Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and its Relationship to Weighted Least-squares Tree Fitting

Richard Desper, Olivier Gascuel

Theoretical Foundation of the Balanced Minimum Evolution Method of Phylogenetic Inference and Its Relationship to Weighted Least-Squares Tree Fitting

*Richard Desper**[1] *and Olivier Gascuel*†[1]

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland; and †Equipes Méthodes et Algorithmes pour la Bioinformatique, LIRMM, Montpellier, France

Due to its speed, the distance approach remains the best hope for building phylogenies on very large sets of taxa. Recently (R. Desper and O. Gascuel, J. Comp. Biol. **9**:687–705, 2002), we introduced a new "balanced" minimum evolution (BME) principle, based on a branch length estimation scheme of Y. Pauplin (J. Mol. Evol. **51**:41–47, 2000). Initial simulations suggested that FASTME, our program implementing the BME principle, was more accurate than or equivalent to all other distance methods we tested, with running time significantly faster than Neighbor-Joining (NJ). This article further explores the properties of the BME principle, and it explains and illustrates its impressive topological accuracy. We prove that the BME principle is a special case of the weighted least-squares approach, with biologically meaningful variances of the distance estimates. We show that the BME principle is statistically consistent. We demonstrate that FASTME only produces trees with positive branch lengths, a feature that separates this approach from NJ (and related methods) that may produce trees with branches with biologically meaningless negative lengths. Finally, we consider a large simulated data set, with 5,000 100-taxon trees generated by the Aldous beta-splitting distribution encompassing a range of distributions from Yule-Harding to uniform, and using a covarion-like model of sequence evolution. FASTME produces trees faster than NJ, and much faster than WEIGHBOR and the weighted least-squares implementation of PAUP*. Moreover, FASTME trees are consistently more accurate at all settings, ranging from Yule-Harding to uniform distributions, and all ranges of maximum pairwise divergence and departure from molecular clock. Interestingly, the covarion parameter has little effect on the tree quality for any of the algorithms. FASTME is freely available on the web.

Introduction

Distance-based methods for phylogeny reconstruction represent the best hope for accurately building phylogenies on very large sets of taxa. Distance methods have been shown to be statistically consistent in all settings, as opposed to parsimony methods, and have a huge speed advantage over parsimony and likelihood methods. This advantage in speed allows the user to build larger trees and/or use bootstrapping methods. For many years, the gold standards in this subdiscipline of phylogeny estimation have been, for speed, Neighbor-Joining (Saitou and Nei 1987) and its offshoots BIONJ (Gascuel 1997*a*) and WEIGHBOR (Bruno, Socci, and Halpern 2000); and, for accuracy, the Fitch-Margoliash weighted least-squares algorithm (Fitch and Margoliash 1967), as implemented by Felsenstein (1997). Recently, we introduced a heuristic implementation of a new "balanced" minimum evolution approach to phylogeny estimation (Desper and Gascuel 2002), based on a branch length estimation scheme of Pauplin (2000). Initial simulations on a 2,000-tree data set suggested that our program, FASTME, was at least as accurate as the Fitch-Margoliash approach to tree fitting, and we proved that FASTME uses an algorithm whose running time was significantly better than Neighbor-Joining (NJ).

The current work is divided into two parts: an investigation of the theoretical underpinnings of the balanced minimum evolution (BME) approach, and a discussion of extensive simulations comparing the BME approach to three popular distance methods. First, we demonstrate that the balanced minimum evolution branch lengths represent, in fact, a special type of weighted least-squares tree fitting, where the variances for each leaf-to-leaf distance estimate are assumed to be exponentially related to the topological distance in the tree between the pair of leaves. Next, we demonstrate that this approach is consistent: as distance estimates converge to true evolutionary distances, the FASTME tree converges to the true tree. Our proof is modeled on the proof of Rzhetsky and Nei (1993) that demonstrated consistency for a minimum evolution approach when branch lengths were assigned by ordinary least-squares fitting. Next, we also note a feature of FASTME trees: whereas many distance algorithms produce branches with confusing negative branch lengths, FASTME only produces positive branch lengths.

The second major section of the paper is an expanded simulation over a generalized model of tree topology selection, branch length assignment, and DNA evolution. We used the Aldous (1996) model for random tree topology selection, a generalization of the two most common random distributions on tree topologies: the uniform distribution and the biologically relevant Yule-Harding (Yule 1925; Harding 1971) distribution. We created a departure from the molecular clock using another random factor, and evolved 600 base-pair DNA sequences for each tree topology, using a covarion model analogous to Galtier (2001) to determine the evolutionary rate changes of the sites. The resulting 5,000 data sets cover a wide variety of tree topologies, model parameters, and evolutionary conditions. We used FASTME, NJ, WEIGHBOR, and PAUP's heuristic weighted least-squares topology search to determine a tree for each data set. Our simulations

[1] Both authors contributed equally to this work.

Key words: minimum evolution, least-squares, distance-based phylogenetic inference, consistency, method comparison using simulations.

E-mail: gascuel@lirmm.fr.

Mol. Biol. Evol. 21(3):587–598. 2004
DOI: 10.1093/molbev/msh049
Advance Access publication December 23, 2003
Molecular Biology and Evolution vol. 21 no. 3

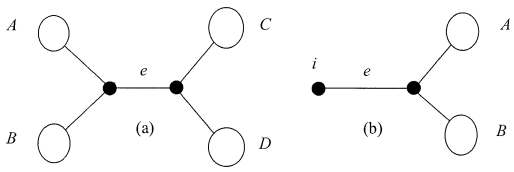

FIG. 1.—*a*, Internal branch; *b*, external branch.

demonstrate the superiority of FASTME at estimating the original topology, as measured by counting the number of branches in the output tree that correspond to branches of the model tree.

Theoretical Foundation of the BME Approach

Notation and Background

Let δ_{ij} be the estimate of the evolutionary distance between taxa i and j, obtained from sequences or any other data, and $\Delta = (\delta_{(ij)})$ be a vector containing all the δ_{ij} estimates, with (ij) denoting the index of the pair (i, j). Let T be the tree being studied, d_{ij} the distance induced by T between taxa i and j (i.e., d_{ij} is equal to the length of the path connecting i to j in T), and $D = (d_{(ij)})$ a vector containing all of the inter-taxa distances d_{ij}. Using matrix notation, the branch lengths of T can be represented by a vector $L = (l_k)$, with l_k denoting the length of branch k, while the topology (shape) of T can be represented by a 0–1 matrix $S = (s_{(ij)k})$, such that $s_{(ij)k}$ is equal to 1 if the branch k lies on the path connecting i and j, and $s_{(ij)k}$ equals 0 otherwise.

Using this notation, we observe that $D = SL$, and the branch lengths are estimated by minimizing the difference between the observation Δ and D. The ordinary least-squares (OLS) approach involves selecting branch lengths $\hat{L}$ minimizing the squared Euclidean fit between Δ and D, i.e., $(D - \Delta)^t(D - \Delta)$. This yields $\hat{L} = (S^tS)^{-1}S^t\Delta$. However, this approach implicitly assumes that each estimate δ_{ij} has the same variance and is independent, a supposition not generally true because large distances are much more variable than short distances, and because the sequences in question share a common evolutionary history. To address this problem, Fitch and Margoliash (1967), Felsenstein (1997), and others have proposed using a weighted least-squares approach—i.e., minimizing $(D - \Delta)^tV^{-1}(D - \Delta)$, where V is the diagonal matrix containing the variances of the δ_{ij} estimates. This approach yields $\hat{L} = (S^tV^{-1}S)^{-1}S^tV^{-1}\Delta$. The weighted least-squares (WLS) approach accounts for the variable variances of the distance estimates, but not for their dependencies. The generalized least-squares approach (Bulmer 1991; Susko 2003) uses the full variance-covariance matrix V, and then accounts for dependencies, but it is rarely used because the covariances of the distance estimates are usually poorly known, and because this approach requires a lot of computing time.

In the minimum-evolution framework, we select the topology with the shortest estimated length. The tree length of T is equal to the sum of its branch lengths, and it may be written as $l(T) = \mathbf{1}^tL$, where $\mathbf{1}$ is a vector of 1's. The standard WLS tree length estimate is then equal to:

$$\hat{l}(T) = \mathbf{1}^t(S^tV^{-1}S)^{-1}S^tV^{-1}\Delta. \quad (1)$$

However, computing this formula in its matrix form is expensive, and faster methods have been proposed. In the OLS framework, Vach (1989), Rzhetsky and Nei (1993), Gascuel (1997*b*), and others, have exhibited analytical formulae for branch length estimates that considerably accelerate the computations. In the WLS case, Bryant and Waddell (1998) have shown that calculations can be simplified thanks to the special form of the matrices, and their solution is implemented in version 4.0 of PAUP* (Swofford 1996).

Pauplin (2000) followed another approach, modifying the OLS analytical formulae. Consider the two possible branch configurations in figure 1.

When e is internal (fig. 1*a*), we have:

$$\hat{l}(e) = \frac{1}{4}(\delta^T_{AC} + \delta^T_{BD} + \delta^T_{AD} + \delta^T_{BC}) - \frac{1}{2}(\delta^T_{AB} + \delta^T_{CD}), \quad (2)$$

and when e is external (fig. 1*b*):

$$\hat{l}(e) = \frac{1}{2}(\delta^T_{iA} + \delta^T_{iB} - \delta^T_{AB}). \quad (3)$$

In these formulae, δ^T_{AB} represents the weighted or "balanced" average distance between the taxa of A and B. If A and B each contain only one taxon, denoted as a and b, respectively, then $\delta^T_{AB} = \delta_{ab}$, while if one of the two, say B, is made of two subtrees denoted as B_1 and B_2, then $\delta^T_{AB} = (\delta^T_{AB_1} + \delta^T_{AB_2})/2$. Note that in this scheme the distance between A and B depends not only on distances between pairs of taxa contained in A and B, but also on the topology of T. In Appendix 1, we demonstrate that these formulae consistently estimate the branch lengths; i.e., if Δ exactly corresponds to a tree T, then the estimated length of e is equal to its real length in T. Also, Pauplin (2000) showed that in this framework, the tree length itself can be expressed analytically by the simple and elegant following formula:

$$\hat{l}(T) = \sum_{i,j} 2^{1-p_{ij}}\delta_{ij}, \quad (4)$$

where p_{ij} denotes the topological distance between i and j, i.e., the number of branches in the path from i to j in T. The consistency of the tree length estimate (eq. 4) has been shown by Semple and Steel (2003), and its minimization forms the basis of the balanced minimum evolution principle (BME).

Finally, we demonstrated (Desper and Gascuel 2002) that in the BME framework, computing the tree length, as expressed by equations 2, 3, and 4, is unnecessary for tree inference. Indeed, our tree building and swapping algorithms only exploit the difference in tree length corresponding to a "nearest neighbor interchange" (NNI). Assume that T is the tree of figure 1*a*, and that T' is obtained from T by exchanging subtrees B and C. We then have:

$$\hat{l}(T) - \hat{l}(T') = \frac{1}{4}[(\delta^T_{AB} + \delta^T_{CD}) - (\delta^T_{AC} + \delta^T_{BD})]. \quad (5)$$

The speed of our algorithms is explained by the simplicity of this equation, notably regarding equation (1). Moreover,

we observed in simulations that BME is at least as accurate as the traditional WLS approaches, and it is much more accurate than OLS, a fact that was unexplained and is the subject of the following sections.

From Balanced Tree Length to Minimum Variance Tree Length Estimation

Fitch and Margoliash (1967) assumed that the variances of the δ_{ij} estimates are proportional to δ_{ij}^2, the default option in both the FITCH (Felsenstein 1997) and PAUP* (Swofford 1996) programs. Another common approximation (e.g., Gascuel 1997*b*) is simply to set the variance of δ_{ij} to be proportional to δ_{ij}. Better approximations have been found (Nei, Stephens, and Saitou 1985; Nei and Jin 1989; Bulmer 1991), basically indicating that the variance grows exponentially as a function of δ_{ij}. WEIGHBOR (Bruno, Socci, and Halpern 2000), for example, uses this latter approximation.

We assume here that the variance of δ_{ij} is proportional to $2^{p_{ij}}$, where p_{ij} is the topological distance between i and j. In other words, we have:

$$Var(\delta_{ij}) = k2^{p_{ij}}, \quad (6)$$

where k is a constant and should be thought of as incorporating the inverse of the sequence length that is contained in all variance formulae (Sasko 2003). Even when topological and evolutionary distances differ, they are strongly correlated, especially when the taxa are homogeneously sampled. Our approximation (eq. 6) is then likely capturing most of the exponential approximations by Bulmer (1991) and others. Under these assumptions, we show that there exists a strong relationship between balanced tree length and the weighted least-squares framework, as expressed by the following theorem.

Theorem 1: Let T be the tree being studied and use the notation above defined. Assuming that the variances of the δ_{ij} estimates are defined by equation 6, and assuming the WLS framework (the covariances are null), the balanced tree length estimation of T in equation 4 is then: (1) the minimum variance tree length estimator of T; (2) identical to the length defined by equation (1).

The proof of Theorem 1 is given in Appendix 2. The first part of this theorem implies that, assuming equation 6, the tree length given by BME is as reliable as possible. Because we select the shortest tree, reliability in tree length estimation is of great importance and tends to minimize the probability of selecting a wrong tree. Moreover, it is well known in statistics that rough variance values such as our approximation (eq. 6) are usually sufficient. We thus expect that BME computations will provide quite reliable branch and tree length estimates.

The second part of Theorem 1 indicates that FASTME should be close to FITCH and PAUP*, which use equation 1 to define the tree length, and, to a lesser extent, to WEIGHBOR, which is based on a different tree-building strategy but is also a WLS approach. However, because not all these programs use the same approximations for the variances nor the same criteria and algorithms, some differences between them can still be expected.

Consistency of the BME Principle of Phylogenetic Inference

Statistical consistency is a central issue in phylogenetic inference. In the case of distance-based methods, it is defined as follows: Let T be the correct tree, D the associated tree distance matrix, and Δ the matrix of estimated distances. Assuming that Δ is a consistent estimate of D, the more data we have (e.g., the longer the sequences used to estimate the pairwise distances), the closer Δ is to D. Statistical consistency of tree inference then means that T is obtained with certainty as soon as Δ is sufficiently close to D. In other words, assuming that the model used to estimate the pairwise distance matrix is satisfied, the more data we have, the higher the probability to recover the correct tree. This property is essential and has been discussed at length in the past (e.g., Felsenstein 1978). Consistent methods contrast with inconsistent ones (e.g., parsimony in some cases), which may converge toward a wrong tree when the amount of data increases.

The ordinary least-squares version of the minimum evolution principle was shown to be consistent by Rzhetsky and Nei (1993), a result generalized by Denis and Gascuel (2003). However, as explained in the previous section, BME is a weighted least-squares version of the minimum evolution principle, and it was demonstrated (Gascuel, Bryant, and Denis 2001) that in some cases, depending on the variance matrix, this version might be inconsistent. So our aim in this section is to verify the consistency of BME.

A direct consequence of statistical consistency is that when $\Delta = D$ the correct tree T has the shortest length among all possible tree topologies. This shortest length property is necessary for consistency, but also sufficient. Indeed, the length associated with a tree topology relative to a distance matrix is a continuous function of this matrix. Therefore, when Δ is sufficiently close to D, the estimated tree lengths relative to Δ and to D become close, and T becomes the shortest tree for Δ as it already is for D. When this occurs, T is then inferred with certainty from Δ. To this end, we prove the following theorem:

Theorem 2: Let T be the correct tree and D the corresponding distance matrix, and assume that the matrix Δ of distance estimates is equal to D. Let W be any tree topology, and define its length estimate $\hat{l}(W)$ by equation 4 or, equivalently, by combining equations 1 and 6. Then, $\hat{l}(W) > \hat{l}(T) = l(T)$, whenever $W \neq T$.

Theorem 2 demonstrates the consistency of BME. The equality $\hat{l}(T) = l(T)$ simply results from the consistency of equation 1 or equation 4 in estimating the tree length. The rest of the proof ($\hat{l}(W) > l(T)$) is given in Appendix 3, and it follows a line similar to Rzhetsky and Nei's (1993) for the OLS version of ME. A difference between the two proofs is that ours is very closely tied to our balanced nearest-neighbor interchange (BNNI) algorithm: we show that when $W \neq T$ we can apply to W a nearest-neighbor interchange (NNI) that makes its length shorter. This seems to indicate that our simple BNNI algorithm is itself consistent, as confirmed by numerous computer simulations (not

shown), but a complete formal proof remains to be done. See Appendix 3 for more details and technical comments.

Branch Length Positivity

Neighbor-Joining (Saitou and Nei 1987) and related algorithms often output trees with negative branch length estimates, which are biologically meaningless (Swofford 1996) and have to be interpreted as irresolutions (in the case of internal branches). In contrast, the FITCH and PAUP* weighted least-squares algorithms impose positivity (as the default option), but this constraint often yields branches with zero length, which also correspond to irresolutions. Moreover, when the positivity constraint is removed, the topological accuracy of these algorithms tends to be lower (Kuhner and Felsenstein 1994).

Our BNNI algorithm is used in FASTME software to improve a starting tree by performing NNIs, based on equation 5, until no more NNI decreases the length of the current tree. The following theorem indicates that, after running BNNI, all branches in the output tree have positive length. Thus BNNI trees tend to be better resolved than NJ or FITCH trees, and this fact partially explains the good performance of FASTME, as we shall see in the simulation section.

Theorem 3: Let Δ be any evolutionary distance matrix, and suppose T is a tree that is a local minimum for a BNNI topology search. Moreover, assume that the branch lengths in T are obtained from equation 2 or 3. Then, the length estimate of every branch in T is positive.

The proof is in Appendix 4. We have assumed that Δ is a distance; i.e., it satisfies the triangle inequality. This should be the case for any practical data set, except when some sites are unknown for some of the sequences, or correspond to gaps. Then, depending on the distance computation options, the matrix Δ might (slightly) violate the triangle inequality and FASTME might output trees with negative external branches. In any case the internal branches are always positive.

Simulation Results

In Desper and Gascuel (2002) we considered simulated data with trees generated by a Yule-Harding (Yule 1925; Harding 1971) process, with random variation of branch lengths and random perturbation from a molecular clock. In the current work, we consider a broader simulation using a more general model of tree topology generation and branch length assignment, and also using a covarion-like model of sequence evolution. Because distance methods mostly address the reconstruction of large data sets, we used trees with 100 taxa. Moreover, the parameters defining the generation process were chosen from the study of numerous recently published phylogenies and should cover most practical situations. These data sets, as well as FASTME, can be downloaded from http://www.ncbi.nlm.nih.gov/CBBresearch/Desper/FastME.html or from http://www.lirmm.fr/w3ifa/MAAS/. We first describe the tree and sequence generation processes, then the topological error measures we used to compare the inferred and true trees, and, lastly, we discuss the performances of various distance methods with respect to these error measures.

Random Tree Generation

The Yule-Harding branching process (Yule 1925; Harding 1971) is a standard, biologically relevant method for generating phylogenetic trees. However, the uniform distribution on phylogenies is another natural approach for generating topologies when comparing tree inference methods. It has been argued (Gascuel 2000, Nakhleh et al. 2001) that method performance could vary depending on the tree-generation scheme. We thus chose the Aldous (1996) beta-splitting model, which generalizes both of the aforementioned distributions. In this model, topology generation is directed by a parameter denoted as β: $\beta = -1.5$ corresponds to the uniform distribution, whereas $\beta = 0$ defines the Yule-Harding distribution. In our experiments β was uniformly drawn from the interval $[-1.5, 0]$ for each new tree, which was then generated using this value. One advantage of this approach is that it provides a much larger variety of trees than solely using either a Yule-Harding or a uniform distribution, as the Yule-Harding distribution diverges strongly from the uniform distribution when considering trees with 100 taxa (Gascuel 2000). Notably, Yule-Harding trees have a moderate topological diameter (about 22 branches in average), while uniform tree diameter is much larger (about 38 branches in average).

After topology generation each branch was assigned a length. We first used the standard coalescent model (Kuhner and Felsenstein 1994) to assign branch lengths, yielding a molecular clock on the tree. We then perturbed this molecular clock by multiplying every branch length (independently) by $(1 + X)$, where X was an exponential variable with parameter λ. The factor $(1 + X)$ was used (as opposed to, say, X) to avoid an excessive number of very small branches. The parameter λ was identical within each tree, but it varied from tree to tree. It was selected as $\lambda = 0.3/(0.01 + U)$, where U was uniformly drawn from the interval [0, 1]. If $U = 0$, λ becomes very large and then $X \approx 0$; i.e., the tree remains close to the molecular clock. If $U = 1$, the variance of X is large, and the tree tends to clearly depart from a molecular clock. The observed departure from the molecular clock, as measured by the ratio between the longest and shortest root-to-leaf lineages, was in the range [1.4, 6.0] (fig. 4), with a median value of 3.1 (where 1.0 represents the perfect molecular clock). Finally, trees were rescaled so that their total length would be uniform between 0.5 and 8.0. The maximum pairwise divergence was then in the range [0.1, 1.1] (fig. 5), with a median value of 0.55.

Sequence Evolution Model

Covarion-like models have been advocated by several authors (Fitch 1971; Galtier 2001; Lopez, Casane, and Philippe 2002; Huelsenbeck 2002) to accurately represent sequence evolution. In these models, evolutionary rates differ from site to site, and the rate of a given site can

change along the course of evolution. Some sites are slow in some parts of the tree but fast in the other parts, to account for structural or functional changes of the sequences being studied in certain clades.

We used nucleotide sequences with 600 sites, evolving under a model analogous to Galtier's (2001). Four evolutionary rates were considered, defined by a discrete gamma distribution (Yang 1994) with parameter 1.0, which corresponds to moderate rate heterogeneity. The rate of each site was drawn at the root of each tree with equal probability, and then changed at each new speciation event (independently) with probability ξ as one proceeded from root to each leaf. When a rate changed, it changed to each of the other three rates with equal probability. The parameter ξ was identical within each tree, but it differed from one tree to another. Its value was uniformly drawn from the interval [0, 1/98], where the number 98 was chosen to correspond to the number of speciation events (other than the root) in a tree with 100 taxa. When $\xi = 0$, the data set corresponds to the standard four-rate model of Yang (1994), with a gamma parameter equal to 1.0. When $\xi = 1/98$, the expected number of rate changes per site is equal to 1.0; this means that some sites would witness no rate changes, while other sites (among the 600 in the simulation) could witness up to 5–6 rate changes in the whole tree.

Once site rates have been determined as explained above, making branches shorter or longer depending on the site considered, every site evolved under the standard Kimura (1981) two-parameter model with a transition/transversion ratio of 2.0. The number of substitutions that effectively occurred along every branch was stored; we obtained this way the "observed" tree, whose topology was identical to that of the true tree, but whose branch lengths equaled the actual number of substitutions along each branch, divided by the sequence length.

Finally, a distance matrix was computed from the sequences using the Nei and Jin (1989) estimate for gamma distributed rates with parameter 1.0. When $\xi = 0$ this estimate is almost unbiased (we only neglect the sparseness of rates); but when $\xi \neq 0$, the covarion effect is not taken into account by the distance estimation (and no distance correction allows for this). Thus, we can measure the robustness of each method to model violation: as the value of ξ grows, the violation of the model grows worse, leading to increasing problems with distance estimation, and (we expect) less accuracy with any tree-reconstruction method.

Topological Error Measures

To quantify the topological gap between the inferred tree and the correct tree, and also to compare tree inference methods, most authors use the Robinson and Foulds (1981) topological distance (RF). Let T be the observed tree, $\hat{T}$ the inferred tree, $S(X, \delta)$ the set of internal branches of tree X whose lengths are greater than or equal to δ, n the number of taxa, and s the sequence length. In the context of comparing tree topologies, the Type I error is the number of inferred branches that do not belong to the correct tree, and the Type II error is the number of branches in the correct tree that are missing from the inferred tree. The standard RF distance between T and $\hat{T}$ is the sum of the Type I and Type II errors, which are denoted and computed as follows:

$$E_1(\hat{T}, T, \delta) = |S(\hat{T}, \delta) - S(T, \delta)|,$$
$$E_2(\hat{T}, T, \delta) = |S(T, \delta) - S(\hat{T}, \delta)|,$$

where, for the RF distance, $\delta = 0$.

However, branches that are not supported by any substitution in observed tree T cannot be recovered except by chance. We define the irresolution of T, $I(T)$, to be the number of such branches. We should then have approximately:

$$E_2(\hat{T}, T, 0) \approx E_2(\hat{T}, T, \delta) + I(T)$$

when $0 < \delta \leq 1/s$, e.g., $\delta = 1/2s$. $E_2(\hat{T}, T, 1/2s)$ represents the true Type II error of $\hat{T}$ (Kumar 1996). Moreover, $E_2(\hat{T}, T, 0)$ is always less that the sum of the two right-hand terms.

Consider an internal branch e in $\hat{T}$. If the length estimate of e is small or negative, $\hat{T}$ does not indicate any substitution on e, and e should then be considered to be an irresolution. The difference with T is that branch lengths in $\hat{T}$ are not restricted to sparse values of the form m/s, where m is an integer. So, we have to fix a threshold and the simplest choice, analogous to interval estimation in statistics, is to decide that branches with length less than $1/2s$ are not supported by any substitution, whereas longer branches are likely to have undergone one or more substitutions. The irresolution of $\hat{T}$, $I(\hat{T})$, is defined to be the number of non-supported branches. We cannot exclude the possibility that branches with length less than $1/2s$ are true branches (any threshold is imperfect), but we should still have, to some extent, an approximation of the form:

$$E_1(\hat{T}, T, 0) \approx E_1(\hat{T}, T, 1/2s) + I(\hat{T}).$$

$E_1(\hat{T}, T, 1/2s)$ and $E_2(\hat{T}, T, 1/2s)$ are clearly more appropriate than $E_1(\hat{T}, T, 0)$ and $E_2(\hat{T}, T, 0)$ to evaluate the performance of $\hat{T}$ in estimating the topology of T. Moreover, the above equations indicate that the standard topological error (RF) can be approximately decomposed into the sum of the true Type I and Type II errors plus the irresolution of the true and estimated trees. Figure 2 provides an illustration of these error functions for NJ and FASTME trees when using the data sets described above. It can be seen that, with low divergence, inferred and observed trees are poorly resolved and most of the RF error is due to irresolution. With higher divergence, the trees are much better resolved, but the fundamental errors of the inference methods become higher because of saturation, and thus the RF distance is also high. The best results are obtained with moderate divergence. Finally, it appears that for any divergence rate and error measure, and also regarding the resolution of the inferred tree, BME is vastly superior to NJ. We detail this observation in the next section.

Results

Our simulations yielded 5,000 data sets generated as described above, each with 100 sequences of 600 sites, and we used the PHYLIP program DNADIST to calculate

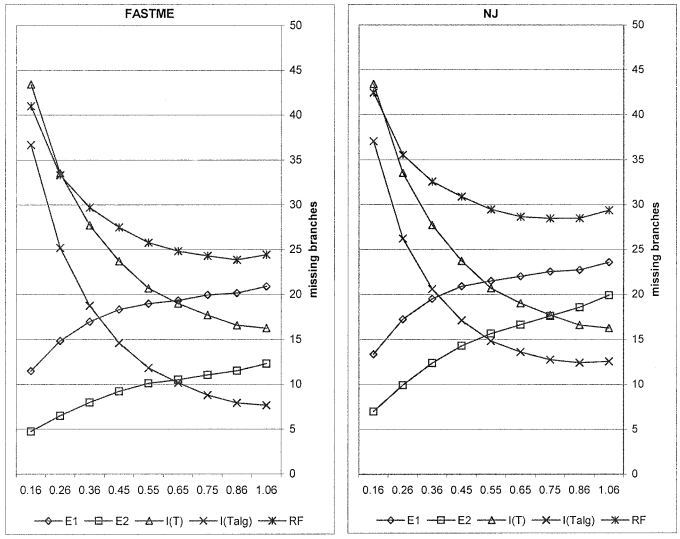

FIG. 2.—Topological error measures versus maximum pairwise divergence. NOTE.—E_1 and E_2 are the true Type I and Type II errors, respectively; $I(T)$ and I(Talg) are the fundamental irresolution of the observed and inferred trees, respectively; RF is equal to half (for the sake of readability) of the standard topological error.

the Nei and Jin (1989) distance estimates. Four distance methods were compared:

- NJ (Saitou and Nei 1987) which was taken from PAUP* (Swofford 1996).
- WEIGHBOR version 1.2, available at http://www.t10.lanl.gov/billb/weighbor/.
- PAUP*'s *hsearch* command (WLS in the following), which was used to find a locally optimal topology, starting by iteratively adding taxa to partial trees, and then using tree bisection-reconnection (TBR) transformations to optimize the weighted least-squares criterion, with a positivity constraint on branch lengths imposed. TBR topology searches separate a tree into two parts, and they try every pair of branches as possible reconnection points. Variances were assumed to be proportional to the square of the observed distances. We allowed 60 s of searching for each iteration of *hsearch*, as this appeared to be enough to produce a solution that was a local optimum with regard to TBR topology searching.
- FASTME with default settings, which involves calculating an initial tree using a greedy minimum evolution algorithm, and using BNNI postprocessing to improve this initial tree in the sense of the BME criterion.

Table 1
Summary Statistics

Algorithm	*RF*	E_1	E_2	$I(\hat{T})$	$I(T)$
FASTME	58.06	17.65	9.25	16.75	25.15
WEIGHBOR	61.50	18.10	11.59	18.64	25.15
WLS	62.08	18.91	11.28	17.52	25.15
NJ	64.99	20.09	14.49	19.56	25.15

NOTE.—*RF* is the standard topological error; E_1 and E_2 correspond to the true Type I and Type II errors, respectively, using $\delta = 1/2s$; $I(\hat{T})$ and $I(T)$ are the irresolutions of the inferred and observed trees, respectively, using the same δ threshold. All these measures are expressed in number of missing branches: *RF* is in the range $[0, 2 \times 97 = 194]$, while other measures are within [0, 97], where 97 corresponds to the number of internal branches in a 100-taxon tree.

Table 1 summarizes the results for the entire data set of 5,000 trees, using the error measures defined above. The irresolution of the observed tree, $I(T)$, which corresponds to the number of branches not supported by any substitution, is quite high (25.15) and explains almost half of the RF topological error. The Type I error (E_1) is always more than the Type II error (E_2), indicating that a more severe threshold could be used to discard branches in the inferred tree. This phenomenon is particularly sensitive for FASTME ($E_1 = 18.64$, $E_2 = 9.25$), which is explained by the fact that this program provides trees better resolved than other methods we considered. However, regarding all error measures FASTME is best. Although the WEIGHBOR trees are approximately as good as the FASTME trees when comparing by E_1, FASTME has a clear advantage over all other algorithms with respect to E_2.

We also considered this data set with regard to four variables:

- The Aldous β shape parameter used while generating the initial topology.
- The observed departure from the molecular clock, measured by the ratio between the longest and shortest root-to-leaf lineages.
- The maximum pairwise divergence measured on the true tree.
- The covarion parameter, corresponding to the expected number of rate changes per site.

The first three parameters control the topology and the branch lengths of the tree, and the fourth controls the substitution process. We sorted the 5,000 data sets with respect to each of the four variables. We used this sorted list to create nine subsets of the data set for each parameter: For $i = 0, \ldots, 8$, we considered the subset of the data defined by those data sets whose parameter values lay ordinally in the interval of the form $[500i + 500i + 1{,}000]$.

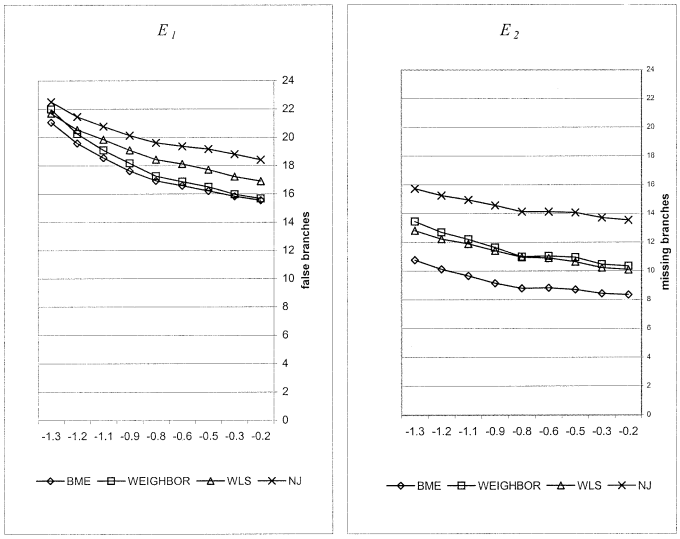

FIG. 3.—Type I and Type II error versus Aldous β shape parameter. NOTE.—Type I error: E_1; Type II error: E_2; $\beta = -1.5$ corresponds to uniform distribution on phylogenies, whereas $\beta = 0$ corresponds to Yule-Harding distribution.

Figures 3–6 show the error rates E_1 and E_2 for each algorithm over each data subset.

In figure 3, we see that all tree topologies are easier to recover as the tree distribution moves from a uniform distribution ($\beta = -1.5$) to a Yule-Harding distribution ($\beta = 0$). This is very likely explained by the fact that, with the uniform distribution, trees have, on average, much larger topological diameters, and thus larger maximum pairwise divergences, than with the Yule-Harding distribution (see also figure 5).

In figure 4, we see the relationship between the departure from the molecular clock and the error measures. As the data sets diverge from a molecular clock, reconstruction of the observed tree becomes more difficult. However the difference remains slight, at about 3.0 branches for E_1 and 1.5 branches for E_2 between two extreme subintervals for all methods.

Figure 5 shows the relationship between the topological error measures and the maximum pairwise divergence. Unsurprisingly, the error increases with divergence, because of saturation. This parameter is very sensitive, the difference between extreme subintervals being about 10 for all methods, considering both E_1 and E_2.

Figure 6 shows the values of the error measures for each interval when the data sets are sorted according to the parameter of the covarion model. Quite surprisingly, the presence of a high covarion effect (the expected number of rate changes per site is equal to 1.0) has almost no influence on topological accuracy. The difference between extreme subintervals is below 1.5 branches for all methods and for both E_1 and E_2, even when the general tendency is, as expected, that the trees with a large covarion effect are (slightly) more difficult to recover than those without such an effect. This finding is quite reassuring given the fact that (DNA, RNA, or protein) sites were certainly subjected to different evolutionary pressures in different parts of the Tree of Life, and thus observed changes in the substitution rate during the course of evolution. According to our results, distance-based reconstructions then seem to be robust regarding this phenomenon.

In all of figures 3–6, we see that the BME algorithm outperforms the WLS approach, which outperforms the NJ algorithm. This ordering is true for all of the intervals looked at, over all of the parameters selected, and for both types of error. The WEIGHBOR algorithm is superior to WLS by E_1, and in some intervals it approaches FASTME; but with respect to E_2, WEIGHBOR is worse than WLS and FASTME.

It must be underscored that the good topological accuracy of FASTME does not correspond to an increase in computing times relative to the other algorithms, but rather a decrease. Indeed, the decrease is quite sharp in comparison to all of the other algorithms except for NJ. Using a two-way 2.2 GHz DELL PE2650 running Linux 2.4 to handle 1,000 (as commonly used in bootstrap studies) data sets with 100 taxa and 600 sites (as those above described), FASTME requires 34.2 s, NJ requires 2 min, WEIGHBOR requires 451 min, and (PAUP*) WLS requires 560 min. DNADIST from the PHYLIP package (Felsenstein 1989) requires 44 min to compute the distance matrices. (For more comparisons, see Desper and Gascuel 2002.)

Discussion

Speed and accuracy are the goals of any distance algorithm for phylogeny reconstruction or estimation. The need for accuracy is self-evident, whereas the need for speed is caused by the increase in the size of data sets, which leads to an explosion in the number of possible topologies and often renders difficult or infeasible other, slower phylogenetic estimation approaches, such as maximum parsimony and maximum likelihood. Consistency is also a necessity for any distance algorithm, as we wish to know that better estimates of evolutionary

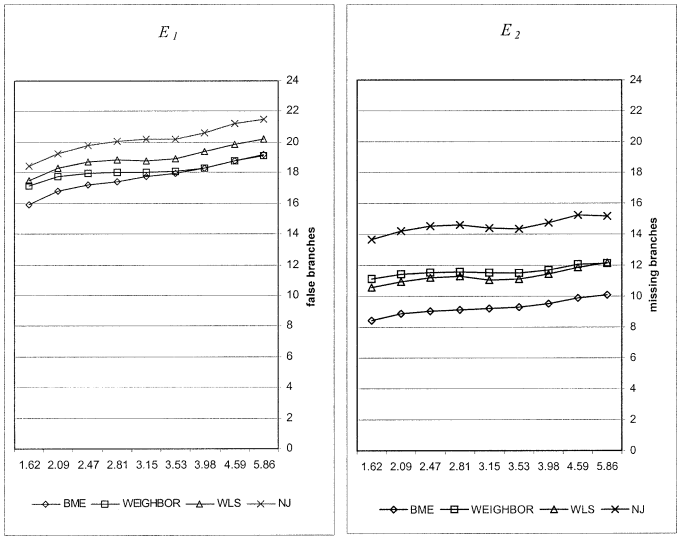

FIG. 4.—Type I and Type II error versus departure from molecular clock. NOTE.—See note to figure 3. The departure from molecular clock is measured by the ratio between the longest and shortest root-to-leaf lineages, the perfect molecular clock then corresponding to 1.

distances can provide a more accurate picture of the actual evolutionary history for any set of species. Also, it is desirable that a distance algorithm provide an output tree free of internal branches with meaningless negative lengths.

The trees produced by the FASTME program, using BNNI postprocessing, make strides towards all of these goals. Earlier work has demonstrated that the running time for FASTME far outstrips that of the leading traditional algorithms, even those of the NJ family. The current simulations show that BME trees are more accurate than even traditional WLS trees. The finding that the BME principle is actually a novel form of the WLS approach, with biologically realistic weights, may explain this advantage. It is unsurprising that the FASTME approach to tree reconstruction is consistent, but the discovery that FASTME will never output negative branch lengths represents another advantage over popular and fast methods such as NJ, BIONJ, and WEIGHBOR.

Appendix 1

Assume that the matrix Δ exactly corresponds to the tree T. We demonstrate that equations 2 and 3 of Pauplin (2000) are consistent in estimating branch lengths—i.e., that the estimated length of any branch using Δ is equal to its true length in T. The balanced average distances between subtrees are calculated using relatively less weight on pairs of taxa that are separated by numerous branches. Let A and B be two non-intersecting subtrees of T whose roots are separated by r_{AB} branches. It is easily seen that the following equations hold:

$$\delta^T_{AB} = \sum_{a,b} 2^{r_{AB}-p_{ab}}\delta_{ab}, \qquad \sum_{a,b} 2^{r_{AB}-p_{ab}} = 1, \qquad (7)$$

where a and b are any taxa from A and B, respectively. Equation 2 can then be rewritten as:

$$\hat{l}(e) = \sum_{a,b,c,d} 2^{4-p_{ab}-p_{cd}} \times \left[\frac{1}{4}(\delta_{ac}+\delta_{bd}+\delta_{ad}+\delta_{bc}) - \frac{1}{2}(\delta_{ab}+\delta_{cd})\right],$$

where c and d also are any taxa from C and D, respectively. In this expression the inner bracket is equal to $l(e)$ for any values of a, b, c, and d (e.g., Rzhetsky and Nei 1993), whereas the sum of the weights is equal to 1, as can be seen from equation 7. In the same way, equation 3 can be rewritten as:

$$\hat{l}(e) = \sum_{a,b} 2^{4-p_{ia}-p_{ib}}\left[\frac{1}{2}(\delta_{ia}+\delta_{ib}-\delta_{ab})\right], \qquad (8)$$

where the inner bracket is also equal to $l(e)$, and the sum of the weights is again 1. Thus, the result holds for external branches as well as for internal branches.

Appendix 2

We first demonstrate part (1) of Theorem 1. Any linear tree length estimator can be written as $F^t\Delta$, where $F=(f_{(ij)})$ is a vector of $f_{(ij)}$ coefficients. For F to be a consistent estimator of $l(T)$, we must have $F^tD = F^tSL = \mathbf{1}^tL$. Because this property must hold for any L, this implies:

$$S^tF = \mathbf{1} \qquad (9)$$

and, clearly, this property is not only necessary but also sufficient to ensure the consistency of F. Within the WLS framework, the variance of any linear estimator F satisfies:

$$Var(F^t\Delta) = \sum_{i,j} v_{ij}f^2_{ij}, \qquad (10)$$

where the notation $f_{ij}=f_{(ij)}$ and $v_{ij}=Var(\delta_{ij})$ is used for the sake of simplicity. Combining equations 9 and 10, we see that finding the minimum variance tree length estimator of T is equivalent to solving:

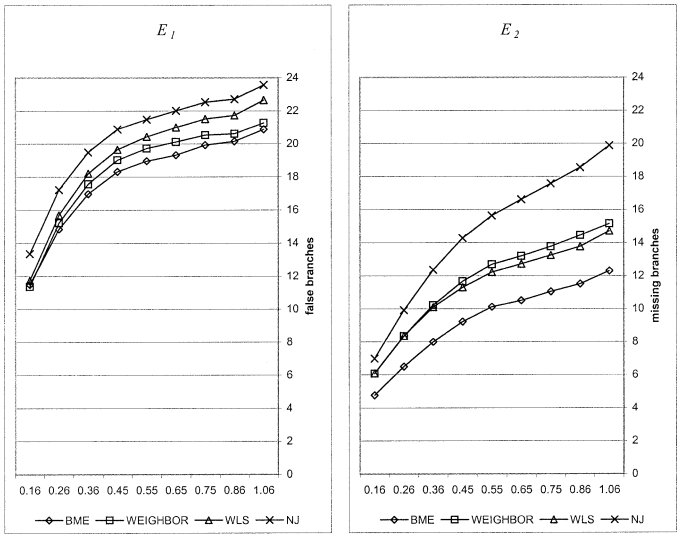

FIG. 5.—Type I and Type II error versus maximum pairwise divergence. NOTE.—See note to figure 3. The maximum pairwise divergence is measured on the true tree in expected number of substitutions per site.

$$\text{Minimize}: \sum_{i,j} v_{ij} f_{ij}^2, \qquad (11)$$

$$\text{subject to: } S^t F = \mathbf{1}.$$

Each row of S^t corresponds to a branch of T. Letting e be any branch of T and $[i, j]$ be the path from i to j in T, the constraints may then be written as:

$$\forall\, e \in T: \sum_{i,j:\, e \in [i,j]} f_{ij} = 1.$$

Because the constraints are linear and the cost function is quadratic, the minimization problem (11) admits a unique solution. Letting μ_e be the Lagrange multiplier associated with e, this solution is defined by the following linear system:

$$\forall\, i,j: 2 v_{ij} f_{ij} = \sum_{e \in [i,j]} \mu_e, \qquad \forall\, e \in T: \sum_{i,j:\, e \in [i,j]} f_{ij} = 1. \quad (12)$$

The balanced tree length estimation of T is defined by $f_{ij} = 2^{1-p_{ij}}$, and it has been demonstrated (Semple and Steel 2003) that this formula consistently estimates the length of T. Using the property of equation 9, this implies that the f_{ij}s so defined satisfy the second set of constraints. Now, assuming equation 6, we have $2 v_{ij} f_{ij} = 4k$, and the first set of equations becomes:

$$\forall\, i,j: \sum_{e \in [i,j]} \mu_e = 4k. \qquad (13)$$

This system (equation 13) corresponds to fitting the branch lengths of T so that the distance between any taxon pair is equal to $4k$; the unique solution is $\mu_e = 2k$ when e is an external branch, and $\mu_e = 0$ otherwise. In other words, we have found the solution of equation 12 and $f_{ij} = 2^{1-p_{ij}}$ defines the minimum variance tree length estimator of T, thus finishing the proof of part (1) of the theorem.

Let us now turn our attention to part (2). We demonstrate that coefficients from equation (1) also satisfy the linear system (equation 12) and are then identical to BME coefficients. First, it is well known that equation 1 consistently estimates the tree length; corresponding coefficients then satisfy the second set of constraints, as expressed by:

$$F^t S = (\mathbf{1}^t (S^t V^{-1} S)^{-1} S^t V^{-1}) S = \mathbf{1}^t.$$

Second, the first set of equations can be rewritten as $VF = SM$, where M is a vector containing the Lagrange multipliers. We then have:

$$\begin{aligned} VF &= V(\mathbf{1}^t (S^t V^{-1} S)^{-1} S^t V^{-1})^t, \\ &= S(S^t V^{-1} S)^{-1} \mathbf{1} \end{aligned}$$

and then:

$$M = (S^t V^{-1} S)^{-1} \mathbf{1}.$$

In other words, we have found the values of the Lagrange multipliers. This concludes the proof of part (2).

Given part (1), part (2) of Theorem 1 seems natural. Indeed, equation 1 defines the minimum variance branch length estimators. But those estimators are non-independent, and it is not trivial that the minimum tree length variance estimator is equal to the sum of minimum variance branch length estimators. Moreover, it is easily seen from above equations that this result extends to any diagonal V matrix.

Appendix 3

In this appendix, we prove Theorem 2. The line of reasoning is similar to Rzhetsky and Nei's (1993) for the OLS version of ME, but, to provide an independent proof, we include all of the relevant details.

We first introduce some more notation and definitions. Removing any branch in T induces a split (or bipartition) of the taxon set, which is denoted as $X \mid Y$, where X and Y are the two induced subsets. The set of

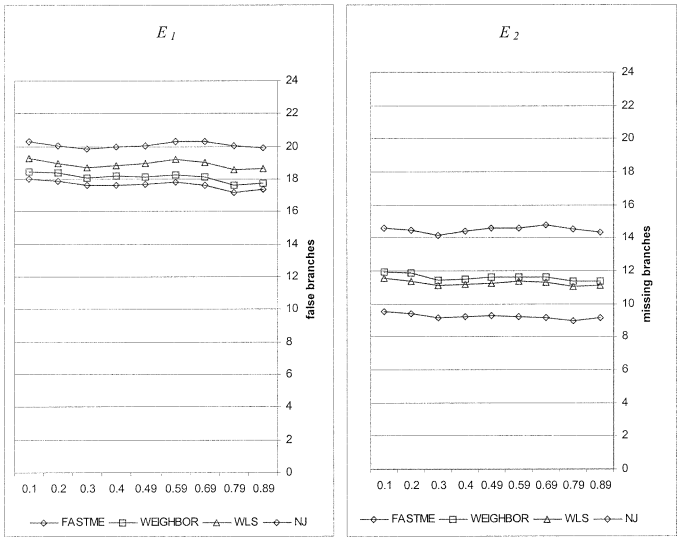

FIG. 6.—Type I and Type II error versus covarion parameter. NOTE.—See note to figure 3. The covarion parameter is the expected number of rate changes per site in the whole tree; 0.0 corresponds to the standard discrete gamma distribution of rates.

splits induced by T is denoted as $S(T)$. Letting $X \mid Y$ be any split, we define the split metric $D^{X \mid Y} = (d_{ij}^{X \mid Y})$ by: $d_{ij}^{X \mid Y} = 1$ when $i \neq j$ and $|\{i, j\} \cap X| = 1$, and $d_{ij}^{X \mid Y} = 0$ otherwise. $D^{X \mid Y}$ is the metric that is obtained from T by having all branch lengths equal to zero, except for the branch corresponding to $X \mid Y$, which has length 1. Let D be the metric corresponding to T and $l(X \mid Y)$ be the length in T of the branch inducing $X \mid Y$. It is easily seen (Bandelt and Dress 1992) that:

$$D = \sum_{X \mid Y \in S(T)} l(X \mid Y) D^{X \mid Y}. \qquad (14)$$

Balanced tree length is a linear function of D (see equation 1). Let $\hat{l}(W, D)$ denote the estimated length of any tree topology W when $\Delta = D$. We then have from equation 14:

$$\hat{l}(W, D) = \sum_{X \mid Y \in S(T)} l(X \mid Y) \hat{l}(W, D^{X \mid Y}).$$

To demonstrate $\hat{l}(W, D) > \hat{l}(T, D)$ it is sufficient to demonstrate that this inequality holds for any split metric of T. If $X \mid Y$ is in both $S(T)$ and $S(W)$, one sees from equations 2 and 3 that $\hat{l}(W, D^{X \mid Y}) = \hat{l}(T, D^{X \mid Y}) = 1$. Because W and T are different, at least one split of T is not in W. Let us suppose $X \mid Y$ belongs to $S(T)$ but not $S(W)$. In this case, we will demonstrate:

$$\hat{l}(W, D^{X \mid Y}) > \hat{l}(T, D^{X \mid Y}) = 1. \qquad (15)$$

We color the leaves of W according to $X \mid Y$: let X be colored white and Y be colored black. Following the proof of Rzhetsky and Nei, we shall change the tree topology W to a tree topology W' that splits X from Y via a series of topological transformations, $W = W_0 \rightarrow W_1 \rightarrow \cdots W_t = W'$. Each transformation $W_i \rightarrow W_{i+1}$ (1) merges two disjoint monochromatic clusters into one (or four into two), and (2) decreases the estimated length of the corresponding tree. Because of (1), the number of clusters decreases until we have one branch with all the black leaves on one side and all the white ones on the other side; i.e., the corresponding topology splits X from Y and has length 1 (see above). Because of (2), we have a guarantee that the estimated length of W is larger than 1.

There are two types of transformations that we shall use as we move from W to W'. For consistency with Rzhetsky and Nei (1993), we shall refer to them as transformations of Type I (fig. 7) and Type II (fig. 8), respectively. It is easily seen that any black-and-white leaf coloring of any binary tree either splits the black-and-white leaves with one branch, or contains a Type I or Type II configuration.

Consider the Type I transformation, with A_1, A_2, B, and C as in figure 7. We use equation 5, which expresses the change in the tree length as a result of an NNI:

$$\begin{aligned} \hat{l}(w_i) - \hat{l}(w_{i+1}) &= \frac{1}{4}(\delta_{A_1B}^{W_i} + \delta_{A_2C}^{W_i} - \delta_{BC}^{W_i} - \delta_{A_1A_2}^{W_i}) \\ &= \frac{1}{4}(1 + \delta_{A_2C}^{W_i} - \delta_{BC}^{W_i} - 0) > 0 \qquad (16) \end{aligned}$$

The inequality follows from the assumption that C is not monochromatically white, and thus $\delta_{BC}^{W_i} < 1$.

Now consider the Type II transformation. Because the second step corresponds to two Type I transformation (swapping B_1 and A_2, and then C and A_2), we need only to show that the first step detailed in figure 8 decreases tree size. Without loss of generality, we assume that C is not monochromatically black (otherwise, we could perform a Type I transformation at this juncture). Let p_C be the *distance* from C to either of the white subtrees; i.e., $p_C = \delta_{A_1C}^{W_i} = \delta_{A_2C}^{W_i}$. Because C is not monochromatically black, we note that $p_C < 1$. Consider W_i and W_i^* (fig. 8) and apply equation 5 to the swap between B_1 and C. We obtain:

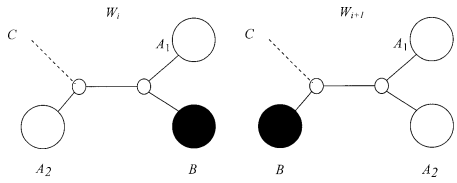

FIG. 7.—A Type I transformation. A Type I transformation uses a NNI to consolidate two monochromatic subtrees. C is not monochromatically white, otherwise the initial configuration splits the white-and-black leaves.

$$\begin{aligned}\hat{l}(W_i) - \hat{l}(W_i^*) &= \frac{1}{4}(\delta^{W_i}_{A_1B_1} + \delta^{W_i}_{(A_2B_2)C} - \delta^{W_i}_{A_1C} - \delta^{W_i}_{B_1(A_2B_2)}) \\ &= \frac{1}{4}\left(1 + \frac{1}{2}p_C + \frac{1}{2}\delta^{W_i}_{B_2C} - p_C - \frac{1}{2}\right) \\ &= \frac{1}{8}(1 + \delta^{W_i}_{B_2C} - p_C) > 0,\end{aligned} \tag{17}$$

where the inequality follows from our assumption.

Equations 16 and 17 demonstrate that the lengths of the trees W_i are monotonically decreasing as $i = 0, \ldots, t$, and thus $\hat{l}(W) > \hat{l}(W')$. The split $X|Y$ was chosen arbitrarily from $S(T) - S(W)$, and thus Inequality (15) holds for all of the splits in this set. This concludes the proof of Theorem 2. It has to be noted that only NNIs are used in this proof; thus, we have also demonstrated the consistency of our BNNI algorithm when restricted to split metrics.

Appendix 4

In this appendix, we prove Theorem 3. Consider a tree T that is a local minimum under the BNNI topology search. We first demonstrate that every internal branch e has positive length estimate. Consider figure 1 and equation 5: Because T is a local minimum, swapping subtrees B and C increases the tree length, and therefore:

$$\delta^T_{AB} + \delta^T_{CD} < \delta^T_{AC} + \delta^T_{BD}.$$

A similar argument regarding the swap of B and D proves that:

$$\delta^T_{AB} + \delta^T_{CD} < \delta^T_{AD} + \delta^T_{BC}.$$

It follows from the two above inequalities and from equation 2 that $\hat{l}(e)$ is positive for any internal branch. Now let e be an external branch and consider equation 8. The inner bracket is positive as long as Δ satisfies the triangle inequality, i.e., is a distance. This concludes the proof of Theorem 3.

Acknowledgments

Thanks to David Bryant and Mike Steel for helpful discussions. Part of this study was achieved when the second author was invited to the Biomathematics Research Center (Canterbury University, New Zealand). This work was supported by the Montpellier-LR Genopole of interEPST Bioinformatics program.

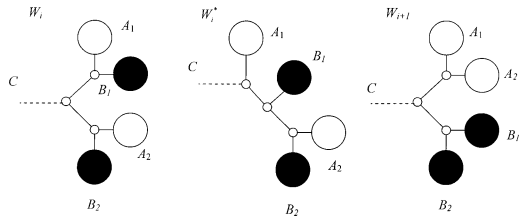

FIG. 8.—A Type II transformation. A Type II transformation is achieved by three NNIs, first transforming W_i into W_i^* and then transforming W_i^* into W_{i+1} via two Type I transformations. C is assumed to be neither monochromatically white nor black; otherwise we could perform a Type I transformation.

Literature Cited

Aldous, D. 1996. Probability distributions of cladograms. Pp. 1–18 *in* D. Aldous and R. Pemantle, eds. Random Discrete Structures, Springer-Verlag, New York.

Bandelt, H.-J., and A. W. M. Dress. 1992. A canonical decomposition theory for metrics on a finite set. Adv. Math. **92**: 47–105.

Bruno, W. M., N. D. Socci, and A. L. Halpern. 2000. Weighted Neighbor Joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. **17**:189–197.

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

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

Denis, F., and O. Gascuel. 2003. On the consistency of the minimum evolution principle of phylogenetic inference. Discrete Appl. Math. **127**:63–77.

Desper, R., and O. Gascuel. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. J. Comp. Biol. **9**:687–705.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. **27**:401–410.

———. 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics **5**:164–166.

———. 1997. An alternating least-squares approach to inferring phylogenies from pairwise distances. Syst. Biol. **46**:101–111.

Fitch, W. M. 1971. Rate of change of concomitantly variable codons. J. Mol. Evol. **1**:84–96.

Fitch, W. M., and E. Margoliash. 1967. Construction of phylogenetic trees. Science **155**:279–284.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. **18**:866–873.

Gascuel, O. 1997*a*. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

———. 1997*b*. Concerning the NJ algorithm and its unweighted version, UNJ. Pp. 149–170 *in* B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, eds. Mathematical hierarchies. American Mathematical Society, Providence, R. I.

———. 2000. Evidence for a relationship between algorithmic scheme and shape of inferred trees. Pp. 157–168 *in* W. Gaul, O. Opitz, and M. Schader, eds. Data analysis, scientific Modeling and practical applications. Springer-Verlag, Berlin.

Gascuel, O., D. Bryant, and F. Denis. 2001. Strengths and limitations of the minimum evolution principle. Syst. Biol. **50**:621–627.

Harding, E. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Probab. **3**:44–77.

Huelsenbeck J. P. 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. **19**:698–707.

Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458.

Kuhner, M. K., and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

Kumar, S. 1996. A stepwise algorithm for finding minimum evolution trees. Mol. Biol. Evol. **13**:584–593.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. **19**:1–7.

Nakhleh L., U. Roshan, K. St John, J. Sun, and T. Warnow. 2001. Designing fast converging phylogenetic methods. Bioinformatics **17**(Suppl 1):190–198.

Nei, M., and L. Jin. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. **6**:290–300.

Nei, M., J. C. Stephens, and N. Saitou. 1985. Methods for computing the standard errors branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2**:66–85.

Pauplin, Y. 2000. Direct calculation of a tree length using a distance matrix. J. Mol. Evol. **51**:41–47.

Robinson, D., and L. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

Rzhetsky, A., and M. Nei. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. **10**:1073–1095.

Saitou, N., and M. Nei. 1987. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:1073–1095.

Semple, C., and M. Steel. 2003. Cyclic permutations and evolutionary trees. Adv. Appl. Math. In Press.

Susko, E. 2003. Confidence regions and hypothesis tests for topologies using generalized least squares. Mol. Biol. Evol. **20**:862–868.

Swofford, D. 1996. PAUP—phylogenetic analysis using parsimony (and other methods), Version 4.0. Sinauer Associates, Sunderland, Mass.

Vach, W. 1989. Least squares approximation of additive trees. Pp. 230–238 *in* O. Opitz, ed. Conceptual and numerical analysis of data. Springer-Verlag, Berlin.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

Yule, G. 1925. A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis. Philos. Trans. R. Soc. London Ser. B, Biol. Sci. **213**:21–87.

Manolo Gouy, Associate Editor

Accepted November 3, 2003