Fast recovery of evolutionary trees with thousands of nodes

Miklós Csűrös

Department of Computer Science
Yale University
New Haven, CT 06520, USA

Current address:
Département d'informatique et recherche opérationnelle
Université de Montréal
CP 6128 succ. Centre-Ville
Montréal QC H3C3J7, Canada
E-mail: csuros@iro.umontreal.ca
Fax: +1 (514) 343-5834

Abstract

We present a novel distance-based algorithm for evolutionary tree reconstruction. Our algorithm reconstructs the topology of a tree with n leaves in $\mathcal{O}(n^2)$ time using $\mathcal{O}(n)$ working space. In the general Markov model of evolution, the algorithm recovers the topology successfully with $(1-o(1))$ probability from sequences with polynomial length in n. Moreover, for almost all trees, our algorithm achieves the same success probability on polylogarithmic sample sizes. The theoretical results are supported by simulation experiments involving trees with 500, 1895, and 3135 leaves. The topologies of the trees are recovered with high success from 2000 bp DNA sequences.

1 Introduction

What is the largest evolutionary tree we can derive today? The limits of large-scale phylogeny reconstruction are determined by the availability of useful molecular sequences, and by the availability of useful reconstruction methods. With current advances in bioinformatics, DNA sequencing is now both fast and reliable enough that efficiency is becoming a major concern for large-scale problems in phylogeny reconstruction. Ambitious projects such as the Green Plant Phylogeny project (GPP) (Brown 1999) and the Ribosomal Database Project (RDP) (Maidak et al. 2000) involve phylogenies with hundreds and thousands of homologous DNA sequences. When reconstructing a large tree, primary considerations for efficiency are computational speed and statistical accuracy. For instance, algorithms with exponential running time in the tree size cannot be used with trees that have more than a few tens of leaves. In fact, even algorithms that build trees with n leaves in $\mathcal{O}(n^4)$ time may be too slow if n is in the order of thousands. On the other hand, algorithms that fail to extract topology information efficiently enough from the input sequences may require inordinately large amounts of data, preventing successful reconstruction of large trees. Recent theoretical results on the statistical efficiency of distance-based algorithms (Erdős et al. 1999b; Huson et al. 1999; Csűrös and Kao 2001) suggest that these latter are ideal candidates for large-scale phylogeny reconstruction. Nonetheless, simulation studies corroborating the theoretical predictions for trees of sizes comparable with those in GPP and RDP are still needed.

This paper has two goals. First, it presents a novel distance-based algorithm with provably high statistical and computational efficiency. Secondly, it reports the results of experiments conducted with large, biologically-motivated model trees with various ranges of mutation probabilities. In the experiments, we simulated DNA sequence evolution in the Jukes-Cantor (1969) model on trees with 500, 1895, and 3135 leaves. These trees are unusually large for simulation studies. To our knowledge, the largest trees reconstructed from simulated data have 256 leaves (Kim 1998). The methods we compare include Neighbor-Joining of Saitou and Nei (1987), BioNJ of Gascuel (1997), Weighbor of Bruno et al. (2000), our Harmonic Greedy Triplets algorithm, and parsimony. Our results establish that even such large trees can be successfully recovered from DNA sequences with 2000 nucleotides. The experimental results support our theoretical results on the efficiency of our algorithm, called Harmonic Greedy Triplets with the Four-Point Condition (HGT/FP).

1.1 The general Markov model of sequence evolution

Mathematical models of sequence evolution play a fundamental role in providing a framework for developing evolutionary tree reconstruction algorithms, and for analyzing the algorithms' computational and statistical characteristics. A widely studied model is the general Markov model (Steel 1994), in which sequence characters evolve independently. The model is formulated as follows. Let $\mathcal{A} = \{1, 2, \ldots, r\}$ be a finite alphabet of size $r \geq 2$, and for every $\ell > 0$, let $\mathcal{A}^\ell$ denote the set of sequences over $\mathcal{A}$ with length ℓ. An evolutionary tree T is defined by an underlying tree and a mutation model. The underlying tree is a rooted binary tree representing the evolutionary ancestor-descendant relationships between its nodes. For every length $\ell > 0$, the mutation model randomly associates sequences of $\mathcal{A}^\ell$ with the nodes. The vector formed by the characters in the same position of the sequences is called a *site*. In the general Markov model, the sites are independent and identically distributed.

Let E be the set of tree edges, let $V = \langle u_1, \ldots, u_k \rangle$ be the ordered set of tree nodes, and let $\boldsymbol{\xi} = \langle \xi^{(u_1)}, \ldots, \xi^{(u_k)} \rangle$ be a site. For each node $u \in V$, the random variable $\xi^{(u)}$ is called the *label* of u. The Markov assumption in the model is that the label of a non-root node u, conditional on the label of its parent, is independent from the labels of other nodes that are not descendants of u. Together with the i. i. d. assumption, this implies (Steel 1994) that the distribution of $\boldsymbol{\xi}$ is determined by a *root label distribution* $\boldsymbol{\pi} = \langle \pi_1, \ldots, \pi_r \rangle$ and by a set of *mutation matrices* $\{\mathbf{M}_e \colon e \in E\}$ assigned to the tree edges. For each edge e, $\mathbf{M}_e$ is an $r \times r$ stochastic matrix. The root label is distributed according to $\boldsymbol{\pi}$. For all nodes $u, v \in V$, if v is the child of u on edge e, then for all characters $i, j \in V$,

$$\mathbb{P}\Big\{\xi^{(v)} = j \;\Big|\; \xi^{(u)} = i\Big\} = \mathbf{M}_e[i, j].$$

In other words, labels evolve along the tree from the root towards the leaves. For simplicity's sake, we assume that for every node u and character $i \in \mathcal{A}$, $\mathbb{P}\Big\{\xi^{(u)} = i\Big\} \neq 0$, and that for every edge e, $0 < |\det \mathbf{M}_e| < 1$, which are standard identifiability assumptions (Steel 1994; Erdős et al. 1999b).

1.2 Efficient evolutionary tree reconstruction

For an evolutionary tree T, the topology $\Psi(T)$ is the unrooted binary tree that is obtained from the underlying tree by removing the direction of the edges, and by replacing the root and its two incident edges with one single

edge connecting the root's children. The problem of evolutionary tree reconstruction is that of finding $\Psi(T)$ from sequences associated with the leaf set L, called a *sample*. An evolutionary tree reconstruction algorithm outputs an unrooted tree Ψ^* with the same leaf set L for an input sample. The algorithm succeeds if $\Psi^* = \Psi(T)$, i.e., if the topology is recovered. The *success rate* of an algorithm on sample sequences of length ℓ is the probability that T generates a sample for which the algorithm succeeds. The minimum sample length required to achieve a given success rate $(1 - \delta)$, where $0 < \delta < 1$ is the error probability sought, defines the algorithm's *statistical efficiency* (Kim 1998). Let n denote the number of leaves in $\Psi(T)$. An algorithm is statistically efficient if the minimum sample length is polynomial in n and $(1/\delta)$. An algorithm is *computationally efficient* if its running time is polynomial in n and ℓ.

Most popular evolutionary tree reconstruction algorithms today fall short of achieving provable computational or statistical efficiency. The HGT/FP algorithm, however, is both computationally and statistically efficient. In fact, it is the fastest statistically efficient algorithm to date, running in $\mathcal{O}(n^2)$ time. Previous theoretical results include those of Erdős et al. (1999a, 1999b), who started the study of statistical efficiency in the context of topology recovery and who also devised the first algorithms with provable computational and statistical efficiency. Their algorithms run in $\mathcal{O}(n^5 \log n)$ and $\mathcal{O}(n^4 \log n)$ time. Farach and Kannan (1999) introduced the study of sample sizes required by evolutionary tree reconstruction algorithms in probabilistic models of sequence evolution, but their problem is slightly different from ours since their primary focus was to estimate the distribution of leaf labels based on a sample. Cryan et al. (1998) gave a polynomial-time solution to the problem and proved that their algorithm is statistically efficient. The recently developed Disc Covering Method of Huson et al. (1999) is statistically efficient but needs to solve an NP-hard problem at its core, and its heuristic implementation runs in $\mathcal{O}(n^4)$ time. The statistically efficient Fast Harmonic Greedy Triplets algorithm of Csűrös and Kao (2001) also runs in $\mathcal{O}(n^2)$ time but its efficiency is contingent upon knowing the highest mutation rate on a tree edge. A major advantage of HGT/FP is that it does not rely on any such input parameter. Warnow et al. (2001) describe an algorithm that turns a statistically efficient algorithm needing such a parameter into an algorithm that is statistically efficient without it. The transformation, however, increases the running time of the original algorithm by $\mathcal{O}(n^4)$.

1.3 Distance-based algorithms

A number of evolutionary tree reconstruction algorithms calculate an $n \times n$ matrix in a preprocessing step from the input sequences, and build the tree based on the matrix. The input matrix estimates a matrix of *evolutionary distances* between leaves, defined as follows. Let T be an evolutionary tree with n leaves. Evolutionary distances between the nodes of $\Psi(T)$ arise by equipping the edges of $\Psi(T)$ with positive weights. The edge weights are also called *edge lengths*. The distance between two nodes is the sum of the edge lengths on the path between them. A *tree metric* $\mathbf{D}$ over $\Psi(T)$ is the $n \times n$ matrix of pairwise distances between leaves. The tree metric $\mathbf{D}$ is a functional of the site distribution and uniquely determines the topology $\Psi(T)$.

Common tree metrics in the general Markov model include paralinear distance (Lake 1994) and LogDet distance (Steel 1994; Lockhart et al. 1994). Define the $r \times r$ matrix $\mathbf{M}_{uv}$ for all nodes u, v of $\Psi(T)$ by its entries as

$$\mathbf{M}_{uv} = \left[\mathbb{P}\left\{\xi^{(v)} = j \;\middle|\; \xi^{(u)} = i\right\} : i, j \in \mathcal{A}\right].$$

The paralinear distance is defined as

$$\mathbf{D}[u, v] = -\ln \sqrt{(\det \mathbf{M}_{uv})(\det \mathbf{M}_{vu})}, \tag{1}$$

for all leaves u, v. Since the expression on the right-hand side is additive along any path in $\Psi(T)$, the paralinear distance is a tree metric. Specifically, if the labels form a time-reversible Markov chain along any path in $\Psi(T)$, which is a frequent assumption in molecular evolutionary studies (Lanave et al. 1984; Tavaré 1986), then the paralinear distance is realized by setting the length of each edge e to $\left(-\ln|\det \mathbf{M}_e|\right)$. For all leaves u, v, let

$$\mathbf{J}_{uv} = \left[\mathbb{P}\left\{\xi^{(u)} = i, \xi^{(v)} = j\right\} : i, j \in \mathcal{A}\right]$$

be the joint probability matrix of the leaf labels. The LogDet distance between leaves $u \neq v$ is defined by $\mathbf{D}[u, v] = -\ln|\det \mathbf{J}_{uv}|$.

There are many other tree metrics within various subclasses of the general Markov model restricting the set of mutation matrices. For example, Neyman's model (1971) imposes that for every mutation matrix $\mathbf{M}_e$ there exists a mutation probability $0 < p_e < (1 - 1/r)$, such that

$$\mathbf{M}_e[i, j] = \begin{cases} 1 - p_e & \text{if } i = j; \\ p_e/(r-1) & \text{if } i \neq j. \end{cases}$$

Subsequently, $\mathbf{D}[u,v] = -\ln(1 - \frac{r}{r-1}\mathbb{P}\{\xi^{(u)} \neq \xi^{(v)}\})$ is a tree metric. This tree metric is known as the Jukes-Cantor distance (1969) in the case $r = 4$.

Distance-based algorithms thus have to estimate the tree metrics in a preprocessing step, which typically entails substituting probabilities in the tree metric's definition with relative frequencies calculated from the sample. We call such estimators *empirical tree metrics* and discuss them further in §2.4. An important feature of empirical tree metrics is that estimation error increases with evolutionary distance between the leaves in question. In order to achieve statistical efficiency, a topology reconstruction algorithm has to strive to use leaves that are close to each other. The HGT/FP algorithm is designed with that goal in mind. The techniques we use to achieve that goal are based on analyzing the convergence rate of empirical tree metrics.

1.4 Recovering the topology from a tree metric

If an exact tree metric is known, then the problem of reconstructing the topology can be reduced to the problem of obtaining an unrooted tree with positive edge weights from distances between its leaves. A basic technique for that purpose uses *triplets*. A triplet uvw comprises three leaves u, v, and w of $\Psi(T)$. Every triplet defines an internal node at which the three pairwise paths between the leaves intersect, with the four nodes forming a star. This internal node is the *center* of the triplet. Using the tree metric's definition, the distance between the center o and a leaf u in the triplet uvw can be calculated as

$$D_{uo} = \Delta(u, uvw) = \frac{\mathbf{D}[u,v] + \mathbf{D}[u,w] - \mathbf{D}[v,w]}{2},$$

where $\Delta(u, uvw)$ denotes the triangle-star transformation formula on the right-hand side. This formula can be used repeatedly to reconstruct the topology with the edge lengths by adding one leaf and one internal node at a time (Waterman et al. 1977). The main idea of such a reconstruction is fairly simple. Let Ψ^* be a subtree of $\Psi(T)$ spanned by a subset of the leaves, and let each edge of Ψ^* be weighted by the distance between its endpoints in $\Psi(T)$. If u and v are two leaves in Ψ^* and w is a leaf of $\Psi(T)$ missing from Ψ^*, then the center o of uvw in Ψ^* is on the path P between u and v, and its exact location can be found by comparing $\Delta(u, uvw)$ to the distances between u and the nodes on P. If that location falls properly on an edge e in Ψ^*, then o can be added on e, and w can be connected to it with an edge of length $\Delta(w, uvw)$. The edge lengths for the newly created edges between o

and the endpoints of e can be calculated as the difference between $\Delta(u, uvw)$ and the distances from u to the endpoints. This approach is complicated only by the fact that the center of uvw may be a node that is already in Ψ^*. In other words, there may be a node z on the path P between u and v that is at distance $\Delta(u, uvw)$ from u. In this case w should be connected to Ψ^* through an internal node in the subtree rooted at w that contains neither u nor v, by using a different triplet. The reconstruction starts by selecting an arbitrary triplet uvw and initializing Ψ^* as the star formed by uvw and its center.

2 The HGT/FP algorithm

Using the algorithm outlined in §1.4 with an estimated tree metric $\hat{\mathbf{D}}$ almost certainly leads to failure. The main reason is that $\hat{\mathbf{D}}$ is usually not a tree metric, due to random estimation errors. As a consequence, $\Psi(T)$ is not determined by $\hat{\mathbf{D}}$. We describe two specific measures to deal with the fact that $\hat{\mathbf{D}}$ may not be a tree metric. These general measures are helpful for any algorithm following the outline of §1.4 and do not make assumptions about the exact way $\hat{\mathbf{D}}$ is calculated. We address the problem of estimating edge lengths in §2.1. In §2.2 we address the problem of determining whether a triplet defines a new internal node in Ψ^*. These measures do not ensure statistical efficiency on their own, and we analyze their error in §2.4 after studying the convergence of empirical tree metrics in §2.3. The results of the analysis suggest a greedy selection of triplets, which is employed in HGT/FP. It is this greedy selection that leads not only to statistical efficiency but also to the $\mathcal{O}(n^2)$ running time. The techniques for achieving the fast running time are described in §2.5. Figure 2 shows the algorithm.

2.1 Estimating edge lengths

The HGT/FP algorithm follows the general outline of the algorithm in §1.4 with specific techniques for dealing with estimated tree metrics. Let $\hat{\mathbf{D}}$ be the estimated tree metric and let

$$\hat{\Delta}(u, uvw) = \frac{\hat{\mathbf{D}}[u,v] + \hat{\mathbf{D}}[u,w] - \hat{\mathbf{D}}[v,w]}{2}$$

denote the corresponding triangle-star transformation for every triplet uvw. In order to prevent the accumulation of error in edge length estimates, the HGT/FP algorithm stores a triplet $\mathrm{def}(z)$ for each internal node z in Ψ^*, which is the triplet used for adding z to Ψ^*. For notational uniformity, let $\mathrm{def}(z) = \{z\}$ if z is a leaf. In order to add a new internal node o on an edge z_1z_2 in Ψ^*, o must be the center of a triplet u_1v_2w for which the following conditions hold: $u_1 \in \mathrm{def}(z_1)$, $v_2 \in \mathrm{def}(z_2)$, w is not in Ψ^*, and the edge z_1z_2 is on the path between u_1 and v_2 in Ψ^*. Such an edge-triplet pair $\langle z_1z_2, u_1v_2w\rangle$ is called *relevant*. Assume that z_1 is an internal node, and $\mathrm{def}(z_1) = u_1v_1w_1$. The value $d_1 = \left|\hat{\Delta}(u_1, u_1v_1w_1) - \hat{\Delta}(u_1, u_1v_2w)\right|$ is an estimate of the distance between the centers of $u_1v_1w_1$ and u_1v_2w in $\Psi(T)$. Similarly, $d_2 = \left|\hat{\Delta}(v_2, u_2v_2w_2) - \hat{\Delta}(v_2, u_1v_2w)\right|$ estimates the distance between the centers of $u_2v_2w_2$ and u_1v_2w. Let $D^*_{z_1z_2}$ be the length of the

edge z_1z_2 in Ψ^*. The edge lengths for inserting o on z_1z_2 are calculated by

$$D^*_{oz_1} = (d_1 + D^*_{z_1z_2} - d_2)/2;$$
$$D^*_{oz_2} = (d_2 + D^*_{z_1z_2} - d_1)/2.$$

If z_i is a leaf for $i = 1$ or for $i = 2$, then $d_i = 0$ but otherwise the calculations are the same.

The theoretical importance of this procedure is that it results in edge length estimation errors that depend only on the error in estimating the center of individual triplets. Assume that Ψ^* is correct (i.e., it is topologically equivalent to the subtree of $\Psi(T)$ spanned by the leaves of Ψ^*) and the center of u_1v_1w falls onto the path between z_1 and z_2 in $\Psi(T)$. Further assume that the triplet centers are estimated within ϵ error, i.e., that for every triplet $x_1x_2x_3 \in \{u_1v_1w_1, u_2v_2w_2, u_1v_2w\}$ and leaf x_i,

$$\left|\hat{\Delta}(x_i, x_1x_2x_3) - \Delta(x_i, x_1x_2x_3)\right| < \epsilon.$$

If $|D^*_{z_1z_2} - D_{z_1z_2}| = 4\epsilon'$, where $D_{z_1z_2}$ is the distance between z_1 and z_2 in $\Psi(T)$, then $|D^*_{oz_1} - D_{oz_1}| < 2\epsilon' + 2\epsilon$ and $|D^*_{oz_2} - D_{oz_2}| < 2\epsilon' + 2\epsilon$. Similar bounds hold if z_1 or z_2 is a leaf. If $\epsilon' \leq \epsilon$, then the error of the newly created edge lengths is bounded by 4ϵ. Consequently, if the maximum error in estimating triplet centers used by the algorithm is bounded by ϵ, then all edge lengths are estimated within 4ϵ error, given that the topology is recovered correctly.

2.2 Finding triplet centers

While in the case of tree metrics, we can always tell whether a triplet defines a new internal node in the partially built topology Ψ^*, this is not so in the case of estimated tree metrics, where triplet centers may appear to define a new internal node due to estimation error. For example, even if the triplets uvw and $uv'w'$ have the same center in $\Psi(T)$, it is possible that $\hat{\Delta}(u, uvw) \neq \hat{\Delta}(u, uv'w')$ and thus the techniques in §1.4 for choosing triplets are likely to be inadequate.

A safeguarding measure in HGT/FP for dealing with an estimated tree metric is based on the *four-point condition* (Buneman 1971), which is defined as follows. An evolutionary tree with four leaves $\{u, v, w, z\}$ has three possible topologies denoted by $uv|wz$, $uw|vz$ and $uz|vw$, depending on which leaf pairs are separated by the internal edge in the topology. The four-point condition states that by distance additivity, the topology is $uv|wz$ if and

only if

$$\mathbf{D}[u,v] + \mathbf{D}[w,z] < \mathbf{D}[u,w] + \mathbf{D}[v,z] = \mathbf{D}[u,z] + \mathbf{D}[v,w].$$

Since the equality of the two larger sums is unlikely when using an estimated tree metric, the HGT/FP algorithm employs the *relaxed four-point condition* (Bandelt and Dress 1986), which for $uv|wz$ is defined as

$$\begin{aligned} \hat{\mathbf{D}}[u,v] + \hat{\mathbf{D}}[w,z] &< \hat{\mathbf{D}}[u,w] + \hat{\mathbf{D}}[v,z]; \\ \hat{\mathbf{D}}[u,v] + \hat{\mathbf{D}}[w,z] &< \hat{\mathbf{D}}[u,z] + \hat{\mathbf{D}}[v,w]. \end{aligned} \tag{2}$$

Let $\langle z_1 z_2, uvw \rangle$ be a relevant pair. The relaxed four-point condition is used to determine whether the center of uvw falls onto $z_1 z_2$ in the following manner. Let z_1 be an internal node in Ψ^*, let $\mathrm{def}(z_1) = u_1 v_1 w_1$, and assume that z_2 lies on the path between u_1 and z_1 in Ψ^* without loss of generality (see Figure 1). Recall that w is not a leaf in Ψ^*. HGT/FP tests whether the relaxed four-point condition holds for $u_1 w | v_1 w_1$. If so, then for the center o of uvw, the paths from z_1 to o and to z_2 overlap. The condition is used similarly with z_2 if it is an internal node, in order to decide if the paths from z_2 to o and to z_1 overlap. If z_i is a leaf, then the condition for z_i is not tested. If the tested conditions hold for the pair $\langle z_1 z_2, uvw \rangle$, then this latter is called a *good relevant pair*. If $\langle z_1 z_2, uvw \rangle$ is a good relevant pair, then HGT/FP concludes that the center of uvw can be inserted on the edge $z_1 z_2$. HGT/FP uses only good relevant pairs for adding new nodes. This way it tolerates some error in the estimated tree metrics, since Equation (2) may hold for the correct topology even if the distances between the leaves are estimated within a small error. In particular, if the distances between the leaves in the triplets uvw and $u_1 v_1 w_1$ are estimated within $D_{z_1 o}/2$ error, where $D_{z_1 o}$ is the distance between the triplet centers in $\Psi(T)$, and Ψ^* is correct, then the relaxed four-point condition holds for $u_1 w | v_1 w_1$ if and only if the paths from z_1 to o and to z_1 overlap.

2.3 Empirical tree metrics

The *Harmonic Greedy Triplets* (HGT) principle provides a guideline for the triplet selection mechanism when empirical tree metrics are used. The empirical tree metrics for the discussed distances are calculated as follows. The empirical Jukes-Cantor distance is computed by $\hat{\mathbf{D}}[u,v] = -\ln(1 - \frac{r}{r-1}\hat{p}_{uv})$ where $\hat{p}_{uv}$ is the relative frequency of the event $\{\xi^{(u)} \neq \xi^{(v)}\}$ observed in the sample. (If the relative frequency is larger than $(1 - 1/r)$, then the distance is set to ∞ or a large positive constant.)

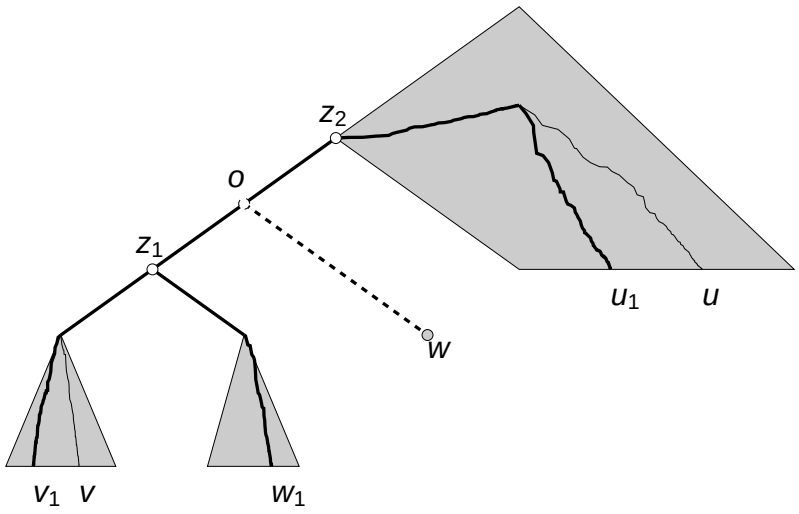

Figure 1: *Using the four-point condition with relevant triplets.*

For the empirical LogDet distance, calculate the matrices $\hat{\mathbf{J}}_{uv} = [\hat{p}_{uv,ij}]$ where $\hat{p}_{uv,ij}$ is the relative frequency of the event $\{\xi^{(u)} = i, \xi^{(v)} = j\}$ in the sample.

Lemma 1. *Let $\hat{\mathbf{J}}_{uv}$ be calculated from a sample of length $\ell \geq r$. Then*

$$\mathbb{E} \det \hat{\mathbf{J}}_{uv} = \Big(1 - \frac{1}{\ell}\Big)\Big(1 - \frac{2}{\ell}\Big) \cdots \Big(1 - \frac{r-1}{\ell}\Big) \det \mathbf{J}_{uv}. \tag{3}$$

Proof. By definition of the determinant,

$$\det \hat{\mathbf{J}}_{uv} = \sum_{j_1,\ldots,j_r}{}' (-1)^{\kappa(j_1,\ldots,j_r)} \prod_{i=1}^{r} \hat{\mathbf{J}}_{uv}[i, j_i],$$

where $\sum'$ denotes the sum over permutations and $\kappa(\cdot)$ equals ± 1 depending on whether the number of switched pairs in the permutation is odd or even. Since the vector $\Big\langle \ell \hat{\mathbf{J}}_{uv}[i,j] \colon i,j = 1,\ldots,r \Big\rangle$ is multinomially distributed, for every term of the equation,

$$\mathbb{E} \prod_{i=1}^{r} \hat{\mathbf{J}}_{uv}[i, j_i] = \frac{1}{\ell^r} \mathbb{E} \prod_{i=1}^{r} (\ell \hat{\mathbf{J}}_{uv}[i, j_i]) = \frac{\ell(\ell-1)\cdots(\ell-r+1)}{\ell^r} \prod_{i=1}^{r} \mathbf{J}_{uv}[i, j_i].$$

Consequently,

$$\begin{aligned}\mathbb{E}\det\hat{\mathbf{J}}_{uv} &= \sum_{j_1,\ldots,j_r}{}' (-1)^{\kappa(j_1,\ldots,j_r)}\frac{\ell(\ell-1)\cdots(\ell-r+1)}{\ell^r}\prod_{i=1}^{r}\mathbf{J}_{uv}[i,j_i]\\ &=\Big(1-\frac{1}{\ell}\Big)\Big(1-\frac{2}{\ell}\Big)\cdots\Big(1-\frac{r-1}{\ell}\Big)\det\mathbf{J}_{uv},\end{aligned}$$

which is tantamount to Equation (3). □

Based on Lemma 1, define the *empirical LogDet distance* as

$$\hat{\mathbf{D}}[u,v] = -\ln|\det\hat{\mathbf{J}}_{uv}| + \sum_{k=1}^{r-1}\ln(1-k/\ell). \tag{4}$$

(If $\det\hat{\mathbf{J}}_{uv}=0$, then the distance is set to ∞ or a large positive constant.)

Lemma 2. *Let* $\mathbf{D}$ *be the LogDet distance, let* $\hat{\mathbf{D}}$ *be the empirical LogDet distance calculated from a sample of length* ℓ*, and define* $\gamma = (1-\frac{1}{\ell})(1-\frac{2}{\ell})\cdots(1-\frac{r-1}{\ell})$*. For all leaves* u, v*, and* $\epsilon>0$*,*

$$\mathbb{P}\Big\{\hat{\mathbf{D}}[u,v]-\mathbf{D}[u,v]\geq-\ln(1-\epsilon)\Big\}\leq\exp\Big(-\frac{\gamma^2(r-1)^{2(r-1)}}{2}\ell\det{}^2\mathbf{J}_{uv}\epsilon^2\Big);$$

$$\mathbb{P}\Big\{\hat{\mathbf{D}}[u,v]-\mathbf{D}[u,v]\leq-\ln(1+\epsilon)\Big\}\leq\exp\Big(-\frac{\gamma^2(r-1)^{2(r-1)}}{2}\ell\det{}^2\mathbf{J}_{uv}\epsilon^2\Big).$$

Proof. Define the i. i. d. random variables $\boldsymbol{\eta}_k = \langle\xi_k^{(u)},\xi_k^{(v)}\rangle$, for $k=1,\ldots,\ell$, where $\xi_k^{(u)}$ is the label of u and $\xi_k^{(v)}$ is the label of v at site k. Furthermore, define $\hat{S}_{uv}=\exp(-\hat{\mathbf{D}}[u,v])$, and $S_{uv}=\exp(-\mathbf{D}[u,v])=\det\mathbf{J}_{uv}$. Then $\hat{S}$ can be written as

$$\hat{S}_{uv} = f(\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_\ell),$$

where f is defined by Equation (4). Notice that by Lemma 1, $\mathbb{E}\hat{S}_{uv}=S_{uv}$. If the labels of u and v change in the k-th labeling, then f changes by at most

$$\frac{2}{\ell\gamma}(r-1)^{-(r-1)},$$

based on the fact that if two entries in $\hat{\mathbf{J}}_{uv}$ change by $\pm\frac{1}{\ell}$, then $\det\hat{\mathbf{J}}$ changes by at most $2(r-1)^{-(r-1)}/\ell$. Subsequently, McDiarmid's inequality (1989)

is used to bound the probability of large deviations of $\hat{S}$:

$$\begin{aligned}\mathbb{P}\Big\{\hat{\mathbf{D}}[u,v] - \mathbf{D}[u,v] \geq -\ln(1-\epsilon)\Big\} \\ &= \mathbb{P}\Big\{\frac{\hat{S}_{uv}}{S_{uv}} \leq 1-\epsilon\Big\} = \mathbb{P}\Big\{\hat{S}_{uv} - \mathbb{E}\hat{S}_{uv} \leq -\epsilon S_{uv}\Big\} \\ &\leq \exp\Big(-\frac{\gamma^2(r-1)^{2(r-1)}}{2}\ell S_{uv}^2\epsilon^2\Big),\end{aligned}$$

proving one half of the lemma. The other half is proven analogously. □

Finally, in the case of empirical paralinear distance, we calculate $\hat{\mathbf{M}}_{uv} = [\hat{p}_{uv,ij}/\hat{p}_{u,i}]$ where $\hat{p}_{u,i}$ is the relative frequency of the event $\{\xi^{(u)} = i\}$, with the convention that if $\hat{p}_{u,i} = 0$, then $\hat{\mathbf{M}}_{uv}[i,j] = 1$ for $i = j$ and $\hat{\mathbf{M}}_{uv}[i,j] = 0$ for $i \neq j$. The matrices $\hat{\mathbf{M}}_{uv}$ and $\hat{\mathbf{M}}_{vu}$ are used in place of $\mathbf{M}_{uv}$ and $\mathbf{M}_{vu}$ in Equation (1) to compute the empirical paralinear distance.

Lemma 3. *Let u and v be two leaves of $\Psi(T)$. Define $\pi_i^{(u)} = \mathbb{P}\Big\{\xi^{(u)} = i\Big\}$, the probability of labeling node u with symbol i. For arbitrary symbols $i, j \in \mathcal{A}$ and $\epsilon > 0$,*

$$\mathbb{P}\Big\{\hat{\mathbf{M}}_{uv}[i,j] - \mathbf{M}_{uv}[i,j] \geq \epsilon\Big\} \leq \exp\Big(-(1-e^{-2})\ell\pi_i^{(u)}\epsilon^2\Big); \tag{5a}$$

$$\mathbb{P}\Big\{\hat{\mathbf{M}}_{uv}[i,j] - \mathbf{M}_{uv}[i,j] \leq -\epsilon\Big\} \leq \exp\Big(-(1-e^{-2})\ell\pi_i^{(u)}\epsilon^2\Big). \tag{5b}$$

Proof. As a shorthand notation, define

$$p = \mathbf{M}_{uv}[i,j]; \qquad \hat{p} = \hat{\mathbf{M}}_{uv}[i,j]; \qquad q = \pi_i^{(u)}.$$

We first prove Equation (5a). Assume that $\epsilon < 1 - p = 1 - \mathbf{M}_{uv}[i,j]$. Otherwise the equation is trivial because $\hat{\mathbf{M}}_{uv}[i,j]$ is never larger than one. The random variable $\hat{p}_{uv,ij}$ has a binomial distribution with parameters $\hat{p}_{u,i}$ and p. Thus for every $k = 0, 1, \ldots, \ell$, Hoeffding's inequality (1963) implies that

$$\mathbb{P}\Big\{\hat{p}_{uv,ij} \geq k(p+\epsilon) \;\Big|\; \hat{p}_{u,i} = k\Big\} \leq e^{-2k\epsilon^2}. \tag{6}$$

The inequality holds vacuously even for $k = 0$. Since the random variable $\hat{p}_{u,i}$ has a binomial distribution with parameters ℓ and q, by Equation (6),

$$\begin{aligned}\mathbb{P}\{\hat{p} \geq p+\epsilon\} &= \sum_{k=0}^{\ell} \mathbb{P}\Big\{\hat{p}_{uv,ij} \geq k(p+\epsilon) \;\Big|\; \hat{p}_{u,i} = k\Big\}\mathbb{P}\Big\{\hat{p}_{u,i} = k\Big\} \\ &\leq \sum_{k=0}^{\ell}\binom{\ell}{k}q^k(1-q)^{\ell-k}e^{-2k\epsilon^2} = \Big(1-q+qe^{-2\epsilon^2}\Big)^{\ell}.\end{aligned} \tag{7}$$

Define

$$\phi(x) = -\ln\Big(1 - q + qe^{-2x}\Big).$$

Since the function ϕ is concave and $\phi(0) = 0$, for every $x < x'$,

$$\phi(x) \geq x\frac{\phi(x')}{x'}.$$

In particular,

$$\phi(\epsilon^2) \geq \epsilon^2\frac{\phi\Big((1-p)^2\Big)}{(1-p)^2}, \tag{8}$$

since $\epsilon < 1 - p$. Therefore,

$$\begin{aligned}
\mathbb{P}\{\hat{p} \geq p + \epsilon\} &\leq \exp\Big(-\ell\phi(\epsilon^2)\Big) && \text{by Eq. (7)} \\
&\leq \exp\left(\ell\epsilon^2\frac{\ln(1 - q + qe^{-2(1-p)^2})}{(1-p)^2}\right) && \text{by Eq. (8)} \\
&\leq \exp\left(-\ell q\epsilon^2\frac{1 - e^{-2(1-p)^2}}{(1-p)^2}\right) && x \leq -\ln(1-x) \\
&\leq \exp\Big(-(1 - e^{-2})\ell q\epsilon^2\Big) && \min_{x\in[0,1]}\frac{1 - e^{-2x^2}}{x^2} = 1 - e^{-2}
\end{aligned}$$

corresponding to Equation (5a). The proof of Equation (5b) is analogous. □

Lemma 4. *Define*

$$\pi_{\min} = \min_{u \text{ is a leaf}, i\in\mathcal{A}} \mathbb{P}\Big\{\xi^{(u)} = i\Big\}.$$

Let $\mathbf{D}$ *be the paralinear distance, let* $\hat{\mathbf{D}}$ *be the empirical paralinear distance, and define* $S_{uv} = \exp(-\mathbf{D}[u, v])$ *and* $\hat{S}_{uv} = \exp(-\hat{\mathbf{D}}[u, v])$*. For all leaves* $u, v \in \Psi(T)$*, sample length* ℓ *and* $\epsilon > 0$*,*

$$\mathbb{P}\left\{\left|\frac{\hat{S}_{uv}}{S_{uv}} - 1\right| \geq \epsilon\right\} \leq 4r^2\exp\left(-\frac{1 - e^{-2}}{(r-1)^2}\ell\pi_{\min}^r r^{r-3}S_{uv}^2\epsilon^2\right). \tag{9}$$

Proof. First we claim that for arbitrary $\epsilon > 0$,

$$\mathbb{P}\Big\{\Big|\det\hat{\mathbf{M}}_{uv} - \det\mathbf{M}_{uv}\Big| \geq \epsilon\Big\} \leq 2r^2\exp\left(-\frac{1 - e^{-2}}{r^2(r-1)^2}\ell\pi_{\min}\epsilon^2\right). \tag{10}$$

By the definition of the determinant, and the fact that $\mathbf{M}_{uv}$ and $\hat{\mathbf{M}}_{uv}$ are stochastic matrices,

$$\Big|\det \hat{\mathbf{M}}_{uv} - \det \mathbf{M}_{uv}\Big| \leq r(r-1) \max_{i,j} \Big|\det \hat{\mathbf{M}}_{uv}[i,j] - \det \mathbf{M}_{uv}[i,j]\Big|$$

Thus, by Lemma 3,

$$\begin{aligned}
&\mathbb{P}\Big\{\Big|\det \hat{\mathbf{M}}_{uv} - \det \mathbf{M}_{uv}\Big| \geq \epsilon\Big\} \\
&\qquad \leq \mathbb{P}\Big\{\exists i,j \colon \Big|\det \hat{\mathbf{M}}_{uv}[i,j] - \det \mathbf{M}_{uv}[i,j]\Big| > \frac{\epsilon}{r(r-1)}\Big\} \\
&\qquad \leq 2r^2 \exp\Big(-\frac{1-e^{-2}}{r^2(r-1)^2}\ell \pi_{\min} \epsilon^2\Big),
\end{aligned}$$

proving Equation (10).

By definition of the paralinear distance,

$$\begin{aligned}
\Big(\det \mathbf{M}_{uv}\Big)^2 &= S_{uv}^2 \frac{\prod_{i=1}^r \pi_i^{(v)}}{\prod_{i=1}^r \pi_i^{(u)}}; \\
\Big(\det \mathbf{M}_{vu}\Big)^2 &= S_{uv}^2 \frac{\prod_{i=1}^r \pi_i^{(u)}}{\prod_{i=1}^r \pi_i^{(v)}}.
\end{aligned}$$

Now,

$$\frac{\prod_{i=1}^r \pi_i^{(v)}}{\prod_{i=1}^r \pi_i^{(u)}} \geq \frac{\pi_{\min}^{r-1}\Big(1-(r-1)\pi_{\min}\Big)}{r^{-r}} \geq \big(r\pi_{\min}\big)^{r-1}.$$

Similarly,

$$\frac{\prod_{i=1}^r \pi_i^{(u)}}{\prod_{i=1}^r \pi_i^{(v)}} \geq \big(r\pi_{\min}\big)^{r-1}.$$

Consequently, by Equation (10),

$$\begin{aligned}
\mathbb{P}\Big\{\Big|\frac{\hat{S}_{uv}}{Suv} - 1\Big| \geq \epsilon\Big\} \leq& \mathbb{P}\Big\{\Big|\frac{\det \hat{\mathbf{M}}_{uv}}{\det \mathbf{M}_{uv}} - 1\Big| \geq \epsilon\Big\} \\
&+ \mathbb{P}\Big\{\Big|\frac{\det \hat{\mathbf{M}}_{vu}}{\det \mathbf{M}_{vu}} - 1\Big| \geq \epsilon\Big\} \\
\leq& 4r^2 \exp\Big(-\frac{1-e^{-2}}{r^2(r-1)^2}\ell \pi_{\min}^r r^{r-1} S_{uv}^2 \epsilon^2\Big) \\
=& 4r^2 \exp\Big(-\frac{1-e^{-2}}{(r-1)^2}\ell \pi_{\min}^r r^{r-3} S_{uv}^2 \epsilon^2\Big),
\end{aligned}$$

proving the lemma. □

Theorem 1. *Let* $\mathbf{D}$ *be one of the tree metrics over* $\Psi(T)$ *discussed, i.e., let* $\mathbf{D}$ *be the paralinear, the LogDet, or the Jukes-Cantor distance. Let* $\hat{\mathbf{D}}$ *be the corresponding empirical tree metric. There exist constants* $a, b > 0$ *such that for all leaves* u, v *and* $0 < \epsilon < 1$*,*

$$\begin{aligned} \mathbb{P}\Big\{\hat{\mathbf{D}}[u,v] - \mathbf{D}[u,v] \geq -\ln(1-\epsilon)\Big\} &\leq a e^{-b\ell\epsilon^2 S_{uv}^2}; \\ \mathbb{P}\Big\{\hat{\mathbf{D}}[u,v] - \mathbf{D}[u,v] \leq -\ln(1+\epsilon)\Big\} &\leq a e^{-b\ell\epsilon^2 S_{uv}^2}, \end{aligned} \tag{11}$$

where $S_{uv} = e^{-\mathbf{D}[u,v]}$*.*

Proof. The result for the Jukes-Cantor distance is proven by Farach and Kannan (1999) among others. Lemma 2 proves the theorem for the LogDet distance. Lemma 4 proves the theorem for the paralinear distance. □

Definition 1. *An estimated tree metric* $\hat{\mathbf{D}}$ *for which Equation* (11) *holds is called an* (a, b)-regular estimator *for* $\mathbf{D}$*.*

Remark. Csűrös (2000) proves that Kimura's three parameter distance (Kimura 1981) is also (a, b)-regular. Erdős et al. (1999b) prove a similar result for a different estimator of the LogDet distance.

2.4 The Harmonic Greedy Triplets principle

Definition 2. *Let* $\mathbf{D}$ *be a tree metric. The value* $S_{uv} = \exp(-\mathbf{D}[u,v])$ *is called the* similarity *between the leaves* u *and* v*.*

The HGT principle originates from Theorem 2 below, which relates the error in triplet center estimation with regular estimators to a harmonic average of similarities. For every triplet uvw, define the *average similarity*

$$S_{uvw} = \frac{3}{S_{uv}^{-1} + S_{uw}^{-1} + S_{vw}^{-1}} = \frac{3}{e^{\mathbf{D}[u,v]} + e^{\mathbf{D}[u,w]} + e^{\mathbf{D}[v,w]}}.$$

Recall the triangle-star transformation formulas for calculating triplet centers: $\Delta(u, uvw) = \frac{\mathbf{D}[u,v]+\mathbf{D}[u,w]-\mathbf{D}[v,w]}{2}$ and $\hat{\Delta}(u, uvw) = \frac{\hat{\mathbf{D}}[u,v]+\hat{\mathbf{D}}[u,w]-\hat{\mathbf{D}}[v,w]}{2}$.

Theorem 2. *Let* $\hat{\mathbf{D}}$ *be an* (a, b)*-regular estimator for the tree metric* $\mathbf{D}$*. For every triplet* uvw*, and* $0 < \epsilon < 1$*,*

$$\mathbb{P}\left\{\hat{\Delta}(u, uvw) - \Delta(u, uvw) \geq \frac{-\ln(1-\epsilon)}{2}\right\} \leq 3a \exp\Big(-\frac{b}{9}\ell\epsilon^2 S_{uvw}^2\Big).$$

Proof. The proof is virtually identical to the one given by Csűrös and Kao (2001) for the case of Jukes-Cantor distance. For completeness' sake, we prove here a similar claim without explaining the appearance of the harmonic average. Since

$$\begin{aligned}&\Big|\hat{\Delta}(u,uvw)-\Delta(u,uvw)\Big|\\&\qquad\le\frac{\Big|\hat{\mathbf{D}}[u,v]-\mathbf{D}[u,v]\Big|+\Big|\hat{\mathbf{D}}[u,w]-\mathbf{D}[u,w]\Big|+\Big|\hat{\mathbf{D}}[v,w]-\mathbf{D}[v,w]\Big|}{2},\end{aligned}$$

$$\begin{aligned}\mathbb{P}\Big\{\Big|\hat{\Delta}(u,uvw)-\Delta(u,uvw)\Big|\ge\frac{-\ln(1-\epsilon)}{2}\Big\}&\\\le\mathbb{P}\Big\{\Big|\hat{\mathbf{D}}[u,v]-\mathbf{D}[u,v]\Big|\ge\frac{-\ln(1-\epsilon)}{3}\Big\}&\\+\mathbb{P}\Big\{\Big|\hat{\mathbf{D}}[u,w]-\mathbf{D}[u,w]\Big|\ge\frac{-\ln(1-\epsilon)}{3}\Big\}&\\+\mathbb{P}\Big\{\Big|\hat{\mathbf{D}}[v,w]-\mathbf{D}[v,w]\Big|\ge\frac{-\ln(1-\epsilon)}{3}\Big\}.&\end{aligned}\tag{12}$$

Since $S_{uv}>0$ for $u\neq v$,

$$S_{\min}(u,v,w)=\min\Big\{S_{uv},S_{uw},S_{vw}\Big\}\ge\frac{S_{uvw}}{3}.\tag{13}$$

Subsequently, using Equation (12) and the fact that $\hat{\mathbf{D}}$ is (a,b)-regular,

$$\begin{aligned}&\mathbb{P}\Big\{\Big|\hat{\Delta}(u,uvw)-\Delta(u,uvw)\Big|\ge\frac{-\ln(1-\epsilon)}{2}\Big\}\\&\qquad\le 6a\exp\Big(-\frac{b}{9}\ell\Big(S_{\min}(u,v,w)\Big)^2\epsilon^2\Big)\le 6a\exp\Big(-\frac{b}{81}\ell S_{uvw}^2\epsilon^2\Big),\end{aligned}\tag{14}$$

which is a two-sided bound with slightly worse constants than those of the theorem. □

The novel principle of HGT is that the selection of triplets in an algorithm following the outline of §1.4 with regular estimators should be a greedy selection of the triplet uvw with the largest *average estimated similarity* defined by

$$\hat{S}_{uvw}=\frac{3}{e^{\hat{\mathbf{D}}[u,v]}+e^{\hat{\mathbf{D}}[u,w]}+e^{\hat{\mathbf{D}}[v,w]}}.$$

Algorithm Harmonic Greedy Triplets with Four Point Condition

Input: An $n \times n$ estimated tree metric.

Output: Ψ^*.

F1 Select an arbitrary leaf u and find a triplet uvw with the maximum $\hat{S}_{uvw}$.

F2 Let Ψ^* be the star with three edges formed by uvw and its center o.

F3 Initialize $\mathcal{R}$ using the good relevant pairs for edges uo, vo, wo.

F4 **repeat**

F5 Find $\langle z_1 z_2, uvw \rangle \in \mathcal{R}$ with the maximum $\hat{S}_{uvw}$.

F6 Add a new internal node z on $z_1 z_2$ and connect w to it.

F7 Delete the pairs from $\mathcal{R}$ that contain the edge $z_1 z_2$.

F8 Update $\mathcal{R}$ using the good relevant pairs for edges $z_1 z$, $z_2 z$, wz.

F9 **until** all leaves are inserted to Ψ^*; i.e., this loop has iterated $(n-3)$ times.

F10 Output Ψ^*.

Figure 2: *The HGT/FP algorithm.* Calculations pertaining to edge length estimation are sketched in §2.1. Good relevant pairs are discussed in §2.2. The set $\mathcal{R}$ of good relevant pairs is discussed in §2.5.

2.5 Fast topology reconstruction

The HGT/FP algorithm uses good relevant pairs to add new nodes to Ψ^*. For every edge $e \in \Psi^*$ and leaf $w \notin \Psi^*$, there are $\mathcal{O}(1)$ relevant pairs of the form $\langle e, uvw \rangle$. Maintaining the set of all $\mathcal{O}(n^2)$ good relevant pairs while constructing Ψ^* would be possible: whenever a new leaf is added, $\mathcal{O}(n)$ relevant pairs are eliminated, $\mathcal{O}(n)$ new relevant pairs are created, and the new relevant pairs can be tested as described in §2.2. By the HGT principle, however, it is enough to consider one good relevant pair for every leaf $w \notin \Psi^*$, namely, the one in which the triplet has the largest average similarity. Denote the set of those pairs by $\mathcal{R}$. The HGT/FP algorithm maintains $\mathcal{R}$ by updating it every time new nodes are added. The set $\mathcal{R}$ is implemented as a vector of size n indexed by the leaves. Each entry of $\mathcal{R}$ contains either null or a good relevant pair. In order to add a new internal node and a new leaf, the HGT/FP algorithm uses the relevant pair from $\mathcal{R}$ in which the triplet has the largest average empirical similarity.

Figure 2 shows the HGT/FP algorithm. The use of the HGT principle and relevant triplets results in the following theorem.

Theorem 3. *The running time of the HGT/FP algorithm on a tree with n leaves is $\mathcal{O}(n^2)$. The algorithm uses $\mathcal{O}(n)$ work space.*

Proof. The algorithm stores the tree Ψ^*, the vector $\mathcal{R}$, and $\mathcal{O}(1)$ local variables, resulting in the $\mathcal{O}(n)$ space requirement.

The topology Ψ^* is stored by the implementation as a directed binary tree, where the edges are directed from the first leaf u selected in Line F1. The orientation of the edges is maintained throughout the algorithm, which is achieved by adding a new leaf w in Line F6 as a left child if $z_1 z_2$ is a right edge, or adding w as a right child if $z_1 z_2$ is a left edge. Furthermore, for every internal node z, the triplet $\mathrm{def}(z)$ is stored with the orientation determined by the implementation, so that for every leaf $u \in \mathrm{def}(z)$, it can be decided in $\mathcal{O}(1)$ time whether it is in the left or right subtree rooted at z, or neither. With this technique, it takes $\mathcal{O}(1)$ time to collect the relevant pairs for every edge edge $e \in \Psi^*$ and leaf $w \notin \Psi^*$, and thus Lines F3 and F8 take $\mathcal{O}(n)$ time. Since there are $\mathcal{O}(n)$ entries in $\mathcal{R}$, Lines F5 and F7 take $\mathcal{O}(n)$ time also. Line F1 runs in $\mathcal{O}(n^2)$ time. Lines F2 and F6 update Ψ^* in $\mathcal{O}(1)$ time.

Thus, initialization in Lines F1–F3 takes $\mathcal{O}(n^2)$ time, and the repeat loop of Line F4 is executed $(n-3)$ times, taking $\mathcal{O}(n)$ time in each step, which results in the $\mathcal{O}(n^2)$ total running time. □

3 Statistical efficiency of the algorithm

In order to obtain sample length bounds for the HGT/FP algorithm, we bound the algorithm's success probability. Let $0 < S_{\mathrm{sm}} < S_{\mathrm{lg}} < 1$ be two threshold values on similarities with $S_{\mathrm{sm}} = S_{\mathrm{lg}}/\sqrt{2}$ (we specify S_{lg} later). The thresholds define three sets of triplets: every triplet uvw is either a *large* triplet, if $S_{uvw} \geq S_{\mathrm{lg}}$, or a *medium* triplet if $S_{\mathrm{sm}} < S_{uvw} < S_{\mathrm{lg}}$, or a *small* triplet if $S_{uvw} \leq S_{\mathrm{sm}}$. We show that with high probability, the HGT/FP algorithm recovers the tree correctly using only large and medium triplets. Let Ψ_k^* be the version of Ψ^* with k leaves at the beginning of the repeat loop in line F4 for $k = 3, \ldots, n-1$, and let $\Psi_n^* = \Psi^*$, the algorithm's output. We prove for all k by induction that with high probability, Ψ_k^* is built correctly by using only large and medium triplets. Establishing the base case and the induction step relies on the following three arguments.

1. With high probability, the greedy selection favors large triplets over small triplets.

2. With high probability, HGT/FP correctly determines whether any relevant pair $\langle z_1 z_2, uvw \rangle$, for which uvw is not small is a good relevant pair.

3. With high probability, the relevant pair $\langle z_1 z_2, uvw \rangle$ selected from $\mathfrak{R}$ is such that the triplet uvw is not small and its center falls on the path between z_1 and z_2 in $\Psi(T)$.

(In order to make the third argument more precise, observe that there is a mapping from the nodes of Ψ^* to those of $\Psi(T)$ defined by $\mathrm{def}(\cdot)$: if $z \in \Psi^*$ is a leaf, it corresponds to leaf z in $\Psi(T)$, otherwise it corresponds to the center of $\mathrm{def}(z)$ in $\Psi(T)$. Thus, the path between z_1 and z_2 in $\Psi(T)$ is well-defined: it is the path between the *images* of z_1 and z_2 as defined by the mapping.)

The first argument follows from Lemma 5 below, which states that if a regular estimator is used, then the probability of a small triplet appearing better than a large triplet is exponentially small in the sample sequence lengths.

Lemma 5. *Assume that* $\hat{\mathbf{D}}$ *is an* (a, b)*-regular estimator for* $\mathbf{D}$ *calculated from sample sequences of length* ℓ*. Let* $\mathcal{E}_3$ *denote the random event that* $\hat{S}_{uvw} < \hat{S}_{u'v'w'}$ *for every small triplet* uvw *and every large triplet* $u'v'w'$*.*

$$\mathbb{P}\Big\{\mathcal{E}_3\Big\} \geq 1 - \frac{a}{6} n^3 \exp\left(-b \frac{(\sqrt{2}-1)^2}{72} \ell S_{\mathrm{lg}}^2\right)$$

Proof. We claim that for every small triplet uvw,

$$\mathbb{P}\Big\{\hat{S}_{uvw} \geq S_{\mathrm{md}}\Big\} \leq a\exp\Big(-b\frac{(\sqrt{2}-1)^2}{72}\ell S_{\mathrm{lg}}^2\Big), \tag{15}$$

and for every large triplet $u'v'w'$,

$$\mathbb{P}\Big\{\hat{S}_{u'v'w'} \leq S_{\mathrm{md}}\Big\} \leq a\exp\Big(-b\frac{(\sqrt{2}-1)^2}{72}\ell S_{\mathrm{lg}}^2\Big). \tag{16}$$

We use the following basic inequality.

$$\min\Big\{\frac{\hat{S}_{uv}}{S_{uv}}, \frac{\hat{S}_{uw}}{S_{uw}}, \frac{\hat{S}_{vw}}{S_{vw}}\Big\} \leq \frac{\hat{S}_{uvw}}{S_{uvw}} \tag{17}$$

We first prove Equation (15). Without loss of generality, we may suppose

$$\min\Big\{\frac{\hat{S}_{uv}}{S_{uv}}, \frac{\hat{S}_{uw}}{S_{uw}}, \frac{\hat{S}_{vw}}{S_{vw}}\Big\} = \frac{\hat{S}_{uv}}{S_{uv}}.$$

Then by Equations (13), and (17), and the fact that $\hat{\mathbf{D}}$ is (a,b)-regular,

$$\begin{aligned}
\mathbb{P}\Big\{\hat{S}_{uvw} \leq S_{\mathrm{md}}\Big\} &= \mathbb{P}\Big\{\frac{\hat{S}_{uvw}}{S_{uvw}} \leq \frac{S_{\mathrm{md}}}{S_{uvw}}\Big\} \\
&\leq \mathbb{P}\Big\{\frac{\hat{S}_{uv}}{S_{uv}} \leq \frac{S_{\mathrm{md}}}{S_{uvw}}\Big\} \leq a\exp\left(-b\ell S_{uv}^2\Big(1-\frac{S_{\mathrm{md}}}{S_{uvw}}\Big)^2\right) \\
&\leq a\exp\left(-\frac{b\Big(1-\frac{S_{\mathrm{md}}}{S_{\mathrm{lg}}}\Big)^2}{9}\ell S_{\mathrm{lg}}^2\right).
\end{aligned}$$

By the choice of S_{md},

$$\frac{\Big(1-\frac{S_{\mathrm{md}}}{S_{\mathrm{lg}}}\Big)^2}{9} = \frac{(\sqrt{2}-1)^2}{72},$$

and thus Equation (15) holds. Equation (16) is proven similarly.

The probability of the complementary event $\bar{\mathcal{E}}_3$ is bounded by the probability that there is a small triplet uvw for which $\hat{S}_{uvw} \geq S_{\mathrm{md}}$ or there is a large triplet $u'v'w'$ for which $\hat{S}_{u'v'w'} \leq S_{\mathrm{md}}$. Hence Equations (15) and (16) imply the lemma, since there are $\binom{n}{3} < n^3/6$ triplets. □

For the second argument, assume that Ψ_k^* has the correct topology, and that $\mathrm{def}(z)$ is a large or a medium triplet for every internal node z. Assume furthermore that $\langle z_1 z_2, uvw\rangle$ is a relevant pair in Ψ_k^*. We take advantage of the fact that if uvw is not small, then the leaves for which the relaxed four-point condition is tested cannot be arbitrarily far from each other. Specifically, the distance between two leaves within the quartets is bounded from above by $-2\ln\left(\frac{2}{3}S_{\mathrm{sm}}\right)$, based on the fact that for every triplet uvw with center o, $S_{uo} \geq \frac{2}{3}S_{uvw}$.

Definition 3. *Let $D_{\min}$ be the minimum, and $D_{\max}$ be the maximum distance between endpoints of edges in $\Psi(T)$, and define $S_0 = e^{-D_{\max}}$, $S_1 = 1 - e^{-D_{\min}}$.*

Lemma 6. *A quartet is a* short quartet *if each pairwise distance between its leaves is less than $-2\ln\left(\frac{2}{3}S_{\mathrm{sm}}\right)$. Let $\mathcal{E}_4$ denote the random event that for every short quartet with topology $uv|wz$, the relaxed four-point condition of Equation* (2) *holds.*

$$\mathbb{P}\Big\{\mathcal{E}_4\Big\} \geq 1 - an^2 \exp\Big(-\frac{b}{81}\ell S_{\mathrm{lg}}^4 S_1^2\Big).$$

Proof. Using the technique of Erdős et al. (1999a), the probability of the complementary event $\bar{\mathcal{E}}_4$ is bounded by the probability that there is a leaf pair (u, v) for which $\left|\mathbf{D}[u,v] - \hat{\mathbf{D}}[u,v]\right| \geq (-\ln(1-S_1))/2$. The lemma follows from the fact that $\hat{\mathbf{D}}$ is (a,b)-regular, and that there are $\binom{n}{2} < \frac{n^2}{2}$ leaf pairs. □

Notice that $\mathcal{E}_4$ implies also that for every short quartet for which the relaxed four-point condition of Equation (2) holds, the correct quartet topology is $uv|wz$. Lemma 6 leads immediately to the following corollary.

Corollary 7. *Assume that Ψ_k^* has the correct topology, and that for every internal node $z \in \Psi_k^*$, $\mathrm{def}(z)$ is not small. The event $\mathcal{E}_4$ implies that if $\langle z_1 z_2, uvw\rangle$ is a good relevant pair and the triplet uvw is not small, then the center of uvw falls on the path between z_1 and z_2 in $\Psi(T)$. Conversely, if $\langle z_1 z_2, uvw\rangle$ is a relevant pair, the triplet uvw is not small, and the center of uvw falls on the path between z_1 and z_2 in $\Psi(T)$, then $\langle z_1 z_2, uvw\rangle$ is a good relevant pair.*

The third argument is based on Lemma 8 below. Lemma 8 depends on how large the defining triplets are for the internal nodes of $\Psi(T)$, and determines the value of S_{lg}.

Definition 4. *Define the* tree depth ϱ *as the smallest number such that for every edge* $e \in \Psi(T)$*, there is a path from each endpoint to a leaf with at most* ϱ *edges, which does not go through* e*.*

Remark. The value ϱ was first studied in the context of evolutionary tree reconstruction by Erdős et al. (1999a) under different topology distributions. They proved that for all topologies, $\varrho \leq 1 + \log_2(n-1)$, and for almost every tree topology under the uniform or Yule-Harding distributions, $\varrho = \mathcal{O}(\log \log n)$.

Lemma 8. *Assume that* Ψ_k^* *has the correct topology, and that for every internal node* $z \in \Psi_k^*$*,* $\mathrm{def}(z)$ *is not small. If*

$$S_{\mathrm{lg}} \leq \frac{3\sqrt{2}}{2}\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^2 S_0^{2\varrho+4} \qquad \left(\approx \frac{S_0^{2\varrho+4}}{16}\right),$$

then the following statement holds. For every edge $z_1 z_2 \in \Psi_k^*$*, if* z_1 *and* z_2 *are not connected by one edge in* $\Psi(T)$*, then there exists a relevant pair* $\langle z_1 z_2, uvw\rangle$ *such that* uvw *is large, and the center of* uvw *falls onto the path between* z_1 *and* z_2 *in* $\Psi(T)$*.*

Proof. The proof is identical to the one given by Csűrös and Kao (2001) for the Jukes-Cantor distance. □

The statistical efficiency of the HGT/FP algorithm is stated by the following theorem.

Theorem 4. *Let* T *be an arbitrary evolutionary tree that has* n *leaves. Let* $\mathbf{D}$ *be a tree metric over the topology* $\Psi(T)$*, and* $\hat{\mathbf{D}}$ *be an* (a, b)*-regular estimator. For every error probability* $0 < \delta < 1$*, there exists*

$$\ell = \mathcal{O}\left(\frac{\log \frac{a}{\delta} + \log n}{b S_0^{\mathcal{O}(\varrho)} S_1^2}\right), \tag{18}$$

such that the success rate of HGT/FP is at least $(1-\delta)$ *on samples of length* ℓ*.*

Proof. Let $\mathcal{A}_k$ denote the event that Ψ_k^* is correct, and $\mathcal{B}_k$ denote the event that for every internal node $z \in \Psi_k^*$, $\mathrm{def}(z)$ is a large or medium triplet. Let $\mathcal{R}_k$ denote the version of $\mathcal{R}$ in Line F5 when Ψ^* has k leaves. Let $\mathcal{C}_k$ denote the event that the following two statements hold.

(i) $\mathcal{R}_k$ contains a good relevant pair $\langle z_1 z_2, uvw\rangle$ for which $\hat{S}_{uvw} > S_{\mathrm{md}}$.

(ii) For every pair $\langle z_1 z_2, uvw \rangle$, if uvw is not small, then the center of uvw falls on the path between z_1 and z_2 in $\Psi(T)$.

We prove by induction that $\mathcal{E}_3$ and $\mathcal{E}_4$ imply that for all k, $\mathcal{A}_k$, $\mathcal{B}_k$, and $\mathcal{C}_k$ hold. By $\mathcal{E}_3$, a medium or large triplet is selected to initialize Ψ^*, and thus $\mathcal{A}_3$, and $\mathcal{B}_3$ hold. By $\mathcal{E}_3$, Lemma 8, and Corollary 7, $\mathcal{C}_3$ holds also.

Assume that $\mathcal{A}_k$, $\mathcal{B}_k$, and $\mathcal{C}_k$ hold for some value of $3 \leq k < n$. By $\mathcal{E}_3$ and $\mathcal{C}_k$, the pair selected in Line F5 has a non-small triplet, and thus $\mathcal{B}_{k+1}$ holds. By Corollary 7, $\mathcal{A}_{k+1}$ holds also. If $k < n$, then by Lemma 8, $\mathcal{C}_{k+1}$ holds as well.

Consequently, $\mathcal{E}_3$ and $\mathcal{E}_4$ imply $\mathcal{A}_n$. By Lemmas 5 and 6, if

$$\ell \geq \max\left\{ \frac{72}{(\sqrt{2}-1)^2} \frac{3\ln n + \frac{a}{3\delta}}{bS_{\text{lg}}^2}, 81 \frac{2\ln n + \frac{2a}{\delta}}{bS_1^2 S_{\text{lg}}^4} \right\},$$

then both $\mathcal{E}_3$ and $\mathcal{E}_4$ hold with probability at least $(1-\delta)$. The theorem follows from setting $S_{\text{lg}} = S_0^{2\varrho+4}/17$ as suggested by Lemma 8. □

4 Simulation experiments

We simulated DNA sequence evolution with 2000 nucleotides along three large trees in the Jukes-Cantor model. The 500-leaf tree has the topology of a seed plant phylogeny by Chase et al. (1993). The 1895-leaf tree is derived from the evolutionary tree of Eukaryotes in RDP (Maidak et al. 2000). The 3135-leaf tree is based on the subtree of Proteobacteria within the phylogeny of Prokaryotes in RDP. We scaled the edge lengths of the original trees using a linear transformation. We evaluated the accuracy of the reconstruction by the Robinson-Foulds error (Robinson and Foulds 1981) RF%, which measures the percentage of misplaced internal edges in the tree.

The distance-based algorithms we used included BioNJ (Gascuel 1997), Weighbor (Bruno et al. 2000), and Neighbor-Joining (NJ) (Saitou and Nei 1987). The two former algorithms were recently developed, and are related to NJ. The NJ algorithm is arguably the most popular distance-based algorithm to date. All three algorithms run in $\mathcal{O}(n^3)$ time (Studier and Keppler 1988), and their statistical efficiency is not proven. Atteson (1999) derives sample length bounds for Neighbor-Joining and BioNJ similar to those of Equation (18), but the sample length bounds use the diameter of the topology instead of ϱ, and are thus exponential in the tree size. For Weighbor and BioNJ, we used the implementations provided by their authors; for NJ, we used its implementation in qclust (Brzustowski 1998). We also included a heuristic parsimony method, called DNAPARS (Felsenstein 1993). Parsimony algorithms aim at deriving a topology that gives rise to sample sequences with a minimal number of character changes along the edges, which is an NP-hard optimization problem (Graham and Foulds 1982). It is known that the exact optimization is not statistically efficient for certain trees (Felsenstein 1978). In simulation experiments, however, they often perform very well (Rice and Warnow 1997).

Figures 3–5 show the simulation results. In our experiments parsimony performs the best on the 500-leaf tree. This phenomenon may be potentially due to the fact that the model tree was built using parsimony, and thus there is a certain bias in favor of pasimony methods in simulations involving the 500-leaf tree. Similar systematic bias was also experienced by Rannala et al. (1998) while performing experiments with a 228-leaf model tree. Parsimony's running time increases rapidly with the tree size and the mutation probabilities, so that in some cases it takes several hours on a desktop computer[1] to recover the topology of the 500-leaf tree. In contrast, HGT/FP

[1]We used a PC with Pentium III 500 MHz CPU and 256M memory, running Windows

takes less than two minutes to reconstruct the topology of the 3135-leaf tree. Weighbor proves to be even slower than parsimony despite its good asymptotic running time. We omitted Weighbor and DNAPARS from the experiments with the larger trees because their running time increased to more than a day, and also omitted BioNJ because its performance is very similar to that of NJ. In the case of high mutation probabilities, HGT/FP performs better than NJ and BioNJ, while the neighbor-joining methods are better for low mutation probabilities. Specifically, on the 500-leaf tree, Weighbor is almost as successful as parsimony, while NJ and BioNJ do not recover the topologies completely.

Figure 6 shows the results of a different set of experiments comparing the effect of sample length on the recovery for HGT/FP and NJ. We simulated sequence evolution along the 1895-leaf tree with sample lengths ranging from 200 to 10000, for two edge scalings. In the case of high mutation probabilities, HGT/FP recovers the tree completely from 5000 bp sequences, while NJ misses more than 200 edges even for 10000 bp sequences. In the case of low mutation probabilities, NJ performs better than HGT/FP, but the difference is not so striking between the two algorithms as in the case of high mutation probabilities. In particular, the convergence rate of HGT/FP seems to be close to that of NJ.

NT 4.0.

5 Concluding remarks

When working with trees with over one thousand leaves, the algorithms' running time becomes crucial. Existing $\mathcal{O}(n^4)$-time evolutionary tree building algorithms may take days to finish on today's desktop computers and slower algorithms are virtually unusable without having considerable insight into biological features of the data set at hand.

In addition to the computational issues, statistical characteristics of algorithms also become more stressed as one builds larger trees. Neighbor-Joining and most other algorithms have not been proven to require asymptotically polynomial sample sizes to correctly recover the topology, while HGT/FP is provably statistically efficient. Neighbor-Joining calculates edge lengths from an average that involves distances between arbitrarily remote nodes in T, which may cause the estimation error to be very large. When the mutation probabilities are small, the averaging approach may be justified, as shown in the experiments with low mutation probabilities. On the other hand, the error committed while calculating the average is governed by the error in estimating the largest distance in the expression, which may be significant when mutation probabilities are large. In this case a greedy algorithm such as HGT/FP is more successful, as shown by our experimental results.

Many possible applications of evolutionary tree building algorithms may need to build large trees. Examples include large projects in evolutionary biology such as the ones cited and problems in molecular epidemiology. It will be of practical importance to determine which of the existing algorithms are the most suitable for the ranges of mutation probabilities and tree topologies defined by the application at hand. Our work indicates that recovering phylogenies with thousands of nodes is not as daunting a task as previously assumed. While we do not foresee that future algorithms will significantly increase the speed of recovery, future work in the field will undoubtedly lead to algorithms with even greater efficiency.

Acknowledgments

I am very thankful to Ming-Yang Kao and Dana Angluin for discussions leading to many of the results. The design and discussion of the experiments benefited from conversations with Bill Bruno and Aaron Halpern, who also kindly provided version 1.2 of their Weighbor program.

References

Atteson, K. (1999). The performance of neighbor-joining methods of phylogeny reconstruction. *Algorithmica 25*, 251–278.

Bandelt, H.-J. and A. Dress (1986). Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics 7*, 309–343.

Brown, K. S. (1999). Deep Green rewrites evolutionary history of plants. *Science 285*, 990.

Bruno, W. J., N. D. Socci, and A. L. Halpern (2000). Weighted Neighbor-Joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution 17*, 189–197.

Brzustowski, J. (1998). *qclust V0.2*. ⟨http://www.biology.ualberta.ca/jbrzusto/⟩.

Buneman, P. (1971). The recovery of trees from dissimilarity matrices. In F. R. Hodson, D. G. Kendall, and P. Tautu (Eds.), *Mathematics in the Archaelogical and Historical Sciences*, pp. 387–395. Edinburgh University Press.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y.-L. Qiu, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedrn, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. M. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn, Jr., S. W. Graham, S. C. H. Barrett, S. Dayanandan, and V. A. Albert (1993). Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene rbcL. *Annals of the Missouri Botanical Garden 80*, 528–580.

Cryan, M., L. A. Goldberg, and P. W. Goldberg (1998). Evolutionary trees can be learned in polynomial time in the two-state general Markov model. In *39th Annual Symposium on Foundations of Computer Science*, pp. 436–445. IEEE.

Csűrös, M. (2000). *Reconstructing Phylogenies in Markov Models of Sequence Evolution*. Ph. D. thesis, Yale University. ⟨http://www.iro.umontreal.ca/~csuros/papers/⟩.

Csűrös, M. and M.-Y. Kao (2001). Provably fast and accurate recovery of evolutionary trees through Harmonic Greedy Triplets. *SIAM Journal on Computing 31*, 306–322.

Erdős, P. L., M. A. Steel, L. A. Székely, and T. J. Warnow (1999a). A few logs suffice to build (almost) all trees (I). *Random Structures and Algorithms 14*, 153–184.

Erdős, P. L., M. A. Steel, L. A. Székely, and T. J. Warnow (1999b). A few logs suffice to build (almost) all trees (II). *Theoretical Computer Science 221*, 77–118.

Farach, M. and S. Kannan (1999). Efficient algorithms for inverting evolution. *Journal of the ACM 46*, 437–449.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology 22*, 240–249.

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author*. Seattle, Wash.: University of Washington Department of Genetics.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution 14*(7), 685–695.

Graham, R. L. and L. R. Foulds (1982). Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Mathematical Biosciences 60*, 133–142.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Huson, D. H., S. M. Nettles, and T. J. Warnow (1999). Disk-Covering, a fast-converging method for phylogenetic tree reconstruction. *Journal of Computational Biology 6*, 369–386.

Jukes, T. H. and C. R. Cantor (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian Protein Metabolism*, Volume III, Chapter 24, pp. 21–132. New York: Academic Press.

Kim, J. (1998). Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Systematic Biology 47*, 43–60.

Kimura, M. (1981). Estimation of evolutionary differences between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the USA 78*, 454–458.

Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences of the USA 91*, 1455–1459.

Lanave, C., G. Preparata, C. Saccone, and G. Serio (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution 20*, 86–93.

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution 11*, 605–612.

Maidak, B. L., J. R. Cole, T. G. Lilburn, J. Charles T. Parker, P. R. Saxman, J. M. Stredwick, G. M. Garrity, B. Li, G. J. Olsen, S. Pramanik, T. M. Schmidt, and J. M. Tiedje (2000). The RDP (Ribosomal Database Project) continues. *Nucleic Acids Research 28*, 173–174.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, pp. 148–188. Cambridge: Cambridge University Press.

Neyman, J. (1971). Molecular studies of evolution: a source of novel statistical problems. In S. S. Gupta and J. Yackel (Eds.), *Statistical Decision Theory and Related Topics*, pp. 1–27. New York: Academic Press.

Rannala, B., J. P. Huelsenbeck, Z. Yang, and R. Nielsen (1998). Taxon sampling and the accuracy of large phylogenies. *Systematic Biology 47*, 702–710.

Rice, K. and T. Warnow (1997). Parsimony is hard to beat! In *Computing and Combinatorics, Third Annual International Conference*, Volume 1276 of *Lecture Notes in Computer Science*, pp. 124–133. Springer-Verlag.

Robinson, D. F. and L. R. Foulds (1981). Comparison of phylogenetic trees. *Mathematical Biosciences 53*, 131–147.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution 4*(4), 406–425.

Steel, M. A. (1994). Recovering a tree from the leaf colourations it generates under a Markov model. *Applied Mathematics Letters 7*, 19–24.

Studier, J. A. and K. J. Keppler (1988). A note on the neighbor-joining method of Saitou and Nei. *Molecular Biology and Evolution 5*, 729–731.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Lectures on mathematics in the life sciences*, Volume 17, Providence, RI, pp. 57–86. AMS.

Warnow, T., B. M. Moret, and K. S. John (2001). Absolute convergence: true trees from short sequences. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 186–195. ACM.

Waterman, M. S., T. F. Smith, M. Singh, and W. A. Beyer (1977). Additive evolutionary trees. *Journal of Theoretical Biology 64*, 199–213.

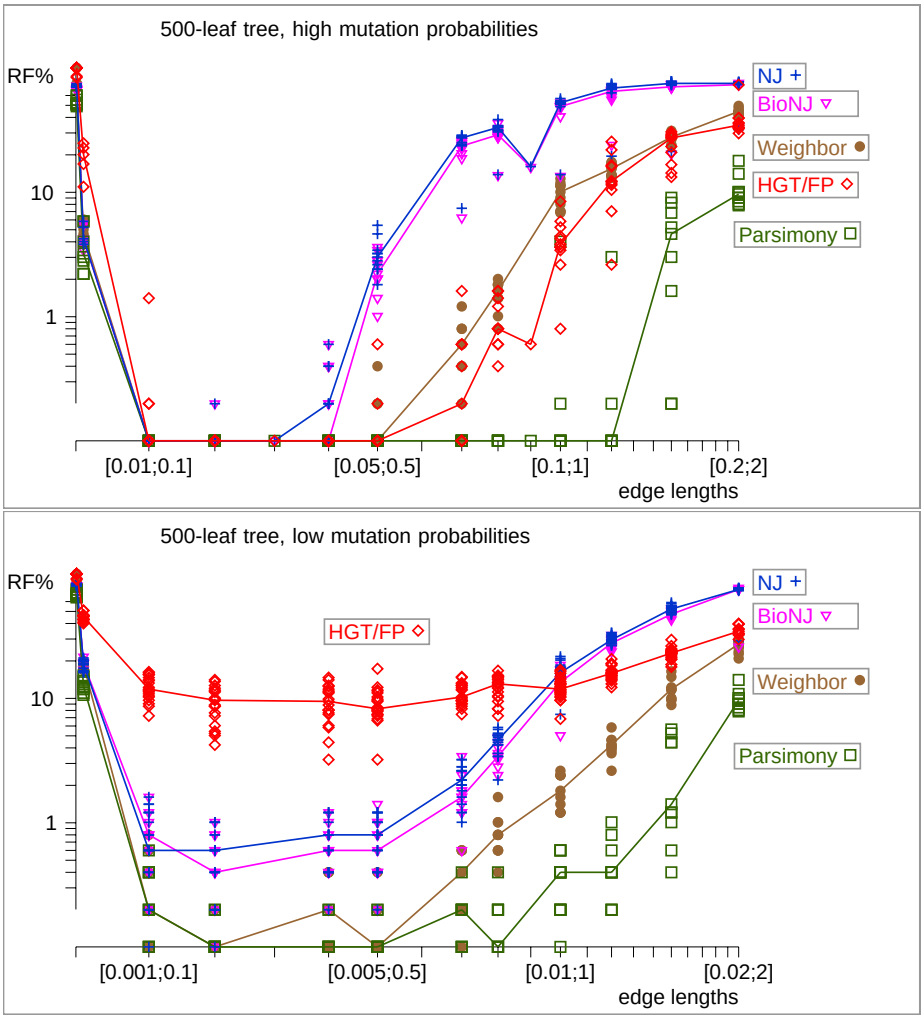

Figure 3: *Simulation of DNA sequence evolution in the Jukes-Cantor model along the 500-leaf tree with different mutation probabilities.* The plots show the percentage of misplaced internal edges (Robinson-Foulds error) as a function of the largest edge length $D_{\max}$ in the tree after linear scaling. The graphs are calculated from generating ten sets of samples with 2000 bp long sequences. The graphs go through the median values. On the top, the minimum edge length equals $D_{\max}/10$. On the bottom, it is set to $D_{\max}/100$. For reference, $D_{\max} = 0.5$ corresponds to maximum mutation probability 0.30; on the top the minimum mutation probability equals 0.037 and on the bottom it equals 0.0037 for that scaling. Most edge lengths in the trees are very close to the minimum edge length.

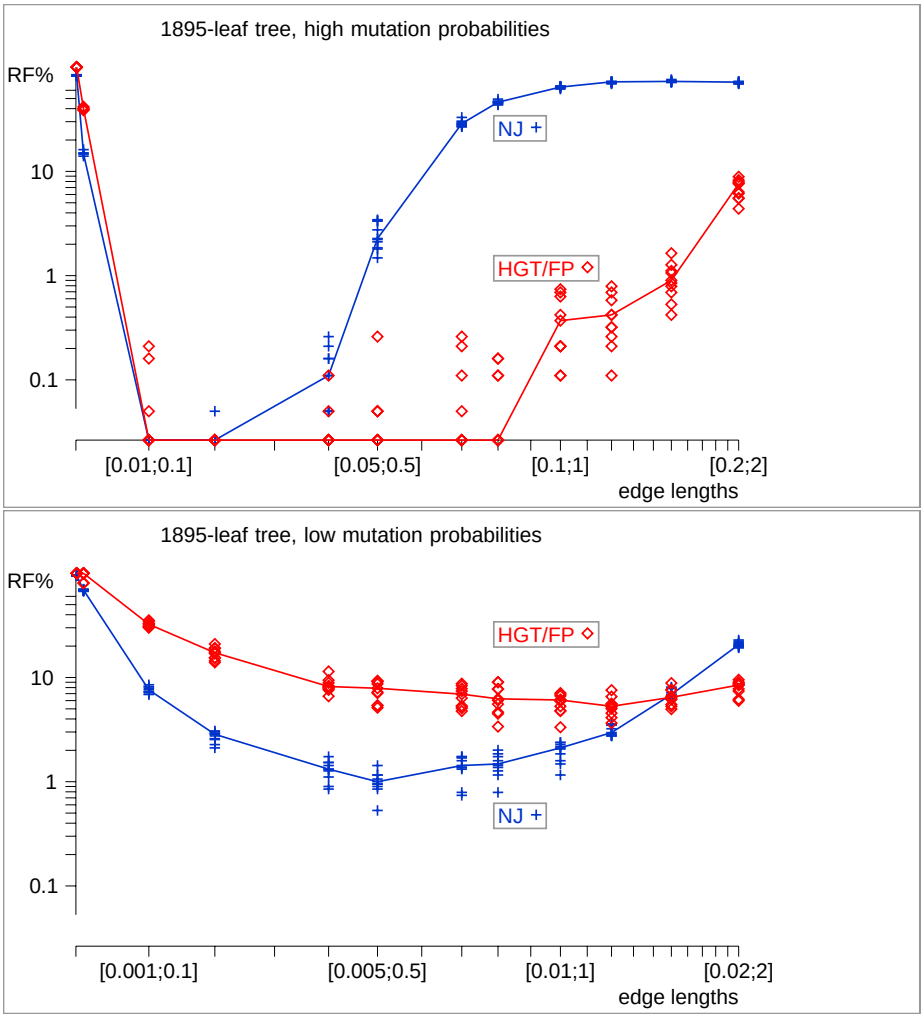

Figure 4: *Simulation of DNA sequence evolution in the Jukes-Cantor model along the 1895-leaf tree with different mutation probabilities.*

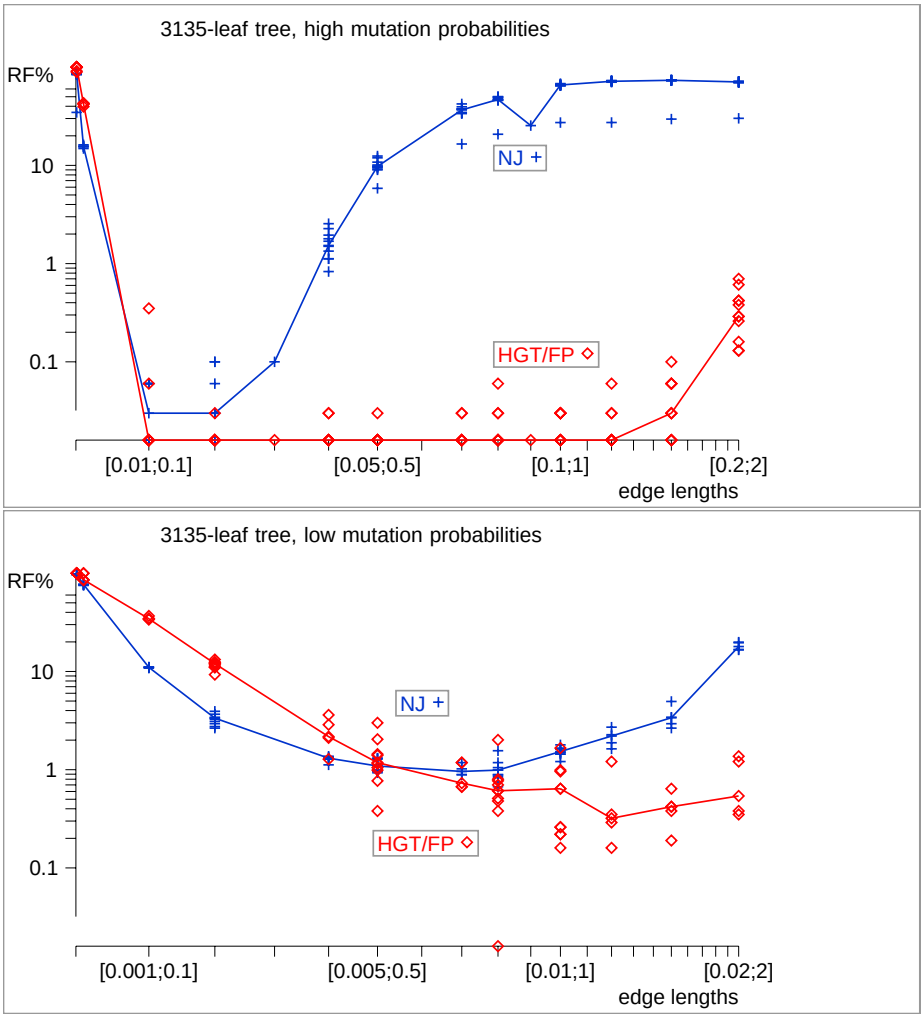

Figure 5: *Simulation of DNA sequence evolution in the Jukes-Cantor model along the 3135-leaf tree with different mutation probabilities.*

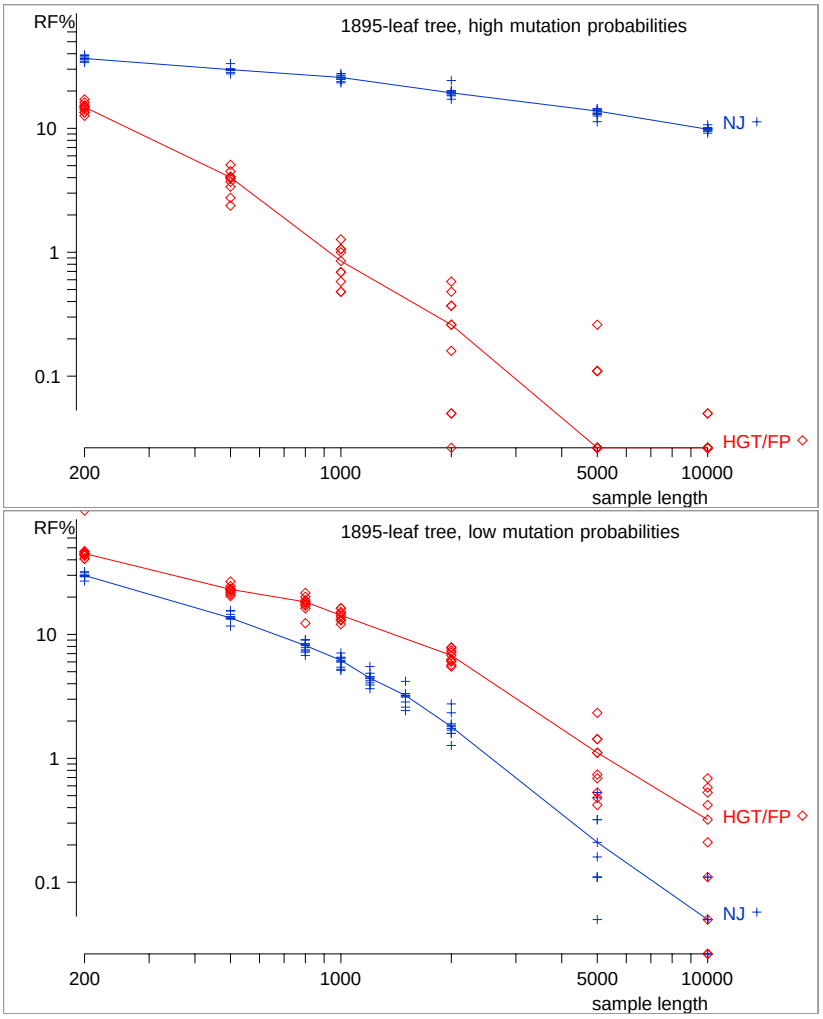

Figure 6: *Simulation of DNA sequence evolution in the Jukes-Cantor model along the 1895-leaf tree with different sample lengths.* The plots show the percentage of misplaced internal edges (Robinson-Foulds error) as a function of the sample lengths. The graphs are calculated from generating ten sets of samples for each sequence length. The graphs go through the median values. The edge lengths on the top are linearly scaled to fall into the interval $[0.1, 1]$. The edge lengths on the bottom fall into the interval $[0.01, 1]$.