

Robustness of Phylogenetic Inference Based on Minimum Evolution

Fabio Pardi, Sylvain Guillemot, Olivier Gascuel

▶ **To cite this version:**

Fabio Pardi, Sylvain Guillemot, Olivier Gascuel. Robustness of Phylogenetic Inference Based on Minimum Evolution. Bulletin of Mathematical Biology, 2010, 72 (7), pp.1820-1839. 10.1007/s11538-010-9510-y . lirmm-00511779v2

HAL Id: lirmm-00511779

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00511779v2

Submitted on 5 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bulletin of Mathematical Biology (2010) 72: 1820–1839
DOI 10.1007/s11538-010-9510-y

ORIGINAL ARTICLE

Robustness of Phylogenetic Inference Based on Minimum Evolution

Fabio Pardi[a,b,*], Sylvain Guillemot[a], Olivier Gascuel[a]

[a] *Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS—Université de Montpellier, 161 rue Ada, 34392 Montpellier, France*
[b] *EMBL—European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK*

Received: 4 September 2009 / Accepted: 19 January 2010 / Published online: 7 May 2010
© Society for Mathematical Biology 2010

Abstract Minimum evolution is the guiding principle of an important class of distance-based phylogeny reconstruction methods, including neighbor-joining (NJ), which is the most cited tree inference algorithm to date. The minimum evolution principle involves searching for the tree with minimum length, where the length is estimated using various least-squares criteria. Since evolutionary distances cannot be known precisely but only estimated, it is important to investigate the robustness of phylogenetic reconstruction to imprecise estimates for these distances. The *safety radius* is a measure of this robustness: it consists of the maximum relative deviation that the input distances can have from the correct distances, without compromising the reconstruction of the correct tree structure. Answering some open questions, we here derive the safety radius of two popular minimum evolution criteria: balanced minimum evolution (BME) and minimum evolution based on ordinary least squares (OLS + ME). Whereas BME has a radius of $\frac{1}{2}$, which is the best achievable, OLS + ME has a radius tending to 0 as the number of taxa increases. This difference may explain the gap in reconstruction accuracy observed in practice between OLS + ME and BME (which forms the basis of popular programs such as NJ and FastME).

Keywords Phylogenetics · Distance methods · Minimum evolution · Least squares · Safety radius

1. Introduction

Minimum evolution methods for reconstructing phylogenetic trees (Kidd and Sgaramella-Zonta, 1971; Rzhetsky and Nei, 1992) are based upon the following informal idea: given a matrix of distances between each pair of taxa in a set, reconstruct the phylogenetic tree for these taxa that implies the minimum amount of evolution in order to explain the given

*Corresponding author.
E-mail address: pardi@lirmm.fr (Fabio Pardi).

distances. In practice, a method to estimate the length of any possible tree structure is specified and minimum evolution aims to reconstruct the tree with the minimum length estimate.

Several types of minimum evolution can then be defined based on the different methods for estimating the tree length. In this paper, we deal with two important versions of minimum evolution: the one based on an ordinary least squares technique for length estimation, and the one based on an increasingly popular "balanced" technique. We call these versions OLS + ME and BME, respectively.

OLS + ME is based on a method for estimating branch lengths which originated from the work by Cavalli-Sforza and Edwards (1964, 1967) and was the first version of minimum evolution to be proposed (Kidd and Sgaramella-Zonta, 1971; Rzhetsky and Nei, 1992). It is introduced in more detail in the next section. Interestingly, neighbor-joining (NJ)—one of the most popular algorithms for tree reconstruction—has been shown to have some connections with OLS + ME (Saitou and Nei, 1987; Nei and Kumar, 2000).

However, further mathematical investigation showed that NJ is much more tightly related to the other criterion we study, BME: In fact, NJ turns out to be a greedy agglomerative algorithm aiming to construct the optimal tree with respect to this criterion (Desper and Gascuel, 2005; Gascuel and Steel, 2006). The vast popularity of NJ (more than 20,000 citations for the original paper (Saitou and Nei, 1987) on Google Scholar), the fact that BME length estimates are biologically meaningful variants of the least squares estimates (Desper and Gascuel, 2004), the computational efficiency of the proposed heuristics for BME (Desper and Gascuel, 2002), and last but not least, the high accuracy of these heuristics observed in simulation studies (Desper and Gascuel, 2002, 2004; Vinh and von Haeseler, 2005) justify the growing importance of BME as a criterion for reconstructing phylogenetic trees. Again, a more detailed account of BME is given in the next section.

BME and OLS + ME share an important property: if the input distances perfectly correspond to the distances between leaves in a tree with branch lengths, then these criteria identify this tree as the correct one. Evolutionary distances are estimated using genetic sequences or any other comparative data from the taxa under consideration. Ideal estimation procedures should ensure that, as the amount of compared data increases, the estimated distances converge to those in the correct phylogenetic tree, which means that the property just stated is equivalent to the statistical *consistency* of these tree reconstruction methods. In fact, consistency is a desirable property for *any* minimum evolution method (Willson, 2005a), and more generally for all tree reconstruction methods.

Note, however, that the amount of data is usually limited, and the models used to estimate the distances are usually only rough approximations of the reality. As a consequence, the estimated distances $\boldsymbol{\delta}$ will somehow deviate from the distances $\boldsymbol{d}^{\mathcal{T}}$ in the correct tree $\mathcal{T}$ and the accuracy of tree reconstruction will depend on its robustness to such deviations. Define the L_∞ difference between $\boldsymbol{\delta}$ and $\boldsymbol{d}^{\mathcal{T}}$ as $\|\boldsymbol{\delta} - \boldsymbol{d}^{\mathcal{T}}\|_\infty = \max_{i,j} |\delta_{ij} - d_{ij}^{\mathcal{T}}|$, where i and j denote two taxa; then a measure of this robustness, proposed by Atteson (1999), is the safety radius defined in the following way, where $\ell_{\min}^{\mathcal{T}}$ denotes the length of the shortest branch in $\mathcal{T}$.

Definition 1. A distance-based tree reconstruction method has *safety radius* α [*over n taxa*] if, for every weighted bifurcating tree $\mathcal{T}$ [over a set of size n] and any distances $\boldsymbol{\delta}$

such that

$$\left\| \boldsymbol{\delta} - \boldsymbol{d}^{\mathcal{T}} \right\|_{\infty} < \alpha \ell_{\min}^{\mathcal{T}},$$

the method reconstructs the topology of $\mathcal{T}$.

Note the distinction between the safety radius over a given number of taxa, and safety radius *tout-court*.

An important remark is that there are distances $\boldsymbol{\delta}$ which lie at $\frac{1}{2}\ell_{\min}^{\mathcal{T}}$ not only from $\boldsymbol{d}^{\mathcal{T}}$, but also from the distances $\boldsymbol{d}^{\mathcal{T}'}$ in a tree $\mathcal{T}'$ with a different topology from that of $\mathcal{T}$ (Atteson, 1999). This implies that: (1) robustness must be measured relative to the length of the shortest branch in $\mathcal{T}$, as no maximum value for the difference between $\boldsymbol{\delta}$ and $\boldsymbol{d}^{\mathcal{T}}$ can guarantee a correct tree reconstruction if nothing is assumed regarding $\ell_{\min}^{\mathcal{T}}$; (2) no method reconstructing a unique tree can have a safety radius greater than $\frac{1}{2}$.

Atteson (1999) also proved that a number of agglomerative algorithms, including NJ, have optimal safety radius $\frac{1}{2}$. A related result was recently shown by Bordewich et al. (2009) who proved that another heuristic aimed at minimizing BME (based on *subtree pruning and regrafting*) has *at least* radius $\frac{1}{3}$.

Note that these results are about *heuristics* for minimum evolution, not the criterion itself. Here, we identify the safety radius of a tree reconstruction criterion as the safety radius of any algorithm returning the optimal tree with respect to that criterion. Although there is no version of minimum evolution for which we have a practical algorithm for this task, it would still be interesting to know the safety radius that such an algorithm would have, as this may give some insight about the effectiveness of the heuristic algorithms used in practice. (We note that, at least in the case of BME, this optimization problem is NP-hard: manuscript in preparation.)

What is then known about the safety radii of minimum evolution criteria? For BME, Bordewich et al. (2009) have extended their result and proved that also the BME criterion has radius *at least* $\frac{1}{3}$. As for OLS + ME, Willson (2005b) proved that the safety radius over n taxa is limited above by a function tending to $\frac{1}{4}$, as n tends to infinity. Another relevant result was obtained by Gascuel and McKenzie (2004) for the case where both the estimation of branch lengths and the selection of a tree structure are based on a least squares criterion. They proved that when this criterion is constrained to produce ultrametric trees (i.e., rooted trees in which all leaves have the same distance from the root), its radius tends to 0 as the number of taxa increases. This result has relevance for the practice of hierarchical clustering, where the ultrametric constraint is usually required, but it does not immediately imply anything regarding the radius of OLS + ME, of relevance in phylogenetics.

The present paper concludes the debate regarding the radii of BME and OLS + ME. We prove that (as conjectured Bordewich et al., 2009) BME has radius $\frac{1}{2}$ and that the result by Willson on the relative sensitiveness of OLS + ME can be strengthened: its safety radius actually tends to 0, as n tends to infinity. There is then a large gap between the safety radius of BME (the maximum possible) and that of OLS + ME (the minimum possible), which is consistent with the greater reconstruction accuracy of BME observed in several simulation studies (Saitou and Imanishi, 1989; Kumar, 1996; Gascuel, 2000; Desper and Gascuel, 2002).

The paper is organized as follows: after some necessary notational preliminaries (Section 2), we proceed to prove the stated results about BME (Section 3) and OLS + ME (Section 4). In Section 5, we discuss the implications of these results on the heuristics that are used in practice for BME and OLS + ME. Finally, in the Appendix, we discuss the possibility of extending our positive results on the robustness of BME to other ME methods: in particular, we prove that such an extension is not possible for an obvious candidate class generalizing BME.

2. Preliminaries

A *phylogenetic tree* over a set X is a tree (in the graph-theoretic sense) whose leaves are bijectively labeled by the elements of X, called *taxa*. Here, for convenience of notation, we assume $X = \{1, 2, \ldots, n\}$. A *weighted phylogenetic tree* is a phylogenetic tree whose branches (edges) are assigned *lengths* (usually nonnegative) typically representing the amount of evolution that has occurred between their endpoints. A phylogenetic tree that is not weighted is also called a *topology* and we say that a weighted phylogenetic tree $\mathcal{T}$ has topology T if $\mathcal{T}$ can be obtained from T by assigning lengths to the branches of T. Note that, in this paper, calligraphic letters (such as $\mathcal{T}$ and $\mathcal{W}$) denote weighted phylogenetic trees and their corresponding italic symbols (such as T and W) denote their topologies. The length of branch e in a tree $\mathcal{T}$ is denoted by $\ell_e^{\mathcal{T}}$. The length of the shortest branch in $\mathcal{T}$ is denoted by $\ell_{\min}^{\mathcal{T}}$.

A phylogenetic tree is *rooted* if one of its nodes is designated as its *root*, representing the ancestor of all the taxa in the tree. Except in one occasion (Lemma 8), we assume that the trees we deal with do not have a root. A phylogenetic tree is said to be *bifurcating* if every internal (i.e., nonleaf) node has degree three, with the exception of the root, if there is one, which is required to have degree two. For simplicity, in this paper we only deal with bifurcating trees.

Every branch e of a phylogenetic tree over X induces a bipartition of X consisting of the two sets of taxa in the two connected components obtained by deleting e. The bipartitions $\{A, B\}$ induced in this way by the branches of a phylogenetic tree are called its *splits* and are denoted by $A|B$. A *clade* of a phylogenetic tree over X is a subset $A \subseteq X$ such that $A|(X \setminus A)$ is one of the tree's splits. A classic result in phylogenetics is that the topology of a phylogenetic tree is determined, up to isomorphism, by the set of its splits (Buneman, 1971), or equivalently by the set of its clades. For this reason, we here identify a topology with the set of its splits, which allows us to write propositions such as $A|B \in T$ and $T \neq W$. We also make no distinction between branches of a tree and the splits they induce, which justifies the use of expressions such as $\ell_{A|B}^{\mathcal{T}}$ and $e \in T$ (where e is a branch).

Given a weighted phylogenetic tree $\mathcal{T}$, the *distance in* $\mathcal{T}$ between taxa i and j, denoted $d_{ij}^{\mathcal{T}}$, is the sum of the branch lengths in the path between i and j in $\mathcal{T}$. The matrix containing the distances in $\mathcal{T}$ between each pair of taxa is denoted by $\boldsymbol{d}^{\mathcal{T}} = (d_{ij}^{\mathcal{T}})$. Distance methods for phylogenetic reconstruction are based upon the assumption that a *distance matrix* $\boldsymbol{\delta} = (\delta_{ij})$, somewhat approximating $\boldsymbol{d}^{\mathcal{T}}$ in an unknown tree, is given in input. Here, the only assumptions on $\boldsymbol{\delta}$ will be that, for every i and j in $\{1, 2, \ldots, n\}$, $\delta_{ii} = 0$ and $\delta_{ij} = \delta_{ji} \geq 0$. (Although the δ_{ij} are not, strictly speaking, necessarily distances, the use of this term is standard in phylogenetics.) In the practice of phylogenetic tree reconstruction, the distances $\boldsymbol{\delta}$ are obtained (from molecular or morphological data) with various

techniques aiming to estimate the distances $\boldsymbol{d}^{\mathcal{T}}$ in the "true" evolutionary tree $\mathcal{T}$ that generated the taxa of interest. The aim of distance methods is to reconstruct this unknown "true" tree.

Minimum evolution (ME) is a class of distance methods based upon the two following logical components:

(1) Establish how to assign branch lengths to any given topology T so that the distances $d_{ij}^{\hat{T}}$ in the resulting weighted tree $\hat{T}$ "fit" the distances δ_{ij} given in input.
(2) "Look for" the bifurcating topology T that results in the weighted tree $\hat{T}$ of minimum "length."

Clearly, these two steps are loosely defined. Different ways of defining the terms in quotes lead to different versions of ME.

Although other approaches are possible (Kidd and Sgaramella-Zonta, 1971; Swofford et al., 1996; Gascuel et al., 2001) (differing in the way they deal with negative branch lengths), we here define the *length* of a weighted tree as the sum of all its branches' lengths (Rzhetsky and Nei, 1992). In symbols:

$$\mathcal{L}(\mathcal{T}) = \sum_{e \in \mathcal{T}} \ell_e^{\mathcal{T}}.$$

Here, we study two versions of ME that differ for the chosen approach for the first component, branch length estimation. Possibly the simplest method for this step, called OLS (*ordinary least squares*), is to look for the branch length assignment that minimizes the sum $\sum_{ij}(\delta_{ij} - d_{ij}^{\hat{T}})^2$. In this case, the optimal branch lengths are linear functions of $\boldsymbol{\delta}$ (see, e.g., the book chapter by Desper and Gascuel, 2005). Therefore, also the total length of the fitted tree $\hat{T}$ is a linear function of $\boldsymbol{\delta}$, which we denote by $S^T(\boldsymbol{\delta})$. We call OLS + ME the version of ME using $S^T(\boldsymbol{\delta})$ to estimate the length of T. Formally, OLS + ME aims to reconstruct $T^* = \text{argmin}_T S^T(\boldsymbol{\delta})$, where T ranges over all bifurcating topologies over the input taxa.

The second method for branch length estimation that we study here is the *balanced* method. The branch length estimates, in this case, are optimal with respect to a weighted version of the least squares approach sketched above. We refer the reader to other sources for a more detailed introduction (Pauplin, 2000; Desper and Gascuel, 2004; Semple and Steel, 2004; Gascuel and Steel, 2006). For our purposes, it suffices to know that the length of the fitted tree of bifurcating topology T is given by a simple, again linear, function of $\boldsymbol{\delta}$ (due to Pauplin, 2000):

$$B^T(\boldsymbol{\delta}) := \sum_{i<j} 2^{1-t_{ij}} \delta_{ij}, \tag{1}$$

where t_{ij} indicates the topological distance between i and j, that is, the number of branches in the path between i and j in T. We call BME the version of ME using $B^T(\boldsymbol{\delta})$ to estimate the length of T. Formally, BME aims to reconstruct $T^* = \text{argmin}_T B^T(\boldsymbol{\delta})$, where T ranges again over all bifurcating topologies over the input taxa.

2.1. Characterization of general linear formulae for estimating tree length

The fact that $S^T(\boldsymbol{\delta})$ and $B^T(\boldsymbol{\delta})$ are both linear functions of $\boldsymbol{\delta}$ is not surprising, as it can be shown that optimizing branch lengths with respect to a large class of generalizations of

the least squares approach always results in linear functions of the input distances (Bulmer, 1991). Furthermore, $S^T(\boldsymbol{\delta})$ and $B^T(\boldsymbol{\delta})$ have a number of properties that make them suitable for use in combination with minimum evolution (Willson, 2005a). We briefly summarize these properties here, as some of the statements in the rest of this paper are best expressed in terms of the general class of functions satisfying these properties. We start by defining a generalization of the notion of split metric (Bandelt and Dress, 1992):

Definition 2. Let A and B be disjoint subsets of $\{1, 2, \ldots, n\}$; $\boldsymbol{\sigma}^{A|B}$ is defined by

$$\sigma_{ij}^{A|B} = \begin{cases} 1 & \text{if one between } i \text{ and } j \text{ belongs to } A \text{ and the other to } B, \\ 0 & \text{otherwise.} \end{cases}$$

Note that if $\mathcal{T}$ is a weighted tree with topology T, then

$$\boldsymbol{d}^{\mathcal{T}} = \sum_{A|B \in T} \ell_{A|B}^{\mathcal{T}} \boldsymbol{\sigma}^{A|B}. \tag{2}$$

This follows from the fact that $\sigma_{ij}^{A|B} = 1$ if and only if the branch corresponding to $A|B$ is on the path joining i to j in T.

The most basic requirement of any linear function $L^T(\boldsymbol{\delta})$ for estimating the length of a tree T is that it should give its correct length given perfect data. As remarked below, the functions satisfying this requirement are precisely those in the following class (Willson, 2005a).

Definition 3. Let T be a topology over $\{1, 2, \ldots, n\}$. We denote by $\mathcal{U}(T)$ the class of linear functions L^T of any $n \times n$ distance matrix, such that $L^T(\boldsymbol{\sigma}^{A|B}) = 1$ for every split $A|B$ of T.

Remark. By using the decomposition in (2) and the linearity of L^T it is easy to see that the functions in $\mathcal{U}(T)$ are precisely the linear functions such that $L^T(\boldsymbol{d}^{\mathcal{T}}) = \mathcal{L}(\mathcal{T})$, for any weighted tree $\mathcal{T}$ with topology T.

Note that if we write $L^T(\boldsymbol{\delta}) = \sum_{i<j} c_{ij}^T \delta_{ij}$ then the quantity above can simply be expressed as the sum of the coefficients for pairs of taxa in $A \times B$:

$$L^T\left(\boldsymbol{\sigma}^{A|B}\right) = \sum_{\substack{i \in A \\ j \in B}} c_{ij}^T,$$

where c_{ij}^T with $i > j$ is the same as c_{ji}^T.

The following definition (Willson, 2005a) further restricts the set of linear functions suitable to be used in combination with minimum evolution.

Definition 4. Let T be a topology over $\{1, 2, \ldots, n\}$. We denote by $\mathcal{U}^+(T)$ the class of linear functions $L^T \in \mathcal{U}(T)$ such that $L^T(\boldsymbol{\sigma}^{A|B}) > 1$ for all bipartitions $A|B$ of $\{1, 2, \ldots, n\}$ that are not splits of T.

Remark. Imagine that for every bifurcating topology T we have a function from $\mathcal{U}^+(T)$ that we use to estimate its length. Then when the input distances $\boldsymbol{d}^{\mathcal{T}}$ coincide with those in a weighted bifurcating tree $\mathcal{T}$ with positive branch lengths, minimum evolution correctly identifies the topology of $\mathcal{T}$ as the unique optimal one. This is because every "wrong" bifurcating topology W (i.e., not coinciding with the topology of $\mathcal{T}$) contains at least one split not belonging to the correct topology and, therefore, is such that

$$L^W\left(\boldsymbol{d}^{\mathcal{T}}\right) = \sum_{A|B\in\mathcal{T}} \ell^{\mathcal{T}}_{A|B} L^W\left(\boldsymbol{\sigma}^{A|B}\right) > \mathcal{L}(\mathcal{T}),$$

whereas, as noted before, $L^T(\boldsymbol{d}^{\mathcal{T}}) = \mathcal{L}(\mathcal{T})$ when T is the topology of $\mathcal{T}$.

Minimum evolution using formulae from $\mathcal{U}^+(T)$ is then consistent. As a matter of fact, the proofs of consistency for OLS + ME (Rzhetsky and Nei, 1993) and BME (Desper and Gascuel, 2004) were essentially equivalent to proving that S^T and B^T belong to $\mathcal{U}^+(T)$.

3. BME has safety radius $\frac{1}{2}$

In this section, we prove that BME has a safety radius of $\frac{1}{2}$, the best achievable by any distance method. The proof proceeds as follows: first (Section 3.1, Theorem 5), we derive a sufficient condition on the linear functions L^T guaranteeing that ME has safety radius (at least) α; second (Section 3.2, Corollary 9), we show that this condition is satisfied by B^T with $\alpha = \frac{1}{2}$.

3.1. A general condition for ME to have safety radius α

We here show a condition on the formulae used by ME, providing a minimum guaranteed radius. Although here we only prove that this is a sufficient condition, it can be shown that this is in fact also a necessary condition for ME to have at least safety radius α. This theorem has similarities with results presented by Willson (2005b), which however were only stated for the OLS tree length estimates $S^T(\boldsymbol{\delta})$.

Theorem 5. *Assume that for each bifurcating topology T, we use a linear function $L^T \in \mathcal{U}^+(T)$ to estimate its length, where $L^T(\boldsymbol{\delta}) = \sum_{i<j} c^T_{ij}\delta_{ij}$. Suppose that for any bifurcating topologies T and W over the same set of taxa the following holds*:

$$\sum_{i<j}\left(c^W_{ij}(t_{ij} - w_{ij}) - \alpha\left|c^W_{ij} - c^T_{ij}\right|\right) \geq 0, \tag{3}$$

where t_{ij} and w_{ij} denote the number of branches in the path between i and j in T and W, respectively. Then minimum evolution has safety radius (at least) α.

The proof of Theorem 5 relies on the following two lemmas.

Lemma 6. *Let $L(\boldsymbol{\delta}) = \sum_{i<j} c_{ij}\delta_{ij}$ be a linear function and $\boldsymbol{\delta}$ and $\boldsymbol{\delta}'$ two distance matrices whose components differ by less than $\epsilon > 0$, that is, $\|\boldsymbol{\delta} - \boldsymbol{\delta}'\|_\infty < \epsilon$. Then*

$$\left|L(\boldsymbol{\delta}) - L\left(\boldsymbol{\delta}'\right)\right| < \epsilon \sum_{i<j} |c_{ij}|. \tag{4}$$

Proof:

$$\left|L(\boldsymbol{\delta}) - L\left(\boldsymbol{\delta}'\right)\right| = \left|\sum_{i<j} c_{ij}\left(\delta_{ij} - \delta'_{ij}\right)\right| \leq \sum_{i<j} |c_{ij}|\left|\delta_{ij} - \delta'_{ij}\right| < \epsilon \sum_{i<j} |c_{ij}|.$$

□

Lemma 7. *Let $\mathcal{T}$ be a weighted tree with nonnegative branch lengths and a bifurcating topology T and let W be another bifurcating topology over the same set of taxa $\{1, 2, \ldots, n\}$. Assuming $L^W \in \mathcal{U}^+(W)$, then*

$$L^W\left(\boldsymbol{d}^{\mathcal{T}}\right) - \mathcal{L}(\mathcal{T}) \geq \ell^{\mathcal{T}}_{\min} L^W(\boldsymbol{t} - \boldsymbol{w}),$$

where $\boldsymbol{t} = (t_{ij})$ and $\boldsymbol{w} = (w_{ij})$ contain the number of branches in the paths between i and j in T and W, respectively.

Proof: Using the decomposition of $\boldsymbol{d}^{\mathcal{T}}$ given by (2), we obtain

$$L^W\left(\boldsymbol{d}^{\mathcal{T}}\right) - \mathcal{L}(\mathcal{T}) = \sum_{A|B \in T} \ell^{\mathcal{T}}_{A|B}\left(L^W\left(\boldsymbol{\sigma}^{A|B}\right) - 1\right).$$

Since $L^W \in \mathcal{U}^+(W)$, each $L^W(\boldsymbol{\sigma}^{A|B}) - 1$ term above is nonnegative. This, together with $\ell^{\mathcal{T}}_{A|B} \geq \ell^{\mathcal{T}}_{\min} \geq 0$, implies

$$\begin{aligned} L^W\left(\boldsymbol{d}^{\mathcal{T}}\right) - \mathcal{L}(\mathcal{T}) &\geq \ell^{\mathcal{T}}_{\min} \sum_{A|B \in T} \left(L^W\left(\boldsymbol{\sigma}^{A|B}\right) - 1\right) \\ &= \ell^{\mathcal{T}}_{\min}\left(L^W\left(\sum_{A|B \in T} \boldsymbol{\sigma}^{A|B}\right) - (2n-3)\right). \end{aligned}$$

Now note that $\boldsymbol{t} = \sum_{A|B \in T} \boldsymbol{\sigma}^{A|B}$ and that $L^W(\boldsymbol{w}) = 2n - 3$. Then

$$\begin{aligned} L^W\left(\boldsymbol{d}^{\mathcal{T}}\right) - \mathcal{L}(\mathcal{T}) &\geq \ell^{\mathcal{T}}_{\min}\left(L^W(\boldsymbol{t}) - L^W(\boldsymbol{w})\right) \\ &= \ell^{\mathcal{T}}_{\min} L^W(\boldsymbol{t} - \boldsymbol{w}). \end{aligned}$$

□

Proof of Theorem 5: We wish to prove that for every weighted bifurcating tree $\mathcal{T}$ with topology T, if $||\boldsymbol{\delta} - \boldsymbol{d}^{\mathcal{T}}||_\infty < \alpha \ell^{\mathcal{T}}_{\min}$, then $\Delta L(\boldsymbol{\delta}) = L^W(\boldsymbol{\delta}) - L^T(\boldsymbol{\delta}) > 0$, for every bifurcating topology $W \neq T$.

First, application of Lemma 6 with $L = \Delta L$, $\boldsymbol{\delta}' = \boldsymbol{d}^{\mathcal{T}}$ and $\epsilon = \alpha \ell^{\mathcal{T}}_{\min}$, shows that

$$\Delta L(\boldsymbol{\delta}) > \Delta L\left(\boldsymbol{d}^{\mathcal{T}}\right) - \alpha \ell^{\mathcal{T}}_{\min} \sum_{i<j} \left|c^W_{ij} - c^T_{ij}\right|. \tag{5}$$

Using the fact that $L^T \in \mathcal{U}^+(T)$ correctly estimates the length of $\mathcal{T}$ given perfect data, $\Delta L(\boldsymbol{d}^{\mathcal{T}}) = L^W(\boldsymbol{d}^{\mathcal{T}}) - L^T(\boldsymbol{d}^{\mathcal{T}}) = L^W(\boldsymbol{d}^{\mathcal{T}}) - \mathcal{L}(\mathcal{T})$. Lemma 7 shows that a lower bound for this quantity is given by

$$\ell^{\mathcal{T}}_{\min} L^W(\boldsymbol{t} - \boldsymbol{w}) = \ell^{\mathcal{T}}_{\min} \sum_{i<j} c^W_{ij}(t_{ij} - w_{ij}).$$

Plugging this expression into (5), we obtain

$$\Delta L(\boldsymbol{\delta}) > \ell^{\mathcal{T}}_{\min} \sum_{i<j} \left(c^W_{ij}(t_{ij} - w_{ij}) - \alpha \left|c^W_{ij} - c^T_{ij}\right|\right).$$

It is then clear that if the sum in the right-hand side above is nonnegative for every choice of T and W, then ME has safety radius α. □

3.2. The length estimates used by BME satisfy Theorem 5

We now prove that the inequality in Theorem 5 is satisfied for $L^T = B^T$ and $\alpha = \frac{1}{2}$. This will follow as a simple corollary (Corollary 9) of the next lemma, where we need some additional definitions: the *depth* of a taxon i in a rooted topology T is the number of branches in the path, in T, from the root to i; also, a *cherry* is a clade of size 2.

Lemma 8. *Let T and W be bifurcating rooted topologies over $\{1, 2, \ldots, n\}$. Let d_i and d'_i denote the depths of $i \in \{1, 2, \ldots, n\}$ in T and W, respectively. Then*

$$\sum_{i=1}^{n} 2^{-d'_i} \left(d_i - d'_i - \frac{1}{2}\left|2^{d'_i - d_i} - 1\right|\right) \geq 0.$$

Proof: By induction on the number of taxa n. Denote by $f(T, W)$ the sum in the lemma's statement. If $n = 1$ or 2, then T and W must be the same topology and, therefore, $f(T, W) = 0$. For the case $n \geq 3$, we consider two scenarios: either T and W have a common cherry, or they do not.

If T and W have a common cherry, say $\{x, y\}$, then define T' and W' as the trees that are obtained from T and W, respectively, by removing the cherry $\{x, y\}$ and replacing it with a new taxon z (see Fig. 1(a)). Since T' and W' have $n - 1$ taxa, we can apply the induction hypothesis and have that $f(T', W') \geq 0$. The difference $f(T, W) - f(T', W')$ is easy to calculate, as the two sums only differ for the terms corresponding to taxa x, y, and z. Calling d and d' the depths of x (and, therefore, y) in T and W, respectively, and noting that therefore the depths of z in T' and W' equal $d - 1$ and $d' - 1$ (see Fig. 1(a)), we have that

$$\begin{aligned} f(T, W) - f\left(T', W'\right) &= 2 \cdot 2^{-d'} \left(d - d' - \frac{1}{2}\left|2^{d' - d} - 1\right|\right) \\ &\quad - 2^{-(d'-1)} \left(d - d' - \frac{1}{2}\left|2^{d' - d} - 1\right|\right) = 0. \end{aligned}$$

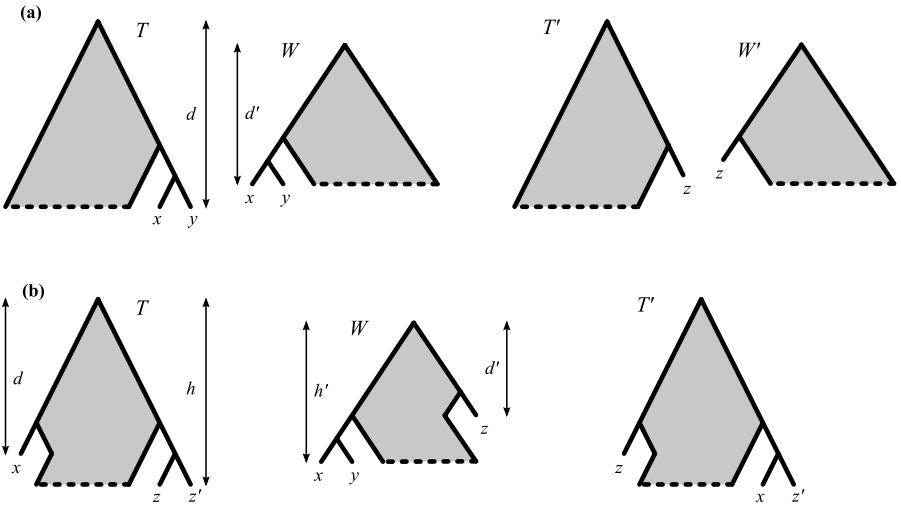


Fig. 1 Topologies in the proof of Lemma 8.

Therefore, $f(T, W) = f(T', W') \geq 0$.

If T and W do not have a common cherry, let $\{x, y\}$ be a cherry with maximum depth in W. We will show that if we swap taxa in T so as to form $\{x, y\}$ in the resulting tree T'', this tree is such that $f(T, W) \geq f(T'', W)$. Since T'' and W do have a common cherry, for the arguments above, we have that $f(T'', W)$—and, therefore, $f(T, W)$—is nonnegative.

Let $\{z, z'\}$ be a cherry with maximum depth in T. Because T and W do not have common cherries, at least one of z and z' must not be in $\{x, y\}$. Without loss of generality, assume $z \notin \{x, y\}$. Swap x and z in T and call the resulting tree T' (see Fig. 1(b)). Let d and h be the depths in T of x and z, respectively, and note that this implies that, in T', x and z have depths h and d, respectively. Also, let d' and h' be the depths in W of z and x, respectively.

We now show that $f(T, W) \geq f(T', W)$. The two sums only differ for the terms corresponding to x and z:

$$\begin{aligned} f(T, W) - f\big(T', W\big) &= 2^{-h'}\left(d - h' - \frac{1}{2}\left|2^{h'-d} - 1\right|\right) \\ &\quad - 2^{-h'}\left(h - h' - \frac{1}{2}\left|2^{h'-h} - 1\right|\right) \\ &\quad + 2^{-d'}\left(h - d' - \frac{1}{2}\left|2^{d'-h} - 1\right|\right) \\ &\quad - 2^{-d'}\left(d - d' - \frac{1}{2}\left|2^{d'-d} - 1\right|\right) \\ &= \big(2^{-d'} - 2^{-h'}\big)(h - d) + \frac{1}{2}\big(\left|2^{-h} - 2^{-h'}\right| - \left|2^{-d} - 2^{-h'}\right| \\ &\quad + \left|2^{-d} - 2^{-d'}\right| - \left|2^{-h} - 2^{-d'}\right|\big). \end{aligned}$$

Using the fact that $|x - y| = x + y - 2\min\{x, y\}$, the expression above simplifies into

$$\begin{aligned}&\left(2^{-d'} - 2^{-h'}\right)(h - d) - \min\left\{2^{-h}, 2^{-h'}\right\} + \min\left\{2^{-d}, 2^{-h'}\right\} \\ &\quad - \min\left\{2^{-d}, 2^{-d'}\right\} + \min\left\{2^{-h}, 2^{-d'}\right\}.\end{aligned}$$

Note that if $h = d$ then $f(T, W) - f(T', W) = 0$. We then assume $h \geq d + 1$ (remember that h and d are integers, as they represent depths in T).

Considering the three cases $d \leq d' \leq h'$, $d' \leq d \leq h'$ and $d' \leq h' \leq d$, it is easy to see that

$$\min\left\{2^{-d}, 2^{-h'}\right\} - \min\left\{2^{-d}, 2^{-d'}\right\} \geq 2^{-h'} - 2^{-d'}.$$

Similarly, considering the three possible positions of h relative to d' and h', it is easy to see that

$$\min\left\{2^{-h}, 2^{-d'}\right\} - \min\left\{2^{-h}, 2^{-h'}\right\} \geq 0.$$

Therefore,

$$\begin{aligned}f(T, W) - f\left(T', W\right) &\geq \left(2^{-d'} - 2^{-h'}\right)(h - d) + 2^{-h'} - 2^{-d'} \\ &= \left(2^{-d'} - 2^{-h'}\right)(h - d - 1) \geq 0.\end{aligned}$$

This completes the proof that $f(T, W) \geq f(T', W)$. Now T' has $\{x, z'\}$ as a cherry. If $z' = y$, then define $T'' = T'$; otherwise let T'' be obtained by swapping z' with y in T'. In any case, by the same arguments as above, $f(T', W) \geq f(T'', W)$.

Since T'' has a cherry in common with W, we have that $f(T'', W) \geq 0$. But then,

$$f(T, W) \geq f\left(T', W\right) \geq f\left(T'', W\right) \geq 0. \qquad \square$$

Remark. The inequality of Lemma 8 holds more generally for any integer sequences (d_i) and (d_i') of equal length for which $\sum_i 2^{-d_i} = \sum_i 2^{-d_i'} = 1$, because such sequences coincide precisely with the taxon depths in two suitably defined bifurcating rooted trees (not shown).

Corollary 9. *Let T and W be bifurcating topologies over $\{1, 2, \ldots, n\}$. Let t_{ij} and w_{ij} denote the number of branches in the paths between i and j in T and W, respectively. Then*

$$\sum_{i<j} 2^{1-w_{ij}} \left(t_{ij} - w_{ij} - \frac{1}{2}\left|2^{w_{ij}-t_{ij}} - 1\right|\right) \geq 0.$$

Proof: The above sum can be reexpressed in the following way:

$$\frac{1}{2}\sum_{j=1}^{n}\sum_{\substack{i=1 \\ i\neq j}}^{n} 2^{1-w_{ij}} \left(t_{ij} - w_{ij} - \frac{1}{2}\left|2^{w_{ij}-t_{ij}} - 1\right|\right).$$

Now define T_j and W_j as the rooted topologies that are obtained from T and W by rooting in taxon j and removing the branch at the root. Then

$$\sum_{\substack{i=1\\ i\neq j}}^{n} 2^{1-w_{ij}}\left(t_{ij}-w_{ij}-\frac{1}{2}\left|2^{w_{ij}-t_{ij}}-1\right|\right)=\sum_{\substack{i=1\\ i\neq j}}^{n} 2^{-d_i'}\left(d_i-d_i'-\frac{1}{2}\left|2^{d_i'-d_i}-1\right|\right),$$

where $d_i = t_{ij} - 1$ and $d_i' = w_{ij} - 1$ are the depths of i in T_j and W_j, respectively. Because of Lemma 8, this sum is nonnegative and, therefore, the whole sum in the statement—which is a sum of sums with this form—is also nonnegative. □

3.3. Putting it all together

We are now ready to complete the proof of the main result of this section.

Theorem 10. *BME has safety radius* $\frac{1}{2}$.

Proof: Because $B^T \in \mathcal{U}^+(T)$ (Desper and Gascuel, 2004), Theorem 5 provides a sufficient condition for BME to have radius $\frac{1}{2}$. Using (1), it is easy to verify that this condition precisely coincides with that proven in Corollary 9. □

Given the generality of our arguments in Section 3.1—holding for ME using generic formulae in $\mathcal{U}^+(T)$—one can wonder whether our proof can be extended to other formulae apart from the B^T used by BME. However, the next section shows that OLS + ME, which uses $S^T \in \mathcal{U}^+(T)$ to estimate tree length, has safety radius equal to 0. In order to exclude even more strongly the possibility to extend our arguments beyond BME, in the Appendix, we show that even if we only use members of $\mathcal{U}^+(T)$ that share with B^T the property of being linear functions with positive coefficients, ME can have an arbitrarily small safety radius.

4. OLS + ME has safety radius 0

In the example that follows we construct, for any fixed n ($n \geq 4$), two bifurcating topologies T and W over the same set of n taxa, a weighted tree $\mathcal{T}$ with topology T, and distances $\boldsymbol{\delta}$ such that

$$\left\|\boldsymbol{\delta}-\boldsymbol{d}^{\mathcal{T}}\right\|_\infty=\frac{3}{\lfloor\sqrt{n}\rfloor+1}\,\ell^{\mathcal{T}}_{\min}\quad\text{and}\quad S^W(\boldsymbol{\delta})<S^T(\boldsymbol{\delta}).$$

Clearly, this implies that the safety radius of OLS + ME over n taxa cannot be greater than $3/(\lfloor\sqrt{n}\rfloor+1)$. Because this function converges to 0, then also the safety radius of OLS + ME over n taxa tends to 0 as the number of taxa grows. This result completes the work by Willson (2005b), who previously showed that this safety radius is limited above by a function tending to $\frac{1}{4}$, but left open the question of how tight this bound was.

The following example makes use of a standard notation that we now introduce. Let A and B be two disjoint clades. Given the distances in $\boldsymbol{\delta}$, the average distance between A

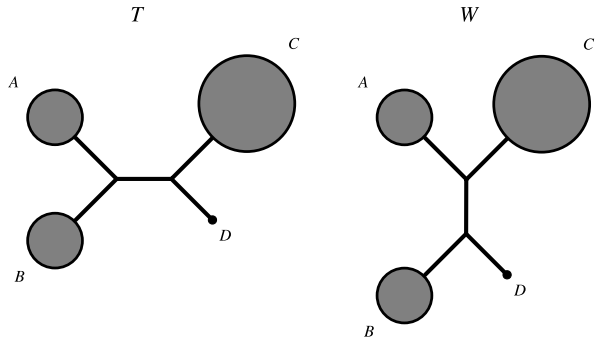


Fig. 2 Topologies referred to in Example 11.

and B is defined by:

$$\delta_{AB} = \frac{1}{|A||B|} \sum_{\substack{i \in A \\ j \in B}} \delta_{ij}.$$

Example 11. Let $\mathcal{T}$ be a weighted bifurcating tree whose branch of minimum length separates clades A and B on one side from C and D on the other side (as in Fig. 2, left). Let $|A| = |B| = \lfloor \sqrt{n} \rfloor$, $|C| = n - 2\lfloor \sqrt{n} \rfloor - 1$ and $|D| = 1$. Let T be the topology of $\mathcal{T}$, and let W be the topology that is obtained by swapping the positions of B and C in T (as in Fig. 2, right).

Now define $\boldsymbol{\delta}$ in the following way:

$$\begin{aligned}
\delta_{ij} &= d^{\mathcal{T}}_{ij} + \epsilon \quad && \text{if } (i, j) \in (A \times B) \cup (C \times D) \cup (A \times D) \cup (B \times C), \\
\delta_{ij} &= d^{\mathcal{T}}_{ij} - \epsilon \quad && \text{if } (i, j) \in (A \times C) \cup (B \times D), \\
\delta_{ij} &= d^{\mathcal{T}}_{ij} && \text{if } (i, j) \in A^2 \cup B^2 \cup C^2 \cup D^2.
\end{aligned}$$

Clearly, $\|\boldsymbol{\delta} - \boldsymbol{d}^{\mathcal{T}}\|_\infty = \epsilon$. If we express the average distances between the clades A, B, C, and D, we then have:

$$\begin{aligned}
\delta_{AB} &= d^{\mathcal{T}}_{AB} + \epsilon, \qquad & \delta_{CD} &= d^{\mathcal{T}}_{CD} + \epsilon, \\
\delta_{AD} &= d^{\mathcal{T}}_{AD} + \epsilon, \qquad & \delta_{BC} &= d^{\mathcal{T}}_{BC} + \epsilon, \\
\delta_{AC} &= d^{\mathcal{T}}_{AC} - \epsilon, \qquad & \delta_{BD} &= d^{\mathcal{T}}_{BD} - \epsilon.
\end{aligned}$$

Desper and Gascuel (2002) have shown how to calculate the difference $\Delta S(\boldsymbol{\delta}) = S^W(\boldsymbol{\delta}) - S^T(\boldsymbol{\delta})$, for a generic $\boldsymbol{\delta}$, when W is obtained from T with a transformation such as that in Fig. 2 (a *nearest neighbor interchange, NNI*). Applying their formula (Eq. (9) in Desper and Gascuel, 2002), we have that

$$\Delta S(\boldsymbol{\delta}) = \frac{1}{2}\big[(1-\lambda)(\delta_{AC} + \delta_{BD}) - \big(1-\lambda'\big)(\delta_{AB} + \delta_{CD}) + \big(\lambda - \lambda'\big)(\delta_{AD} + \delta_{BC})\big], \quad (6)$$

where

$$\lambda = \frac{|A||D| + |B||C|}{(|A| + |B|)(|C| + |D|)} = \frac{1}{2}$$

and

$$\lambda' = \frac{|A||D| + |B||C|}{(|A| + |C|)(|B| + |D|)} = \frac{n\lfloor\sqrt{n}\rfloor - 2\lfloor\sqrt{n}\rfloor^2}{(n - \lfloor\sqrt{n}\rfloor - 1)(\lfloor\sqrt{n}\rfloor + 1)}.$$

For the calculations that follow it suffices to note that, as $\lfloor\sqrt{n}\rfloor^2 \leq n$,

$$\lambda' \geq \frac{n(\lfloor\sqrt{n}\rfloor - 2)}{(n - \lfloor\sqrt{n}\rfloor - 1)(\lfloor\sqrt{n}\rfloor + 1)} > \frac{\lfloor\sqrt{n}\rfloor - 2}{\lfloor\sqrt{n}\rfloor + 1}. \tag{7}$$

Using the expressions for the average distances between A, B, C, and D and (6), we obtain:

$$\begin{aligned}
\Delta S(\boldsymbol{\delta}) &= \frac{1}{2}\big[(1-\lambda)\big(d^{\mathcal{T}}_{AC} + d^{\mathcal{T}}_{BD} - 2\epsilon\big) - \big(1-\lambda'\big)\big(d^{\mathcal{T}}_{AB} + d^{\mathcal{T}}_{CD} + 2\epsilon\big) \\
&\quad + \big(\lambda - \lambda'\big)\big(d^{\mathcal{T}}_{AD} + d^{\mathcal{T}}_{BC} + 2\epsilon\big)\big] \\
&= \frac{1}{2}\big[(1-\lambda)\big(d^{\mathcal{T}}_{AC} + d^{\mathcal{T}}_{BD}\big) - \big(1-\lambda'\big)\big(d^{\mathcal{T}}_{AB} + d^{\mathcal{T}}_{CD}\big) + \big(\lambda - \lambda'\big)\big(d^{\mathcal{T}}_{AD} + d^{\mathcal{T}}_{BC}\big)\big] \\
&\quad + \big[(\lambda - 1) - \big(1-\lambda'\big) + \big(\lambda - \lambda'\big)\big]\epsilon \\
&= \Delta S\big(\boldsymbol{d}^{\mathcal{T}}\big) - 2(1-\lambda)\epsilon \\
&= \Delta S\big(\boldsymbol{d}^{\mathcal{T}}\big) - \epsilon.
\end{aligned}$$

We are now going to express $\Delta S(\boldsymbol{d}^{\mathcal{T}})$. We use the following relationship:

$$d^{\mathcal{T}}_{AC} + d^{\mathcal{T}}_{BD} = d^{\mathcal{T}}_{AD} + d^{\mathcal{T}}_{BC} = d^{\mathcal{T}}_{AB} + d^{\mathcal{T}}_{CD} + 2\ell^{\mathcal{T}}_{\min},$$

which can be proved by noting that $d^{\mathcal{T}}_{ac} + d^{\mathcal{T}}_{bd} = d^{\mathcal{T}}_{ad} + d^{\mathcal{T}}_{bc} = d^{\mathcal{T}}_{ab} + d^{\mathcal{T}}_{cd} + 2\ell^{\mathcal{T}}_{\min}$ holds for any $(a, b, c, d) \in A \times B \times C \times D$. Applying again (6), we then obtain:

$$\begin{aligned}
\Delta S\big(\boldsymbol{d}^{\mathcal{T}}\big) &= \frac{1}{2}\big[(1-\lambda)\big(d^{\mathcal{T}}_{AC} + d^{\mathcal{T}}_{BD}\big) - \big(1-\lambda'\big)\big(d^{\mathcal{T}}_{AB} + d^{\mathcal{T}}_{CD}\big) \\
&\quad + \big(\lambda - \lambda'\big)\big(d^{\mathcal{T}}_{AD} + d^{\mathcal{T}}_{BC}\big)\big] \\
&= \frac{1}{2}\big[\big((1-\lambda) - \big(1-\lambda'\big) + \big(\lambda - \lambda'\big)\big)\big(d^{\mathcal{T}}_{AB} + d^{\mathcal{T}}_{CD}\big) \\
&\quad + 2\big((1-\lambda) + \big(\lambda - \lambda'\big)\big)\ell^{\mathcal{T}}_{\min}\big] \\
&= \big(1-\lambda'\big)\ell^{\mathcal{T}}_{\min}.
\end{aligned}$$

We therefore have

$$\Delta S(\boldsymbol{\delta}) = \big(1-\lambda'\big)\ell^{\mathcal{T}}_{\min} - \epsilon.$$

Using the bound in (7), we then have

$$\Delta S(\boldsymbol{\delta}) < \frac{3}{\lfloor\sqrt{n}\rfloor + 1}\ell^{\mathcal{T}}_{\min} - \epsilon$$

from which it becomes clear that if $\epsilon = \frac{3}{\lfloor\sqrt{n}\rfloor+1}\ell^{\mathcal{T}}_{\min}$, then $S^W(\boldsymbol{\delta}) - S^T(\boldsymbol{\delta}) < 0$, which is what we wanted to show.

5. Discussion

The performance of minimum evolution strongly depends on the chosen method for estimating branch lengths. Not only length estimation determines whether or not ME is consistent (Rzhetsky and Nei, 1993; Gascuel et al., 2001; Desper and Gascuel, 2004; Willson, 2005a), but also whether it is robust to noise in the input data. A measure of this robustness is provided by the safety radius studied here.

Previously, it was shown that OLS estimation causes ME to have radius at most $\frac{1}{4}$ (Willson, 2005b); here, we have shown that this version of ME in fact has safety radius 0 (Section 4). As for BME, the fact that NJ has optimal radius $\frac{1}{2}$ (Atteson, 1999) led some to conjecture that also BME has radius $\frac{1}{2}$ (Bordewich et al., 2009); we have proved this here for the first time (Section 3).

The result on the safety radius of OLS + ME (0) may at first surprise, as it is very different from the safety radius of NJ ($\frac{1}{2}$), an algorithm that was originally designed to approximate OLS + ME (Saitou and Nei, 1987). However, it has previously been noted that further optimizing for OLS + ME the trees produced by NJ actually results in a decrease—not an increase—of their accuracy (Gascuel, 2000; Desper and Gascuel, 2002). This observation, together with the lack of robustness of OLS + ME shown here, demonstrates that the merits of NJ lie in the fact that it is a heuristic for BME, and do not come from its relationship with OLS + ME.

In fact, two heuristics precisely aimed at optimizing OLS + ME have been proposed (Desper and Gascuel, 2002) and it is easy to see that also these heuristics have safety radius (converging to) 0. The first of these heuristics (called GME by Desper and Gascuel, 2002), is based on a *greedy sequential addition* strategy: starting from the only possible topology on the first three taxa, sequentially add each of the remaining taxa by attaching it (with a new terminal branch) onto the branch that minimises the OLS length $S^{T'}(\boldsymbol{\delta})$ of the resulting tree T'. Now imagine applying this greedy algorithm to the distances $\boldsymbol{\delta}$ constructed in Example 11 where taxon n (the last to be added by the algorithm) coincides with the only taxon in clade D. It is then clear that, if ϵ is set to the value derived in that example, this algorithm will not reconstruct the correct topology: even if all the taxa up to $n-1$ are added in the right places, because $S^W(\boldsymbol{\delta}) < S^T(\boldsymbol{\delta})$, taxon n would be added in the position producing the wrong tree W rather than that producing T. The other heuristic for OLS + ME (called FASTNNI; Desper and Gascuel, 2002), consists of repeatedly swapping the positions of clades separated by three branches (i.e., applying an NNI), again greedily choosing at each step the swap that produces the largest decrease in OLS length, until a local minimum is reached. Again this algorithm does not reconstruct the correct topology when applied to Example 11: because W is only one swap away from T and has a smaller OLS length, T is not a local minimum. Since they do not reconstruct

the correct tree when confronted with distances deviating by ϵ (with $\epsilon/\ell^{\mathcal{T}}_{\min} \overset{n\to\infty}{\longrightarrow} 0$) from the correct distances, both these heuristics have safety radius (converging to) 0.

A related question is that of the safety radius of heuristic algorithms for BME. In addition to NJ, two other such heuristics can simply be obtained by adapting to BME the two algorithms described above for OLS + ME. A simple inductive reasoning shows that the greedy sequential addition algorithm for BME has radius $\frac{1}{2}$ (an alternative proof of this result has also been found by R.Mihaescu and M.Bordewich; unpublished, personal communication). Suppose $\|\boldsymbol{\delta} - \boldsymbol{d}^{\mathcal{T}}\|_\infty < \frac{1}{2}\ell^{\mathcal{T}}_{\min}$. If $n = 4$, then all the three possible bifurcating topologies are taken into account by this algorithm, which then returns the topology with the minimum balanced length. Because of Theorem 10, this is precisely the topology of $\mathcal{T}$. If $n > 4$, we can assume by inductive hypothesis that the greedy sequential algorithm applied up to taxon $n-1$ reconstructs the correct topology of $\mathcal{T}$ restricted to the first $n-1$ taxa. The complete topology of $\mathcal{T}$ is therefore one of the possible topologies that can be produced when adding the last taxon, n. Because of Theorem 10, this topology is the one with minimum balanced length and is therefore the one that gets reconstructed.

As for the NNI-based heuristic applied to BME (called BNNI; Desper and Gascuel, 2002), the question of whether this algorithm is consistent (and therefore that of determining its radius) is still open, although empirical evidence leads us to conjecture that consistency does in fact hold.

In conclusion, there seems to be a clear difference in the robustness of methods based on OLS + ME and those based on BME. Interestingly, although the safety radius is here defined in terms of the maximum norm $\|\boldsymbol{\delta} - \boldsymbol{d}^{\mathcal{T}}\|_\infty$, the high sensitivity of OLS + ME and the relative robustness of BME to noise in the input data are independent of the choice of a norm: For instance, whatever the definition of the difference between $\boldsymbol{\delta}$ and $\boldsymbol{d}^{\mathcal{T}}$, the example in Section 4 shows that this difference can become arbitrarily small, with OLS + ME still favoring a wrong tree over the correct one.

BME and OLS + ME therefore have *qualitatively* different properties when confronted with realistic (noisy) data. This may explain the lower reconstruction accuracy of (heuristics for) OLS + ME as compared to (heuristics for) BME, a difference in performance that was observed in several simulation studies (Saitou and Imanishi, 1989; Kumar, 1996; Gascuel, 2000; Desper and Gascuel, 2002).

Acknowledgements

We wish to thank Mike Steel for helpful suggestions about the original manuscript and Radu Mihaescu for proposing an alternative proof for Lemma 8, based on the observation that the sum in its statement can be seen as the difference between the Kullback–Leibler divergence and the total variation distance between the discrete probability distributions defined by $(2^{-d'_i})_i$ and $(2^{-d_i})_i$.

Appendix

Recently, it has been shown that S^T and B^T can be seen as special cases of larger classes of length estimators (Willson, 2005a; Mihaescu and Pachter, 2008). $\mathcal{U}^+(T)$, defined in the Preliminaries, is one of these classes (and, in fact, it can be proven to contain the class by

Mihaescu and Pachter, 2008; result not shown). Given our results of Sections 3 and 4, a reasonable question is whether there is a subclass of $\mathcal{U}^+(T)$, containing B^T but not S^T, such that ME in combination with the members of this subclass is robust to noisy data.

A good candidate for such a subclass is the set of all the linear functions $L^T \in \mathcal{U}(T)$ (i.e., those giving the correct length of a tree given perfect data) having positive coefficients: that is, assuming $L^T(\boldsymbol{\delta}) = \sum_{i<j} c_{ij}^T \delta_{ij}$, such that $c_{ij}^T > 0$ for every $i < j$. It is easy to verify that B^T belongs to this set—as its coefficients are powers of two and therefore always positive—whereas S^T, in general, does not—for example, some of the coefficients of S^T are negative when T is a caterpillar tree with 8 taxa.

The interest in this class comes from an elegant result which we prove now. We say that two clades X and Y of T are *separated* by a branch e if any path in T between a leaf in X and a leaf in Y contains e. Note that clades separated by at least one branch must be disjoint.

Theorem A.1. *Assume that for each bifurcating topology T we use a linear function $L^T \in \mathcal{U}(T)$ to estimate its length, where $L^T(\boldsymbol{\delta}) = \sum_{i<j} c_{ij}^T \delta_{ij}$ and $c_{ij}^T > 0$ for every T and $i < j$. Then minimum evolution using the functions in $\{L^T\}_T$ is consistent.*

Proof: Since $L^T \in \mathcal{U}(T)$, we just need to prove that $L^T(\boldsymbol{\sigma}^{A|B}) > 1$ for all bipartitions $A|B$ of $\{1, 2, \ldots, n\}$ that are not splits of T to conclude that $L^T \in \mathcal{U}^+(T)$ and, therefore, ME is consistent.

Let then $A|B$ be a bipartition of $\{1, 2, \ldots, n\}$ but not a split of T. It is easy to realize that there must be in T four disjoint clades A_1, B_1, A_2, and B_2, such that: (1) A_1 and B_1 are separated by exactly two branches, (2) A_2 and B_2 are separated by exactly two branches, (3) A_1 and A_2 are contained in A and (4) B_1 and B_2 are contained in B (see Fig. 3). It is easy to see that

$$\boldsymbol{\sigma}^{A|B} = \boldsymbol{\sigma}^{A_1|B_1} + \boldsymbol{\sigma}^{A_2|B_2} + \boldsymbol{r},$$

where $\boldsymbol{r} = (r_{ij})$ has some entries such that $r_{ij} = 1$ and the remaining ones such that $r_{ij} = 0$.

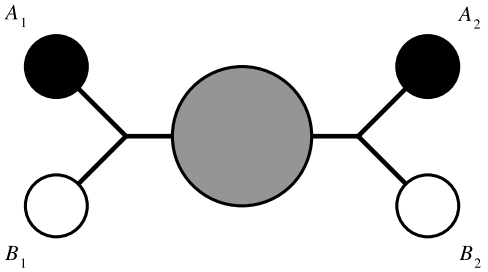


Fig. 3 Illustration of the proof of Theorem A.1: all the taxa in A_1 and A_2 also belong to A, which we denote by coloring these clades in black. All the taxa in B_1 and B_2 also belong to B, which we denote by coloring these clades in white. The rest of the tree is colored in grey to denote that some of its taxa are in A and some in B.

Because L^T is linear,

$$L^T\left(\boldsymbol{\sigma}^{A|B}\right) = L^T\left(\boldsymbol{\sigma}^{A_1|B_1}\right) + L^T\left(\boldsymbol{\sigma}^{A_2|B_2}\right) + L^T(\boldsymbol{r}).$$

Lemma A.2 (proved below) shows that $L^T(\boldsymbol{\sigma}^{A_1|B_1}) = L^T(\boldsymbol{\sigma}^{A_2|B_2}) = \frac{1}{2}$. Because L^T only has positive coefficients and $\boldsymbol{r}$ is binary with some positive entries, we also have that $L^T(\boldsymbol{r}) > 0$. We can then conclude that $L^T(\boldsymbol{\sigma}^{A|B}) > 1$. □

Lemma A.2. *Let T be a bifurcating topology and $L^T \in \mathcal{U}(T)$. For every two clades, A and B, separated by exactly two branches in T,*

$$L^T\left(\boldsymbol{\sigma}^{A|B}\right) = \frac{1}{2}.$$

Proof: Any internal node of a bifurcating tree T is attached to three branches, defining three clades A, B, and C each separated from the other two by exactly two branches. Because $A|B \cup C$, $B|A \cup C$, and $C|A \cup B$ are all splits of T, and since $L^T \in \mathcal{U}(T)$, the following equalities must hold:

$$L^T\left(\boldsymbol{\sigma}^{A|B\cup C}\right) = L^T\left(\boldsymbol{\sigma}^{A|B}\right) + L^T\left(\boldsymbol{\sigma}^{A|C}\right) = 1,$$
$$L^T\left(\boldsymbol{\sigma}^{B|A\cup C}\right) = L^T\left(\boldsymbol{\sigma}^{A|B}\right) + L^T\left(\boldsymbol{\sigma}^{B|C}\right) = 1,$$
$$L^T\left(\boldsymbol{\sigma}^{C|A\cup B}\right) = L^T\left(\boldsymbol{\sigma}^{A|C}\right) + L^T\left(\boldsymbol{\sigma}^{B|C}\right) = 1.$$

Solving the system above for the unknowns $L^T(\boldsymbol{\sigma}^{A|B})$, $L^T(\boldsymbol{\sigma}^{A|C})$, and $L^T(\boldsymbol{\sigma}^{B|C})$ leads to the following solution:

$$L^T\left(\boldsymbol{\sigma}^{A|B}\right) = L^T\left(\boldsymbol{\sigma}^{A|C}\right) = L^T\left(\boldsymbol{\sigma}^{B|C}\right) = \frac{1}{2}.$$

Since any two clades separated by exactly two branches must be in the configuration outlined above, our proof is complete. □

Can then our results on the robustness of BME be extended more generally to all versions of ME using functions from $\mathcal{U}(T)$ with positive coefficients? Unfortunately, the following example shows that if we estimate the length of each tree using an arbitrary element of this class of functions, ME may have an arbitrarily small safety radius.

Example A.3. Let T and W be defined as in Fig. 2 except that now we assume that the four corner clades only contain one taxon each, with $A = \{1\}$, $B = \{2\}$, $C = \{3\}$, $D = \{4\}$. Give the branches of T the following (positive) lengths: $\ell_1, \ell_2, \ell_3, \ell_4$ for the terminal branches incident with $1, 2, 3, 4$, respectively, and ℓ_0 for the internal branch. Call the resulting weighted tree $\mathcal{T}$ and assume $\ell_0 = \ell^{\mathcal{T}}_{\min}$.

Assume ME uses the following formulae to estimate the lengths of T and W:

$$L^T(\boldsymbol{\delta}) = \frac{1}{2}(\delta_{12} + \delta_{34}) + \frac{1}{4}(\delta_{13} + \delta_{14} + \delta_{23} + \delta_{24}),$$
$$L^W(\boldsymbol{\delta}) = L^W_\epsilon(\boldsymbol{\delta}) = \frac{1}{2}(\delta_{13} + \delta_{24}) + \left(\frac{1}{2} - \epsilon\right)(\delta_{12} + \delta_{34}) + \epsilon(\delta_{14} + \delta_{23}),$$

with ϵ being a small positive number. (It is unimportant what formula ME uses for the third bifurcating topology over $\{1, 2, 3, 4\}$.) It is easy to check that $L^T \in \mathcal{U}(T)$, $L^W_\epsilon \in \mathcal{U}(W)$ and that they both have positive coefficients. We now show that by making ϵ arbitrarily small, we can also make the safety radius of ME arbitrarily small.

Let α be such that $0 < \alpha \leq \frac{1}{2}$. Define $\boldsymbol{\delta}$ in the following way:

$$\begin{aligned}
\delta_{12} &= d^{\mathcal{T}}_{12} + \alpha\ell_0 = \ell_1 + \ell_2 + \alpha\ell_0,\\
\delta_{34} &= d^{\mathcal{T}}_{34} + \alpha\ell_0 = \ell_3 + \ell_4 + \alpha\ell_0,\\
\delta_{14} &= d^{\mathcal{T}}_{14} + \alpha\ell_0 = \ell_1 + \ell_4 + (1+\alpha)\ell_0,\\
\delta_{23} &= d^{\mathcal{T}}_{23} + \alpha\ell_0 = \ell_2 + \ell_3 + (1+\alpha)\ell_0,\\
\delta_{13} &= d^{\mathcal{T}}_{13} - \alpha\ell_0 = \ell_1 + \ell_3 + (1-\alpha)\ell_0,\\
\delta_{24} &= d^{\mathcal{T}}_{24} - \alpha\ell_0 = \ell_2 + \ell_4 + (1-\alpha)\ell_0.
\end{aligned}$$

Clearly, $||\boldsymbol{\delta} - \boldsymbol{d}^{\mathcal{T}}||_\infty = \alpha\ell_0$. We also have

$$L^T(\boldsymbol{\delta}) = \ell_1 + \ell_2 + \ell_3 + \ell_4 + (1+\alpha)\ell_0$$

and

$$L^W_\epsilon(\boldsymbol{\delta}) = \ell_1 + \ell_2 + \ell_3 + \ell_4 + (1+2\epsilon)\ell_0.$$

These formulae show that, if $\epsilon < \frac{1}{2}\alpha$, then $L^W_\epsilon(\boldsymbol{\delta}) < L^T(\boldsymbol{\delta})$, that is ME will favor a wrong topology W over the true topology T.

Therefore, the safety radius of ME using $L^T(\boldsymbol{\delta})$ and $L^W_\epsilon(\boldsymbol{\delta})$ (with ϵ sufficiently small) is limited above by α (if the radius were greater than α, ME should return T on the $\boldsymbol{\delta}$ defined above). Since this holds for an arbitrarily small α, we conclude that using ME with correct linear formulae with positive coefficients does not guarantee any positive safety radius.

The example above suggests that in order to guarantee a certain safety radius, the coefficients of the used linear functions must not only be positive, but also "large enough." This is an interesting idea for further research, but outside the scope of the present paper.

References

Atteson, K., 1999. The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25, 251–278.

Bandelt, H.J., Dress, A.W., 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. 1(3), 242–252.

Bordewich, M., Gascuel, O., Huber, K., Moulton, V., 2009. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. IEEE/ACM Trans. Comput. Biol. Bioinf. 6, 110–117.

Bulmer, M., 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. 8(6), 868.

Buneman, P., 1971. The recovery of trees from measures of dissimilarity. In: Hodson, F. (Ed.), Mathematics in the Archaeological and Historical Sciences, pp. 387–395. Edinburgh University Press, Edinburgh.

Cavalli-Sforza, L.L., Edwards, A.W.F., 1964. Analysis of human evolution. In: Proceedings 11th International Congress of Genetics, vol. 3, pp. 923–933.
Cavalli-Sforza, L.L., Edwards, A.W.F., 1967. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. 19(3 part 1), 233–257.
Desper, R., Gascuel, O., 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J. Comput. Biol. 9, 687–705.
Desper, R., Gascuel, O., 2004. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. 21, 587–598.
Desper, R., Gascuel, O., 2005. The minimum evolution distance-based approach to phylogenetic inference. In: Gascuel, O. (Ed.), Mathematics of Evolution & Phylogeny, pp. 1–32. Oxford University Press, Oxford.
Gascuel, O., 2000. On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. Mol. Biol. Evol. 17, 401–405.
Gascuel, O., McKenzie, A., 2004. Performance analysis of hierarchical clustering algorithms. J. Classif. 21(1), 3–18.
Gascuel, O., Steel, M., 2006. Neighbor-joining revealed. Mol. Biol. Evol. 23, 1997–2000.
Gascuel, O., Bryant, D., Denis, F., 2001. Strengths and limitations of the minimum evolution principle. Syst. Biol. 50, 621–627.
Kidd, K.K., Sgaramella-Zonta, L.A., 1971. Phylogenetic analysis: concepts and methods. Am. J. Hum. Genet. 23, 235–252.
Kumar, S., 1996. A stepwise algorithm for finding minimum evolution trees. Mol. Biol. Evol. 13, 584–593.
Mihaescu, R., Pachter, L., 2008. Combinatorics of least squares trees. Proc. Natl. Acad. Sci. USA 105, 13206–13211.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Oxford.
Pauplin, Y., 2000. Direct calculation of a tree length using a distance matrix. J. Mol. Evol. 51, 41–47.
Rzhetsky, A., Nei, M., 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9, 945–967.
Rzhetsky, A., Nei, M., 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. 10, 1073–1095.
Saitou, N., Imanishi, T., 1989. Relative efficiencies of the Fitch–Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. 6, 514–525.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Semple, C., Steel, M., 2004. Cyclic permutations and evolutionary trees. Adv. Appl. Math. 32, 669–680.
Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic inference. In: Hillis, D., Moritz, C., Mable, B. (Eds.), Molecular Systematics, pp. 407–514. Sinauer, Sunderland.
Vinh, L.S., von Haeseler, A., 2005. Shortest triplet clustering: reconstructing large phylogenies using representative sets. BMC Bioinf. 6, 92.
Willson, S.J., 2005a. Consistent formulas for estimating the total lengths of trees. Discrete Appl. Math. 148(3), 214–239.
Willson, S.J., 2005b. Minimum evolution using ordinary least-squares is less robust than neighbor-joining. Bull. Math. Biol. 67(2), 261–279.