-2}\sum_{i=1}^{k}\frac{s_i}{m_i^2}\right] + s_{k+1}\cdot 4n\cdot\left[\left(\sum_{i=0}^{k-3}\frac{2^i}{2m_{k-i}^2}\right)+\frac{2^{k-2}-1}{2m_2^2}+\frac{2^{k-1}-2}{2m_1^2}\right]+\frac{s^*}{2^{2k-7}}$$

Proof. First note that, by the construction of the distances $s_1, s_2, \ldots, s_{k+1}, s^*$ and Lemma 6, the upper part of any optimal binary rooted phylogenetic tree T on X must be a fully balanced binary tree. This upper part has $k+1$ levels. Level $i \in \{0, 1, \ldots, k\}$ consists of 2^i subtrees, each of which containing precisely one of the sets $Y_{i,j}$, $j \in \{0, 1, \ldots, 2^i - 1\}$ in its set of leaves. In particular, the lowest level consists of 2^k subtrees $T_1, T_2, \ldots, T_{2^k}$ and the leaf set of each of these subtrees consists of precisely one element from Y and n elements from $V(G) \cup W$. Thus, in view of $y_l^* \in Y_{2,l}$, $l \in \{0, 1, 2, 3\}$, we can choose the numbering of the subtrees so that $T_{l2^{k-2}+1}$ is the subtree that contains leaf y_l^* (cf. Figure 5).

Now consider any vertex $v \in V(G)$. First assume that v is a leaf in one of the subtrees $T_{l2^{k-2}+1}$, $l \in \{0, 1, 2, 3\}$. Then, because the backbone tree is fully balanced, the contribution to $\sigma_D(T)$ of the distances from v to the elements in Y is

$$s_{k+1}\cdot\left[\left(\sum_{i=0}^{k-3}\frac{2^i}{2m_{k-i}^2}\right)+\frac{2^{k-2}-1}{2m_2^2}+\frac{2^{k-1}-2}{2m_1^2}\right].$$

Next assume that v is a leaf in some subtree T_q, $q \neq l2^{k-2}+1$ for all $l \in \{0, 1, 2, 3\}$. Then one summand of the form $\frac{s_{k+1}}{2m_{k-i}^2}$, $i \in \{0, 1, \ldots, k-3\}$, is replaced by a summand that contributes, by Equation (3), at least $\frac{1}{2(m_k/2)^2}\cdot s_{k+1} = \frac{2s_{k+1}}{m_k^2} = \frac{2s_{k+1}}{(n+1)^2}$. Thus, the increase in the contribution is at least

$$\left[\frac{2}{m_k^2}-\frac{1}{2m_k^2}\right]\cdot s_{k+1} = \frac{3s_{k+1}}{2(n+1)^2} > 2^{k-1}(n+1)s^* \geq \sum_{\{x,y\}\in\binom{X}{2}}\alpha_{\{x,y\}}^T s^*,$$

that is, using Lemma 1(i) to obtain the last inequality, it is strictly larger than the total contribution of all distances that equal s^*. Hence, the contribution of the distances s_{k+1} to the score is minimized if and only if each vertex $v \in V(G)$ is a leaf in one of the subtrees T_l, $1 \leq l \leq 4$, inducing a partition of $V(G)$ into 4 subsets V_1, V_2, V_3, V_4 each of size n.

Summarizing the contribution of the distances $s_1, s_2, \ldots, s_{k+1}$ to the score of the tree, we obtain:

$$\lambda(n,k) = \left[2^{k-2}\sum_{i=1}^{k}\frac{s_i}{m_i^2}\right] + s_{k+1}\cdot 4n\cdot\left[\left(\sum_{i=0}^{k-3}\frac{2^i}{2m_{k-i}^2}\right) + \frac{2^{k-2}-1}{2m_2^2} + \frac{2^{k-1}-2}{2m_1^2}\right]$$

Note that this contribution does *not* depend on the structure of the graph G.

It remains to calculate the contribution of the distances that equal s^* to the score of the tree. Note that $16(n+1)^2$ is a trivial upper bound on the number of edges in G. Thus, if the partition V_1, V_2, V_3, V_4 induced by the tree is a 4-coloring of G, then the total contribution to the score of the tree is at most

$$\frac{16(n+1)^2 s^*}{2m_2^2} = \frac{16s^*}{2^{2k-3}} = \frac{s^*}{2^{2k-7}}.$$

In contrast, a single edge with both endpoints in one of the sets V_l, $1 \leq l \leq 4$, contributes, by Equation (3), at least $\frac{1}{2((n+1)/2)^2}s^* = \frac{2s^*}{(n+1)^2}$. Hence, noting that $k = \lfloor \log_2(n+1) + 4 \rfloor$ implies $\frac{2}{(n+1)^2} > \frac{1}{2^{2k-7}}$, we obtain that G has a 4-coloring if and only if there exists a binary rooted phylogenetic tree T on X with

$$\sigma_D(T) \leq \lambda(n,k) + \frac{s^*}{2^{2k-7}}.$$

■

Note that the dissimilarity $D = D(G)$ constructed above need not satisfy the triangle inequality. However, putting $D'(x,x') = D(x,x') + 1$ for all $x, x' \in X$, $x \neq x'$, and $D'(x,x) = 0$ for all $x \in X$, we obtain a dissimilarity D' on X that satisfies the triangle inequality. Moreover, by Lemma 1(i), for every binary rooted phylogenetic tree T on X, we have $\sigma_{D'}(T) = \sigma_D(T) + \frac{1}{2}(|X|-1)$, that is, a tree T is optimal for D if and only if T is optimal for D'. Thus we have the main result of this section:

Theorem 8. *Computing a binary rooted phylogenetic tree with minimum NEME score for a dissimilarity D' on a set X is NP-hard even if D' satisfies the triangle inequality and D' takes on only $O(\log_2(|X|))$ different values.*

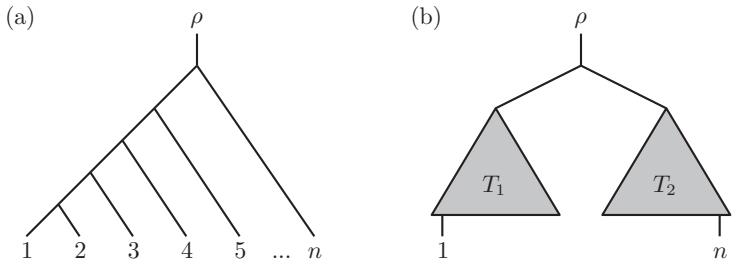


Figure 6: Two rooted phylogenetic trees used in the proof of Lemma 9.

6. Approximating the minimum NEME score

Note that Lemma 1(iii) states that *any* tree in $\mathbf{R}_X$ approximates the minimum NEME score over all trees in $\mathbf{R}_X$ up to a factor that is in $O(n^2)$, $n = |X|$. It is not hard to see that this bound is asymptotically tight and in this section we explore ways to obtain better approximation guarantees.

We first look at the approximation guarantees that can be achieved with existing algorithms. We start with UPGMA and establish a lower bound of $\Omega(n)$ on the approximation guarantee achieved by it.

Lemma 9. *For every non-empty finite set X with $n \geq 3$ elements there exists a dissimilarity D on X such that*

$$\sigma_D(T) \geq \frac{n}{4} \min\{\sigma_D(T') : T' \in \mathbf{R}_X\}$$

holds for the tree T produced by UPGMA.

Proof. Let $X = \{1, 2, \ldots, n\}$ and consider the ultrametric D' on X defined by putting, for all $1 \leq i < j \leq n$, $D'(i,j) = D'(j,i) = 1 + \frac{(j-2)}{n-2}$. Note that the unique rooted phylogenetic tree T on X with $D_{(T,\omega)} = D'$ for some interior positive normalized equidistant edge weighting ω is the rooted caterpillar depicted in Figure 6(a). This is also the tree constructed by UPGMA on input D'.

From D' we construct the dissimilarity D by putting $D(x,y) = D'(x,y)$, for all $\{x,y\} \in \left(\binom{X}{2} - \{\{1,n\}\}\right)$, and $D(1,n) = D(n,1) = s$ for some constant $s > 2$. Note that UPGMA will still generate the tree T on input D and that $\sigma_D(T) \geq \frac{s}{2(n-1)}$ holds.

Now consider any binary rooted phylogenetic tree T' on X whose structure is as depicted in Figure 6(b). More specifically, the rooted subtree T_1 has $\lfloor \frac{n}{2} \rfloor$ leaves, one of them being

1, and T_2 has $\lceil \frac{n}{2} \rceil$ leaves, one of them being n. Using Lemma 1(i) and the fact that $D'(x,y) \leq 2$ for all $\{x,y\} \in \binom{X}{2}$, we have

$$\sigma_D(T') \leq \alpha^{T'}_{\{1,n\}} \cdot D(1,n) + \sum_{\{x,y\} \in \binom{X}{2}} \alpha^{T'}_{\{x,y\}} \cdot D'(x,y) \leq \frac{2s}{n^2-1} + (n-1),$$

implying that $\frac{n}{4}\sigma_D(T') \leq \frac{s}{2(n-1)} \leq \sigma_D(T)$ for $s \geq \frac{1}{2}n(n+1)(n-1)^2$. But this implies $\sigma_D(T) \geq \frac{n}{4} \min\{\sigma_D(T') : T' \in \mathbf{R}_X\}$, as required. ■

Next we briefly touch upon another potential approach from the literature for approximating the minimum NEME score. To describe this approach, note that the NEME problem is related to the problem of finding a *sparsest cut*, that is, given a dissimilarity D on a finite set X compute a *split* $A|B$ of X, that is, a bipartition of X into two non-empty subsets A and B, such that

$$\frac{1}{|A| \cdot |B|} \sum_{a \in A,\, b \in B} D(a,b)$$

is minimum. This problem is usually stated in terms of edge weighted graphs and known to be NP-hard. Recent work on this problem mainly concentrated in finding good approximations of a sparsest cut (see e.g. [30]).

Interestingly, a greedy top-down analogy of UPGMA based on recursively splitting X by sparsest cuts has been proposed for detecting hierarchical community structure of social networks [31]. Using again the dissimilarity D described in the proof of Lemma 2, it can be seen, however, that this approach can yield trees whose NEME score is worse by a quadratic factor in $|X|$ than the minimum NEME score.

In view of the fact that the approaches we explored so far have not led to good approximation guarantees, we apply in the following a generic approach from the literature to establish a polylogarithmic approximation guarantee for dissimilarities that satisfy the triangle inequality, that is, *metrics*. This approach relies on two ingredients:

(i) The existence of a polynomial time algorithm with polylogarithmic approximation guarantee for treelike metrics.

(ii) The fact [32] that there exists a polynomial time algorithm that computes, for any metric D on a set X with n elements, a collection $D_1, D_2, \ldots, D_k$ of treelike metrics on X along with positive coefficients $\beta_1, \beta_2, \ldots, \beta_k$, $\beta_1 + \beta_2 + \cdots + \beta_k = 1$, such that

(1) $D(x,y) \leq D_i(x,y)$ for all $i \in \{1,2,\ldots,k\}$ and all $x,y \in X$, and

(2) there exists a constant $c > 0$ such that

$$\sum_{i=1}^{k} \beta_i D_i(x,y) \leq c \cdot \log_2(n) \cdot D(x,y)$$

for all $x,y \in X$.

We shall first establish (i). To this end, we rely on the following fact that, phrased in various guises, seems to be mathematical folklore. For the convenience of the reader we provide a short proof and phrase it in terms of splits in *unrooted* binary phylogenetic trees, that is, trees obtained from rooted binary phylogenetic trees by removing the root and the edge incident with it, and then suppressing the resulting degree two vertex. Recall that each edge e in an unrooted binary phylogenetic tree T on X induces a split $S_e = A_e|B_e$ of X in which A_e and B_e are the leaf sets of the two connected components resulting from removing e from T.

Lemma 10. *In every unrooted binary phylogenetic tree T on X with $|X| \geq 2$ there exists an edge e such that the split $A_e|B_e$ of X satisfies*

$$\frac{1}{3}n \leq \min\{|A_e|,|B_e|\} \leq \max\{|A_e|,|B_e|\} \leq \frac{2}{3}n. \tag{7}$$

Proof. Assume that such an edge does not exist. Replace all edges e of T by a directed edge in such a way that this directed edge points to the larger of the two sets A_e and B_e. Then every directed edge incident with a leaf of T is directed away from that leaf and, in view of the fact that all other vertices of T have degree three, one of those vertices must be such that all three directed edges incident with this vertex v are directed towards it. But this implies that, while all edges e incident with v must clearly satisfy $\min\{|A_e|,|B_e|\} \leq \frac{n}{2}$, by the pigeonhole principle, at least one of these edges must also satisfy $\min\{|A_e|,|B_e|\} \geq \frac{n}{3}$, contradicting our assumption. ■

Note that Lemma 10 does not hold for non-binary trees. The next lemma establishes (i). Recall that a *treelike metric* on X is a metric for which there exists an unrooted phylogenetic tree T' on X with a non-negative edge-weighting ω' with $D = D_{(T',\omega')}$.

Lemma 11. *Let D be a treelike metric on a set X with n elements. Then a binary rooted phylogenetic tree T on X with*

$$\sigma_D(T) \leq c \cdot \log_2(n) \cdot \min\{\sigma_D(T') : T' \in \mathbf{BR}_X\}$$

for some positive constant c can be computed in time $O(n^2)$.

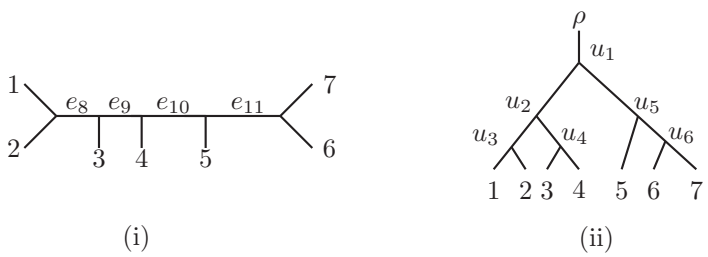
Proof. Let $T' = (V', E')$ be a binary unrooted phylogenetic tree on X and ω' a non-negative edge weighting of T' with $D = D_{(T',\omega')}$. For every edge $e \in E'$ we denote by D_e the metric that assigns 1 to a pair $(x, y) \in X \times X$ if the path from x to y in T contains edge e. Otherwise D_e assigns the value 0 to (x, y). Then $D = \sum_{e \in E'} \omega'(e) \cdot D_e(x, y)$ and hence by Equation (5) we have, for any rooted phylogenetic tree T on X,

$$\sigma_D(T) = \sum_{e \in E'} \omega'(e)\sigma_{D_e}(T) \geq \sum_{e \in E'} \frac{\omega'(e)}{2},$$

where the last inequality follows from the fact that, for all $e \in E'$, D_e is an ultrametric and, therefore, $\min\{\sigma_{D_e}(T'') : T'' \in \mathbf{BR}_X\} = \frac{1}{2}$ by Theorem 4.

Hence, it suffices to show how to construct in polynomial time a rooted phylogenetic tree T on X with $\sigma_D(T) \leq c \cdot \log_2(n) \cdot \sum_{e \in E'} \frac{\omega'(e)}{2}$ for some constant $c > 0$. This is done recursively as follows. Using Lemma 10, we find an edge $e \in E'$ such that the split $S_e = A_e|B_e$ of X induced by e satisfies (7). We require that the resulting rooted phylogenetic tree T on X will have the clusters A_e and B_e. Then we remove e from T'. This yields, after suppressing the two vertices of degree 2, two unrooted phylogenetic trees T'_{A_e} on A_e and T'_{B_e} on B_e which represent the restriction of D to A_e and B_e, respectively. If $|A_e| > 1$ ($|B_e| > 1$) we construct a binary rooted phylogenetic tree T_{A_e} on A_e (T_{B_e} on B_e) recursively. Otherwise T_{A_e} (T_{B_e}) is the unique rooted phylogenetic tree on A_e (B_e). Then we glue the roots of T_{A_e} and T_{B_e} together and add a new root ρ to obtain a binary rooted phylogenetic tree $T = (V, E, \rho)$ on X. Note that this construction can clearly be done in time $O(n^2)$ and, since we choose the edges for recursively partitioning T' in such a way that Inequalities (7) hold, it follows that every path in T contains $O(\log_2(n))$ vertices.

Note that the construction of T induces the canonical map $\varphi : E' \to (V - (X \cup \{\rho\}))$ that assigns to $e \in E'$ the internal vertex v of T that was constructed as the root in a recursive step when edge e was removed from T' (see Fig. 7 for an example). Also note that φ is, by construction, surjective. It is, however, not injective because suppressing



(i) (ii)

Figure 7: An illustration of the canonical map φ used in the proof of Lemma 11: (i) An unrooted binary phylogenetic tree T' on $X = \{1, 2, \cdots, 7\}$ (for simplicity, the label e_i of the edge incident with leaf i is omitted). (ii) A binary rooted phylogenetic tree on X constructed from T' by consecutively removing the following sets of edges $\{e_{10}\}, \{e_8\}, \{e_5, e_{11}\}, \{e_3, e_4, e_9\}, \{e_6, e_7\}, \{e_1, e_2\}$. Then the canonical map φ is as follows: $\varphi(e_{10}) = u_1, \varphi(e_8) = u_2, \varphi(e_5) = \varphi(e_{11}) = u_5, \varphi(e_3) = \varphi(e_4) = \varphi(e_9) = u_4, \varphi(e_6) = \varphi(e_7) = u_6$, and $\varphi(e_1) = \varphi(e_2) = u_3$.

degree 2 vertices leads to clusters of original edges in T' that are removed together at a single recursive step.

Next, for each edge $e \in E'$ let $\omega_e = \omega_{(D_e,T)}$ and let Γ_e be the set of those vertices $v \in (V - (X \cup \{\rho\}))$ for which there exist $x, y \in X$ such that (i) $v = lca(x, y)$ and (ii) e lies on the path from x to y in T'. By construction of T, the set Γ_e consists precisely of those vertices in $V - (X \cup \{\rho\})$ that lie on the path from $\varphi(e)$ to ρ in T, implying that $|\Gamma_e| \in O(\log_2(n))$. In addition, Equation (2) implies that D_e contributes at most $1/2$ to $h_{(T,\omega_e)}(v)$ for all $v \in \Gamma_e$ and, for all $v \in V - \Gamma_e$ we have $h_{(T,\omega_e)}(v) = 0$. Therefore, using Equation (1) to obtain the second equality below, we have

$$\begin{aligned}
\sigma_D(T) &= \sum_{e \in E'} \omega'(e)\sigma_{D_e}(T) \\
&= \sum_{e \in E'} \omega'(e)\Big(\sum_{v \in (V-(X \cup \{\rho\}))} h_{(T,\omega_e)}(v)\Big) \\
&= \sum_{e \in E'} \omega'(e)\Big(\sum_{v \in \Gamma_e} h_{(T,\omega_e)}(v)\Big) \\
&\leq c \cdot \log_2(n) \cdot \sum_{e \in E'} \frac{\omega'(e)}{2} \qquad (8)
\end{aligned}$$

for some constant $c > 0$, as required. ■

Note that there are treelike metrics D on X for which a binary rooted phylogenetic tree with minimum NEME score cannot be obtained by rooting the unrooted tree representing D somewhere. That means that the structure of the unrooted tree representing D need not

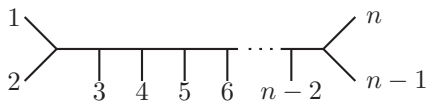


Figure 8: An unrooted caterpillar tree on $X = \{1, 2, \ldots, n\}$. The metric D induced by this tree on X (all edges have weight 1) is such that a rooted phylogenetic tree on X with minimum NEME score for D cannot by obtained by rooting the caterpillar somewhere.

reflect much the structure of the rooted trees with minimum NEME score for D. Consider, for example, the metric D on $X = \{1, 2, \ldots, n\}$, $n \geq 6$, induced by the unrooted caterpillar in Figure 8. All edges are assigned weight 1. Then the recursive algorithm in the proof of Lemma 11 yields a binary rooted phylogenetic tree T on X with $\sigma_D(T) \in O(n \log_2 n)$ by the upper bound in (8). In contrast, for any binary rooted phylogenetic tree T' obtained by rooting the caterpillar, there must exist, for all $k \in \{3, 4, \ldots, \lfloor \frac{n}{2} \rfloor\}$, at least one vertex v in T' with $|ch(v)| = k$ and $h_{(T',\omega)}(v) \geq \frac{1}{2(k-1)}(k + (k-1) + \cdots + 4 + 3) \geq \frac{k}{4}$, where $\omega = \omega_{(D,T')}$. This implies that $\sigma_D(T') \geq \sum_{k=3}^{\lfloor \frac{n}{2} \rfloor} \frac{k}{4} \in \Omega(n^2)$. The next theorem summarizes our results on approximating the NEME score for metrics.

Theorem 12. *Let D be a metric on a finite set X with n elements. Then a rooted binary phylogenetic tree T on X with*

$$\sigma_D(T) \leq c \cdot \log^2(n) \cdot \min\{\sigma_D(T') : T' \in \mathbf{BR}_X\}$$

for some positive constant c can be computed in polynomial time.

Proof. Let $D_1, D_2, \ldots, D_k$ be a collection of treelike metrics for D together with coefficients $\beta_1, \beta_2, \ldots, \beta_k$ as described in (ii) above. In addition, let T^* be a binary rooted phylogenetic tree on X with minimum NEME score for D and, similarly, T_i^* be a binary rooted phylogenetic tree on X with minimum NEME score for D_i, $1 \leq i \leq k$. Assume that $\sigma_{D_1}(T_1^*) \leq \sigma_{D_i}(T_i^*)$ for all $i \in \{1, 2, \ldots, k\}$. Finally, let T_1^o be the binary rooted phylogenetic tree on X constructed for D_1 using the recursive algorithm in the proof of Lemma 11.

Then we have, using repeatedly the linearity from Equation (5):

$$\begin{aligned}
\sigma_D(T_1^o) &\leq \sigma_{D_1}(T_1^o) \\
&\leq c' \cdot \log_2(n) \cdot \sigma_{D_1}(T_1^*) \\
&\leq c' \cdot \log_2(n) \cdot \sum_{i=1}^{k} \beta_i \cdot \sigma_{D_i}(T_i^*) \\
&\leq c' \cdot \log_2(n) \cdot \sum_{i=1}^{k} \beta_i \cdot \sigma_{D_i}(T^*) \\
&= c' \cdot \log_2(n) \cdot \sigma_{(\sum_{i=1}^{k} \beta_i \cdot D_i)}(T^*) \\
&\leq c'c'' \cdot \log_2^2(n) \cdot \sigma_D(T^*),
\end{aligned}$$

where c' and c'' are positive constants that come from the upper bound (8) and property (2) of the collection $D_1, D_2, \ldots, D_k$, respectively. Hence $\sigma_D(T_1^o) \leq c \cdot \log_2^2(n) \cdot \sigma_D(T^*)$ for some constant $c > 0$ and the tree T_1^o can be constructed in polynomial time. ■

Interestingly, the above approach can be applied to *any* variant of minimum evolution as long as the objective function is a linear combination of the input distances and the variant is consistent (the latter trivially implies ingredient (i) above and, thus, saves a factor of $\log n$ in the approximation guarantee). In particular, the original unrooted ME problem [11] has these properties and can, therefore, be approximated for metrics within a factor of $O(\log_2 n)$. To the best of our knowledge, this is the first non-trivial approximation result for the unrooted ME problem.

7. Concluding remarks and open problems

In this paper, we have highlighted some properties of the UPGMA method. We now conclude by pointing out two possible directions for future work. The first direction concerns improving the approximation guarantee for the NEME problem presented in the last section. Recall that the interpretation of the balanced minimum evolution score of an unrooted tree as an average over spanning cycles has been used (as one ingredient amongst others) in [17] to design a constant-factor polynomial time approximation algorithm for the balanced minimum evolution problem in case the given dissimilarity satisfies the triangle inequality. We expect that the results presented in this paper can similarly serve as the

basis for a better understanding of the approximation properties of the NEME problem. A concrete conjecture we have in this direction is that UPGMA always generates a tree whose NEME score is within a factor in $O(n)$ of the minimum score.

The second direction for future work concerns the so-called safety radius of the NEME approach for computing rooted phylogenetic trees. The safety radius concept was introduced to quantify how much distortion of the input distance matrix a method can tolerate and still return the correct tree (see e.g. [33]). For example, it is known that in the rooted-tree setting UPGMA has a safety radius of 1 [34], and that both neighbor joining and BME-based tree construction have a safety radius of $\frac{1}{2}$ (see [33] and [35], respectively). We conjecture that the safety radius of NEME-based tree construction tends to 0 as the number of leaves of the trees tends to infinity. In this context, it might also be of interest to study the stochastic safety radius of the NEME problem, a concept that was recently introduced [36], and which aims to understand consistency within a probabilistic setting.

References

[1] P. D'haeseleer, How does gene expression clustering work?, Nature biotechnology 23 (12) (2005) 1499–1501.

[2] R. Sokal, C. Michener, A statistical method for evaluating systematic relationships, University of Kansas Science Bulletin 38 (1958) 1409–1438.

[3] Y. Loewenstein, E. Portugaly, M. Fromer, M. Linial, Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space, Bioinformatics 24 (13) (2008) i41–i49.

[4] J. Felsenstein, Inferring phylogenies, Sinauer associates Sunderland, 2004.

[5] P. Legendre, L. Legendre, Numerical ecology, Vol. 24, Elsevier, 2012.

[6] W. Li, L. Fu, B. Niu, S. Wu, J. Wooley, Ultrafast clustering algorithms for metagenomic sequence analysis, Briefings in bioinformatics 13 (6) (2012) 656–668.

[7] C. Romesburg, Cluster analysis for researchers, Lulu. com, 2004.

[8] R. Davidson, S. Sullivant, Polyhedral combinatorics of UPGMA cones, Advances in Applied Mathematics 50 (2) (2013) 327–338.

[9] O. Gascuel, M. Steel, Neighbor-Joining revealed, Molecular Biology and Evolution 23 (2006) 1997–2000.

[10] N. Saitou, M. Nei, The Neighbor-Joining method: a new method for reconstructing phylogenetictrees, Molecular Biology and Evolution 4 (1987) 406–425.

[11] A. Rzhetsky, M. Nei, Theoretical foundation of the minimum-evolution method of phylogenetic inference, Molecular Biology and Evolution 10 (1993) 1073–1095.

[12] R. Agarwala, V. Bafna, M. Farach, M. Paterson, M. Thorup, On the approximability of numerical taxonomy (fitting distances by tree metrics), SIAM Journal on Computing 28 (1998) 1073–1085.

[13] W. Day, Computational complexity of inferring phylogenies from dissimilarity matrices, Bulletin of Mathematical Biology 49 (1987) 461–467.

[14] M. Farach, S. Kannan, T. Warnow, A robust model for finding optimal evolutionary trees, Algorithmica 13 (1995) 155–179.

[15] D. Catanzaro, The minimum evolution problem: overview and classification, Networks 53 (2009) 112–125.

[16] R. Desper, O. Gascuel, Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting, Molecular Biology and Evolution 21 (3) (2004) 587–598.

[17] S. Fiorini, G. Joret, Approximating the balanced minimum evolution problem, Operation Research Letters 40 (2012) 31–35.

[18] S. Bastkowski, V. Moulton, A. Spillner, T. Wu, The minimum evolution problem is hard: a link between tree inference and graph clustering problems, Bioinformatics 32 (4) (2016) 518–522.

[19] C. Semple, M. Steel, Cyclic permutations and evolutionary trees, Advances in Applied Mathematics 32 (2004) 669–680.

[20] C. Fahey, S. Hosten, N. Krieger, L. Timpe, Least squares methods for equidistant tree reconstruction, available online at arXiv:0808.3979 (2008).

[21] F. Pardi, O. Gascuel, Combinatorics of distance-based tree inference, Proceedings of the National Academy of Sciences of the USA 109 (2012) 16443–16448.

[22] M. Bordewich, O. Gascuel, K. Huber, V. Moulton, Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference, IEEE/ACM Transactions on Computational Biology and Bioinformatics 6 (2009) 110–117.

[23] M. Bordewich, R. Mihaescu, Accuracy guarantees for phylogeny reconstruction algorithms based on balanced minimum evolution, IEEE/ACM Transactions on Computational Biology and Bioinformatics 10 (2013) 576–583.

[24] J. Farris, On the cophenetic correlation coefficient, Systematic Biology 18 (1969) 279–285.

[25] W. Vach, Least squares approximation of addititve trees, in: Conceptual and Numerical Analysis of Data, Springer, 1989, pp. 230–238.

[26] S. Willson, Consistent formulas for estimating the total lengths of trees, Discrete Applied Mathematics 148 (2005) 214–239.

[27] C. Semple, M. Steel, Phylogenetics, Oxford University Press, 2003.

[28] K. Eickmeyer, P. Huggins, L. Pachter, R. Yoshida, On the optimality of the Neighbor-Joining algorithm, Algorithms for Molecular Biology 3 (2008) 1–11.

[29] M. Garey, D. Johnson, Computers and intractability: a guide to the theory of NP-completeness, Freeman, 1979.

[30] S. Arora, E. Hazan, S. Kale, $o(\sqrt{\log n})$ approximation to SPARSEST CUT in $\tilde{O}(n^2)$ time, SIAM Journal on Computing 39 (2010) 1748–1771.

[31] C. Mann, D. Matula, E. Olinick, The use of sparsest cuts to reveal the hierarchical community structure of social networks, Social Networks 30 (2008) 223–234.

[32] J. Fakcharoenphol, S. Rao, K. Talwar, A tight bound on approximating arbitrary metrics by tree metrics, Journal of Computer and System Sciences 69 (2004) 485–497.

[33] K. Atteson, The performance of neighbor-joining methods of phylogenetic reconstruction, Algorithmica 25 (2-3) (1999) 251–278.

[34] O. Gascuel, A. McKenzie, Performance analysis of hierarchical clustering algorithms, Journal of Classification 21 (2004) 3–18.

[35] F. Pardi, S. Guillemot, O. Gascuel, Robustness of phylogenetic inference based on minimum evolution, Bulletin of Mathematical Biology 72 (2010) 1820–1839.

[36] O. Gascuel, M. Steel, A 'stochastic safety radius' for distance-based tree reconstruction, Algorithmica (2015) 1–18.