# Combinatorics of distance-based tree inference

Fabio Pardi[1] and Olivier Gascuel

Institut de Biologie Computationelle, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, Centre National de la Recherche Scientifique, Université Montpellier II, 34095 Montpellier Cedex 5, France

Edited by* Joseph Felsenstein, University of Washington, Seattle, WA, and approved May 23, 2012 (received for review November 29, 2011)

**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.**

phylogenetics | distance-based methods | statistical consistency | tree rearrangement

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

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

where the $\delta_{ij}$ 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 $\delta_{ij}$: 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

Author contributions: F.P. and O.G. designed research; F.P. performed research; and F.P. and O.G. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1] To whom correspondence should be addressed. E-mail: pardi@lirmm.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1118368109/-/DCSupplemental.

APPLIED MATHEMATICS

EVOLUTION

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 (8, 14, 15); 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, because it can be shown that the variances of the distance estimates are approximately multiplicative for large distances (9, 22, 23).

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, 21, 24–26)]. This addresses the question by Mihaescu and Pachter (ref. 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 (2, 4, 27) (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 $\delta_{ij}$ 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, because 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 $d_{ij}^T$ 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 $T$) (28).

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 $\gamma_{ef}$ and $\gamma_{eg} = 1 - \gamma_{ef}$, respectively. We require $0 < \gamma_{ef}, \gamma_{eg} < 1$ (note the strict inequalities). These parameters define a (nonzero) probability of reaching any branch of $T$ from any other branch of $T$.

This also defines a probability distribution over the leaves of any clade $A$: For any $i \in A$, let $p_{i|A} = \gamma_{e_0 e_1} \cdot \gamma_{e_1 e_2} \cdot \ldots \cdot \gamma_{e_{k-1} e_k}$, 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 Fig. S1) Clearly, the probabilities $\{p_{i|A} | i \in A\}$ form a distribution over $A$. We can then define the average distance $\delta_{AB}$ 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 $\delta_{AB}$ so defined depend on the $\gamma_{ef}$ 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 $\delta_{iA}$ (or $\delta_{Ai}$) instead of $\delta_{\{i\}A}$.

In addition to the $\gamma_{ef}$ probabilities defined for each pair of adjacent branches $(e, f)$, we introduce a parameter $\lambda_{XY}$ 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

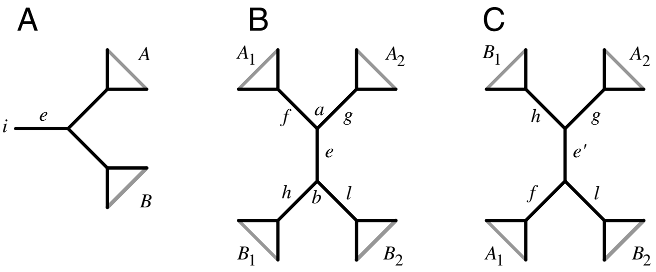

**Fig. 1.** Standard naming of clades and branches when (*A*) *e* is external, and (*B*) *e* is internal. (*C*) An NNI-neighbor (around *e*) of the tree in *B*.

$A = A_1 \cup A_2$ 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 $\lambda_{XY}$ 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 $T$ ($\gamma_{ef}, \gamma_{fg}$ , and $\gamma_{ge}$ determine $\gamma_{eg}, \gamma_{fe}$ , and $\gamma_{gf}$) 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 $T$:

$(\gamma^T, \boldsymbol{\lambda}^T)$**-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 $T$:

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

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 $\gamma_{ef}$ 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, 29) 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, because 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 $\ell_e$ 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 $\mathrm{Var}[\delta_{ij}]^{-1}$, because 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) (30). 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 (25) (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 $b = \frac{1}{2}$, 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 $\gamma_{ef}$ parameters satisfying P1. This condition implies that any of $\gamma_{ef}$, $\gamma_{fg}$ and $\gamma_{ge}$ can be determined from the other two and P2 implies that all the $\lambda_{XY}$ parameters are determined by the $\gamma_{ef}$ 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, 31). Second, in the case of distances estimated from molecular sequences, we note that for many models of sequence evolution [for instance Jukes–Cantor (32); see ref. 1 or appendix B in ref. 2 for the general technique], the variance of $\delta_{ij}$ 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 (33), or its variant with variances proportional to $d_{ij}^2$ (13), are used in practice with $\delta_{ij}$ 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 $\delta_{ij}$ 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 $\delta_{ij}$ 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 (34–36).

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 $\mathrm{ME}_{-1}$ (14), $\mathrm{ME}_{+1}$ (15, 37), and $\mathrm{ME}_0$ (8). Assuming that a tree has been assigned the branch lengths $\hat{\ell}_e$, $\mathrm{ME}_i$ defines its length as

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

The three versions of ME then differ in how they deal with negative branch lengths when calculating tree length: $\mathrm{ME}_{+1}$ adds together all branch lengths irrespective of their sign, whereas $\mathrm{ME}_0$ ignores negative branch lengths and $\mathrm{ME}_{-1}$ takes their absolute value. Gascuel et al. (24) previously named $\mathrm{ME}_{+1}$, $\mathrm{ME}_0$ and $\mathrm{ME}_{-1}$, "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, $\mathrm{ME}_{+1}$, $\mathrm{ME}_0$ *and* $\mathrm{ME}_{-1}$ *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 $w_{ij}$) (24). Furthermore, Theorem 3 generalizes all previously known cases of consistency for the ME principle (17, 25, 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 (21, 27).

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 $\boldsymbol{\gamma}^T = (\gamma_{e_1e_2})$ and $\boldsymbol{\gamma}^{T'} = (\gamma'_{e_1e_2})$, defined for $T$ and $T'$, respectively, are *almost identical*, if $\gamma_{e_1e_2} = \gamma'_{e_1e_2}$, 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 $\frac{1}{2}$ 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, 31). Theorem 4 also predicts the efficiency of hill climbing for OLS: The $\gamma_{ef}$ 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 $\gamma_{ef}$ 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 $\text{ME}_{+1}$), 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 (29)] 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 $\text{ME}_{-1}$ and $\text{ME}_0$ above), in combination with any of the classic branch length estimation schemes (e.g., OLS or that used in BME). Alternatively, our framework enables the use of 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.

**ACKNOWLEDGMENTS.** This work was supported by the Labex NUMEV. We wish to thank Radu Mihaescu for discussions on the consistency of minimum evolution, and Joe Felsenstein, together with the anonymous referees, for constructive comments on the manuscript.

1. Felsenstein J (2004) *Inferring Phylogenies* (Sinauer Associates, Sunderland, MA), Chaps 13, 14.
2. Yang Z (2006) *Computational Molecular Evolution* (Oxford Univ Press, Oxford, UK).
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 Associates, Sunderland, MA), Chap 11.
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 Comput Biol* 9:687–705.
8. Swofford D, Olsen G, Waddell P, Hillis D (1996) *Molecular Systematics*, eds D Hillis, C Moritz, and B Mable (Sinauer Associates, Sunderland, MA), 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 Associates, Sunderland, MA).
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) *Conceptual and Numerical Analysis of Data*, ed O Opitz (Springer, 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 Bioinf* 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. Gascuel O, Bryant D, Denis F (2001) Strengths and limitations of the minimum evolution principle. *Syst Biol* 50:621–627.
25. Denis F, Gascuel O (2003) On the consistency of the minimum evolution principle of phylogenetic inference. *Discrete Appl Math* 127:63–77.
26. Willson S (2005) Consistent formulas for estimating the total lengths of trees. *Discrete Appl Math* 148:214–239.
27. Semple C, Steel M (2003) *Phylogenetics* (Oxford Univ Press, Oxford, UK).
28. Buneman P (1971) *Mathematics in the Archaelogical and Historical Sciences*, ed F Hodson (Edinburgh Univ Press, Edinburgh, UK), pp 387–395.
29. Gascuel O, Steel M (2006) Neighbor-joining revealed. *Mol Biol Evol* 23:1997–2000.
30. Aitken AC (1935) On least squares and linear combinations of observations. *Proc R Soc Edinburgh A* 55:42–48.
31. Vinh S, von Haeseler A (2005) Shortest triplet clustering: Reconstructing large phylogenies using representative sets. *BMC Bioinf* 6:92.
32. Jukes T, Cantor C (1969) *Mammalian Protein Metabolism*, ed H Munro (Academic, Waltham, MA), pp 21–132.
33. Beyer W, Stein M, Smith T, Ulam S (1974) A molecular sequence metric and evolutionary trees. *Math Biosci* 19:9–25.
34. Atteson K (1999) The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica* 25:251–278.
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.