

Performance Analysis of Hierarchical Clustering Algorithms

January 19, 2004

Olivier Gascuel and Andy McKenzie

Equipe Méthodes et algorithmes pour la bioinformatique,

LIRMM, 161 rue Ada, 34392 - France

<http://www.lirmm.fr/w3ifa/MAAS/>

{gascuel, andy}@lirmm.fr

Abstract

Let \mathbf{T} be an ultrametric tree with corresponding ultrametric matrix \mathbf{D} . Given an observed matrix $\mathbf{\Delta} = \mathbf{D} + \mathbf{E}$, where \mathbf{E} is a matrix of error terms, one of the goals of hierarchical clustering algorithms is reconstructing \mathbf{T} from $\mathbf{\Delta}$. The performance of any algorithm then depends on how “close” $\mathbf{\Delta}$ is to the true matrix \mathbf{D} , and one measure of this is $\|\mathbf{D} - \mathbf{\Delta}\|_{\infty} = \|\mathbf{E}\|_{\infty}$. The safety radius of an algorithm, introduced by Atteson (1999) for the more general additive case, is the largest value of $\|\mathbf{E}\|_{\infty}$ below which the algorithm correctly reconstructs the tree structure with certainty.

In this paper we use the safety radius approach to analyze the performance of several algorithms for hierarchical clustering. We show that there is a maximal size the safety radius can have, that is l^* , where l^* is the minimum internal edge length of the tree, while in the additive case the best possible safety radius is $l^*/2$. Any algorithm that obtains this maximal size for the safety radius is called optimal, and

we show that several commonly used agglomerative algorithms are optimal, such as the single or average linkage algorithms. Furthermore, we demonstrate that the safety radius of ADDTREE and Neighbor Joining is not improved in the ultrametric case, and remains $l^*/2$. In addition, we show that as the number of objects increases, the safety radius tends to zero under any least-squares optimization approach. Finally, practical and theoretical implications of these findings are discussed.

1 Introduction


Let \mathbf{T} be an ultrametric tree on n objects and its corresponding matrix be $\mathbf{D} = [d_{ij}]$, where d_{ij} is the distance between objects i and j . From \mathbf{D} the tree \mathbf{T} can be reconstructed in terms of both structure and edge lengths. However, commonly one only has the observed matrix $\Delta = [\delta_{ij}]$ available, where $\Delta = \mathbf{D} + \mathbf{E}$, and $\mathbf{E} = [\epsilon_{ij}]$ are error terms. One of the main problems that hierarchical clustering algorithms address is, starting from the observed matrix Δ , finding a good estimate of the tree \mathbf{T} , with the primary interest being in the structure of \mathbf{T} rather than on its edge lengths. This problem is encountered in domains where attempts are made to reconstruct an inheritance pattern as, for example, the history of manuscripts in archaeology, the evolution of species in biology, or the filiation of languages.

The two main approaches used to solve this problem are agglomerative methods and direct optimization techniques. In agglomerative methods, n clusters containing single objects are started with, two of these are joined to give a single cluster, and the distance from this cluster to the rest of the clusters is calculated. The process then starts again and repeats until there is a single cluster left. Most agglomerative methods choose to join the two clusters with the smallest distance between them, and the main difference between the methods is how distances from a newly formed cluster to other

clusters is calculated. The four most common agglomerative methods are single-linkage (Florek et al. 1951; Sneath 1957), complete-linkage (Sørensen 1948; McQuitty 1960), weighted pair group method using arithmetic averages (WPGMA) (McQuitty 1966), and unweighted pair group method using arithmetic averages (UPGMA or average-linkage) (Sokal and Michener 1958). However, numerous other methods exist (see Gordon (1996) for extensive references).

Direct optimization methods attempt to minimize $\sum |\hat{d}_{ij} - \delta_{ij}|$, or more commonly the sum of squares $\sum (\hat{d}_{ij} - \delta_{ij})^2$, subject to the constraint that the estimated matrix $\hat{\mathbf{D}}$ must be ultrametric. More generally the quantity $\|\hat{\mathbf{D}} - \Delta\|_p$ can be minimized, where $\|\cdot\|_p$ is the p -norm, and often the weighted least squares quantity $\sum w_{ij}(\hat{d}_{ij} - \delta_{ij})^2$ is used as well. For large n , and $p = 1$ or $p = 2$, direct optimization methods of $\|\hat{\mathbf{D}} - \Delta\|_p$ have to be heuristic as it has been shown that this optimization problem is NP-hard (Křivánek and Morávek 1986; Křivánek 1986; Day 1987). Indeed, optimal branch-and-bound algorithms for least-squares are limited to about twenty objects (Chandon et al. 1980). In contrast, for $p = \infty$ polynomial time algorithms exist (Farach et al. 1995; Chepoi and Fichet 2000). Hartigan (1967) presented a method that started with a putatively good estimated tree, which was then altered by the movement of edges and insertion of vertices to attempt to obtain a tree that was even better in the weighted least-squares sense. Carroll and Pruzansky (1980) and Soete (1984) introduced a penalty function that measured the extent to which the ultrametric condition held for triplets of leaves. Mathematical programming techniques were then used to minimize the sum of the weighted least-squares function and the penalty function. Other approaches are described in Barthélemy and Guénoche (1991); Gordon (1996); Hansen and Jaumard (1997), and Hubert et al. (2001).

In both approaches an important question is, given an observed matrix,



how good is the estimated tree? Clearly this depends on how “close” the observed matrix Δ is to the true matrix \mathbf{D} , and on the algorithm used to estimate the tree. One measure of “closeness” is $\|\mathbf{D} - \Delta\|_\infty = \|\mathbf{E}\|_\infty = \max_{ij} |\epsilon_{ij}|$. This quantity is used in the definition (Atteson 1999) of a measure of the performance of an algorithm we call the “safety radius” (L_∞ radius in Atteson’s article), this being the largest value that $\|\mathbf{E}\|_\infty$ can have, below which it can be guaranteed that the algorithm will correctly reconstruct the tree structure.

There is a maximal size the safety radius can have, beyond which it cannot be guaranteed that a tree will be reconstructed correctly, no matter what reconstruction algorithm is used. If an algorithm has a safety radius that is maximal then we call it an “optimal” algorithm, as its safety radius cannot be exceeded (though it may not be the only optimal algorithm). Atteson (1999) demonstrated that in the case of additive (or tree) distances, which generalize ultrametric distances (Barthélemy and Guénoche 1991), a certain class of methods, which includes ADDTREE (Sattath and Tversky 1977) and Neighbor Joining (Saitou and Nei 1987), are optimal with a safety radius of $l^*/2$, where l^* is the minimal internal edge length of the additive tree. In this paper we use the same approach to analyze the performance of several algorithms for ultrametric trees.

In the following we first provide notation and formal definitions (Section 2), and then find the optimal safety radius for ultrametric trees (Section 3). Next we introduce a class of agglomerative algorithms, and show that their safety radius is optimal (Section 4). We then show that the safety radius for ADDTREE and Neighbor Joining (NJ) is unimproved when dealing with ultrametric data (Section 5). Following this, we investigate the safety radius under a least-squares optimization approach, and show that the safety radius decreases to zero as the number of leaves increase (Section 6). Lastly, we discuss the safety radius approach and its practical and

theoretical impact (Section 7).

2 Notation and Definitions

We use in this section commonly known definitions and results which can be found in textbooks such as Barthélemy and Guénoche (1991) or Semple and Steel (2003). We refer the reader to these for detailed explanations and bibliographic references.

Let \mathbf{T} be the ultrametric tree we aim at reconstructing. \mathbf{T} has n leaves which are bijectively labeled by the n objects, and the edges of \mathbf{T} are positively valued with $l(e)$ being the valuation (length) of edge e . \mathbf{T} defines a distance matrix $\mathbf{D} = [d_{ij}]$ between the objects, where d_{ij} is equal to the length of the path in \mathbf{T} from i to j . \mathbf{T} can be reconstructed from \mathbf{D} . $\Delta = [\delta_{ij}]$ is the observed matrix with δ_{ij} being an estimate of d_{ij} . $\mathbf{E} = \mathbf{D} - \Delta = [\epsilon_{ij}]$ is the matrix of errors terms and we assume, as usual, that Δ is a dissimilarity: (i) $\delta_{ij} \geq 0$; (ii) $\delta_{ij} = \delta_{ji}$; (iii) $\delta_{ij} = 0 \Leftrightarrow i = j$.

\mathbf{D} satisfies the ultrametric inequality: for all leaves i, j, k we have $d_{ij} \leq \max(d_{ik}, d_{jk})$. This inequality is a special case of the more general additive inequality (Zaretskii 1965; Buneman 1974) also called the four-point condition: for all leaves i, j, k, l we have $d_{ij} + d_{kl} \leq \max(d_{ik} + d_{jl}, d_{jk} + d_{il})$, which corresponds to an additive distance being defined by a tree with positive valuations on the edges, without the ultrametric constraint.

Since \mathbf{T} is ultrametric it is also spherical: there exists an edge on which a vertex r can be inserted such that the length d_{ir} of the path from r to any leaf i is constant and independent of i . We interpret r as a root and \mathbf{T} is then best seen as a rooted tree. Moreover, for any pair of leaves i, j we have: $d_{iv} = d_{jv} = d_{ij}/2$, where v is the least common ancestor of i and j . Figure 1 displays an illustration of the spherical representation of ultrametric trees. The dendrogram is another popular (equivalent) representation, but it will not be used here as it makes comparison of results with the additive case

awkward. For the rest of the paper we implicitly consider \mathbf{T} as a rooted tree, unless otherwise stated

The internal vertices of \mathbf{T} are the non-leaf vertices and include the tree root. We assume that \mathbf{T} is fully resolved, which means that every internal vertex of \mathbf{T} has two descendants. An unresolved (internal) vertex has three descendants or more, and an unresolved tree contains unresolved vertices. The internal edges of \mathbf{T} join internal vertices and the set of internal edges is denoted as E . l^* is the length of the shortest edge in E . We assume $l^* > 0$; having $l^* = 0$ would be equivalent to having an unresolved vertex in \mathbf{T} . We preclude unresolved vertices to avoid having different but equivalent trees and more complicated definitions and proofs.

The clusters of \mathbf{T} correspond to the subsets of leaf labels induced by the subtrees of \mathbf{T} , e.g. in Figure 1 we have: $\{i\}, \{j\}, \{k\}, \dots \{k, j\}, \dots \{i, k, j\} \dots \{i, k, j, l, m\}$. The structure of \mathbf{T} is defined by its graphical representation using vertices and edges as well as by its cluster set. Our aim is to reconstruct from Δ the structure of \mathbf{T} , but not the edge lengths, which is not achievable unless $\Delta = \mathbf{D}$.

3 Optimal Safety Radius in the Ultrametric Case

Theorem 1 *For ultrametric distances the maximal safety radius is at most l^* , where l^* is the minimum internal edge length of the ultrametric tree that has been perturbed.*

Proof. Let \mathbf{T} be the resolved ultrametric tree shown in Figure 2. We now introduce an unresolved ultrametric tree \mathbf{T}' (Figure 2), with associated matrix \mathbf{D}' , which is a perturbed version of \mathbf{T} . We then have $\Delta = \mathbf{D}'$ and for \mathbf{D}' the dissimilarities are

$$d'_{ab} = d'_{ac} = d'_{bc} = 2x + l^* \quad a \in A, b \in B, c \in C$$

which implies that

$$d'_{ab} = d_{ab} + l^*; \quad d'_{ac} = d'_{bc} = d_{ac} - l^* = d_{bc} - l^*,$$

while all other distances are identical in \mathbf{D} and \mathbf{D}' . In other words $\|\mathbf{E}\|_\infty = \|\mathbf{D} - \mathbf{D}'\|_\infty = l^*$.

The pertinent feature of \mathbf{T}' is that the cluster $\{A, B, C\}$ is unresolved, so there are three resolved ultrametric trees that are compatible with it, and it is not possible to recover the structure of \mathbf{T} from that of \mathbf{T}' with certainty. Hence, if $\|\mathbf{D} - \Delta\|_\infty \geq l^*$ it cannot be guaranteed that a tree will be correctly reconstructed, no matter what algorithm is used. \square

We have shown that the safety radius is at most l^* in the ultrametric case. In the next section, we will show that there exist algorithms that achieve this safety radius, which is then optimal. So, significantly, for ultrametric distances the optimal safety radius is twice that for additive distances. This implies that when dealing with ultrametric data there exist, as we shall see, algorithms that can be guaranteed under certain circumstances to outperform algorithms that are based on additivity only.

4 Agglomerative Algorithms

4.1 A Class of Agglomerative Algorithms

Single-linkage, complete-linkage, WPGMA, and UPGMA are special cases of the following agglomerative procedure for ultrametric distances:

1. Input the observed matrix Δ .
2. Select the pair $\{x, y\}$ to be joined that minimizes δ_{ij} .
3. Join x and y to form a new cluster u . For clusters $i \neq x, y$ reduce the dissimilarity matrix using $\delta_{ui} = \lambda\delta_{xi} + (1 - \lambda)\delta_{yi}$, while all other distances remain unchanged.

4. Repeat steps 2-3 until there is one cluster left.

For the purposes of this paper we restrict the value of the parameter λ to $[0, 1]$, but allow it to take different values for each joined pair x, y and for each i . In particular, for single-linkage we have $\lambda = 1$ for $\delta_{xi} \leq \delta_{yi}$ and $\lambda = 0$ when $\delta_{xi} > \delta_{yi}$; in complete-linkage $\lambda = 0$ for $\delta_{xi} \leq \delta_{yi}$ and $\lambda = 1$ when $\delta_{xi} > \delta_{yi}$; for WPGMA $\lambda = 1/2$; in UPGMA $\lambda = |x|/(|x| + |y|)$, where $|x|$ is the number of objects in the cluster x . If $\Delta = \mathbf{D}$ then this procedure recovers the correct tree structure, and does so for any value of the parameter λ (including $\lambda \notin [0, 1]$), which implies that this class of algorithms is *consistent*. There exists a wide variety of other reduction formulae (Lance and Williams 1967; Gordon 1996), but they include non-consistent algorithms (see Gascuel (2000) for a discussion of this point). However, there are some consistent agglomerative algorithms that are not part of this class, for example the algorithm described by Degens (1983) that uses median values instead of averages to reduce the dissimilarity matrix. Finally, the algorithm of Farach et al. (1995) and generalized by Chepoi and Fichet (2000), which is optimal for the infinite norm (see the Introduction), does not belong to this class, but at each agglomeration finds the same tree structure as single-linkage and has the same (optimal) safety radius.

4.2 Safety Radius

In this section we prove that the above class of agglomerative methods is not only consistent but also optimal. This implies that, in particular, the single-linkage algorithm, complete-linkage algorithm, WPGMA, and UPGMA are optimal. Before proving the main theorem we prove three preliminary lemmas.

In our first lemma we give a lower bound that will be useful in the proof of the final theorem. Note that two leaves are called neighbors if they are separated by a single vertex.

Lemma 1 *Let $\mathbf{D} = [d_{ij}]$ be an ultrametric matrix. If the leaves x, y are not neighbors, then there is a pair of neighbors i, j :*

$$d_{xy} - d_{ij} \geq 2 l^* .$$

Proof. Since x, y are not neighbors then there must be at least two internal vertices separating these leaves. Without loss of generality, let v be the internal vertex adjacent to x , and l the length of the edge above v . Finally, let C be the cluster neighboring x . See Figure 3 for an illustration. Whether $i, j \in C$, or one of i, j is equal to x with the other being in C , we have $d_{xy} - d_{ij} \geq 2l \geq 2 l^*$. \square

Lemma 2 *Let \mathbf{D} be the ultrametric matrix corresponding to the ultrametric tree \mathbf{T} , and l^* be the shortest internal edge length in \mathbf{T} . Using \mathbf{D} as input to the agglomerative procedure, and agglomerating neighbors i, j , then denote \mathbf{D}' as the reduced matrix at the next step. We have the following two results:*

(a) \mathbf{D}' is an ultrametric matrix.

(b) $\min_{e \in E'} l(e) \geq l^*$, where E' is the internal edge set for the ultrametric tree corresponding to \mathbf{D}' .

Proof. For the tree \mathbf{T} let v be the vertex adjacent to the neighboring leaves i, j which are chosen to be agglomerated, and w be the vertex adjacent to v (but not either of i, j). Form a new ultrametric tree \mathbf{T}' , which has one less leaf than \mathbf{T} , by removing either of i, j and relabelling the remaining leaf as u . In this new tree $d'_{uk} = d_{ik} = d_{jk}$ for $k \neq i, j$, and all other distances are unchanged. The distances for the new tree are the same as those given by the agglomerative algorithm, so \mathbf{D}' is the ultrametric matrix corresponding to \mathbf{T}' . Part (b) of the lemma follows from noting that \mathbf{T}' has the same internal edges as \mathbf{T} , except for the deletion of the edge $\{v, w\}$. \square

For our last lemma we find an upper bound for the difference between the estimated and actual dissimilarities after one iteration of the agglomerative procedure.

Lemma 3 Define $\mathbf{E} = \Delta - \mathbf{D}$. Let Δ be the input to the agglomerative procedure, with leaves i, j been chosen to join, giving the reduced matrix Δ' . Using the same leaves i, j , reduce \mathbf{D} to give \mathbf{D}' (as was done in Lemma 2). We have

$$\|\Delta' - \mathbf{D}'\|_\infty \leq \|\mathbf{E}\|_\infty .$$

Proof. Let x, y be arbitrary indices. If $x, y \neq u$ then the distances are unchanged and we have $|\delta'_{xy} - d'_{xy}| = |\delta_{xy} - d_{xy}| \leq \|\mathbf{E}\|_\infty$. If $x = u$ (and similarly if $y = u$) then we have

$$\begin{aligned} |\delta'_{uy} - d'_{uy}| &= |(\lambda\delta_{iy} + (1-\lambda)\delta_{jy}) - (\lambda d_{iy} + (1-\lambda)d_{jy})| \\ &= |\lambda(\delta_{iy} - d_{iy}) + (1-\lambda)(\delta_{jy} - d_{jy})| \\ &\leq \lambda|\delta_{iy} - d_{iy}| + (1-\lambda)|\delta_{jy} - d_{jy}| \quad (\text{since } \lambda \in [0, 1]) \\ &\leq \|\mathbf{E}\|_\infty . \end{aligned}$$

□

We now use the three lemmas to prove that the class of agglomerative algorithms introduced in Section 4.1 are all optimal.

Theorem 2 Let $\mathbf{D} = [d_{ij}]$ be the dissimilarity matrix for an ultrametric tree \mathbf{T} with minimum internal edge length l^* , and $\Delta = [\delta_{ij}]$ the observed matrix. If $\|\mathbf{D} - \Delta\|_\infty = \|\mathbf{E}\|_\infty < l^*$ then any algorithm from the class introduced in Section 4.1 correctly reconstructs the structure of \mathbf{T} when Δ is used as the input.

Proof. Start at the first step of the agglomerative procedure with input Δ . Let x, y be a pair of non-neighbors and i, j a pair of neighbors that satisfy Lemma 1. We have the following:

$$\begin{aligned} \delta_{xy} - \delta_{ij} &= (\delta_{xy} - d_{xy} + d_{ij} - \delta_{ij}) + (d_{xy} - d_{ij}) \\ &\geq -2\|\mathbf{E}\|_\infty + (d_{xy} - d_{ij}) \quad (\text{by definition}) \\ &\geq -2\|\mathbf{E}\|_\infty + 2l^* \quad (\text{Lemma 1}) \\ &> 0 . \end{aligned}$$

Therefore a pair of neighbors will be chosen to join. Joining the pair of neighbors i, j , let Δ be reduced to Δ' and \mathbf{D} to \mathbf{D}' . By Lemma 2, the matrix \mathbf{D}' is ultrametric with smallest internal edge length (l'^*) greater than or equal to l^* . Combining this with Lemma 3 we have $\|\Delta' - \mathbf{D}'\|_\infty \leq \|\mathbf{E}\|_\infty < l'^*$. Thus using Δ' as input to the agglomerative procedure again, neighbors will be chosen to join again, and by induction this will happen at every iteration of the agglomerative procedure. \square

5 Neighbor Joining and ADDTREE

We have shown (Section 3) that the optimal safety radius in the ultrametric case is l^* , while it is only $l^*/2$ in the more general additive case. Moreover, we have demonstrated (Section 4) that the usual agglomerative algorithms for ultrametric data are optimal, i.e. their safety radius equals l^* . An analogous result was shown by Atteson (1999) in the additive case for ADDTREE, NJ and related algorithms (also agglomerative but using specific pair selection criteria), i.e. their safety radius is $l^*/2$ (only). So the more restricted ultrametric setting seems to be easier to solve than the additive one, and we may wonder whether ADDTREE or NJ possess a better safety radius with ultrametric data than in the general case. The following theorem indicates that the answer is negative.

Theorem 3 *The safety radius of ADDTREE and NJ in the ultrametric case remains equal to $l^*/2$, just as in the more general additive case.*

Proof. We shall see with a simple four objects example that the safety radius of NJ and ADDTREE cannot be better than $l^*/2$. In this case, ADDTREE and NJ are identical and use the additive inequality (Saitou and Nei 1987; Gascuel 1994). Refer to the tree in Figure 4. Considering this tree as unrooted (ADDTREE and NJ infer unrooted trees and do not have the ability to recover the tree root), it contains only one internal edge with

length l^* . For i, j (or x, y) to be selected as a neighboring pair, the observed dissimilarity must satisfy:

$$\delta_{xy} + \delta_{ij} < \min(\delta_{xi} + \delta_{yj}, \delta_{xj} + \delta_{yi}),$$

or

$$4h - 2l^* - 2l + \epsilon_{xy} + \epsilon_{ij} < \min(4h - 2l + \epsilon_{xi} + \epsilon_{yj}, 4h - 2l + \epsilon_{xj} + \epsilon_{yi}).$$

Taking the worst-case for the error terms, then rewriting in terms of the infinity-norm gives $\|\mathbf{E}\|_\infty < l^*/2$. \square

In fact most of algorithms are identical with four objects, e.g.: the convex versions of ADDTREE (Bandelt and Dress 1986) or of NJ (Atteson 1999), or the iterative least-squares projection approach (Gascuel and Levy 1996), which, moreover, is not heuristic but exact with four objects. So Theorem 3 applies to all these methods, all of which compare unfavorably with the simple usual agglomerative algorithms in the ultrametric case.

6 Safety Radius under Least Squares Optimization

Least-squares optimization is the second commonly used approach to build trees from dissimilarities. In this section, we do not examine any special algorithm, but demonstrate that the safety radius tends to zero as the number of objects increases, when assuming that least-squares are effectively minimized. Note, however, that this latter assumption is not that obvious for any practical algorithm, since the minimization task is NP-Hard (Křivánek and Morávek 1986). So it is still conceivable that some poor (regarding minimization) heuristic least-squares algorithm has better safety radius than that predicted here.

The principle of the proof is to bound from the above the safety radius for the reconstruction of a certain set of trees, under ordinary least-squares

(OLS) optimization. We show that for this set of trees, with a fixed minimum edge length, the safety radius tends to zero as the number of leaves increases. Accordingly, since this set of trees is a subset of the possible trees, the safety radius for OLS optimization also tends to zero as the number of leaves increases. Moreover, since OLS is a special case of weighted least-squares and generalized least-squares (Searle 1971), this result implies that the safety radius also tends to zero for these other two cases as well.

OLS approaches can be decomposed into two steps: (1) finding the appropriate tree structure, (2) optimizing the edge lengths associated with this structure. The following lemma provides the basic property of OLS edge length estimates, for the case of ultrametric trees.

Lemma 4 *Let $\Delta = [\delta_{ij}]$ be the observed dissimilarity matrix, and $\hat{\mathbf{T}}$ any ultrametric tree with fixed structure for which the edge lengths have been fitted so that its induced matrix $\hat{\mathbf{D}}$ is as close as possible to Δ in the OLS sense, i.e. $\|\Delta - \hat{\mathbf{D}}\|_2$ is minimal. Let X be a cluster of $\hat{\mathbf{T}}$, which is composed of two neighboring clusters A and B . Then for any $a \in A$ and $b \in B$, we have:*

$$\hat{d}_{ab} = \frac{1}{|A||B|} \sum_{\substack{i \in A \\ j \in B}} \delta_{ij}.$$

Proof. For completeness we outline the proof here - full details can be found in Degens (1986). The fact that \hat{d}_{ab} is constant for any $a \in A$ and $b \in B$ is a direct consequence of the spherical property (Figure 1). Let \hat{d} be this constant. Now $\|\hat{\mathbf{D}} - \Delta\|_2$ can be written as

$$\|\hat{\mathbf{D}} - \Delta\|_2 = 2 \sum_{\substack{a \in A \\ b \in B}} (\hat{d} - \delta_{ab})^2 + \sum_{i \notin A \text{ or } j \notin B} (\hat{d}_{ij} - \delta_{ij})^2.$$

Both summations are independent, and only the first refers to cluster X and to \hat{d} . Thus to minimize $\|\hat{\mathbf{D}} - \Delta\|_2$ it is necessary (and sufficient when considering all clusters of $\hat{\mathbf{T}}$) to minimize the first sum, which implies that \hat{d} is equal to the mean of the terms δ_{ab} . \square

We now state the main result of this section.

Theorem 4 *Under any least-squares approach to reconstructing an ultrametric tree, the safety radius tends to zero as the number of leaves increases.*

Proof. We place ourselves in the OLS framework, which is a special case of weighted least-squares and generalized least-squares. Let \mathbf{T} be the ultrametric tree shown in Figure 5a with corresponding matrix \mathbf{D} . Denote by Δ the dissimilarities when an error term is added and let these terms take the special form:

$$\begin{aligned}\delta_{ij} &= a + \epsilon \\ \delta_{ix} &= b - \epsilon \\ \delta_{jx} &= b + \epsilon \\ \delta_{xy} &= d_{xy}\end{aligned}$$

where ϵ is a constant, and x, y are any leaves from C . The leaves i, j are a neighboring pair and the correctly estimated tree should have the same structure as that of Figure 5a. Under this tree structure, repeated application of Lemma 4, gives for the estimated distances $\hat{d}_{ij} = \delta_{ij}$, $\hat{d}_{xy} = \delta_{xy}$, $\hat{d}_{ix} = \hat{d}_{jx} = b$. With these estimates, the sum of squares error is $2(n-2)\epsilon^2$.

If instead the estimated tree has the incorrect tree structure of Figure 5b, then using Lemma 4 again, the estimated distances are $\hat{d}_{ij} = \hat{d}_{jx} = \beta$, $\hat{d}_{xy} = \delta_{xy}$, $\hat{d}_{ix} = \delta_{ix}$, where $\beta = \frac{(n-2)(b+\epsilon)+(a+\epsilon)}{n-1}$. It follows that the sum of squares error for this incorrect tree structure is

$$(n-2)[(b+\epsilon) - \beta]^2 + [(a+\epsilon) - \beta]^2 = \frac{n-2}{n-1}(b-a)^2.$$

Thus the incorrect tree structure will be preferred under the ordinary least-squares criterion as soon as

$$2(n-2)\epsilon^2 > \frac{n-2}{n-1}(b-a)^2,$$

or in terms of the size of the error term

$$\epsilon > \frac{1}{\sqrt{2(n-1)}}(b-a).$$

Consequently, as the number of leaves (n) increases, but with $(b-a)$ fixed, the safety radius decreases towards zero. Moreover, note that the example of Figure 5 is precisely that encountered by iterative addition algorithms, such as KITSCH from the Phylip package (Felsenstein 1989). Assuming that i is the new object to be added in the current tree, the correct position (a) will not be selected and a wrong tree will be constructed. \square

7 Discussion

Our findings may be summarized as follows: (i) the optimal safety radius in the ultrametric case equals the length of the shortest internal edge of the true tree, instead of half of this length in the more general additive case; (ii) the usual agglomerative algorithms such as single or average linkage are optimal, while ADDTREE, NJ and related algorithms, which are optimal for additive data, do not perform any better when the true tree is ultrametric and are sub-optimal in this case; (iii) the safety radius decreases to zero when the number of objects increases, under any least-squares approach.

These results may be seen as surprising regarding the numerous published computer simulations that have compared those approaches, and which do not reflect the method ordering induced by the safety radius analysis. However, most of these simulations used Gaussian error terms, which perfectly fits least-squares approaches. Moreover, in phylogenetic simulations the error is mostly additive (Bruno et al. 2000), i.e. affects the edge lengths and transforms the true ultrametric tree into a non-ultrametric one, which explains why NJ tends to outperform UPGMA (Saitou and Nei 1987). With non-phylogenetic noise and ultrametric data, UPGMA and related agglomerative algorithms tends to perform better than NJ, as expected from

our analysis (authors results, available on request). But, fundamentally, simulation studies and safety radius analysis differ. The former aims at describing an average behavior, while the latter exhibits worst cases and pathological situations. So our findings must be interpreted carefully. They do not say that least-squares approaches are weak in average, but indicate that they may be extremely weak in some (perhaps unlikely) situations, where the data matrix is very close to an ultrametric tree and these approaches do not recover this tree. Agglomerative algorithms do not have such shortcomings and can then be seen as more robust.

In the additive case, the safety radius approach has proven to be useful for computing the convergence rate of various phylogeny methods (Erdős et al. 1997; Atteson 1999; Berry and Gascuel 2000), and the very same approach was used by Atteson (1999) to demonstrate that ADDTREE systematically recovers the edges which are longer than twice $\|\mathbf{E}\|_\infty$, while NJ does not possess this property. Moreover, Wang and Gusfield (1998) used a closely related approach to determine necessary and sufficient conditions for the uniqueness of optimal tree structures. Similar studies could be conducted in the ultrametric case. On the other hand, it would be of interest to determine whether our findings for least-squares ultrametric methods extend to the additive case.

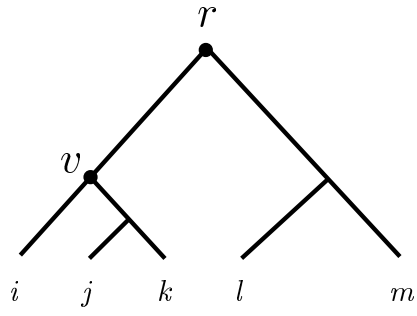


Figure 1: Spherical tree distances. $d_{iv} = d_{jv} = d_{ij}/2$.

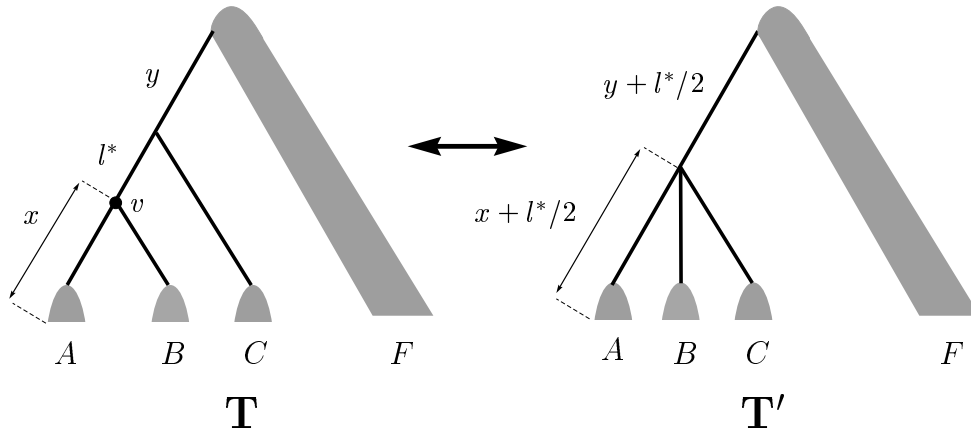


Figure 2: An unresolved ultrametric tree \mathbf{T}' close to the resolved ultrametric tree $\mathbf{T} : \|\mathbf{D}' - \mathbf{D}\|_\infty = l^*$. The quantities l^* and y are edge lengths, while x represents the path length between v and any leaf of A or B . The shaded regions A, B, C, F represent arbitrary ultrametric trees.

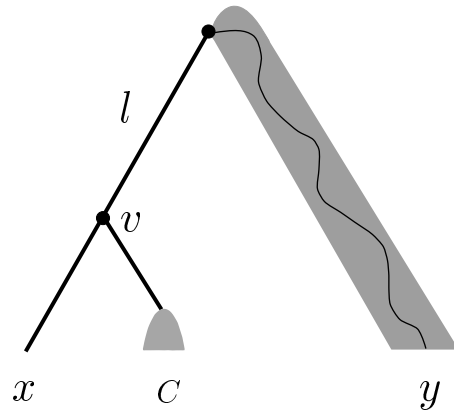


Figure 3: Non-neighbors x, y , separated by at least two vertices (\bullet). The tree structure labelled C contains at least one of the neighbors i, j .

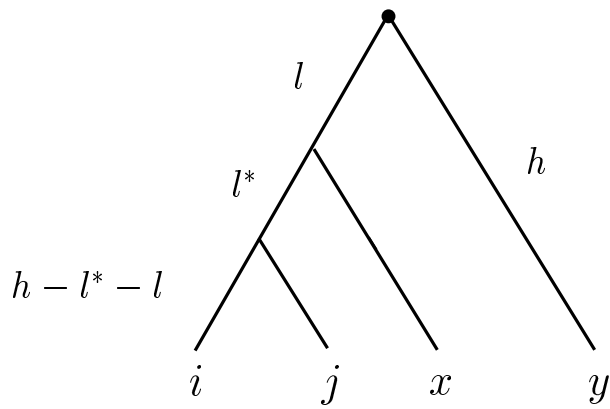


Figure 4: ADDTREE and NJ in the ultrametric case.

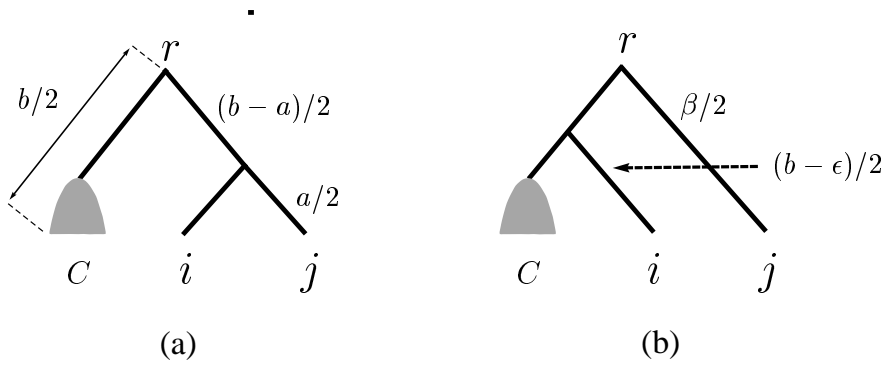


Figure 5: C is an arbitrary ultrametric tree structure with $n - 2$ leaves (a) correct tree structure (b) incorrect tree structure.

References

- ATTESON, K. (1999), “The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction,” *Algorithmica*, 25, 251–278.
- BANDELT, H., and DRESS, A. (1986), “Reconstructing the Shape of a Tree from Observed Dissimilarity Data,” *Advances in Applied Mathematics*, 7, 309–343.
- BARTHÉLEMY, J., and GUÉNOCHE, A. (1991), *Trees and Proximity Representations*, Chichester: John Wiley & Sons, chapter 3.
- BERRY, V., and GASCUEL, O. (2000), “Inferring Evolutionary Trees with Strong Combinatorial Evidence,” *Theoretical Computer Science*, 240, 271–298.
- BRUNO, W., SOCCI, N. D., and HALPERN, A. L. (2000), “Weighted Neighbor Joining: A Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction,” *Molecular Biology and Evolution*, 17 (1), 189–197.
- BUNEMAN, P. (1974), “A Note on Metric Properties of Trees,” *Journal of Combinatorial Theory*, 17, 48–50.
- CARROLL, J. D., and PRUZANSKY, S. (1980), “Discrete and Hybrid Scaling Models,” in E. D. Lantermann and H. Feger, eds., “Similarity and Choice,” Bern: Hans Huber, 108–139.
- CHANDON, J. L., LEMAIRE, J., and POUGET, J. (1980), “Construction de l’ultramétrie la plus proche d’une dissimilarité au sens des moindres carrés,” *Recherche Opérationnelle/Operations Research*, 14, 157–170.
- CHEPOI, V., and FICHET, B. (2000), “ l_∞ -Approximation Via Subdominants,” *Journal of Mathematical Psychology*, 44, 600–616.
- DAY, W. H. E. (1987), “Computational Complexity of Inferring Phylogenies from Dissimilarity Matrices,” *Bulletin of Mathematical Biology*, 49, 461–467.

- DEGENS, P. O. (1983), “Hierarchical Cluster Methods as Maximum Likelihood Estimators,” in J. Felsenstein, ed., “Numerical Taxonomy,” Springer-Verlag, volume G1 of *Nato ASI*, 249–253.
- DEGENS, P. O. (1986), “Ultrametric Approximation to Distances,” *Computational Statistics Quarterly*, 2 (2), 93–101.
- ERDÖS, P., STEEL, M., SZÉKELY, L., and WARNOW, T. (1997), “Constructing Big Trees from Short Sequences,” in “Lecture Notes in Computer Science,” volume 1256, 827–837.
- FARACH, M., KANNAN, S., and WARNOW, T. (1995), “A Robust Model for Finding Optimal Evolutionary Trees,” *Algorithmica*, 13, 155–179.
- FELSENSTEIN, J. (1989), “PHYLP Phylogeny Inference Package (Version 3.2),” *Cladistics*, 5, 164–166.
- FLOREK, K., LUKASZEWICZ, J., PERKAL, J., STEINHAUS, H., and ZUBRZYCKI, S. (1951), “Sur la liaison et la division des points d’un ensemble fini,” *Colloquium Mathematicum*, 2, 282–285.
- GASCUEL, O. (1994), “A Note on Sattath and Tverstky’s, Saitou and Nei’s, and Studier and Keppler’s Algorithms for Inferring Phylogenies from Evolutionary Distances,” *Molecular Biology and Evolution*, 11 (6), 961–963.
- GASCUEL, O. (2000), “Data Model and Classification by Trees: The Minimum Variance Reduction (MVR) Method,” *Journal of Classification*, 17, 67–99.
- GASCUEL, O., and LEVY, D. (1996), “A Reduction Algorithm for Approximating a (Non-Metric) Dissimilarity by a Tree Distance,” *Journal of Classification*, 13, 129–155.
- GORDON, A. D. (1996), “Hierarchical Classification,” in R. Arabie, L. J. Hubert, and G. D. Soete, eds., “Clustering and Classification,” Singapore: World Scientific Publishing Co Pte Ltd, 65–121.
- HANSEN, P., and JAUMARD, B. (1997), “Cluster Analysis and Mathe-

- mathematical Programming,” *Mathematical Programming*, 79, 191–215.
- HARTIGAN, J. (1967), “Representation of Similarity Matrices by Trees,” *Journal of the American Statistical Association*, 62, 1140–1158.
- HUBERT, L., ARABIE, P., and MEULMAN, J. (2001), *Combinatorial Data Analysis: Optimization by Dynamic Programming*, Society for Industrial and Applied Mathematics (SIAM).
- KŘIVÁNEK, M. (1986), “On the Computational Complexity of Clustering,” in E. Diday, Y. Escoufier, L. Lebart, J. Pagès, Y. Schektman, and R. Tomassone, eds., “Data Analysis and Informatics IV,” Amsterdam: North-Holland, 89–96.
- KŘIVÁNEK, M., and MORÁVEK, J. (1986), “NP-Hard Problems in Hierarchical-Tree Clustering,” *Acta Informatica*, 23, 311–323.
- LANCE, G. N., and WILLIAMS, W. T. (1967), “A General Theory of Classificatory Sorting Strategies 1. Hierarchical Systems,” *Computer Journal*, 9, 373–380.
- MCQUITTY, L. (1960), “Hierarchical Linkage Analysis for the Isolation of Types,” *Educational and Psychological Measurement*, 20, 55–67.
- MCQUITTY, L. (1966), “Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data,” *Educational and Psychological Measurement*, 26, 825–831.
- SAITOU, N., and NEI, M. (1987), “The Neighbor-Joining Method: A New Method for Reconstruction of Phylogenetic Trees,” *Molecular Biology and Evolution*, 4, 406–425.
- SATTATH, S., and TVERSKY, A. (1977), “Additive Similarity Trees,” *Psychometrika*, 42 (3), 319–345.
- SEARLE, S. R. (1971), *Linear Models*, New York: Wiley.
- SEMPLE, C., and STEEL, M. (2003), *Phylogenetics*, Oxford Press.
- SNEATH, P. H. (1957), “The Application of Computers to Taxonomy,” *Journal of General Microbiology*, 17, 201–226.

- SOETE, G. D. (1984), "A Least Squares Algorithm for Fitting an Ultrametric Tree to a Dissimilarity Matrix," *Pattern Recognition Letters*, 2, 133–137.
- SOKAL, R. R., and MICHENER, C. D. (1958), "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Science Bulletin*, 38, 1409–1438.
- SØRENSEN, T. (1948), "A Method of Establishing Groups of Equal Amplitude in Plant Biology Based on Similarity of Species Content and its Application to Analyses of the Vegetation on Danish Commons," *Biologiske Skrifter*, 5 (4), 1–34.
- WANG, L., and GUSFIELD, D. (1998), "Constructing Additive Trees When the Error is Small," *Journal of Computational Biology*, 5, 127–133.
- ZARETSKII, K. (1965), "Constructing a Tree Based on a Set of Distances Among its Leaves," *Uspekhi Matematicheskikh Nauk.*, 20, 90–92, in Russian.