COMBINATORICS OF DISTANCE-BASED TREE INFERENCE

FABIO PARDI, OLIVIER GASCUEL

Abstract. Several popular methods for phylogenetic inference (or hierarchical clustering) are based on a matrix of pairwise distances between taxa (or any kind of objects): the objective is to construct a tree with branch lengths so that the distances between the leaves in that tree are as close as possible to the input distances. If we hold the structure (topology) of the tree fixed, in some relevant cases (e.g. ordinary least squares) the optimal values for the branch lengths can be expressed using simple combinatorial formulae. Here we define a general form for these formulae and show that they all have two desirable properties: first, the common tree reconstruction approaches (least squares, minimum evolution), when used in combination with these formulae, are guaranteed to infer the correct tree when given enough data (*consistency*); second, the branch lengths of all the simple (nearest neighbor interchange) rearrangements of a tree can be calculated, optimally, in quadratic time in the size of the tree, thus allowing the efficient application of hill climbing heuristics. The study presented here is a continuation of that by Mihaescu and Pachter on branch length estimation (Mihaescu R, Pachter L (2008) Proc Natl Acad Sci USA 105:13206-13211). The focus here is on the inference of the tree itself, and on providing a basis for novel algorithms to reconstruct trees from distances.

Introduction

A task with several relevant applications is the use of a matrix of distances to construct a tree whose leaves' relative positions somehow reflect the given distances. This is useful both in evolutionary biology, where the tree is intended to represent the evolution of a set of species, populations or genes, and in cluster analysis, where the tree shows the similarities in a collection of objects. In evolutionary biology, the distances are typically estimated from molecular sequences using probabilistic models of sequence evolution [1, 2]. The resulting distances can be expected to be approximately *additive*; that is, there exists a (*phylogenetic*) tree with branch lengths, so that the lengths of the paths between its leaves (sequences) are approximately equal to the input distances. Finding this tree is the goal of several popular *distance-based* tree reconstruction methods.

For phylogenetic reconstruction — which this paper concentrates on — the main advantage of distance-based methods is their speed of execution, which is orders of magnitude faster than that of other (potentially more accurate) approaches. As a consequence, distance methods are used whenever computational efficiency is of critical importance: for the reconstruction of very large trees, or — as in the case of bootstrapping — large collections of trees, or even to construct initial phylogenies for search heuristics based on more sophisticated approaches. In fact, a general trend in bioinformatics and computational biology is the growing demand for methods that can cope with massive datasets of DNA sequences. Distance-based methods are a possible answer to this demand, not only for phylogenetic inference, but also for related tasks such as sequence identification (e.g. in metagenomics) and gene orthology inference (e.g. in functional genomics). A proof of this demand is the continuing success of neighbor-joining (NJ) [3], which to-date remains the most cited algorithm in phylogenetics.

The advantage in speed of distance methods is counterbalanced by a lower accuracy than methods that take full sequence information into account [4], such as maximum likelihood (ML), although it has recently been shown that under a certain measure of statistical efficiency, some distance methods are essentially as good as ML [5]. A limitation of distance-based methods lies in the fact that if the distances are estimated from pairwise sequence comparisons only, then it may be impossible to infer some parameters common to the evolution of all the sequences [6]. However, it is still possible to

estimate these parameters from limited sequence samples by ML, and then use distance methods for the whole sample.

If we consider the estimation of distances as a separate task, virtually all distance methods are based on two components, corresponding to the two main unknowns in a phylogenetic tree: branch lengths and topology. First, (i) we must define a method to assign lengths to the branches of any tree of fixed topology, so that the distances between leaves are as close as possible to the input distances. Second, (ii) we must choose a criterion to discriminate among the trees with different topologies obtained with the step above. Distance-based algorithms then look for the tree that optimizes this criterion, typically using heuristics such as successive agglomeration (e.g. NJ [3]) and hill climbing (e.g. FastME [7]).

For component (i), a *weighted least squares* (WLS) approach is usually adopted: the lengths of the branches in a tree T are set to the values that minimize

$$\sum_{i,j} w_{ij}(\delta_{ij} - d_{ij}^T)^2, \tag{1}$$

where the δ_{ij} are the distance estimates, the d_{ij}^T are the distances between the leaves of T, determined by the lengths assigned to its branches, and the weights $w_{ij} > 0$ are intended to account for the variances of the δ_{ij}: the higher the variance, the lower the weight and the importance given to the corresponding residual $\delta_{ij} - d_{ij}^T$. Ideally, w_{ij} should be proportional to $\mathrm{Var}[\delta_{ij}]^{-1}$ (see *Relationship with WLS and the M&P formulae* below), but in practice setting the weights is a delicate art, because the variances are not known; for example, one trap to avoid is to assume zero variance (and therefore an infinite weight) for the distance between two identical sequences [8]. An even more ideal approach would be to also consider the covariances between distances for different pairs of taxa, which leads a *generalized least squares* (GLS) optimization criterion [9]. The optimal branch lengths with respect to (1), and even GLS, can be expressed succinctly in matrix form. However, despite some progress [10], the matrix calculations involved are computationally expensive and remain a limiting factor for the efficiency of the algorithms that use them (such as those implemented in PAUP* [11]).

As for component (ii), distance methods fall into two broad categories: (*pure*) *least squares* (LS) methods [12, 13] use again a least squares criterion such as (1) to score trees; on the other hand, *minimum evolution* (ME) methods [14, 15] aim to find the tree with minimum total length (which can be defined in a number of different ways [14, 15, 8], see *Statistical consistency* for details), among those whose branch lengths are fitted with component (i). The intuition underlying ME is the same as that of maximum parsimony for character-based tree reconstruction: simpler (i.e. shorter) explanations are preferable to more complicated ones.

An important realization has been that in some relevant cases the branch lengths that minimize (1) can be expressed using simple "combinatorial" formulae, which allow to avoid slow matrix calculations. The best-known cases are that with constant weights w_{ij} (*ordinary least squares*, OLS) [16, 17] and that with weights proportional to $2^{-t_{ij}}$, where t_{ij} is the number of branches in the path between i and j in T (the *balanced* case) [18]. These formulae allow to efficiently calculate branch lengths and to efficiently update the tree length while performing a local search for the optimal tree with respect to ME. For example, the balanced branch length formulae [19] can be used to calculate in $O(n^2)$ time, for any tree with n leaves, not only all its branch lengths, but also the total lengths of all of its NNI (*nearest neighbor interchange*) [7] and SPR (*subtree pruning and regrafting*) rearrangements [20, 21].

A key work on such combinatorial formulae for least squares branch lengths has appeared recently [22]. The authors show that all the known formulae are particular cases of a more general framework: whenever the weights w_{ij} (or, equivalently, the assumed variances w_{ij}^{-1}) have a particular "multiplicative" form (see *Relationship with WLS and the M&P formulae*), then the optimal branch lengths with respect to (1) can be calculated using simple formulae — such as those for OLS or the balanced case — which here we refer to as the *M&P formulae* (from the authors Mihaescu and Pachter or the word "multiplicative"). The multiplicative model is biologically and mathematically meaningful, as it can be

shown that the variances of the distance estimates are approximately multiplicative for large distances [23, 9, 22].

In the following, we use the seminal work by Mihaescu and Pachter [22] as a starting point. Whereas these authors focused on the problem of branch length estimation, here we switch the focus to tree reconstruction itself — namely the statistical and algorithmic consequences of the use of combinatorial branch length formulae on tree reconstruction. Our results can be summarized as follows:

(1) We define a class of formulae for fitting branch lengths that generalizes the M&P formulae and consequently also all known combinatorial formulae.
(2) We prove the statistical consistency of the main distance-based tree reconstruction principles (LS and ME), when combined with our formulae. In other words the optimal tree with respect to any of these principles converges to the correct tree as the input data become more and more abundant and the estimated distances converge to their correct values. Particularly in the case of ME, where it is problematic, this issue has received much attention (e.g., [17, 24, 25, 26, 27, 21]). This addresses the question by Mihaescu and Pachter ([22], p.13211) of "what classes of semimultiplicative [a minor generalization of multiplicative] variance matrices result in consistent tree estimates", by showing that *all* multiplicative variance matrices have this property.
(3) We investigate the computational efficiency of local search heuristics in combination with our class of formulae. In particular, we describe an algorithm that calculates the branch lengths determined by the adopted formulae not only for a fixed tree T, but also for all trees obtained by performing one NNI on T. The entire calculation optimally requires $O(n^2)$ time. This algorithm can be used as the basic component for local searches, and can be combined with any classic tree reconstruction principle.

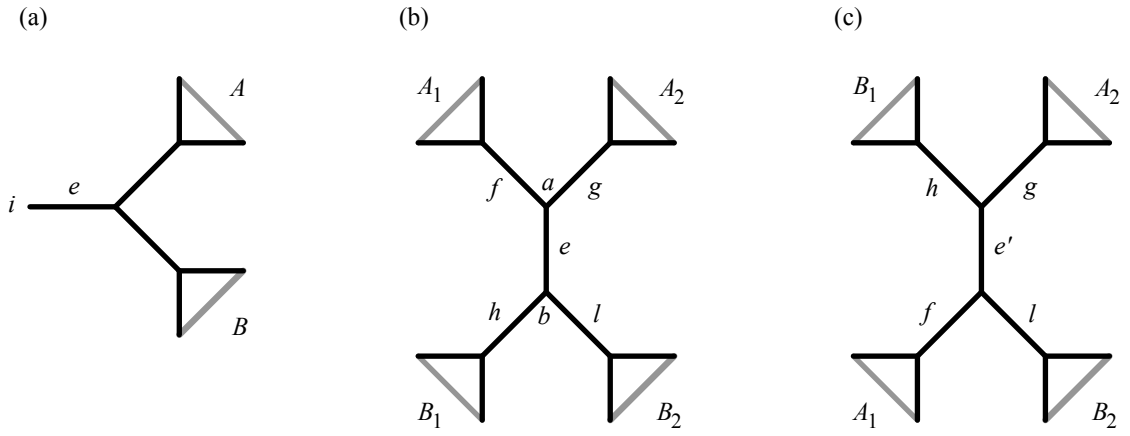
Preliminaries: Branch length formulae

We employ the standard terminology used in the phylogenetics literature [4, 28] (phylogenetic tree, topology, branch lengths, internal and external branches etc.). For simplicity, we identify the leaves of a phylogenetic tree with a set of taxa $\{1, 2, \ldots, n\}$ and we choose to consider only binary trees. We say that two subsets of taxa A and B in a tree are *separated* by a branch e if any path between an element of A and an element of B passes through e. A and B are *k-separated* when they are separated by exactly k distinct branches. A proper subset of taxa $A \subsetneq \{1, 2, \ldots, n\}$ is a *clade* if A and $\{1, 2, \ldots, n\} \setminus A$ are separated by some branch e; in fact, e is unique and is called the *root branch* of A; the endpoint of e to the side of A is called the *root node* of A. A branch *belongs to* clade A if it lies in the path between two elements of A.

We also adopt the following standard conventions for distance-based methods: $\boldsymbol{\delta}$ denotes the $n \times n$ input distance matrix and δ_{ij} its element expressing the distance between taxa i and j (in the following, indices i and j are always assumed to be elements of the set of taxa $\{1, 2, \ldots, n\}$). The distances do not necessarily form a metric, as only $\delta_{ij} = \delta_{ji}$ and $\delta_{ii} = 0$ are assumed. Given a tree T with branch lengths, $\boldsymbol{d}^T$ denotes the distance matrix where d^T_{ij} coincides with the length of the path between i and j in T. When $\boldsymbol{\delta} = \boldsymbol{d}^T$ for some T, we say that $\boldsymbol{\delta}$ is *additive* (with respect to T) [29].

In the rest of this section, we introduce a new class of formulae that express the branch lengths of a generic topology T over $\{1, 2, \ldots, n\}$ as a function of $\boldsymbol{\delta}$. This class is parameterized by some quantities that we present using a probabilistic interpretation (see SI Appendix 1 for more details). Let T be a binary tree topology. Assume that the rules for a random walk on T are defined in the following way: if we enter an internal node from a branch e, we can then exit this node from its two other adjacent branches, f and g, with probabilities γ_{ef} and $\gamma_{eg} = 1 - \gamma_{ef}$, respectively. We require $0 < \gamma_{ef}, \gamma_{eg} < 1$ (note the strict inequalities). These parameters define a (non-zero) probability of reaching any branch of T from any other branch of T.

FIGURE 1. Standard naming of clades and branches when (a) e is external, and (b) e is internal. (c) An NNI-neighbor (around e) of the tree in (b).



This also defines a probability distribution over the leaves of any clade A: for any $i \in A$, let $p_{i|A} = \gamma_{e_0e_1} \cdot \gamma_{e_1e_2} \cdot \ldots \cdot \gamma_{e_{k-1}e_k}$, where e_0 is the root branch of A and $e_1, e_2, \ldots, e_k$ are the branches on the path between the root of A and i. (See SI Appendix 1 for a figure illustrating this.) Clearly, the probabilities $\{p_{i|A} \,|\, i \in A\}$ form a distribution over A. We can then define the average distance δ_{AB} between any two clades A and B as the expected distance between two taxa chosen at random from A and B according to the distributions defined above:

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

Note that the δ_{AB} so defined depend on the γ_{ef} parameters, as well as on the underlying topology T, but for simplicity we do not indicate this in the chosen formalism. Also note that $\delta_{AB} = \delta_{BA}$. For simplicity, we write δ_{iA} (or δ_{Ai}) instead of $\delta_{\{i\}A}$.

In addition to the γ_{ef} probabilities defined for each pair of adjacent branches (e, f), we introduce a parameter λ_{XY} for each unordered pair $\{X, Y\}$ of 3-separated clades in T (recall the definition of k-separated clades above). We constrain these parameters so that, for every internal branch separating clades $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$ (see Fig. 1(b)), $\lambda_{A_1B_1} = \lambda_{A_2B_2} > 0$, $\lambda_{A_1B_2} = \lambda_{A_2B_1} > 0$ and $\lambda_{A_1B_1} + \lambda_{A_1B_2} = 1$, meaning that only one parameter among $\lambda_{A_1B_1}$, $\lambda_{A_2B_2}$, $\lambda_{A_1B_2}$ and $\lambda_{A_2B_1}$ determines all the others. A possible interpretation for λ_{XY} is as the probability of drawing T so that X and Y are consecutive in a clockwise ordering of the taxa (which explains why $\lambda_{A_1B_1} = \lambda_{A_2B_2}$, $\lambda_{A_1B_2} = \lambda_{A_2B_1}$ and $\lambda_{A_1B_1} + \lambda_{A_1B_2} = 1$; see SI Appendix 1 for details).

In summary, we have three free parameters per internal node of T (γ_{ef}, γ_{fg} and γ_{ge} determine γ_{eg}, γ_{fe} and γ_{gf}) and one free parameter per internal branch ($\lambda_{A_1B_1}$ determines $\lambda_{A_2B_2}, \lambda_{A_1B_2}$ and $\lambda_{A_2B_1}$). These parameters determine a set of formulae to estimate the length $\hat{\ell}_e$ of any branch in T:

$(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae. *Let the vectors $\boldsymbol{\gamma}^T = (\gamma_{ef})$ and $\boldsymbol{\lambda}^T = (\lambda_{XY})$ be defined for binary topology T, under the constraints described above. Then, for any branch e in T:*

$$\hat{\ell}_e(\boldsymbol{\delta}) = \begin{cases} \frac{1}{2}(\delta_{iA} + \delta_{iB} - \delta_{AB}) & \text{if } e \text{ is external,} \\ \frac{1}{2}\left[\lambda_{A_1B_1}(\delta_{A_1B_1} + \delta_{A_2B_2}) + (1 - \lambda_{A_1B_1})(\delta_{A_1B_2} + \delta_{A_2B_1}) - \delta_{A_1A_2} - \delta_{B_1B_2}\right] & \text{if } e \text{ is internal,} \end{cases}$$

where, if e is external, we define A, B, i as in Fig. 1(a) and, if e is internal, we define A_1, A_2, B_1, B_2 as in Fig. 1(b).

Note that because $\lambda_{A_1B_2} = \lambda_{A_2B_1} = 1 - \lambda_{A_1B_1} = 1 - \lambda_{A_2B_2}$, the formula above for internal branch lengths is (as desired) independent of how we assign names A_1, A_2 to the two subclades of $A = A_1 \cup A_2$,

and how we assign B_1, B_2 to the two subclades of $B = B_1 \cup B_2$. An interpretation of the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae as averages of simpler formulae is given in SI Appendix 1.

These formulae are a generalization of all the combinatorial formulae proposed in the past to fit the branch lengths of a tree of fixed topology. In particular, the OLS branch lengths [16, 17] can be obtained by setting $\lambda_{A_1B_1} = \frac{|A_1||B_2|+|A_2||B_1|}{|A||B|}$ (same clade naming as above), and by setting $\gamma_{ef} = \frac{|A_1|}{|A_1|+|A_2|}$, for every pair of adjacent branches e and f in the configuration of Fig. 1(b). Note that the γ_{ef} parameters thus defined ensure that $\{p_{i|X} \,|\, i \in X\}$ is uniform for any clade X (in fact the word "unweighted" is often associated to OLS). Similarly, the balanced branch lengths [19] at the basis of the balanced minimum evolution principle [7, 18, 30] are obtained by setting all parameters to $\frac{1}{2}$. The next section shows that the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae also generalize the M&P formulae by Mihaescu and Pachter [22]. It is easy to see that the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae still satisfy the independence of irrelevant pairs (IIP) property introduced by those authors [22] as a basic requirement for their formulae.

Finally, we show that the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae above are *correct*, that is, they calculate the correct values of the branch lengths of any given tree whenever the distances are additive with respect to that tree (proof in SI Appendix 1). Naturally, because the input distances are only estimates of the real evolutionary distances, they are usually only approximately additive. However, this property is an important prerequisite of any branch length formula, as it ensures the statistical consistency of the branch lengths assigned to the correct topology (see *Statistical consistency* below).

Theorem 1. *Let T be a binary topology. For any given branch e in T, assign length ℓ_e to e, and let $\boldsymbol{\delta}$ be additive with respect to the resulting tree. Let $\hat{\ell}_e(\boldsymbol{\delta})$ be the length that is assigned to e by a $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formula. Then, $\hat{\ell}_e(\boldsymbol{\delta}) = \ell_e$.*

Relationship with WLS and the M&P formulae

The choice of the weights in (1) is a key factor for the accuracy of least squares tree estimates. The weights w_{ij} should be proportional to $\mathrm{Var}[\delta_{ij}]^{-1}$, as this implies that the branch lengths that minimize (1) have minimum variance among all linear unbiased estimators of the branch lengths (under the assumption that the distance estimates are unbiased and uncorrelated for different pairs of taxa) [31]. In this section, we consider the case where the weights (and therefore the assumed variances) are "multiplicative": given a tree topology T and a collection of weights $\boldsymbol{w} = (w_{ij})$ associated to pairs of taxa in T, we say that these weights are *multiplicative with respect to* T, if we can assign to each branch e of T a weight $w_e > 0$, so that, for every pair of taxa i and j, $w_{ij} = \prod_{e \in P_{ij}(T)} w_e$, where $P_{ij}(T)$ denotes the set of branches in the path between i and j in T. This condition generalizes several well-known cases: that of constant weights (coinciding with OLS and obtained by setting w_e to 1 for internal branches and to a constant for external ones), that of taxon-specific weights [26] (obtained like for OLS but with w_e free to vary for external branches) and also that of weights exponentially related to the number of branches separating each pair of taxa (which, when the base of the exponent is $b = 1/2$, coincide with the balanced weights [19] and are obtained by setting $w_e = b$ for internal branches and to a constant for external ones).

Mihaescu and Pachter [22] have shown that if the assumed weights $\boldsymbol{w}$ are multiplicative with respect to T, then the optimal branch lengths of T with respect to the WLS criterion (1) are given by their M&P formulae. We refer to SI Appendix 2 for a description of these formulae. The following theorem shows that the class of the M&P formulae is contained in that of the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae and, conversely, it characterizes the values of $\boldsymbol{\gamma}^T$ and $\boldsymbol{\lambda}^T$ corresponding to M&P formulae.

Theorem 2. *Let T be a binary topology. (i) Given any $\boldsymbol{w}$ multiplicative w.r.t. T, the corresponding M&P formulae are also $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae for some choice of $\boldsymbol{\gamma}^T$ and $\boldsymbol{\lambda}^T$ satisfying the properties* P1 *and* P2 *below. (ii) Given any $\boldsymbol{\gamma}^T$ and $\boldsymbol{\lambda}^T$ satisfying the properties* P1 *and* P2 *below, the corresponding $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae are also M&P formulae for some choice of $\boldsymbol{w}$, multiplicative w.r.t. T.*

P1. *For every internal node of T, if e, f and g are the three branches incident to it, then*

$$\gamma_{ef}\gamma_{fg}\gamma_{ge} = (1-\gamma_{ef})(1-\gamma_{fg})(1-\gamma_{ge}).$$

P2. *For every pair of clades A and B separated in T by three branches a, e and b (with a being the root branch of A, and b being the root branch of B), $\lambda_{AB} = \gamma_{ea} + \gamma_{eb} - 2\gamma_{ea}\gamma_{eb}$.*

Theorem 2, proved in SI Appendix 2, not only shows that the M&P formulae are particular types of $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae, but it also provides an alternative set of parameters to represent the M&P formulae: instead of the branch-associated weights w_e, one can use a set of γ_{ef} parameters satisfying P1. This condition implies that any of γ_{ef}, γ_{fg} and γ_{ge} can be determined from the other two and P2 implies that all the λ_{XY} parameters are determined by the γ_{ef} parameters. This reduces the number of free parameters needed to describe the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae that are also M&P formulae to 2 per internal node, that is $2n-4$. This is exactly one less than the $2n-3$ branch-associated parameters w_e describing multiplicative weightings, which corresponds to the fact that multiplying all the w_e for external branches by any positive constant results in equivalent weightings with respect to (1).

Theorem 2 establishes that the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae have enough "expressive power" to optimize the least squares criterion (1), when the weights are multiplicative. It is therefore important to discuss this assumption. First, multiplicative weights generalize the balanced weights, which have been experimentally demonstrated to behave well in combination with ME [18, 32]. Second, in the case of distances estimated from molecular sequences, we note that for many models of sequence evolution (for instance Jukes-Cantor [33]; see Chapter 13 in [4] or Appendix B in [2] for the general technique), the variance of δ_{ij} can be approximated by a function of the correct evolutionary distance d_{ij} that, for small values of d_{ij}, behaves as a linear function of d_{ij}, and, for moderate-to-large d_{ij}, as an exponential of d_{ij}. This means that, for pairs of taxa separated by small d_{ij}, the variances of their distance estimates will tend to be additive, whereas for pairs of taxa separated by moderate-to-large d_{ij}, the variances will tend to be multiplicative. The additive model for the variances [34], or its variant with variances proportional to d_{ij}^2 [13], are used in practice with δ_{ij} in place of d_{ij}, as the latter is unknown. As a result, these approaches need some precautions for very small distance estimates, so as to avoid an overconfidence in these estimates (for δ_{ij} tending to 0, also the assumed variance tends to 0, and w_{ij} tends to infinity): for example, one possibility is to add pseudocounts to the numbers of observed differences between sequences [8] (known as "Laplace smoothing"). In this context, the multiplicative model provides a simple and robust alternative for small distances (for $\delta_{ij} \to 0$, the assumed variance tends to a constant), and is mathematically justified for moderate-to-large distances.

The other important assumption here, common to all WLS methods, is that the δ_{ij} are uncorrelated for different pairs of taxa, which is clearly not true for distances estimated from molecular sequences [9]. As mentioned above, covariances between different distance estimates can be accounted for by adopting a GLS criterion. However, setting the covariances and calculating the resulting branch lengths [10] are difficult problems, which explains the lack (to the best of our knowledge) of practical implementations of GLS for phylogenetic reconstruction.

Statistical consistency

A method for phylogenetic inference is said to be *(statistically) consistent* if the probability that it reconstructs the correct tree (within any given accuracy) converges to 1 as more and more data are analyzed. For distance-based methods, the consistency of tree inference usually depends in turn on the consistency of the distance estimates, that is, the assumption that $\boldsymbol{\delta}$ converges to a matrix $\boldsymbol{d}^T$ containing the distances in the correct phylogenetic tree for the taxa under consideration. Even though in reality the precise consistency of distance estimates cannot be expected to hold — because the models used to obtain these estimates are only approximations of reality — the ability to infer the correct tree in such a best-case scenario is an essential property of any phylogenetic inference method:

it is a prerequisite for robust inference of the correct topology with real distance estimates, subject to sampling errors and not perfectly consistent [24, 35, 36].

In this section, we state our main results on the statistical consistency of the tree reconstruction methods using the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae. We leave the proofs to SI Appendix 3. We assume that, for any binary topology T over the taxa of interest $\{1, 2, \ldots, n\}$, a collection of parameters $\boldsymbol{\gamma}^T = (\gamma_{ef})$ and $\boldsymbol{\lambda}^T = (\lambda_{XY})$ is defined, thus defining in turn, for any such T, a set of $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae for estimating the branch lengths of T. We call this a *branch length estimation scheme based on* $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$*-formulae.* (Note the absence of superscript.) We stress that, for the consistency results here, no connection between $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$ and $(\boldsymbol{\gamma}^{T'}, \boldsymbol{\lambda}^{T'})$ for different topologies T and T' needs to be assumed; in other words, completely unrelated formulae can be used for any pair of topologies.

Now combine a branch length estimation scheme with an optimization principle, such as LS or ME, that allows us to choose among all the topologically-distinct fitted trees over $\{1, 2, \ldots, n\}$. We have already described LS (but also see SI Appendix 3). As for ME, three variants of this principle have been proposed, essentially differing for how tree length is defined in the presence of negative branch lengths (which are allowed by many branch length estimation schemes, including those based on $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$-formulae). We call them ME_{-1} [14], ME_{+1} [15, 37], and ME_0 [8]. Assuming that a tree has been assigned the branch lengths $\hat{\ell}_e$, ME_i defines its length as

$$\sum_{e:\ell_e>0} \hat{\ell}_e + \sum_{e:\ell_e<0} i \cdot \hat{\ell}_e.$$

The three versions of ME then differ in how they deal with negative branch lengths when calculating tree length: ME_{+1} adds together all branch lengths irrespective of their sign, whereas ME_0 ignores negative branch lengths and ME_{-1} takes their absolute value. Gascuel et al. [25] previously named ME_{+1}, ME_0 and ME_{-1}, "all-BL", "positive-BL" and "absolute-BL", respectively. The following theorem shows that for these three versions of ME, as well as for LS, tree inference is consistent when $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae are used.

Theorem 3. *Assume that the input distances* $\boldsymbol{\delta}$ *are consistent estimates of the correct evolutionary distances* $\boldsymbol{d}^{T^*}$*, where* T^* *is a binary tree with positive branch lengths. Adopt a branch length estimation scheme based on* $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$*-formulae. Then, the optimal trees with respect to* LS*,* ME_{+1}*,* ME_0 *and* ME_{-1} *are statistically consistent estimates of* T^**.*

Whereas the consistency of LS is a simple consequence of the correctness of the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae, and is included here for sake of completeness, the result for ME is somewhat surprising, given that ME has been proven to be inconsistent when combined with WLS branch lengths (for some particular values of the weights w_{ij}) [25]. Furthermore, Theorem 3 generalizes all previously known cases of consistency for the ME principle [17, 26, 18]. In particular, it demonstrates the statistical consistency of tree reconstruction when using the formulae by Mihaescu and Pachter, thus answering their fundamental question mentioned in the *Introduction*.

Computational efficiency

While the statistical consistency results above provide a theoretical basis for the use of $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae, we now consider a more practical advantage of these formulae: the fact that they can be efficiently combined with hill climbing heuristics, a pervasive and successful tool for tree reconstruction. Hill climbing consists of repeatedly applying small changes that improve the score of a candidate tree, until no such change is possible anymore. The behavior of hill climbing is essentially determined by the changes allowed at each step, or in other words by a notion of neighborhood defined over tree space. Here, we consider the simplest such changes, known as *nearest neighbor interchanges* (NNIs), which consist of swapping the positions of two 3-separated subtrees in a topology: for example, the topology in Fig. 1(c) can be obtained from that in Fig. 1(b) by swapping clades A_1 and B_1. When topology T' can be obtained from topology T in this way, we say that T and T' are *NNI-neighbors*. An NNI

transforming T into T' is *around* e, if e is the middle branch among the three branches separating the subtrees being swapped in T. While simple, NNIs can be used to efficiently implement more complex changes (such as SPRs) that can be obtained via a series of NNIs [28, 21].

Clearly, the computational efficiency of a hill climbing heuristic depends crucially on the ability to efficiently evaluate some/all neighbors of any candidate topology. For all distance-based optimization principles, the evaluation is essentially done on the basis of some function of the assigned branch lengths. It is then important to calculate efficiently the branch lengths of the neighbors that are considered at each iteration. Here, we show that if $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae are used for computing branch lengths, and a natural relation between the $\boldsymbol{\gamma}^T$ parameters for NNI-neighbors is assumed, then the $O(n^2)$ branch lengths of all the NNI-neighbors of a candidate topology can be calculated in $O(n^2)$ time. This is optimal, because these $O(n^2)$ branch lengths depend on all the $O(n^2)$ input distances.

In order to express the required relation between the $\boldsymbol{\gamma}^T$ parameters for NNI-neighbors, we assume that when performing an NNI around a branch e, all other branches keep their names. (For example, see branches f, g, h and l in Fig. 1(b) and (c).) Then, when T' is obtained from T with an NNI around branch e, we say that parameter sets $\boldsymbol{\gamma}^T = (\gamma_{e_1 e_2})$ and $\boldsymbol{\gamma}^{T'} = (\gamma'_{e_1 e_2})$, defined for T and T', respectively, are *almost identical*, if $\gamma_{e_1 e_2} = \gamma'_{e_1 e_2}$, for every pair of adjacent branches (e_1, e_2) in T such that their common endpoint is not also an endpoint of e (in which case e_1 and e_2 are also adjacent in T'). The intuitive idea is that $\boldsymbol{\gamma}^T$ and $\boldsymbol{\gamma}^{T'}$ may only differ locally around the location of the NNI. This requirement is a prerequisite for the efficient evaluation of T' from that of T. Note the difference here with the approach in the previous section, where we assumed no relationship between parameter sets for different topologies. Our result can now be stated as follows:

Theorem 4. *Let T_0 be a binary topology over taxa $\{1, 2, \ldots, n\}$ and $T_1, T_2, \ldots, T_{2(n-3)}$ all its NNI-neighbors. For all $i \in \{0, 1, \ldots, 2(n-3)\}$, assume that the branch lengths of T_i are defined by the $(\boldsymbol{\gamma}^{T_i}, \boldsymbol{\lambda}^{T_i})$-formulae, with the constraint that $\boldsymbol{\gamma}^{T_i}$ and $\boldsymbol{\gamma}^{T_0}$ are almost identical. Then,*

(i) the branch lengths of T_0 can be calculated in $O(n^2)$ time;

(ii) the branch lengths of all the NNI-neighbors of T_0 can be calculated in $O(n^2)$ time.

We leave the proof of this result to SI Appendix 4. While point (i) merely generalizes further a property already known for all M&P formulae [22], the result in (ii) is novel. It is related to, and somehow explains the existence of a number of efficient hill climbing algorithms for distance-based tree reconstruction. In particular, it predicts the efficiency of hill climbing for balanced minimum evolution (BME), which assumes $\boldsymbol{\gamma}^T$ parameters always equal to $1/2$ and therefore clearly having the property of being almost identical for NNI-neighbors. The existing hill climbing algorithm for BME [7] directly updates the total tree length, rather than the lengths of each branch, but the worst-case time complexity for each iteration is still $O(n^2)$ and results in one of the most accurate and fast distance-based methods [18, 32]. Theorem 4 also predicts the efficiency of hill climbing for OLS: the γ_{ef} parameters for OLS depend in fact on the sizes of the three clades to the sides of e, f and g (where the latter is the branch adjacent to both e and f), and these do not change when performing an NNI around a branch other than e, f and g, which implies the almost identity of the γ_{ef} parameters for NNI-neighbors.

Note that Theorem 4 has very wide applicability, not only because of the generality of the formulae it assumes, but also because it makes no assumption on the optimization criterion used to score trees (apart from its dependence on the branch lengths). This is unlike the hill-climbing algorithms we mentioned above, which were only applicable to the classic version of ME (the one we call ME_{+1}), where all branch lengths are added together, irrespective of their sign.

Discussion

We presented here a framework unifying some of the most successful approaches for distance-based tree reconstruction: for example, ordinary least squares methods for clustering [38] and balanced minimum

evolution (BME, the optimization principle behind Neighbor-Joining [30]) for phylogenetic inference. We have shown that all the methods that fit into this general framework have highly desirable statistical properties (the consistency of the tree estimates) and algorithmic properties (efficiency of hill climbing heuristics).

Our study opens the way for improvements of existing methods and the development of new ones. Novel combinations of branch length formulae and tree optimization principles can be envisaged. For example, our results enable the efficient implementation of hill climbing for the versions of ME discouraging negative branch lengths (or at least not favoring them: see ME_{-1} and ME_0 above), in combination with any of the classic branch length estimation schemes (e.g. OLS or that used in BME). Alternatively, our framework enables to explore novel, biologically-motivated ways of estimating branch lengths, for example assuming multiplicative variance models based on the current tree estimate.

We conclude by noting that although the class of branch length formulae we consider here is inspired by previous work on multiplicative variance models [22], nothing excludes that it may be applicable to least squares criteria other than WLS with multiplicative weights. In fact, it is easy to construct covariance models with nonzero covariances that result in GLS branch length estimators coinciding with $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae. Future research should aim to elucidate the full potential of our class of formulae.

References

[1] Felsenstein, J. (2004) *Inferring Phylogenies*. (Sinauer).

[2] Yang, Z. (2006) *Computational molecular evolution*. (Oxford University Press).

[3] Saitou, N & Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

[4] Felsenstein, J. (2004) *Inferring Phylogenies*. (Sinauer).

[5] Roch, S. (2010) Toward extracting all phylogenetic information from matrices of evolutionary distances. *Science* **327**, 1376–1379.

[6] Steel, M. (2009) A basic limitation on inferring phylogenies by pairwise sequence comparisons. *J. Theor. Biol.* **256**, 467–472.

[7] Desper, R & Gascuel, O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comp. Biol.* **9**, 687–705.

[8] Swofford, D, Olsen, G, Waddell, P, & Hillis, D. (1996) in *Molecular Systematics*, eds. Hillis, D, Moritz, C, & Mable, B. (Sinauer), pp. 407–514.

[9] Bulmer, M. (1991) Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Mol. Biol. Evol.* **8**, 868–883.

[10] Bryant, D & Waddell, P. (1998) Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Mol. Biol. Evol.* **15**, 1346–1359.

[11] Swofford, D. (1998) PAUP* — phylogenetic analysis using parsimony (*and other methods). (Sinauer).

[12] Cavalli-Sforza, L & Edwards, A. (1967) Phylogenetic analysis: models and estimation procedures. *Am. J. Hum. Genet.* **19**, 233–257.

[13] Fitch, W & Margoliash, E. (1967) Construction of phylogenetic trees. *Science* **155**, 279–284.

[14] Kidd, K & Sgaramella-Zonta, L. (1971) Phylogenetic analysis: concepts and methods. *Am. J. Hum. Genet.* **23**, 235–252.

[15] 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.

[16] Vach, W. (1989) in *Conceptual and numerical analysis of data*, ed. Opitz, O. (Springer-Verlag, Berlin), pp. 230–238.

[17] Rzhetsky, A & Nei, M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.* **10**, 1073–1095.

[18] 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.

[19] Pauplin, Y. (2000) Direct calculation of a tree length using a distance matrix. *J. Mol. Evol.* **51**, 41–47.

[20] Hordijk, W & Gascuel, O. (2005) Improving the efficiency of spr moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* **21**, 4338–4347.

[21] 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. Bioinform.* **6**, 110–117.

[22] Mihaescu, R & Pachter, L. (2008) Combinatorics of least squares trees. *Proc. Natl. Acad. Sci. USA* **105**, 13206–13211.

[23] Nei, M & Jin, L. (1989) Variances of the average numbers of nucleotide substitutions within and between populations. *Mol. Biol. Evol.* **6**, 290–300.
[24] Atteson, K. (1999) The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica* **25**, 251–278.
[25] Gascuel, O, Bryant, D, & Denis, F. (2001) Strengths and limitations of the minimum evolution principle. *Syst. Biol.* **50**, 621–627.
[26] Denis, F & Gascuel, O. (2003) On the consistency of the minimum evolution principle of phylogenetic inference. *Discr. Appl. Math.* **127**, 63–77.
[27] Willson, S. (2005) Consistent formulas for estimating the total lengths of trees. *Discr. Appl. Math.* **148**, 214–239.
[28] Semple, C & Steel, M. (2003) *Phylogenetics*. (Oxford University Press).
[29] Buneman, P. (1971) in *Mathematics in the Archaelogical and Historical Sciences*, ed. Hodson, F. (Edinburgh University Press), pp. 387–395.
[30] Gascuel, O & Steel, M. (2006) Neighbor-joining revealed. *Mol. Biol. Evol.* **23**, 1997–2000.
[31] Aitken, A. C. (1935) On least squares and linear combinations of observations. *Proc. Royal Soc. Edinburgh A* **55**, 42–48.
[32] Vinh, S & von Haeseler, A. (2005) Shortest triplet clustering: reconstructing large phylogenies using representative sets. *BMC Bioinformatics* **6**, 92.
[33] Jukes, T & Cantor, C. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. (Academic Press), pp. 21–132.
[34] Beyer, W, Stein, M, Smith, T, & Ulam, S. (1974) A molecular sequence metric and evolutionary trees. *Mathematical Biosciences* **19**, 9–25.
[35] Susko, E, Inagaki, Y, & Roger, A. (2004) On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. *Mol. Biol. Evol.* **21**, 1629–1642.
[36] Pardi, F, Guillemot, S, & Gascuel, O. (2010) Robustness of phylogenetic inference based on minimum evolution. *Bull. Math. Biol.* **72**, 1820–1839.
[37] Rzhetsky, A & Nei, M. (1992) A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.* **9**, 945–967.
[38] De Soete, G. (1983) A least squares algorithm for fitting additive trees to proximity data. *Psychometrika* **48**, 621–626.

SI Appendix 1

Here, we present an interpretation of the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae as averages over a large number of simple branch length formulae, which allows us to prove Theorem 1: because these simple formulae are correct, that is, they correctly provide the branch lengths of a tree whenever the input distances are additive with respect to that tree, it follows that also the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae are correct. The detailed arguments follow below.

Let $\hat{\ell}_e(\boldsymbol{\delta})$ denote the length that is assigned to branch e by an adopted length estimation method. Let T^* be a tree where e has length ℓ_e. The adopted method is *correct* if, for any such tree T^*, $\hat{\ell}_e(\boldsymbol{d}^{T^*}) = \ell_e$.

Suppose e is an external branch, and define A, B, i as in Fig. 1(a) in the main text. Choose a taxon a from A and a taxon b from B. Then calculate the length of e with:

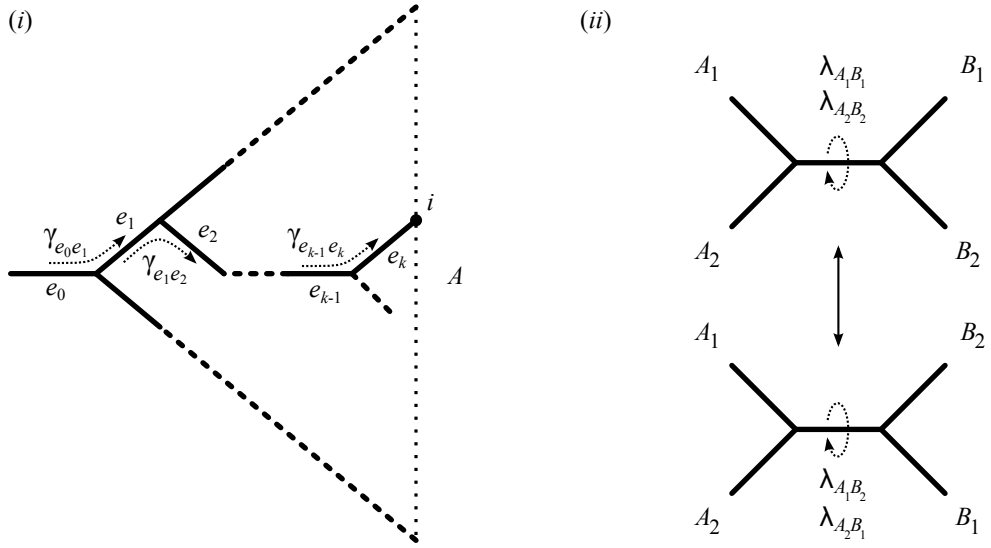
$$\hat{\ell}_e^{ab}(\boldsymbol{\delta}) = \frac{1}{2}(\delta_{ia} + \delta_{ib} - \delta_{ab}).$$

If instead e is an internal branch, let A_1, A_2, B_1 and B_2 be the four clades surrounding it, as in Fig. 1(b) in the main text. Choose taxa a_1, a_2, b_1, b_2 from A_1, A_2, B_1, B_2, respectively. It is clear that any drawing of the tree on the plane either places A_1 to the side of B_1, and therefore A_2 to the side of B_2 (as in Fig. S1(*ii*), top) or alternatively A_1 to the side of B_2 and A_2 to the side of B_1 (Fig. S1(*ii*), bottom). We associate the former drawing with the following formula for the length of e:

$$\hat{\ell}_e^{a_1b_1a_2b_2}(\boldsymbol{\delta}) = \frac{1}{2}(\delta_{a_1b_1} + \delta_{a_2b_2} - \delta_{a_1a_2} - \delta_{b_1b_2}).$$

The alternative drawing is associated to the formula $\hat{\ell}_e^{a_1b_2a_2b_1}(\boldsymbol{\delta})$. Note that $\hat{\ell}_e^{ab}(\boldsymbol{\delta})$, $\hat{\ell}_e^{a_1b_1a_2b_2}(\boldsymbol{\delta})$ and $\hat{\ell}_e^{a_1b_2a_2b_1}(\boldsymbol{\delta})$ are all trivially correct.

FIGURE S1. (*i*) $p_{i|A} = \gamma_{e_0e_1} \cdot \gamma_{e_1e_2} \cdot \ldots \cdot \gamma_{e_{k-1}e_k}$ is the probability of ending up in i when entering clade A from its root and following the random walk rules described in the main text; (*ii*) $\lambda_{A_1B_1} = \lambda_{A_2B_2}$ can be seen as the probability of drawing the tree in the top configuration, whilst $\lambda_{A_1B_2} = \lambda_{A_2B_1} = 1 - \lambda_{A_1B_1}$ can be seen as the probability of drawing the tree in the bottom configuration.



The $\boldsymbol{\gamma}^T$ and $\boldsymbol{\lambda}^T$ parameters introduced in the main text can be interpreted as controlling a probability distribution over all possible such formulae for calculating the length of a given branch e in T. First,

as illustrated in Fig. S1(i), the $\boldsymbol{\gamma}^T$ parameters determine the probability of choosing a given taxon out of any given clade (a from A, b from B in the case of external branches, a_1 from A_1, a_2 from A_2, b_1 from B_1, b_2 from B_2 in the case of internal branches). Second, as illustrated in Fig. S1(ii), the $\boldsymbol{\lambda}^T$ parameters determine the probability of choosing either of the two possible drawings for the clades around an internal branch e, and therefore either of $\hat{\ell}_e^{a_1b_1a_2b_2}(\boldsymbol{\delta})$ or $\hat{\ell}_e^{a_1b_2a_2b_1}(\boldsymbol{\delta})$ for the length of e: $\lambda_{A_1B_1} = \lambda_{A_2B_2}$ is the probability of drawing A_1 to the side of B_1 and A_2 to the side of B_2, whereas its complement $\lambda_{A_1B_2} = \lambda_{A_2B_1} = 1 - \lambda_{A_1B_1}$ is the probability of drawing A_1 to the side of B_2 and A_2 to the side of B_1.

Given this probability distribution, let us take the resulting expected value of the length assigned to e. In the (harder) case of an internal branch, this is given by

$$\sum_{\substack{a_1\in A_1,\ b_1\in B_1\\ a_2\in A_2,\ b_2\in B_2}} p_{a_1|A_1}p_{a_2|A_2}p_{b_1|B_1}p_{b_2|B_2}\left[\lambda_{A_!B_1}\hat{\ell}_e^{a_1b_1a_2b_2}(\boldsymbol{\delta}) + \lambda_{A_!B_2}\hat{\ell}_e^{a_1b_2a_2b_1}(\boldsymbol{\delta})\right]$$

$$= \frac{1}{2}\sum_{\substack{a_1\in A_1,\ b_1\in B_1\\ a_2\in A_2,\ b_2\in B_2}} p_{a_1|A_1}p_{a_2|A_2}p_{b_1|B_1}p_{b_2|B_2}\left[\lambda_{A_!B_1}(\delta_{a_1b_1}+\delta_{a_2b_2}) + (1-\lambda_{A_1B_!})(\delta_{a_1b_2}+\delta_{a_2b_1}) - \delta_{a_1a_2} - \delta_{b_1b_2}\right]$$

$$= \frac{1}{2}\left[\lambda_{A_1B_1}(\delta_{A_1B_1}+\delta_{A_2B_2}) + (1-\lambda_{A_1B_1})(\delta_{A_1B_2}+\delta_{A_2B_1}) - \delta_{A_1A_2} - \delta_{B_1B_2}\right].$$

Thus what we obtain (also in the easier case of an external branch; not shown) are precisely the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae. In other words, these formulae can be seen as providing the expected length of a branch when this is assigned following the random procedure described above. Given this observation, the correctness of the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae follows trivially from the correctness of the base formulae $\hat{\ell}_e^{ab}(\boldsymbol{\delta})$ and $\hat{\ell}_e^{a_1b_1a_2b_2}(\boldsymbol{\delta})$. Theorem 1 is therefore proved.

We note that the approach of expressing a length estimator as the combination of several simple formulae has already been considered by Willson [1]. His base formulae, however, express the length of a *path* in the tree (as a function of the distances between *three* taxa) rather than a *single branch* (which we express as a function of the distances between three or four taxa, for exterior and interior branches, respectively). Moreover, the combination of his base formulae provide an estimate of the total length of the tree (in the ME_{+1} sense).

SI Appendix 2

Here, we prove the relationship between the M&P formulae [2] and our $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae, as stated in Theorem 2. We start by formally defining the M&P formulae (A2.1); then we introduce a few additional formalisms and a useful observation (A2.2) and then prove separately the two parts of Theorem 2 ((i) in A2.3 and (ii) in A2.4). This requires to show how to derive the parameters of each class of formulae from the parameters of the other class (i.e. $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$ from $\boldsymbol{w}$, and vice versa).

A2.1 The M&P formulae. We assume that the weights $\boldsymbol{w} = (w_{ij})$ are multiplicative w.r.t. a binary topology T. Then, for any two clades A and B of T, define

$$Z_{AB} = \sum_{\substack{i\in A\\ j\in B}} w_{ij} \qquad \text{and} \qquad \delta^{\boldsymbol{w}}_{AB} = \frac{1}{Z_{AB}} \sum_{\substack{i\in A\\ j\in B}} w_{ij}\delta_{ij}.$$

Mihaescu and Pachter [2] have shown that the optimal branch lengths of T with respect to the WLS criterion (1) are then given by the following formula, applicable to any branch e in T:

$$\hat{\ell}_e(\boldsymbol{\delta}) = \begin{cases} \frac{1}{2}(\delta^{\boldsymbol{w}}_{iA} + \delta^{\boldsymbol{w}}_{iB} - \delta^{\boldsymbol{w}}_{AB}) & \text{if } e \text{ is external,} \\ \frac{1}{2}\left[\frac{Z_{A_1B_2}+Z_{A_2B_1}}{Z_{AB}}(\delta^{\boldsymbol{w}}_{A_1B_1} + \delta^{\boldsymbol{w}}_{A_2B_2})\right. & \\ \left. +\frac{Z_{A_1B_1}+Z_{A_2B_2}}{Z_{AB}}(\delta^{\boldsymbol{w}}_{A_1B_2} + \delta^{\boldsymbol{w}}_{A_2B_1}) - \delta^{\boldsymbol{w}}_{A_1A_2} - \delta^{\boldsymbol{w}}_{B_1B_2}\right] & \text{if } e \text{ is internal,} \end{cases}$$

where, if e is external, we define A, B, i as in Fig. 1(a) in the main text and, if e is internal, we define A_1, A_2, $A = A_1 \cup A_2$, B_1, B_2, $B = B_1 \cup B_2$ as in Fig. 1(b) in the main text.

A2.2 Decomposition of Z_{XY}. Extend the w_{ij} notation to any pair of nodes x and y (possibly internal) in T:

$$w_{xy} = \prod_{e\in P_{xy}(T)} w_e, \tag{1}$$

where we recall that the w_e are the branch-associated weights that compose the pairwise weights w_{ij}, and $P_{xy}(T)$ is the set of branches on the path between x and y in T. Then define the *multiplicative weight of* X (a clade with root x) as:

$$Z_X = \sum_{i\in X} w_{ix}. \tag{2}$$

We assume $w_{xx} = 1$ for any node x, which implies $Z_{\{i\}} = 1$, for any one-taxon clade $\{i\}$. It is then easy to check that, if X and Y are any two disjoint clades in T, with roots x and y respectively, then

$$Z_{XY} = w_{xy} Z_X Z_Y. \tag{3}$$

A2.3 The M&P formulae are also $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae.

Lemma 1. *Given weights $\boldsymbol{w} = (w_{ij})$ multiplicative w.r.t. a binary topology T, define, for each pair of adjacent branches e and f:*

$$\gamma_{ef} = \frac{w_f Z_{A_1}}{Z_A}, \tag{4}$$

where A_1 and $A = A_1 \cup A_2$ are the clades having f and e as root branches, respectively, as in Fig. 1(b). Then,

(i) The resulting average distances between clades are such that $\delta_{XY} = \delta^{\boldsymbol{w}}_{XY}$, for any two disjoint clades X and Y in T.

(ii) For any internal branch e, let $A = A_1 \cup A_2$ and $B = B_1 \cup B_2$ be the clades in the configuration of Fig. 1(b); then $(Z_{A_1B_2} + Z_{A_2B_1})/Z_{AB} = \gamma_{ef} + \gamma_{eh} - 2\gamma_{ef}\gamma_{eh}$.

Proof. (i) Equation (3) allows us to express $\delta^{\boldsymbol{w}}_{XY}$ in a very similar form to that of δ_{XY}:

$$\delta^{\boldsymbol{w}}_{XY} = \sum_{\substack{i\in X\\ j\in Y}} \frac{w_{ij}}{Z_{XY}}\delta_{ij} = \sum_{\substack{i\in X\\ j\in Y}} \frac{w_{ix}}{Z_X}\frac{w_{jy}}{Z_Y}\delta_{ij},$$

where x and y are the root nodes of X and Y, respectively. In order to have $\delta_{XY} = \delta^{\boldsymbol{w}}_{XY}$, it is then sufficient to prove that, for any clade X with root x, and any taxon $i \in X$,

$$p_{i|X} = \frac{w_{ix}}{Z_X}.$$

Let the path from x to i traverse branches $e_1, e_2, \ldots, e_k$ (in this order), and let X_j be the subclade of X having e_j as root branch ($0 \le j \le k$, with $X_0 = X$ and e_0 being the root branch of X); then, $X_k = \{i\}$ and

$$p_{i|X} = \gamma_{e_0e_1}\cdot\gamma_{e_1e_2}\cdot\ldots\cdot\gamma_{e_{k-1}e_k} = \frac{w_{e_1}Z_{X_1}}{Z_X}\cdot\frac{w_{e_2}Z_{X_2}}{Z_{X_1}}\cdot\ldots\cdot\frac{w_{e_k}Z_{X_k}}{Z_{X_{k-1}}} = \frac{w_{e_1}\cdot w_{e_2}\cdot\ldots\cdot w_{e_k}}{Z_X}Z_{\{i\}} = \frac{w_{ix}}{Z_X}.$$

(ii) Using again (3),

$$\frac{Z_{A_1B_2}+Z_{A_2B_1}}{Z_{AB}} = \frac{w_fw_lZ_{A_1}Z_{B_2}+w_gw_hZ_{A_2}Z_{B_1}}{Z_AZ_B} = \gamma_{ef}\gamma_{el}+\gamma_{eg}\gamma_{eh} = \gamma_{ef}+\gamma_{eh}-2\gamma_{ef}\gamma_{eh}.$$

Points (i) and (ii) are thus both verified. □

Setting γ_{ef} as in (4) has a simple intuitive meaning: if we call g the root branch of A_2 (as in Fig. 1(b)), then $Z_A = w_fZ_{A_1}+w_gZ_{A_2}$. The γ_{ef} above can then be seen as the relative multiplicative weight of the subtree corresponding to clade A_1 in the subtree corresponding to clade A. The random walk defined by these parameters is then "attracted" by the heavier subtrees, in a way that is directly proportional to the weights of the subtrees.

Proof of Theorem 2, part (i). Given $\boldsymbol{w}$ multiplicative w.r.t. T, define the $\boldsymbol{\gamma}^T$ parameters as in (4). Note that because $w_e > 0$ for any branch e, then $Z_X > 0$ for every clade X. Moreover, for any three adjacent branches e, f and g in the configuration of Fig. 1(b), we have $w_fZ_{A_1} + w_gZ_{A_2} = Z_A$ and $w_fZ_{A_1}, w_gZ_{A_2} > 0$, which imply $\gamma_{ef}+\gamma_{eg} = 1$ and $0 < \gamma_{ef}, \gamma_{eg} < 1$. Therefore the definition of the $\boldsymbol{\gamma}^T$ parameters is admissible and implies that $\delta_{XY} = \delta^{\boldsymbol{w}}_{XY}$ (Lemma 1, part (i)).

As for the $\boldsymbol{\lambda}^T$ parameters, set $\lambda_{A_1B_1} = (Z_{A_1B_2}+Z_{A_2B_1})/Z_{AB}$, for every pair of clades A_1 and B_1 separated by 3 branches, and being in the configuration of Fig. 1(b) with $A = A_1\cup A_2$ and $B = B_1\cup B_2$. It is easy to check that this implies $\lambda_{A_1B_1} = \lambda_{A_2B_2} > 0$, $\lambda_{A_1B_2} = \lambda_{A_2B_1} > 0$ and $\lambda_{A_1B_1}+\lambda_{A_1B_2} = 1$, and therefore the definition of the $\boldsymbol{\lambda}^T$ parameters is also admissible.

It is now easy to verify that the resulting $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae coincide with the M&P formulae corresponding to $\boldsymbol{w}$: for the external branches this is an immediate consequence of $\delta_{XY} = \delta^{\boldsymbol{w}}_{XY}$, while for the internal branches we also use the above definition of $\lambda_{A_1B_1}$ and the fact that $1-\lambda_{A_1B_1} = (Z_{A_1B_1}+Z_{A_2B_2})/Z_{AB}$. Finally, the $\boldsymbol{\gamma}^T$ and $\boldsymbol{\lambda}^T$ defined above satisfy properties P1 and P2: the first can be verified by using (4) in P1 and the second is a direct consequence of Lemma 1, part (ii). □

A2.4 Characterization of the $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$-formulae that are also M&P formulae.

Proof of Theorem 2, part (ii).

The proof has the following structure: as an intermediate step, we introduce — for every three clades A_1, A_2 and B whose respective root branches f, g and e are incident to the same internal node (as in Fig. 1(b)) — three values φ_{A_1}, φ_{A_2} and φ_B such that $\varphi_{A_1}+\varphi_{A_2}+\varphi_B = 1$, $0 < \varphi_{A_!}, \varphi_{A_2}, \varphi_B < 1$ and

$$\gamma_{ef} = \frac{\varphi_{A_1}}{\varphi_{A_1}+\varphi_{A_2}}, \quad \gamma_{fg} = \frac{\varphi_{A_2}}{\varphi_{A_2}+\varphi_B}, \quad \gamma_{ge} = \frac{\varphi_B}{\varphi_B+\varphi_{A_1}}. \tag{5}$$

The existence of such values is guaranteed by property P1. In intuitive terms, we are requiring that each clade has somewhat a "preference value", such that the probabilities γ_{ef} and γ_{eg} are proportional to the preference values of the clades that f and g lead to. On the basis of these values, we then define

a set of branch-associated weights w_e specifying a multiplicative model such that the preference values can be obtained as

$$\varphi_B = \frac{w_e Z_B}{w_e Z_B + w_f Z_{A_1} + w_g Z_{A_2}}. \tag{6}$$

(With A_1, A_2, B, f, g, e as above.) It is easy to see (shown below) that this implies that condition (4) of Lemma 1, and therefore its conclusions, hold. This, together with the fact that the $\boldsymbol{\lambda}^T$ parameters satisfy P2, implies that the M&P formulae for $\boldsymbol{w}$ coincide with the given $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae.

Let us now look in detail at each step. First of all, property P1 implies that the determinant of the coefficient matrix for the system of linear equations in φ_{A_1}, φ_{A_2} and φ_B corresponding to (5) is equal to 0; this system is then solved by the following subspace of solutions:

$$(\varphi_B, \varphi_{A_1}, \varphi_{A_2}) \in \left\{ \left(x,\ \frac{1-\gamma_{ge}}{\gamma_{ge}} x,\ \frac{1-\gamma_{ef}}{\gamma_{ef}} \frac{1-\gamma_{ge}}{\gamma_{ge}} x \right) \mid x \in \mathbb{R} \right\}.$$

If furthermore we impose $\varphi_{A_1} + \varphi_{A_2} + \varphi_B = 1$, it is easy to see that a unique solution is determined, such that $\varphi_{A_1}, \varphi_{A_2}, \varphi_B > 0$ (and therefore also < 1).

Given the φ_X parameters, we now show how to define the branch-associated weights w_e. For any internal branch e separating clades A and $\overline{A}$, let

$$w_e = \sqrt{\frac{\varphi_A}{1-\varphi_A} \frac{\varphi_{\overline{A}}}{1-\varphi_{\overline{A}}}}. \tag{7}$$

As for the external branches, which for simplicity we call with the same names as the taxa they are incident to (e.g. branch i being the one incident to taxon i), we assign their weight in the following way: choose arbitrarily $w_1 > 0$ and then, for any other external branch $i \in \{2, 3, \ldots, n\}$ define

$$w_i = w_1 \sqrt{\frac{\gamma^*_{1i}}{\gamma^*_{i1}} \frac{1-\varphi_1}{\varphi_1} \frac{\varphi_i}{1-\varphi_i}}, \tag{8}$$

where $\gamma^*_{ef} = \gamma_{ee_1} \cdot \gamma_{e_1 e_2} \cdot \ldots \cdot \gamma_{e_k f}$, for any pair of branches e, f linked by a path composed of the ordered sequence of branches $(e_1, e_2, \ldots, e_k)$, and for simplicity we write φ_1 and φ_i instead of $\varphi_{\{1\}}$ and $\varphi_{\{i\}}$.

The weights thus defined determine a multiplicative weighting $\boldsymbol{w} = (w_{ij})$ that satisfies (6). In order to show this, we first show that, for every clade A (whose root branch we call e) such that $|A| < n-1$ (i.e. the endpoint of e on the other side of A is not a leaf),

$$w_e Z_A = w_1 \sqrt{\frac{\gamma^*_{1e}}{\gamma^*_{e1}} \frac{1-\varphi_1}{\varphi_1} \frac{\varphi_A}{1-\varphi_A}}. \tag{9}$$

We prove this by induction on the size of A. If A consists of a single taxon, then either this is taxon 1, in which case both sides of the equation reduce to w_1, or this is another taxon i, in which case (9) coincides with (8). In both cases (9) trivially holds. If $|A| > 1$, let A_1, A_2, B, f and g be as in Fig. 1(b). Then, by inductive hypothesis, (9) holds for $w_f Z_{A_1}$ and $w_g Z_{A_2}$ and we have:

$$Z_A = w_f Z_{A_1} + w_g Z_{A_2} = w_1 \sqrt{\frac{1-\varphi_1}{\varphi_1}} \left(\sqrt{\frac{\gamma^*_{1f}}{\gamma^*_{f1}} \frac{\varphi_{A_1}}{1-\varphi_{A_1}}} + \sqrt{\frac{\gamma^*_{1g}}{\gamma^*_{g1}} \frac{\varphi_{A_2}}{1-\varphi_{A_2}}} \right)$$

Now note that

$$\frac{\gamma^*_{1f}}{\gamma^*_{f1}} = \frac{\gamma^*_{1e}}{\gamma^*_{e1}} \frac{\varphi_{A_1}(1-\varphi_{A_1})}{\varphi_B(1-\varphi_B)} \qquad \text{and} \qquad \frac{\gamma^*_{1g}}{\gamma^*_{g1}} = \frac{\gamma^*_{1e}}{\gamma^*_{e1}} \frac{\varphi_{A_2}(1-\varphi_{A_2})}{\varphi_B(1-\varphi_B)}, \tag{10}$$

which can be verified by noting that, depending on the position of taxon 1 (in A_1, A_2 or B), $\gamma^*_{1f}/\gamma^*_{f1}$ can either be equal to $(\gamma^*_{1e}/\gamma^*_{e1})(\gamma_{ef}/\gamma_{fe})$ or $(\gamma^*_{1e}/\gamma^*_{e1})(\gamma_{gf}\gamma_{eg})/(\gamma_{ge}\gamma_{fg})$, and $\gamma^*_{1g}/\gamma^*_{g1}$ can either be equal to $(\gamma^*_{1e}/\gamma^*_{e1})(\gamma_{eg}/\gamma_{ge})$ or $(\gamma^*_{1e}/\gamma^*_{e1})(\gamma_{fg}\gamma_{ef})/(\gamma_{fe}\gamma_{gf})$. The equations in (10) can then be obtained from these expressions by making the substitutions $\gamma_{xy} = \varphi_Y/(1-\varphi_X)$ (equivalent to (5)) for $x, y \in \{e, f, g\}$,

where X, Y are the clades separated by and having x, y as root branches, respectively. If now we use (10) in the expression above for Z_A, we obtain after obvious simplifications

$$Z_A = w_1 \sqrt{\frac{\gamma^*_{1e}}{\gamma^*_{e1}} \frac{1-\varphi_1}{\varphi_1} \frac{1-\varphi_B}{\varphi_B}}. \tag{11}$$

If now we use (7) and (11) to express w_e and Z_A in $w_e Z_e$, what we obtain is precisely (9), which therefore is proven.

We are now ready to prove (6). Note that (11) holds for any 'composite' clade A (i.e. one that that can be decomposed into two other clades A_1 and A_2). Then,

$$\frac{w_e Z_B}{w_e Z_B + w_f Z_{A_1} + w_g Z_{A_2}} = \frac{w_e Z_B}{w_e Z_B + Z_A} = \frac{\sqrt{\frac{\varphi_B}{1-\varphi_B}}}{\sqrt{\frac{\varphi_B}{1-\varphi_B}} + \sqrt{\frac{1-\varphi_B}{\varphi_B}}} = \varphi_B,$$

where for the second equality we have used both (11) and (9).

But this implies that, for every composite clade $A = A_1 \cup A_2$ in the configuration of Fig. 1(b),

$$\frac{w_f Z_{A_1}}{Z_A} = \frac{w_f Z_{A_1}}{w_f Z_{A_1} + w_g Z_{A_2}} = \frac{\varphi_{A_1}}{\varphi_{A_1} + \varphi_{A_2}} = \gamma_{ef}. \tag{12}$$

Condition (4) of Lemma 1 is therefore verified. But this ensures that $\delta_{XY} = \delta^{\boldsymbol{w}}_{XY}$, for any two disjoint clades X and Y (Lemma 1, part (i)), while the fact that the $\boldsymbol{\lambda}^T$ parameters satisfy P2 implies (Lemma 1, part (ii)) that $\lambda_{A_1 B_1} = (Z_{A_1 B_2} + Z_{A_2 B_1})/Z_{AB}$ for every pair of 3-separated clades A_1, B_1 in the configuration of Fig. 1(b). That is, the M&P formulae for $\boldsymbol{w}$ coincide with the given $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae, which is what we set out to prove. □

SI Appendix 3

Here, we prove that the main criteria to score trees, LS and ME (in all their common variants), are statistically consistent when used in combination with our branch length formulae, as stated in Theorem 3. We start by showing that the consistency of any distance-based principle is essentially determined by its behavior on perfect data (A3.1). Next, we move on to proving the consistency of LS (A3.2) and then that of ME: for the latter, first we show a useful dependency property between different variants of ME (A3.3), and then we prove the consistency on perfect data of the classic version of ME (A3.4), which is the key nontrivial result of this appendix and allows us to conclude the proof of Theorem 3 (A.3.5).

We recall that a *branch length estimation scheme* is a method that, for any binary topology over the set of taxa under consideration $\{1, 2, \ldots, n\}$ ($n \geq 3$), determines how to fit the length of its branches on the basis of an $n \times n$ distance matrix $\boldsymbol{\delta}$. We say that a branch length estimation scheme is *continuous* [*linear*] if, for any branch e in any binary topology, the function $\hat{\ell}_e(\boldsymbol{\delta})$ giving its fitted length is continuous [linear] in $\boldsymbol{\delta}$. A branch length estimation scheme is *correct* if, for any branch e in any binary tree with branch lengths, $\hat{\ell}_e(\boldsymbol{\delta})$ returns the length of e whenever $\boldsymbol{\delta}$ is additive with respect to that tree (as in Theorem 1). The branch length estimation schemes that we consider here are those *based on* $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$*-formulae*, whereby a collection of $\boldsymbol{\gamma}^T$ and $\boldsymbol{\lambda}^T$ parameters is chosen for each binary topology T, thus determining a set of $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae for T's branch lengths (with no assumed relation between the values of these parameters across different topologies). It is clear that the resulting branch length estimation schemes are linear (thus continuous) and correct (Theorem 1).

Any branch length estimation scheme can be combined to a number of principles identifying an optimal tree among all the topologically-distinct fitted trees. The optimization principles we consider here are defined by a *tree score function*, which can depend on the topology of the tree, the assigned branch lengths and (in the case of LS, but not ME) the input distances. An optimization principle $\mathcal{M}$ then consists of seeking the fitted tree(s) that minimize this function, and we denote this tree (or set of trees) with $\mathcal{M}(\boldsymbol{\delta})$. We say that $\mathcal{M}$ is statistically consistent if $\mathcal{M}(\boldsymbol{\delta})$ converges (in probability) to the correct tree.

The following assumption (the consistency of the distance estimates and the positive additivity of the correct evolutionary distances) applies to all the propositions that follow, and we state it here so that we do not have to repeat it in every statement.

Assumption 2. *Let the correct phylogenetic tree for the taxa under consideration, T^*, be a binary tree with positive branch lengths. Assume that the input distances $\boldsymbol{\delta}$ converge (in probability) to $\boldsymbol{d}^{T^*}$.*

A3.1 Consistency for perfect data implies statistical consistency.

The following is a well-known sufficient condition for consistency, which has been proven for ME with the same continuity arguments (e.g. [3]). It can be applied to most tree optimization principles (LS, ME, possibly combinations of the two or even totally different criteria).

Proposition 3. *Adopt a continuous branch length estimation scheme. Let $\mathcal{M}$ be an optimization principle based on a tree score function that is continuous in all its continuous parameters (i.e. all but the topology). If $\mathcal{M}(\boldsymbol{d}^{T^*})$ is unique and coincides with T^*, then $\mathcal{M}$ is statistically consistent.*

Proof. To any binary topology, $\mathcal{M}$ assigns a score by first assigning branch lengths to it, and then applying the adopted tree score function. Note that because both the branch length estimation scheme and the tree score function are continuous, then also the score associated to any particular topology is continuous in $\boldsymbol{\delta}$. Because $\mathcal{M}(\boldsymbol{d}^{T^*}) = T^*$ is unique, when $\boldsymbol{\delta} = \boldsymbol{d}^{T^*}$ the score of the topology of T^* must be strictly smaller than the score of all other binary topologies. But then, because the scores of topologies are continuous in $\boldsymbol{\delta}$ and finite in number, this must still hold for every $\boldsymbol{\delta}$ in a neighborhood of $\boldsymbol{d}^{T^*}$; that is, for every $\boldsymbol{\delta}$ in this neighborhood, $\mathcal{M}(\boldsymbol{\delta})$ is unique and has the same topology as T^*.

But, because $\boldsymbol{\delta} \xrightarrow{p} \boldsymbol{d}^{T^*}$, the probability that $\boldsymbol{\delta}$ belongs to this neighborhood, and consequently $\mathcal{M}(\boldsymbol{\delta})$ has the same topology as T^*, converges to 1. Finally, when $\mathcal{M}(\boldsymbol{\delta})$ has the same topology as T^*, the continuity of the branch length estimation scheme implies that the branch lengths in $\mathcal{M}(\boldsymbol{\delta})$ converge in probability to those in $\mathcal{M}(\boldsymbol{d}^{T^*}) = T^*$. We can then conclude that both the topology and branch lengths of $\mathcal{M}(\boldsymbol{\delta})$ consistently converge to T^*. □

A3.2 Consistency of Least Squares.

We have briefly defined LS methods in the Introduction. Here we assume the most general form for LS and prove its consistency under very general conditions (Proposition 4 below). We define LS methods as those that use a tree score function with the following form:

$$Q(T, \boldsymbol{\delta}) = (\boldsymbol{\delta} - \boldsymbol{d}^T)^t W_{T,\boldsymbol{\delta}} (\boldsymbol{\delta} - \boldsymbol{d}^T), \tag{13}$$

where T is the tree fitted using the assumed branch length estimation scheme, $\boldsymbol{\delta}$ is a column vector with the $\binom{n}{2}$ input distances and $W_{T,\boldsymbol{\delta}}$ is a $\binom{n}{2} \times \binom{n}{2}$ matrix which may depend on the topology of T and, continuously, on $\boldsymbol{\delta}$ and the branch lengths of T. Additionally, we assume that, for any T and $\boldsymbol{\delta}$, the matrix $W_{T,\boldsymbol{\delta}}$ is positive-definite. ($W_{T,\boldsymbol{\delta}}$ should be interpreted as the inverse of the assumed covariance matrix for $\boldsymbol{\delta}$.)

Note that whereas the dependence of $W_{T,\boldsymbol{\delta}}$ on $\boldsymbol{\delta}$ is common (e.g. the version of WLS by Fitch and Margoliash [4] uses a diagonal matrix with $W_{ij,ij} = \delta_{ij}^{-2}$), the dependence on T is non-standard, and we have included it here for completeness. (But for example the balanced version of WLS [5] at the basis of the balanced branch lengths [6] does assume a variance model that depends on tree topology.) Also recall that the criterion $Q(T, \boldsymbol{\delta})$ above is used to score trees with already-fitted branch lengths, so the dependence on the branch lengths does not cause any computational problem.

Proposition 4. *Adopt any correct and continuous branch length estimation scheme. Then, LS is consistent.*

Proof. We prove that for LS, the hypotheses of Proposition 3 are satisfied and therefore LS is consistent. First, the branch length estimation scheme is continuous (by hypothesis) and the score function in (13) is a continuous function of both $\boldsymbol{\delta}$ and of the branch lengths assigned to T (note that $\boldsymbol{d}^T$ is linear, and thus continuous, in the branch lengths). It remains then to show that, for $\boldsymbol{\delta} = \boldsymbol{d}^{T^*}$, LS uniquely identifies T^* as optimal. Because $W_{T,\boldsymbol{\delta}}$ is positive definite, $Q(T, \boldsymbol{d}^{T^*}) = 0$ if and only if $\boldsymbol{d}^{T^*} - \boldsymbol{d}^T = \boldsymbol{0}$, that is, if and only if $T = T^*$, whereas for all other trees $W \neq T^*$, $Q(W, \boldsymbol{d}^{T^*}) > 0$. Moreover, because the branch length estimation scheme is correct, T^* is precisely what is obtained when fitting its branch lengths. Therefore T^* uniquely minimizes the score function Q and is returned by LS. □

Corollary 5. *Adopt a branch length estimation scheme based on $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$-formulae. Then, LS is consistent.*

A3.3 Dependency between the consistencies of different variants of Minimum Evolution.

Recall that ME_i reconstructs the fitted tree that minimizes the following tree score function, where $\hat{\boldsymbol{\ell}} = (\hat{\ell}_e)$ denotes the branch lengths in the fitted tree:

$$L_i(\hat{\boldsymbol{\ell}}) = \sum_{e:\ell_e>0} \hat{\ell}_e + \sum_{e:\ell_e<0} i \cdot \hat{\ell}_e. \tag{14}$$

In the main text, we assume $i \in \{-1, 0, +1\}$, but here we consider, more generally, ME_x with x being any real number. We do this not only for the sake of mathematical completeness, but also to include variants of ME that may be considered in the future (e.g. $\mathrm{ME}_{-\infty}$, which corresponds to avoiding at all costs trees which are assigned negative branch lengths). The following proposition shows that if we can prove the consistency of ME_y using Proposition 3, then the same can be done for any ME_x with $x < y$.

Lemma 6. *Adopt any correct branch length estimation scheme and let $x < y$. If $\mathrm{ME}_y(\boldsymbol{d}^{T^*}) = T^*$, then also $\mathrm{ME}_x(\boldsymbol{d}^{T^*}) = T^*$.*

Proof. We adapt a line of reasoning appeared elsewhere [7, 3]. Because of the correctness of the branch length estimation scheme, when the branches of the topology of T^* are fitted using $\boldsymbol{d}^{T^*}$, their lengths are set to their correct values. Because these are all positive, the scores assigned by ME_x and ME_y to T^* equal precisely the sum L^* of all branch lengths in T^*. Now let $\hat{\boldsymbol{\ell}}$ denote the branch lengths assigned to an incorrect topology using $\boldsymbol{d}^{T^*}$; $x < y$ implies that $L_x(\hat{\boldsymbol{\ell}}) \geq L_y(\hat{\boldsymbol{\ell}})$, and because $\mathrm{ME}_y(\boldsymbol{d}^{T^*})$ is unique and equal to T^*, we also have $L_y(\hat{\boldsymbol{\ell}}) > L^*$. But then $L_x(\hat{\boldsymbol{\ell}}) > L^*$ for any incorrect topology. Because L^* coincides with the score assigned to T^* by ME_x, we can then conclude that $\mathrm{ME}_x(\boldsymbol{d}^{T^*})$ is unique and coincides with T^*. □

A3.4 Consistency of the classic version of Minimum Evolution.

We now concentrate on the consistency of ME_{+1}, which because of Lemma 6, implies that of ME_x for any $x < +1$. We use a standard framework for investigating the consistency of ME_{+1} (e.g. [1]), which consists in verifying a property ("Willson's condition") of branch length estimation in the presence of a special kind of binary distance matrix. In the following, we introduce and state Willson's condition (A3.4.1), then prove some properties of the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae that are useful to verify it (A3.4.2), and finally prove the consistency of ME_{+1} via Willson's condition (A3.4.3). For simplicity, we write $\hat{L}^T(\boldsymbol{\delta})$ as a shorthand for $L_{+1}(\hat{\boldsymbol{\ell}}_T(\boldsymbol{\delta}))$, that is, the tree length (sensu ME_{+1}) resulting from fitting the branch lengths of T using $\boldsymbol{\delta}$.

A3.4.1. Willson's condition [1].

In order to state it (Lemma 7 below), we denote by $\boldsymbol{d}^{S|\overline{S}}$ (where $S \subseteq \{1, 2, \ldots, n\}$ and $\overline{S} = \{1, 2, \ldots, n\} \setminus S$) the following collection of $\binom{n}{2}$ distances, indexed by $i \neq j$:

$$d_{ij}^{S|\overline{S}} = \begin{cases} 1 & \text{if } |S \cap \{i, j\}| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 7. *Adopt any linear branch length estimation scheme such that, for any binary topology T and any proper and nonempty subset S of taxa from T,*

$$\hat{L}^T(\boldsymbol{d}^{S|\overline{S}}) \begin{cases} = 1 & \textit{if } S \textit{ is a clade in } T, \\ > 1 & \textit{otherwise.} \end{cases}$$

Then $\mathrm{ME}_{+1}(\boldsymbol{d}^{T^*}) = T^*$.

Informally, this holds because $\boldsymbol{d}^{T^*}$ is a weighed sum of all $\boldsymbol{d}^{S|\overline{S}}$ corresponding to the clades in T^*, the coefficients of this weighted sum being the branch lengths of T^*. The linearity of $\hat{L}^T$ then implies that, in turn, $\hat{L}^T(\boldsymbol{d}^{T^*})$ is a weighted sum of the branch lengths of T^*, where the coefficients are now either 1 or strictly greater than 1, depending on whether or not the corresponding S is a clade in T. Clearly this weighted sum is minimized when all the clades in T^* are also clades in T, that is for $T = T^*$.

A3.4.2 Tools to verify Willson's condition.

Given any subset of taxa $S \subseteq \{1, 2, \ldots, n\}$ and any clade X in a binary topology T for which a set of $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae is defined, define $p_{S|X}$ as the probability of picking an element of S from the random distribution over X defined by the $\boldsymbol{\gamma}^T$ parameters. That is,

$$p_{S|X} = \sum_{i \in X \cap S} p_{i|X}.$$

Moreover, a clade X is *monochromatic* (w.r.t. S) if either $X \subseteq S$ or $X \subseteq \overline{S} = \{1, 2, \ldots, n\} \setminus S$. In this case it is clear that $p_{S|X} \in \{0, 1\}$.

Lemma 8. *Assume that the branch lengths of T are assigned with $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae using the input distances in $\boldsymbol{d}^{S|\overline{S}}$, for some $S \subseteq \{1, 2, \ldots, n\}$. Then,*

(i) For any two clades X and Y in T, the average distance between them $d_{XY}^{S|\overline{S}} = p_{S|X}p_{\overline{S}|Y} + p_{\overline{S}|X}p_{S|Y}$.

(ii) If branch e belongs to a monochromatic clade, it is assigned length $\hat{\ell}_e = 0$.

(iii) If adjacent branches f and g separate (and are the only ones to separate) two monochromatic clades A_1 and A_2, with $A_1 \subseteq S$ and $A_2 \subseteq \overline{S}$, then $\hat{\ell}_f + \hat{\ell}_g = 1$.

Proof. (i) $d_{XY}^{S|\overline{S}} = \sum_{\substack{i\in X\\ j\in Y}} p_{i|X}p_{j|Y}d_{ij}^{S|\overline{S}} = \sum_{\substack{i\in X\cap S\\ j\in Y\cap\overline{S}}} p_{i|X}p_{j|Y} + \sum_{\substack{i\in X\cap\overline{S}\\ j\in Y\cap S}} p_{i|X}p_{j|Y} = p_{S|X}p_{\overline{S}|Y} + p_{\overline{S}|X}p_{S|Y}$.

(ii) Branch e is either internal or external. We consider here only the case where it is internal, as the external case is analogous (and simpler). We assume clades A_1, A_2, B_1 and B_2 are defined as in Fig. 1(b). Because e belongs to a monochromatic clade, at least three clades out of A_1, A_2, B_1 and B_2 are all subsets of S or all subsets of $\overline{S}$. Without loss of generality, we assume that A_1, A_2, B_1 are all subsets of S. Applying part (i), it is easy to see that this implies $d_{A_1A_2}^{S|\overline{S}} = d_{A_1B_1}^{S|\overline{S}} = d_{A_2B_1}^{S|\overline{S}} = 0$ and $d_{A_1B_2}^{S|\overline{S}} = d_{A_1B_2}^{S|\overline{S}} = d_{A_1B_2}^{S|\overline{S}} = p_{\overline{S}|B_2}$. But then,

$$\begin{aligned}\hat{\ell}_e &= \frac{1}{2}\left[\lambda_{A_1B_1}(d_{A_1B_1}^{S|\overline{S}} + d_{A_2B_2}^{S|\overline{S}}) + (1-\lambda_{A_1B_1})(d_{A_1B_2}^{S|\overline{S}} + d_{A_2B_1}^{S|\overline{S}}) - d_{A_1A_2}^{S|\overline{S}} - d_{B_1B_2}^{S|\overline{S}}\right]\\ &= \frac{1}{2}\left[\lambda_{A_1B_1}p_{\overline{S}|B_2} + (1-\lambda_{A_1B_1})p_{\overline{S}|B_2} - p_{\overline{S}|B_2}\right] = 0.\end{aligned}$$

(iii) Let $B = \{1, 2, \ldots, n\} \setminus (A_1 \cup A_2)$, as in Fig. 1(b). In a way analogous to part (ii), it is easy to check that, independently of f being internal or external, $\hat{\ell}_f = \frac{1}{2}(1 + p_{\overline{S}|B} - p_{S|B})$ and, similarly, $\hat{\ell}_g = \frac{1}{2}(1 + p_{S|B} - p_{\overline{S}|B})$. Therefore, $\hat{\ell}_f + \hat{\ell}_g = 1$.

Points (i), (ii) and (iii) are thus all verified. □

A3.4.3 Consistency of ME_{+1} *with perfect data.*

Proposition 9. *Adopt a branch length estimation scheme based on $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$-formulae. Then,* $\mathrm{ME}_{+1}(\boldsymbol{d}^{T^*}) = T^*$.

Proof. We show that any branch length estimation scheme based on $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$-formulae satisfies Willson's condition (i.e. the hypotheses of Lemma 7), and therefore we must have $\mathrm{ME}_{+1}(\boldsymbol{d}^{T^*}) = T^*$.

First, it is trivial to see that any such branch length estimation scheme is linear. Second, it is correct (Theorem 1), which implies $\hat{L}^T(\boldsymbol{d}^{S|\overline{S}}) = 1$ whenever S is a clade of T: in this case, in fact, $\boldsymbol{d}^{S|\overline{S}}$ is additive with respect to a tree with topology T and with all branches of length 0 except the root branch of S, which has length 1; because of their correctness, the $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae result in assigning T precisely these branch lengths and therefore a total length $\hat{L}^T(\boldsymbol{d}^{S|\overline{S}}) = 1$.

It remains to prove that $\hat{L}^T(\boldsymbol{d}^{S|\overline{S}}) > 1$ whenever S is not a clade of T — for any branch length estimation scheme based on $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$-formulae. We do this by induction on the size of T.

For $n = 3$ taxa, this is trivially true, as all proper, nonempty sets of $\{1, 2, 3\}$ are clades of T.

For $n > 3$, if S is not a clade of T, it is always possible to find a pair of 2-separated clades A_1, A_2 such that $A_1 \subsetneq S$ and $A_2 \subsetneq \overline{S}$, i.e. such that A_1 and A_2 are monochromatic, but $A_1 \cup A_2$ and $B = \{1, 2, \ldots, n\} \setminus (A_1 \cup A_2)$ are not monochromatic. To see this, consider the tree $T^{(S)}$ that is obtained by substituting every monochromatic clade in T with a taxon; because S is not a clade of T, then $T^{(S)}$ must have at least two cherries (i.e., pairs of 2-separated taxa); any of these corresponds to a pair of clades A_1, A_2 in the original tree T with the required properties. Let e, f and g be the root branches of B, A_1 and A_2, respectively, and a their common endpoint, as in Fig. 1(b). Because A_1

and A_2 are monochromatic, it is clear that all branches belonging to these clades are assigned length 0 when the input distances are $\boldsymbol{d} = \boldsymbol{d}^{S|\overline{S}}$ (Lemma 8, part (ii)).

Now let T' be the topology that is obtained by deleting from T all branches belonging to $A = A_1 \cup A_2$, so that a is a leaf of T'. It is clear that there is one-to-one correspondence between the branches/clades of T' and a subset of the branches/clades of T. When calculating the branch lengths in T', we assume that the γ_{ef} and λ_{XY} parameters are the same as those for the corresponding branches/clades in T. Now define the following distances over $\{1, 2, \ldots, n\} \cup \{a\} \setminus A$, i.e. the taxa in T':

$$\begin{aligned} \boldsymbol{d}^{(1)} &= \boldsymbol{d}^{(S\cup\{a\}\setminus A_1)|(\overline{S}\setminus A_2)}, \\ \boldsymbol{d}^{(2)} &= \boldsymbol{d}^{(S\setminus A_1)|(\overline{S}\cup\{a\}\setminus A_2)}, \\ \boldsymbol{d}' &= \gamma_{ef}\boldsymbol{d}^{(1)} + \gamma_{eg}\boldsymbol{d}^{(2)}. \end{aligned}$$

Note that $\boldsymbol{d}'$ coincides with $\boldsymbol{d}$ except for the distances that involve a. For these, we have

$$d'_{aj} = \gamma_{ef} d^{(1)}_{aj} + \gamma_{eg} d^{(2)}_{aj},$$

for every $j \in \{1, 2, \ldots, n\} \setminus A$. Note also that $d^{(1)}_{aj} = d_{A_1 j}$ and $d^{(2)}_{aj} = d_{A_2 j}$, which imply $d'_{aj} = d_{Aj}$. This in turn implies that, for any disjoint clades X and Y in T',

$$d'_{XY} = \begin{cases} d_{XY} & \text{if } a \notin X \cup Y, \\ d_{X'Y} & \text{if } a \in X \text{ and where } X' = X \cup A \setminus \{a\}, \end{cases} \tag{15}$$

that is, the average distances between disjoint clades remain the same when going from $(T, \boldsymbol{d})$ to $(T', \boldsymbol{d}')$. To see this, note that the first case is a simple consequence of the fact that $\boldsymbol{d}'$ coincides with $\boldsymbol{d}$ for the distances that do not involve a. As for the second case, it is a consequence of combining $d'_{aj} = d_{Aj}$ (shown above) with the first case (and it can be easily proved by induction on the size of X).

The important consequence of (15) is that the lengths of branches belonging to B — which only depend on average distances between disjoint clades in T' — remain constant when going from $(T, \boldsymbol{d})$ to $(T', \boldsymbol{d}')$. Therefore, the only difference between the two tree lengths will come from the lengths of branches e, f and g:

$$\hat{L}^T(\boldsymbol{d}) = \hat{L}^{T'}(\boldsymbol{d}') + \hat{\ell}_f + \hat{\ell}_g + \hat{\ell}_e - \hat{\ell}'_e, \tag{16}$$

where $\hat{\ell}_b$ and $\hat{\ell}'_b$ represent the lengths assigned to branch b for $(T, \boldsymbol{d})$ and $(T', \boldsymbol{d}')$, respectively.

Because for ME_{+1} the tree length is a linear function of the branch lengths and the branch lengths themselves are linear functions of the input distances, $\hat{L}^{T'}(\boldsymbol{d}')$ is then linear in $\boldsymbol{d}'$ and we can write

$$\hat{L}^{T'}(\boldsymbol{d}') = \gamma_{ef}\hat{L}^{T'}(\boldsymbol{d}^{(1)}) + \gamma_{eg}\hat{L}^{T'}(\boldsymbol{d}^{(2)}).$$

Because $\boldsymbol{d}^{(1)}$ and $\boldsymbol{d}^{(2)}$ are equal to some $\boldsymbol{d}^{S'|\overline{S'}}$ where S' is a proper and nonempty subset of the $n'(< n)$ taxa in T', we can apply the induction hypothesis and infer that $\hat{L}^{T'}(\boldsymbol{d}^{(1)}) \geq 1$ and $\hat{L}^{T'}(\boldsymbol{d}^{(2)}) \geq 1$ (where equality is achieved when S' is a clade of T'). Therefore $\hat{L}^{T'}(\boldsymbol{d}') \geq 1$.

In order to prove that the tree length in (16) is strictly greater than 1, we then just need to prove that $\hat{\ell}_f + \hat{\ell}_g + \hat{\ell}_e - \hat{\ell}'_e > 0$.

First, because of Lemma 8, part (iii), $\hat{\ell}_f + \hat{\ell}_g = 1$. In order to calculate $\hat{\ell}_e$ and $\hat{\ell}'_e$, it is useful to note that $d_{A_1A_2} = 1$ and, for any clade $B' \subseteq B$, $d_{A_1B'} = p_{\overline{S}|B'} = 1 - p_{S|B'}$ and $d_{A_2B'} = p_{S|B'}$ (Lemma 8, part (i)). Then,

$$\begin{aligned} \hat{\ell}_e &= \frac{1}{2}\left[\lambda_{A_1B_1}(1 - p_{S|B_1} + p_{S|B_2}) + (1 - \lambda_{A_1B_1})(1 - p_{S|B_2} + p_{S|B_1}) - 1 - d_{B_1B_2}\right] \\ &= \frac{1}{2}\left[(1 - 2\lambda_{A_1B_1})(p_{S|B_1} - p_{S|B_2}) - d_{B_1B_2}\right]. \end{aligned}$$

Similarly,

$$\begin{aligned}
\hat{\ell}'_e &= \frac{1}{2}\left[d'_{aB_1} + d'_{aB_2} - d'_{B_1B_2}\right] \\
&= \frac{1}{2}\left[d_{AB_1} + d_{AB_2} - d_{B_1B_2}\right] \\
&= \frac{1}{2}\left[\gamma_{ef}(d_{A_1B_1} + d_{A_1B_2}) + (1-\gamma_{ef})(d_{A_2B_1} + d_{A_2B_2}) - d_{B_1B_2}\right] \\
&= \frac{1}{2}\left[\gamma_{ef}(2 - p_{S|B_1} - p_{S|B_2}) + (1-\gamma_{ef})(p_{S|B_!} + p_{S|B_2}) - d_{B_1B_2}\right] \\
&= \gamma_{ef} + \frac{1}{2}\left[(1-2\gamma_{ef})(p_{S|B_!} + p_{S|B_2}) - d_{B_1B_2}\right].
\end{aligned}$$

Then,

$$\hat{\ell}_f + \hat{\ell}_g + \hat{\ell}_e - \hat{\ell}'_e = 1 - \gamma_{ef} + p_{S|B_1}\left(\gamma_{ef} - \lambda_{A_1B_1}\right) + p_{S|B_2}\left(\gamma_{ef} - (1-\lambda_{A_1B_1})\right).$$

But this is a linear function of $(p_{S|B_1}, p_{S|B_2})$ in the square $[0,1]^2$, and is thus minimized in one of its four vertices. In $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$, the function has values $1-\gamma_{ef}$, $\lambda_{A_1B_1}$, $1-\lambda_{A_1B_1}$, γ_{ef}, respectively. Since these are all strictly greater than 0 by hypothesis, then so is $\hat{\ell}_f + \hat{\ell}_g + \hat{\ell}_e - \hat{\ell}'_e$ and therefore $\hat{L}^T(\boldsymbol{d}) > 1$. □

A3.5 Wrapping it all together.

By applying Proposition 9, Lemma 6 and Proposition 3, we then conclude:

Corollary 10. *Adopt a branch length estimation scheme based on $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$-formulae. Then, for any $x \leq +1$, ME_x is consistent.*

Which, together with Corollary 5 and Assumption 2, completes our proof of Theorem 3.

SI Appendix 4

Here, we prove the efficiency of calculating branch lengths with our formulae in hill climbing heuristics, as stated in Theorem 4. We start by showing that efficient branch length calculations essentially depend on the availability of the average distances between (some) clades in the current tree, and that these can be calculated in quadratic time, which allows us to prove Theorem 4, part (i) (A4.1). When performing an NNI, calculating the new branch lengths can be done efficiently by recalculating only some of these average distances, which leads us to prove Theorem 4, part (ii) (A4.2). Finally, we show that updating the average distances following an NNI can also be done efficiently (A4.3), which is not a claim of Theorem 4, but is nevertheless potentially useful. The results and proofs here are inspired by those of Desper and Gascuel [8]. However, their results were specific to the balanced and OLS branch lengths in combination with the ME_{+1} optimization principle. A key property of these estimators is that, when performing an NNI, the sum of the branch lengths in each of the four corner subtrees around the location of the NNI remains constant. Thanks to this property, the difference between the ME_{+1} lengths of any two NNI neighbors T and T' can be efficiently calculated using simple formulae. This property does not hold in general for $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-estimators, and so we have to recalculate all branch lengths every time we perform an NNI. The good news is that (1) the complexity of each iteration in a hill climbing heuristics for ME (computing the length of all NNI neighbors of a given topology and updating the data structures for the new best topology), which for BME was quadratic in the worst case, remains quadratic in the size of the tree, and that (2) recalculating all branch lengths makes it possible to use optimization principles such as ME_0, ME_{-1} and $\mathrm{ME}_{-\infty}$.

In the following, T, T' and T_i always denote binary topologies, and $\boldsymbol{\gamma}^T$, $\boldsymbol{\gamma}^{T'}$ and $\boldsymbol{\gamma}^{T_i}$, collections of γ_{ef} parameters defined for them, in the way described in the main text.

A4.1 Computing the branch lengths of a fixed topology.

Lemma 11. *Adopt a set of $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae for the branch lengths of T. Given $\boldsymbol{\delta}$ and δ_{XY} for every pair of 3-separated clades X, Y in T, the length of any branch in T can be calculated in $O(1)$ time.*

Proof. The $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae are simple linear combinations of average distances δ_{XY} between 2- and 3-separated clades X and Y and can be computed in $O(1)$ once these average distances are available. Since we assume that the average distances between 3-separated clades are given, it remains to show that δ_{XY} between any pair X, Y of 2-separated clades can be obtained in $O(1)$. But this is trivial: either both X and Y consist of one taxon only, i.e. $X = \{i\}$ and $Y = \{j\}$, in which case $\delta_{XY} = \delta_{ij}$, or at least one of the two clades, say X, is such that $X = X_1 \cup X_2$, where both X_1 and X_2 are clades, in which case $\delta_{XY} = \gamma_{ef}\delta_{X_1Y} + \gamma_{eg}\delta_{X_2Y}$, where e, f and g are the root branches of X, X_1 and X_2, respectively, and both δ_{X_1Y} and δ_{X_2Y} are known, as X_i and Y are 3-separated (for $i \in 1, 2\}$). □

Although the one above is a straightforward observation, it determines the minimum amount of information necessary to determine any branch length in T in constant time. Motivated by it, we define $\Delta^T(\boldsymbol{\gamma}^T)$ as a data structure holding all the average distances δ_{XY} between pairs of disjoint clades in T, and make explicit its dependence on $\boldsymbol{\gamma}^T$. Note that $\Delta^T(\boldsymbol{\gamma}^T)$ specifies the average distances between 3-separated clades as a particular case.

Lemma 12. *Given $\boldsymbol{\delta}$, T and $\boldsymbol{\gamma}^T$, the calculation of $\Delta^T(\boldsymbol{\gamma}^T)$ requires $O(n^2)$ time.*

Proof. Consider any total ordering $A_1, A_2, \ldots, A_{2(2n-3)}$ of the clades in T, such that if $A_k = A_i \cup A_j$ then $i < k$ and $j < k$. Finding one such ordering is trivial and can be done in a number of ways, for example by sorting the clades in ascending order of depth [2], or by rooting the tree in one of its leaves and then performing a postorder traversal, listing the clades oriented away from the root, followed by a preorder traversal, listing the clades oriented towards the root. The following procedure then calculates δ_{XY} for all pairs of clades (including non-disjoint ones):

For $i = 2, \ldots, 2(2n-3)$,

for $j = 1, \ldots, i-1$,

$$\delta_{A_iA_j} = \begin{cases} \delta_{xy} & \text{if } A_i = \{x\} \text{ and } A_j = \{y\}, \\ \gamma_{ef}\delta_{A_{i_1}A_j} + \gamma_{eg}\delta_{A_{i_2}A_j} & \text{if } A_i = A_{i_1} \cup A_{i_2} \text{ for some clades } A_{i_1}, A_{i_2}, \\ \gamma_{ef}\delta_{A_iA_{j_1}} + \gamma_{eg}\delta_{A_iA_{j_2}} & \text{if } A_j = A_{j_1} \cup A_{j_2} \text{ for some clades } A_{j_1}, A_{j_2}. \end{cases} \tag{17}$$

In the second case of (17) we assume that e, f, g are the root branches of A_i, A_{i_1}, A_{i_2}, respectively, whereas in the third case they are the root branches of A_j, A_{j_1}, A_{j_2}, respectively. Note that these two cases are not mutually exclusive, and the result is the same independently of which case is applied. Moreover, because of the way the ordering is defined, we must have $i_1, i_2 < i$, in the second case, or $j_1, j_2 < j$, in the third case, which means that $\delta_{A_{i_1}A_j}$ and $\delta_{A_{i_2}A_j}$ (second case), or $\delta_{A_iA_{j_1}}$ and $\delta_{A_iA_{j_2}}$ (third case) have already been calculated and are available when we calculate $\delta_{A_iA_j}$. Because each $\delta_{A_iA_j}$ can be calculated in constant time, the whole calculation requires $O(n^2)$ time. □

The complexity we obtain in Lemma 12 is optimal. Even if we restrict the calculation to 3-separated clades, we still cannot do better than $O(n^2)$, as the average distances between such pairs of clades still depend on $O(n^2)$ input distances.

Proof of Theorem 4, part (i). Combining Lemma 12 with Lemma 11 yields that the branch lengths determined by a set of $(\boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)$-formulae for a binary topology T can be calculated in $O(n^2)$ time. □

A4.2 Computing the branch lengths of the NNI neighbors of a given topology.

Lemma 13. *Let T and T' be NNI-neighbors and let $\boldsymbol{\gamma}^T$ and $\boldsymbol{\gamma}^{T'}$ be almost identical. Let δ_{XY}* and *δ'_{XY}* denote the average clade distances in $\Delta^T(\boldsymbol{\gamma}^T)$ *and* $\Delta^{T'}(\boldsymbol{\gamma}^{T'})$*, respectively. Then, given $\Delta^T(\boldsymbol{\gamma}^T)$, the calculation of δ'_{XY} for every pair of 3-separated clades X, Y in T', requires $O(n)$ time.*

Proof. We assume that T is as in Fig. 1(b) and T' as in Fig. S2. Let the elements of $\boldsymbol{\gamma}^T$ and $\boldsymbol{\gamma}^{T'}$ be denoted by $\gamma_{e_1e_2}$ and $\gamma'_{e_1e_2}$, respectively.

First, we show that δ'_{XY} is straightforward to obtain in the case of pairs of 3-separated clades in T' such that none or 1 of the 3 branches separating X and Y belongs to one of the corner clades A_1, A_2, B_1, B_2. Let $A_1 = A'_1 \cup A''_1$, $A_2 = A'_2 \cup A''_2$, $B_1 = B'_1 \cup B''_1$ and $B_2 = B'_2 \cup B''_2$, where all the sets involved are also clades of T' (and therefore T). It is trivial to verify that

$$\begin{array}{llll} \delta'_{A_1A_2} = \delta_{A_1A_2}, & \delta'_{B_1B_2} = \delta_{B_1B_2}, & \delta'_{A_1B_1} = \delta_{A_1B_1}, & \delta'_{A_2B_2} = \delta_{A_2B_2}, \\ \delta'_{A'_1B_2} = \delta_{A'_1B_2}, & \delta'_{A''_1B_2} = \delta_{A''_1B_2}, & \delta'_{A_1B'_2} = \delta_{A_1B'_2}, & \delta'_{A_1B''_2} = \delta_{A'_1B''_2}, \\ \delta'_{A'_2B_1} = \delta_{A'_2B_1}, & \delta'_{A''_2B_1} = \delta_{A''_2B_1}, & \delta'_{A_2B'_1} = \delta_{A_2B'_1}, & \delta'_{A_2B''_1} = \delta_{A_2B''_1}, \end{array} \tag{18}$$

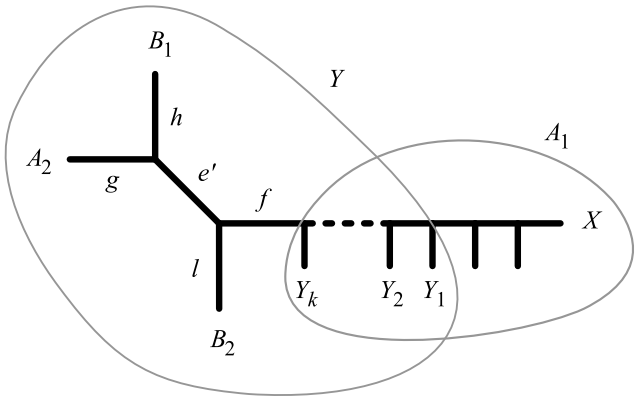
as $\boldsymbol{\gamma}^T$ and $\boldsymbol{\gamma}^{T'}$ are the same within all the clades in the subscripts above. Now observe that $A_1 \cup B_2$ and $A_2 \cup B_1$ are clades in T' but not in T. Their average distances with other 3-separated clades must then be obtained with expressions such as

$$\delta'_{A'_1\,A_2\cup B_1} = \gamma'_{e'g}\delta_{A'_1A_2} + \gamma'_{e'h}\delta_{A'_1B_1}. \tag{19}$$

(Similar formulae are easy to obtain for $\delta'_{A''_1\,A_2\cup B_1}, \delta'_{B'_2\,A_2\cup B_1}, \delta'_{B''_2\,A_2\cup B_1}$ and $\delta'_{A'_2\,A_1\cup B_2}$, $\delta'_{A''_2\,A_1\cup B_2}$, $\delta'_{B'_1\,A_1\cup B_2}$, $\delta'_{B''_1\,A_1\cup B_2}$.) We have therefore proved that these δ'_{XY} can be obtained from one or two corresponding entries in $\Delta^T(\boldsymbol{\gamma}^T)$ in $O(1)$ time.

We still have to show how to derive δ'_{XY} when 2 or all 3 of the branches separating X and Y belong to a corner clade. Without loss of generality, we assume this clade to be A_1. If both $X, Y \subset A_1$, then we trivially have $\delta'_{XY} = \delta_{XY}$. Assume then $Y \supseteq B_1 \cup B_2 \cup A_2$. Let clades $Y_1, Y_2, \ldots, Y_k$ be defined as in Fig. S2 (and note that if $Y = B_1 \cup B_2 \cup A_2$, no such Y_i clade is defined). Also, if clade Y contains clade Y' in T, define $p_{Y'|Y}$ as the probability that the random walk defined by the $\boldsymbol{\gamma}^T$ parameters reaches Y', assuming that it enters Y from its root branch: $p_{Y'|Y} = \gamma_{e_0e_1} \cdot \gamma_{e_1e_2} \cdot \ldots \cdot \gamma_{e_{k-1}e_k}$, where

FIGURE S2. Illustration for the proof of Lemma *13*.



e_0 is the root branch of Y and $e_1, e_2, \ldots, e_k$ are the branches on the path between the roots of Y and Y'. Define $p'_{Y'|Y}$ similarly for T' and $\boldsymbol{\gamma}^{T'}$. Then,

$$\begin{aligned}
\delta'_{XY} &= p'_{Y_1|Y}\delta'_{XY_1} + \cdots + p'_{Y_k|Y}\delta'_{XY_k} + p'_{B_1|Y}\delta'_{XB_1} + p'_{B_2|Y}\delta'_{XB_2} + p'_{A_2|Y}\delta'_{XA_2} \\
&= p'_{Y_1|Y}\delta'_{XY_1} + \cdots + p'_{Y_k|Y}\delta'_{XY_k} + p'_{B_1\cup B_2\cup A_2|Y}(\gamma'_{fe'}\gamma'_{e'h}\delta'_{XB_1} + \gamma'_{fl}\delta'_{XB_2} + \gamma'_{fe'}\gamma'_{e'g}\delta'_{XA_2}) \\
&= p_{Y_1|Y}\delta_{XY_1} + \cdots + p_{Y_k|Y}\delta_{XY_k} + p_{B_1\cup B_2\cup A_2|Y}(\gamma'_{fe'}\gamma'_{e'h}\delta_{XB_1} + \gamma'_{fl}\delta_{XB_2} + \gamma'_{fe'}\gamma'_{e'g}\delta_{XA_2}),
\end{aligned}$$

where the last equality uses the almost identity of $\boldsymbol{\gamma}^T$ and $\boldsymbol{\gamma}^{T'}$. Similarly,

$$\delta_{XY} = p_{Y_1|Y}\delta_{XY_1} + \cdots + p_{Y_k|Y}\delta_{XY_k} + p_{B_1\cup B_2\cup A_2|Y}(\gamma_{fe}\gamma_{eh}\delta_{XB_1} + \gamma_{fe}\gamma_{el}\delta_{XB_2} + \gamma_{fg}\delta_{XA_2}).$$

Therefore,

(20)
$$\delta'_{XY} - \delta_{XY} = p_{B_1\cup B_2\cup A_2|Y}\left[(\gamma'_{fe'}\gamma'_{e'h} - \gamma_{fe}\gamma_{eh})\delta_{XB_1} + (\gamma'_{fl} - \gamma_{fe}\gamma_{el})\delta_{XB_2} + (\gamma'_{fe'}\gamma'_{e'g} - \gamma_{fg})\delta_{XA_2}\right].$$

It is easy to derive similar equations for the cases where (a) $X \subset A_2$, $Y \supseteq B_1 \cup B_2 \cup A_1$, (b) $X \subset B_1$, $Y \supseteq A_1 \cup A_2 \cup B_2$, (c) $X \subset B_2$, $Y \supseteq A_1 \cup A_2 \cup B_1$, which allow us to derive δ'_{XY} in $O(1)$ time from four entries in $\Delta^T(\boldsymbol{\gamma}^T)$ (including δ_{XY}) and $p_{B_1\cup B_2\cup A_1|Y}$, $p_{A_1\cup A_2\cup B_2|Y}$, $p_{A_1\cup A_2\cup B_1|Y}$ for cases (a), (b), (c), respectively. Now consider the following procedure:

(1) For every clade $Y \supseteq B_1 \cup B_2 \cup A_2$, calculate $p_{B_1\cup B_2\cup A_2|Y}$.

(1a, 1b, 1c) Do the same as above, for every clade $Y \supseteq B_1 \cup B_2 \cup A_1$, for every $Y \supseteq A_1 \cup A_2 \cup B_2$ and for every $Y \supseteq A_1 \cup A_2 \cup B_1$.

(2) Use (20), or similar equation, to derive δ'_{XY} for all 3-separated clades X, Y in T' such that 2 or all 3 of the branches separating X and Y belong to a corner clade A_1, A_2, B_1, B_2.

(3) Use the simple equations in (18) and (19) to calculate δ'_{XY} for the remaining 3-separated clades.

Step (1) can be done in $O(n)$ time, by starting with the smallest clades including $B_1 \cup B_2 \cup A_2$ and using the derived values to calculate those for the larger clades. The same holds for steps (1a, 1b, 1c). Then, each δ'_{XY} can be calculated in $O(1)$ time. Since there are $O(n)$ 3-separated pairs of clades, the entire algorithm runs in $O(n)$ time and thus Lemma 13 is proved. □

We are now ready to complete the proof of Theorem 4.

Proof of Theorem 4, part (ii). Recall that all the branch lengths in T_0 and its NNI-neighbors $T_1, T_2, \ldots, T_{2(n-3)}$ are defined by $(\boldsymbol{\gamma}^{T_i}, \boldsymbol{\lambda}^{T_i})$-formulae, with the constraint that $\boldsymbol{\gamma}^{T_i}$ and $\boldsymbol{\gamma}^{T_0}$ are almost identical. We wish to prove that the branch lengths of $T_1, T_2, \ldots, T_{2(n-3)}$ can be calculated in $O(n^2)$ time. Let $\delta^{(i)}_{XY}$ denote the average clade distances in $\Delta^{T_i}(\boldsymbol{\gamma}^{T_i})$. Because of Lemma 12, $\Delta^{T_0}(\boldsymbol{\gamma}^{T_0})$ can be calculated in $O(n^2)$ time. From this, the calculation of $\delta^{(i)}_{XY}$ for every pair of 3-separated clades

in T_i, requires $O(n)$ time (Lemma 13). Combining this to Lemma 11 yields that all $O(n)$ branch lengths in T_i can be calculated in $O(n)$ time. Since there are $O(n)$ neighbors of T and each is treated in $O(n)$ time, the whole calculation requires $O(n^2)$ time. □

A4.3 Updating the accessory information when performing an NNI. The proof of Lemma 13 above suggests a related result which may also be useful for hill climbing, when the best NNI-neighbor T' of T has been identified and we need to calculate $\Delta^{T'}(\boldsymbol{\gamma}^{T'})$ in order to explore efficiently the NNI-neighborhood of T'. Define $\mathrm{diam}(T)$, the diameter of T, as the maximum number of branches separating any two leaves of T.

Proposition 14. *Let T and T' be NNI-neighbors and let $\boldsymbol{\gamma}^T$ and $\boldsymbol{\gamma}^{T'}$ be almost identical. Given $\Delta^T(\boldsymbol{\gamma}^T)$, its update into $\Delta^{T'}(\boldsymbol{\gamma}^{T'})$ requires $O(n \cdot \mathrm{diam}(T))$ time.*

Proof. Let T be as in Fig. 1(b) and T' as in Fig. S2. Let δ_{XY} and δ'_{XY} denote average clade distances from $\Delta^T(\boldsymbol{\gamma}^T)$ and $\Delta^{T'}(\boldsymbol{\gamma}^{T'})$, respectively. In order to obtain $\Delta^{T'}(\boldsymbol{\gamma}^{T'})$ from $\Delta^T(\boldsymbol{\gamma}^T)$, one needs to calculate the entries of $\Delta^{T'}(\boldsymbol{\gamma}^{T'})$ that have no corresponding entry in $\Delta^T(\boldsymbol{\gamma}^T)$ or those that have changed. These are the δ'_{XY} for all pairs of clades X, Y in T' such that some of the branches f, g, h, l belong to X or Y. The only case where both X and Y have at least one of f, g, h, l belonging to them is that where $X = A_1 \cup B_2$ and $Y = A_2 \cup B_1$. In this case, δ'_{XY} can be obtained from $\Delta^T(\boldsymbol{\gamma}^T)$ with $\delta'_{XY} = \gamma'_{e'f}\gamma'_{e'g}\delta_{A_1A_2} + \gamma'_{e'f}\gamma'_{e'h}\delta_{A_1B_1} + \gamma'_{e'g}\gamma'_{e'l}\delta_{A_2B_2} + \gamma'_{e'h}\gamma'_{e'l}\delta_{B_1B_2}$.

All the other cases correspond to a pair of clades X, Y such that one of them, say X, is included in one of the four corner clades A_1, A_2, B_1, B_2 and the other, Y, includes two or three of the other clades (see e.g. Fig. S2, where $X \subset A_1$ and $Y \supseteq B_1 \cup B_2 \cup A_2$). It is clear that for any such X, the number of possible choices for Y equals the number of branches in the path starting with e' and ending in the root of X. In other words, there are $O(n)$ possible choices for X, each of which corresponds to $\mathrm{diam}(T)$ choices for Y. Therefore we need to consider $O(n\,\mathrm{diam}(T))$ pairs of clades. For each of these pairs, we now prove that δ'_{XY} can be calculated in $O(1)$ time from $\Delta^T(\boldsymbol{\gamma}^T)$, once steps (1, 1a, 1b, 1c) from the proof of Proposition 13 have been executed (in $O(n)$ time): if $Y = A_2 \cup B_1$ or $Y = A_1 \cup B_2$, then it is straightforward to obtain δ'_{XY} as $\gamma'_{e'g}\delta_{XA_2} + \gamma'_{e'h}\delta_{XB_1}$ or as $\gamma'_{e'f}\delta_{XA_1} + \gamma'_{e'l}\delta_{XB_2}$, respectively; otherwise, if Y includes three of A_1, A_2, B_1, B_2, it is easy to see that (20), and similar equations for $X \subset A_2, B_1, B_2$, still hold (without the assumption, made in the proof of Lemma 13, that X and Y are 3-separated). It is then possible to calculate each δ'_{XY} in constant time. □

References

[1] Willson, S. (2005) Consistent formulas for estimating the total lengths of trees. *Discr. Appl. Math.* **148**, 214–239.

[2] Mihaescu, R & Pachter, L. (2008) Combinatorics of least squares trees. *Proc. Natl. Acad. Sci. USA* **105**, 13206–13211.

[3] Denis, F & Gascuel, O. (2003) On the consistency of the minimum evolution principle of phylogenetic inference. *Discr. Appl. Math.* **127**, 63–77.

[4] Fitch, W & Margoliash, E. (1967) Construction of phylogenetic trees. *Science* **155**, 279–284.

[5] 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.

[6] Pauplin, Y. (2000) Direct calculation of a tree length using a distance matrix. *J. Mol. Evol.* **51**, 41–47.

[7] Gascuel, O, Bryant, D, & Denis, F. (2001) Strengths and limitations of the minimum evolution principle. *Syst. Biol.* **50**, 621–627.

[8] Desper, R & Gascuel, O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comp. Biol.* **9**, 687–705.