

# PROVABLY FAST AND ACCURATE RECOVERY OF EVOLUTIONARY TREES THROUGH HARMONIC GREEDY TRIPLETS\*

MIKLÓS CSÚRÖS AND MING-YANG KAO

**Abstract.** We give a greedy learning algorithm for reconstructing an evolutionary tree based on a certain harmonic average on triplets of terminal taxa. After the pairwise distances between terminal taxa are estimated from sequence data, the algorithm runs in  $\mathcal{O}(n^2)$  time using  $\mathcal{O}(n)$  work space, where  $n$  is the number of terminal taxa. These time and space complexities are optimal in the sense that the size of an input distance matrix is  $n^2$  and the size of an output tree is  $n$ . Moreover, in the Jukes-Cantor model of evolution, the algorithm recovers the correct tree topology with high probability using sample sequences of length polynomial in (1)  $n$ , (2) the logarithm of the error probability, and (3) the inverses of two small parameters.

**Key words.** evolutionary trees, the Jukes-Cantor model of evolution, computational learning, harmonic greedy triplets

**AMS subject classifications.** 05C05, 05C85, 92D15, 60J85, 92D20

**1. Introduction.** Algorithms for reconstructing evolutionary trees are useful tools in biology [16, 22]. These algorithms usually compare aligned character sequences for the terminal taxa in question to infer their evolutionary relationships. In the past, such characters were often categorical variables of morphological features; newer studies have taken advantage of available biomolecular sequences. This paper focuses on datasets of the latter type.

We present a new learning algorithm, called *Fast Harmonic Greedy Triplets* (Fast-HGT), using a greedy strategy based on a certain harmonic average on triplets of terminal taxa. After the pairwise distances between terminal taxa are estimated from their observed sequences, Fast-HGT runs in  $\mathcal{O}(n^2)$  time using  $\mathcal{O}(n)$  work space, where  $n$  is the number of terminal taxa. These time and space complexities are optimal in the sense that  $n^2$  is the size of an input distance matrix and  $n$  is the size of an output tree. An earlier variant of Fast-HGT takes  $\mathcal{O}(n^5)$  time [5]. In the Jukes-Cantor model of sequence evolution generalized for an arbitrary alphabet [22], Fast-HGT is proven to recover the correct topology with high probability while requiring sample sequences of length  $\ell$  polynomial in (1)  $n$ , (2) the logarithm of the error probability, and (3) the inverses of two small parameters (Theorem 3.8). In subsequent work [6], Fast-HGT and its variants are shown to have similar theoretical performance in more general Markov models of evolution.

Among the related work, there are four other algorithms which have essentially the same guarantee on the length  $\ell$  of sample sequences. These are the Dyadic Closure Method (DCM) [10] and the Witness-Anti-witness Method (WAM) [11] of Erdős, Steel, Székely, and Warnow, the algorithm of Cryan, Goldberg, and Goldberg (CGG) [4], and the DCM-Buneman algorithm of Huson, Nettles, and Warnow [18]. Not all of these results analyzed the space complexity. In terms of time complexity, DCM-Buneman is not a polynomial-time algorithm. CGG runs in polynomial time, whose degree has not been explicitly determined but which appears to be higher than  $n^2$ . DCM takes  $\mathcal{O}(n^5 \log n)$  time to assemble  $\mathcal{O}(n^4)$  quartets using  $\mathcal{O}(n^4)$  space. The two versions of WAM take  $\mathcal{O}(n^6 \log n)$  and  $\mathcal{O}(n^4 \log n \log \ell)$  time, respectively. In the uniform and

---

\*Department of Computer Science, Yale University, New Haven, CT 06520; {csuros-miklos, kao-ming-yang}@cs.yale.edu. Research supported in part by NSF Grant 9531028.

Yule-Harding models of randomly generating trees, with high probability, these two latter running times are reduced to  $\mathcal{O}(n^3 \text{polylog } n)$  and  $\mathcal{O}(n^2 \text{polylog } n)$ , respectively. Under these two tree distributions, Erdős *et al.* [10] further showed that with high probability, the required sample size of DCM is polylogarithmic in  $n$ ; this bound also applies to WAM, CGG, DCM-Buneman, and Fast-HGT.

Among the algorithms with no known comparable guarantees on  $\ell$ , the Neighbor Joining Method of Saitou and Nei [22] runs in  $\mathcal{O}(n^3)$  time and reconstructs many trees highly accurately in practice, although the best known upper bound on its required sample size is exponential in  $n$  [3]. Maximum likelihood methods [14, 15] are not known to achieve the optimal required sample size as such methods are usually expected to [20]; moreover, all their known implementations take exponential time to find local optima, and none can find provably global optima. Parsimony methods aim to compute a tree that minimizes the number of mutations leading to the observed sequences [13]; in general, such optimization is NP-hard [8]. Some algorithms strive to find an evolutionary tree among all possible trees to fit the observed distances the best according to some metric [1]; such optimization is NP-hard for  $L_1$  and  $L_2$  norms [7] and for  $L_\infty$  [1].

A common goal of the above algorithms is to construct a tree with the same topology as that of the true tree. In contrast, the work on PAC-learning the true tree in the  $j$ -State General Markov Model [21] aim to construct a tree which is close to the true tree in terms of the leaf distribution in the sense of Kearns *et al.* [19] but which need not be the same as the true tree. Farach and Kannan [12] gave an  $\mathcal{O}(n^2 \ell)$ -time algorithm (FK) for the symmetric case of the 2-state model provided that all pairs of leaves have a sufficiently high probability of being the same. Ambainis, Desper, Farach, and Kannan [2] gave a nearly tight lower bound on  $\ell$  for achieving a given variational distance between the true tree and the reconstructed tree. As for obtaining the true tree, the best known upper bound on  $\ell$  required by FK is exponential in  $n$ . CGG [4] also improves upon FK to PAC-learn in the general 2-state model without the symmetry and leaf similarity constraints.

The remainder of the paper is organized as follows. Section 2 reviews the generalized Jukes-Cantor model of sequence evolution and discusses distance-based probabilistic techniques. Section 3 gives Fast-HGT. Section 4 concludes the paper with some directions for further research.

**2. Model and techniques.** Section 2.1 defines the model of evolution used in the paper. Section 2.2 defines our problem of recovering evolutionary trees from biological sequences. Sections 2.3 through 2.5 develop basic techniques for the problem.

**2.1. A model of sequence evolution.** This paper employs the generalized Jukes-Cantor model [22] of sequence evolution defined as follows. Let  $m \geq 2$  and  $n \geq 3$  be two integers. Let  $\mathcal{A} = \{a_1, \dots, a_m\}$  be a finite alphabet. An *evolutionary tree*  $T$  for  $\mathcal{A}$  is a rooted binary tree of  $n$  leaves with an *edge mutation probability*  $p_e$  for each tree edge  $e$ . The edge mutation probabilities are bounded away from 0 and  $1 - \frac{1}{m}$ , i.e., there exist  $f$  and  $g$  such that for every edge  $e$  of  $T$ ,  $0 < f \leq p_e \leq g < 1 - \frac{1}{m}$ . Given a sequence  $s_1 \dots s_\ell \in \mathcal{A}^\ell$  associated with the root of  $T$ , a set of  $n$  *mutated sequences* in  $\mathcal{A}^\ell$  is generated by  $\ell$  random labelings of the tree at the nodes. These  $\ell$  labelings are mutually independent. The labelings at the  $j$ -th leaf give the  $j$ -th *mutated sequence*  $s_1^{(j)} \dots s_\ell^{(j)}$ , where the  $i$ -th labeling of the tree gives the  $i$ -th symbols  $s_i^{(1)}, \dots, s_i^{(n)}$ . The  $i$ -th labeling is carried out from the root towards the leaves along the edges. The root is labeled by  $s_i$ . On edge  $e$ , the child's label is the same as the parent's with

probability  $1 - p_e$  or is different with probability  $\frac{p_e}{m-1}$  for each different symbol. Such *mutations* of symbols along the edges are mutually independent.

**2.2. Problem formulation.** The *topology*  $\Psi(T)$  of  $T$  is the unrooted tree obtained from  $T$  by omitting the edge mutation probability and by replacing the two edges  $e_1$  and  $e_2$  between the root and its children with a single edge  $e_0$ . Note that the leaves of  $\Psi(T)$  are labeled with the same sequences as in  $T$ , but  $\Psi(T)$  need not be labeled otherwise. The *weighted topology*  $\Psi_w(T)$  of  $T$  is  $\Psi(T)$  where each edge  $e \neq e_0$  of  $\Psi(T)$  is further weighted by its edge mutation probability  $p_e$  in  $T$  and for technical reasons, the edge  $e_0$  is weighted by  $1 - (1 - p_{e_1})(1 - p_{e_2})$ .

For technical convenience, the weight of each edge  $XY$  in  $\Psi_w(T)$  is often replaced by a certain edge length, such as  $\Delta_{XY}$  in Equation (2.5), from which the weight of  $XY$  can be efficiently determined.

The *weighted evolutionary topology* problem is that of taking  $n$  mutated sequences as input and recovering  $\Psi_w(T)$  with high accuracy and high probability. Fast-HGT is a learning algorithm for this problem.

*Remark.* The special treatment for  $e_1$  and  $e_2$  is due to the fact that the root sequence may be entirely arbitrary and thus, in general, no algorithm can place the root accurately. This is consistent with the fact that the root sequence is not directly observable in practice, and locating the root requires considerations beyond those of general modeling [22]. If the root sequence is also given as input, Fast-HGT can be modified to locate the root and the weights of  $e_1$  and  $e_2$  in a straightforward manner.

**2.3. Probabilistic closeness.** Fast-HGT is based on a notion of probabilistic closeness between nodes. For the  $i$ -th random labeling of  $T$ , we identify each node of  $T$  with the random variable  $X_i$  that gives the labeling at the node. Note that since  $s_1 \cdots s_\ell$  may be arbitrary, the random variables  $X_i$  for different  $i$  are not necessarily identically distributed. For brevity, we often omit the index  $i$  of  $X_i$  in a statement if the statement is independent of  $i$ .

For nodes  $X$  and  $Y \in T$ , let  $p_{XY} = \Pr\{X \neq Y\}$ . The *closeness* of  $X$  and  $Y$  is

$$(2.1) \quad \sigma_{XY} = \Pr\{X = Y\} - \frac{1}{m-1} \Pr\{X \neq Y\} = 1 - \alpha p_{XY}, \text{ where } \alpha = \frac{m}{m-1}.$$

LEMMA 2.1 (folklore). *If node  $Y$  is on the path between two nodes  $X$  and  $Z$  in  $T$ , then  $\sigma_{XZ} = \sigma_{XY}\sigma_{YZ}$ .*

If  $X$  and  $Y$  are leaves, their closeness is estimated from sample sequences as

$$(2.2) \quad \hat{\sigma}_{XY} = \frac{1}{\ell} \sum_{i=1}^{\ell} I_{\hat{X}_i \hat{Y}_i},$$

where  $\hat{X}_1, \dots, \hat{X}_\ell$  and  $\hat{Y}_1, \dots, \hat{Y}_\ell$  are the symbols at positions  $1, \dots, \ell$  of the observed sample sequences for the two leaves, and

$$I_{xy} = \begin{cases} \frac{-1}{m-1} & \text{if } x \neq y; \\ 1 & \text{if } x = y. \end{cases}$$

The next lemma is useful for analyzing the estimation given by Equation (2.2).

LEMMA 2.2. *For  $\epsilon > 0$ ,*

$$(2.3) \quad \Pr\left\{ \frac{\hat{\sigma}_{XY}}{\sigma_{XY}} \leq 1 - \epsilon \right\} \leq \exp\left(-\frac{2}{\alpha^2} \ell \sigma_{XY}^2 \epsilon^2\right);$$

$$(2.4) \quad \Pr\left\{ \frac{\hat{\sigma}_{XY}}{\sigma_{XY}} \geq 1 + \epsilon \right\} \leq \exp\left(-\frac{2}{\alpha^2} \ell \sigma_{XY}^2 \epsilon^2\right).$$

*Proof.* By Equation (2.2),

$$\Pr\left\{\frac{\hat{\sigma}_{XY}}{\sigma_{XY}} \leq 1 - \epsilon\right\} = \Pr\left\{\sum_{i=1}^{\ell} (I_{X_i Y_i} - \sigma_{XY}) \leq -\ell\sigma_{XY}\epsilon\right\};$$

$$\Pr\left\{\frac{\hat{\sigma}_{XY}}{\sigma_{XY}} \geq 1 + \epsilon\right\} = \Pr\left\{\sum_{i=1}^{\ell} (I_{X_i Y_i} - \sigma_{XY}) \geq \ell\sigma_{XY}\epsilon\right\}.$$

Since  $\frac{-1}{m-1} \leq I_{X_i Y_i} \leq 1$  and  $\mathbb{E}[I_{X_i Y_i} - \sigma_{XY}] = 0$ , we use Hoeffding's inequality [17] on sums of independent bounded random variables to have Equations (2.3) and (2.4).  $\square$

**2.4. Distance and harmonic mean.** The *distance* of nodes  $X$  and  $Y \in T$  is

$$(2.5) \quad \Delta_{XY} = -\ln \sigma_{XY}.$$

For an edge  $XY$  in  $T$ ,  $\Delta_{XY}$  is called the *edge length* of  $XY$ .

Fast-HGT uses Statement 3 of the next corollary to locate internal nodes of  $T$ .

**COROLLARY 2.3.** *Let  $X, Y$ , and  $Z$  be nodes in  $T$ .*

1. *If  $X \neq Y$ , then  $\Delta_{XY} = \Delta_{YX} > 0$ . Also,  $\Delta_{XX} = 0$ .*
2. *If  $Y$  is on the path between  $X$  and  $Z$  in  $T$ , then  $\Delta_{XZ} = \Delta_{XY} + \Delta_{YZ}$ .*
3. *For any  $\sigma$  with  $\sigma_{XY} \leq \sigma < 1$ , there is a node  $P$  on the path between  $X$  and  $Y$  in  $T$  such that  $\sigma(1 - \alpha g)^{1/2} \leq \sigma_{XP} \leq \sigma(1 - \alpha g)^{-1/2}$ . Furthermore, if  $\sigma_{XY}(1 - \alpha g)^{1/2} < \sigma < (1 - \alpha g)^{-1/2}$ , then  $P$  is distinct from  $X$  and  $Y$ .*

*Proof.* Statements 1 and 2 follow from Equation (2.1) and Lemma 2.1. Statement 3 becomes straightforward when restated in terms of distance as follows. For any  $\Delta$  with  $\Delta_{XY} \geq \Delta > 0$ , there is a node  $P$  on the path between  $X$  and  $Y$  in  $T$  such that  $\Delta + \frac{-\ln(1-\alpha g)}{2} \geq \Delta_{XP} \geq \Delta - \frac{-\ln(1-\alpha g)}{2}$ . Furthermore, if  $\Delta_{XY} - \frac{-\ln(1-\alpha g)}{2} > \Delta > \frac{-\ln(1-\alpha g)}{2}$ , then  $P$  is distinct from  $X$  and  $Y$ .  $\square$

If  $X$  and  $Y$  are leaves, their distance is estimated from sample sequences as

$$(2.6) \quad \hat{\Delta}_{XY} = \begin{cases} -\ln \hat{\sigma}_{XY} & \text{if } \hat{\sigma}_{XY} > 0; \\ \infty & \text{otherwise.} \end{cases}$$

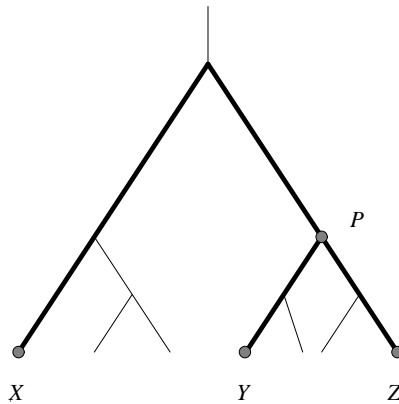


FIG. 2.1.  $P$  is the center of triplet  $XYZ$ .

A *triplet*  $XYZ$  consists of three distinct leaves  $X, Y$ , and  $Z$  of  $T$ . There is an internal node  $P$  in  $T$  at which the pairwise paths between the leaves in  $XYZ$  intersect;

see Figure 2.1.  $P$  is the *center* of  $XYZ$ , and  $XYZ$  *defines*  $P$ . Note that a star is formed by the edges on the paths between  $P$  and the three leaves in  $XYZ$ .

By Corollary 2.3(2), the distance between  $P$  and a leaf in  $XYZ$ , say,  $X$ , can be obtained as  $\Delta_{XP} = \frac{\Delta_{XY} + \Delta_{XZ} - \Delta_{YZ}}{2}$ , which is estimated by

$$(2.7) \quad \hat{\Delta}_{XP} = \frac{\hat{\Delta}_{XY} + \hat{\Delta}_{XZ} - \hat{\Delta}_{YZ}}{2}.$$

The *closeness* of  $XYZ$  is  $\sigma_{XYZ} = \frac{3}{\frac{1}{\sigma_{XY}} + \frac{1}{\sigma_{XZ}} + \frac{1}{\sigma_{YZ}}}$ , which is estimated by  $\hat{\sigma}_{XYZ} = \frac{3}{\frac{1}{\hat{\sigma}_{XY}} + \frac{1}{\hat{\sigma}_{XZ}} + \frac{1}{\hat{\sigma}_{YZ}}}$ .  $XYZ$  is called *positive* if  $\hat{\sigma}_{XY}$ ,  $\hat{\sigma}_{XZ}$ , and  $\hat{\sigma}_{YZ}$  are all positive.

The next corollary relates  $\sigma_{XYZ}$  and the pairwise closenesses of  $X$ ,  $Y$ , and  $Z$ .

**COROLLARY 2.4.** *If  $\sigma_{XP} \leq \sigma_{YP} \leq \sigma_{ZP}$ , then  $\sigma_{XY} \leq \sigma_{XZ} \leq \sigma_{YZ}$ ,  $\sigma_{XZ} \geq \frac{2}{3}\sigma_{XYZ}$ , and  $\sigma_{YP}^2 \geq \frac{1}{3}\sigma_{XYZ}$ .*

*Proof.* This corollary follows from Lemma 2.1 and simple algebra.  $\square$

The next lemma relates  $\sigma_{XYZ}$  to the probability of overestimating the distance between  $P$  and a leaf in  $XYZ$  using Equation (2.7).

**LEMMA 2.5.** *For  $0 < \epsilon < 1$ ,*

$$\Pr \left\{ \hat{\Delta}_{XP} - \Delta_{XP} \geq \frac{-\ln(1-\epsilon)}{2} \right\} \leq 3 \exp \left( -\frac{2}{9\alpha^2} \ell \sigma_{XYZ}^2 \epsilon^2 \right).$$

*Proof.* See §A.1.  $\square$

**2.5. Basis of a greedy strategy.** Let  $d_{XY}$  denote the number of edges in the path between two leaves  $X$  and  $Y$  in  $T$ . By Lemma 2.1,  $\sigma_{XY}$  can be as small as  $(1-\alpha g)^{d_{XY}}$ . Thus, the larger  $d_{XY}$  is, the more difficult it is to estimate  $\sigma_{XY}$  and  $\Delta_{XY}$ . This intuition leads to a natural greedy strategy outlined below that favors leaf pairs with small  $d_{XY}$  and large  $\sigma_{XY}$ .

The *g-depth* of a node in a rooted tree  $T'$  is the smallest number of edges in a path from the node to a leaf. Let  $e$  be an edge between nodes  $u_1$  and  $u_2$ . Let  $T'_1$  and  $T'_2$  be the subtrees of  $T'$  obtained by cutting  $e$  which contain  $u_1$  and  $u_2$ , respectively. The *g-depth* of  $e$  in  $T'$  is the larger of the g-depth of  $u_1$  in  $T'_1$  and that of  $u_2$  in  $T'_2$ . The *g-depth* of a rooted tree is the largest possible g-depth of an edge in the tree. (The prefix *g* emphasizes that this usage of depth is nonstandard in graph theory.)

Let  $d$  be the *g-depth* of  $T$ . Variants of the next lemma have proven very useful and insightful; see, e.g., [9–11].

**LEMMA 2.6.**

1.  $d \leq 1 + \lfloor \log_2(n-1) \rfloor$ .
2. *Every internal node  $P$  of  $T$  except the root has a defining triplet  $XYZ$  such that  $d_{XP}, d_{YP}$ , and  $d_{ZP}$  are all at most  $d+1$  and thus,  $\sigma_{XYZ} \geq (1-\alpha g)^{2(d+1)}$ . Every leaf of  $T$  is in such a triplet.*

*Proof.* The proof is straightforward. Note that the more unbalanced  $T$  is, the smaller its *g-depth* is.  $\square$

In  $T$ , the star formed by a defining triplet of an internal node contains the three edges incident to the internal node. Thus,  $\Psi(T)$  can be reconstructed from triplets described in Lemma 2.6(2) or those with similarly large closenesses. This observation motivates the following definitions. Let

$$\sigma_{\text{lg}} = \frac{3\sqrt{2}}{2} \left( \frac{\sqrt{2}-1}{\sqrt{2}+1} \right)^2 (1-\alpha g)^{2d+4}; \quad \sigma_{\text{sm}} = \frac{\sigma_{\text{lg}}}{\sqrt{2}}; \quad \sigma_{\text{md}} = \frac{\sigma_{\text{lg}} + \sigma_{\text{sm}}}{2}.$$

*Remark.* The choice of  $\sigma_{\text{lg}}$  is obtained by solving Equations (3.2), (3.3) and (3.4).

A triplet  $XYZ$  is *large* if  $\sigma_{XYZ} \geq \sigma_{\text{lg}}$ ; it is *small* if  $\sigma_{XYZ} \leq \sigma_{\text{sm}}$ . Note that by Lemma 2.6(2), each nonroot internal node of  $T$  has at least one large defining triplet.

LEMMA 2.7. *The first inequality below holds for all large triplets  $XYZ$ , and the second for all small triplets.*

$$(2.8) \quad \Pr\{\hat{\sigma}_{XYZ} \leq \sigma_{\text{md}}\} \leq \exp\left(-\frac{(\sqrt{2}-1)^2}{36\alpha^2}\ell\sigma_{\text{lg}}^2\right);$$

$$(2.9) \quad \Pr\{\hat{\sigma}_{XYZ} \geq \sigma_{\text{md}}\} \leq \exp\left(-\frac{(\sqrt{2}-1)^2}{36\alpha^2}\ell\sigma_{\text{lg}}^2\right).$$

*Proof.* See §A.2.  $\square$

A nonroot internal node of  $T$  may have more than one large defining triplet. Consequently, since distance estimates contain errors, we may obtain an erroneous estimate of  $\Psi(T)$  by reconstructing the same internal node more than once from its different large defining triplets. To address this issue, Fast-HGT adopts a threshold  $0 < \Delta_{\min} < \frac{-\ln(1-\alpha f)}{2}$  based on the fact that the distance between two distinct nodes is at least  $-\ln(1-\alpha f)$ ; also let  $c = \frac{\Delta_{\min}}{-\ln(1-\alpha f)}$ . Fast-HGT considers the center  $P$  of a triplet  $XYZ$  and the center  $Q$  of another triplet  $XUV$  to be separate if and only if

$$(2.10) \quad |\hat{\Delta}_{XP} - \hat{\Delta}_{XQ}| \geq \Delta_{\min},$$

where  $\hat{\Delta}_{XP} = (\hat{\Delta}_{XY} + \hat{\Delta}_{XZ} - \hat{\Delta}_{YZ})/2$  and  $\hat{\Delta}_{XQ} = (\hat{\Delta}_{XU} + \hat{\Delta}_{XV} - \hat{\Delta}_{UV})/2$ . Notice that two triplet centers can be compared in this manner only if the triplets share at least one leaf. The next lemma shows that a large triplet's center is estimated within a small error with high probability.

LEMMA 2.8. *Let  $P$  be the center of a triplet  $XYZ$ . If  $XYZ$  is not small, then*

$$(2.11) \quad \Pr\left\{\left|\hat{\Delta}_{XP} - \Delta_{XP}\right| \geq \frac{\Delta_{\min}}{2}\right\} \leq 7 \exp\left(-\frac{c^2}{81}\ell\sigma_{\text{lg}}^2 f^2\right).$$

*Proof.* See §A.3.  $\square$

We next define and analyze two key events  $\mathcal{E}_c$  and  $\mathcal{E}_g$  as follows. The subscripts  $c$  and  $g$  denote the words greedy and center, respectively.

- $\mathcal{E}_c$  is the event that for every triplet  $XYZ$  that is not small,  $|\hat{\Delta}_{XP} - \Delta_{XP}| < \frac{\Delta_{\min}}{2}$ ,  $|\hat{\Delta}_{YP} - \Delta_{YP}| < \frac{\Delta_{\min}}{2}$ , and  $|\hat{\Delta}_{ZP} - \Delta_{ZP}| < \frac{\Delta_{\min}}{2}$ , where  $P$  is the center of  $XYZ$ .
- $\mathcal{E}_g$  is the event that  $\hat{\sigma}_{XYZ} > \hat{\sigma}_{X'Y'Z'}$  for every large triplet  $XYZ$  and every small triplet  $X'Y'Z'$ .

LEMMA 2.9.

$$\Pr\{\bar{\mathcal{E}}_c\} \leq 21 \binom{n}{3} \exp\left(-\frac{c^2}{81}\ell\sigma_{\text{lg}}^2 f^2\right); \quad \Pr\{\bar{\mathcal{E}}_g\} \leq \binom{n}{3} \exp\left(-\frac{(\sqrt{2}-1)^2}{36\alpha^2}\ell\sigma_{\text{lg}}^2\right).$$

*Proof.* The inequalities follow from Equation (2.11) and Lemma 2.7, respectively.

$\square$

**Algorithm** Fast Harmonic Greedy Triplets**Input:**

- $\Delta_{\min}$ ;
- $\hat{\Delta}_{XY}$  for all leaves  $X$  and  $Y$  of  $T$  which are computed via Equations (2.2) and (2.6) from  $n$  mutated length- $\ell$  sequences generated by  $T$ .

**Output:**  $\Psi_w(T)$ .

- F1 Select an arbitrary leaf  $A$  and find a triplet  $ABC$  with the maximum  $\hat{\sigma}_{ABC}$ .  
 F2 **if**  $ABC$  is not positive **then** let  $T^*$  be the empty tree, **fail**, and **stop**.  
 F3 Let  $T^*$  be the star with three edges formed by  $ABC$  and its center  $D$ .  
 F4 Use Equation (2.7) to set  $\Delta_{AD}^* \leftarrow \hat{\Delta}_{AD}$ ,  $\Delta_{BD}^* \leftarrow \hat{\Delta}_{BD}$ ,  $\Delta_{CD}^* \leftarrow \hat{\Delta}_{CD}$ .  
 F5 Set  $\text{def}(D) \leftarrow \{A, B, C\}$ .  
 F6 First set all  $\mathcal{S}[M]$  to null; then for  $Q_1Q_2 \in \{AD, BD, CD\}$ , Update- $\mathcal{S}(Q_1Q_2)$ .  
 F7 **repeat**  
 F8   **if**  $\mathcal{S}[M] = \text{null}$  for all leaves  $M \in T$  **then fail** and **stop**.  
 F9   Find  $\mathcal{S}[N] = \langle P_1P_2, NXY, P, \Delta_{P_1P}^*, \Delta_{P_2P}^*, \Delta_{NP}^* \rangle$  with the maximum  $\hat{\sigma}_{NXY}$ .  
 F10   Split  $P_1P_2$  into two edges  $P_1P$  and  $P_2P$  in  $T^*$  with lengths  $\Delta_{P_1P}^*$  and  $\Delta_{P_2P}^*$ .  
 F11   Add to  $T^*$  a leaf  $N$  and an edge  $NP$  with length  $\Delta_{NP}^*$ .  
 F12   Set  $\text{def}(P) \leftarrow \{N, X, Y\}$ .  
 F13   For every  $M$  with  $\mathcal{S}[M]$  containing the edge  $P_1P_2$ , set  $\mathcal{S}[M] \leftarrow \text{null}$ .  
 F14   For each  $Q_1Q_2 \in \{P_1P, P_2P, NP\}$ , Update- $\mathcal{S}(Q_1Q_2)$ .  
 F15 **until** all leaves of  $T$  are inserted to  $T^*$ ; i.e., this loop has iterated  $n - 3$  times.  
 F16 Output  $T^*$ .

FIG. 3.1. *The Fast-HGT algorithm.***Algorithm** Update- $\mathcal{S}$ **Input:** an edge  $Q_1Q_2 \in T^*$ 

- U1 Find all splitting tuples for  $Q_1Q_2 \in T^*$ .  
 U2 For each  $\langle Q_1Q_2, MUV, Q, \Delta_{Q_1Q}^*, \Delta_{Q_2Q}^*, \Delta_{MQ}^* \rangle$  at line U1, assign it to  $\mathcal{S}[M]$  if  $\hat{\sigma}_{MUV}$  is greater than that of  $\mathcal{S}[M]$ .

FIG. 3.2. *The Update- $\mathcal{S}$  subroutine.*

**3. Fast-HGT.** Section 3.1 details Fast-HGT. Section 3.2 analyzes its running time and work space. Section 3.3 proves technical lemmas for bounding the algorithm's required sample size. Section 3.4 analyzes this sample size.

**3.1. The description of Fast-HGT.** Fast-HGT and its subroutines Update- $\mathcal{S}$  and Split-Edge are detailed in Figures 3.1, 3.2, and 3.3, respectively.

Given  $\Delta_{\min}$  and  $n$  mutated sequences as input, the task of Fast-HGT is to recover  $\Psi_w(T)$ . The algorithm first constructs a star  $T^*$  formed by a large triplet at lines F1 through F3. It then inserts into  $T^*$  a leaf of  $T$  and a corresponding internal node per iteration of the repeat at line F7 until  $T^*$  has a leaf for each input sequence. The  $T^*$  at line F16 is our reconstruction of  $\Psi_w(T)$ . For  $k = 3, \dots, n$ , let  $T_k^*$  be the version of  $T^*$  with  $k$  leaves constructed during a run of Fast-HGT; i.e.,  $T_3^*$  is constructed at line F3, and  $T_k^*$  with  $k \geq 4$  is constructed at line F11 during the  $(k - 3)$ -th iteration of the repeat. Note that  $T_n^*$  is output at line F16.

A node  $Q$  is *strictly between* nodes  $Q_1$  and  $Q_2$  in  $T$  if  $Q$  is on the path between  $Q_1$  and  $Q_2$  in  $T$  but  $Q \neq Q_1$ ,  $Q \neq Q_2$ , and  $Q$  is not the root of  $T$ . At each iteration of the repeat, Fast-HGT finds an edge  $P_1P_2$  in  $T^*$  and a triplet  $NXY$  where  $X, Y \in T^*$ ,  $N \notin T^*$ , and the center  $P$  of  $NXY$  is strictly between on  $P_1$  and  $P_2$  in  $T^*$ . Such  $P_1P_2$

**Algorithm Split-Edge****Input:** an edge  $P_1P_2$  in  $T^*$  and a relevant triplet  $NXY$  with center  $P$ .**Output:** If  $P$  is strictly between  $P_1$  and  $P_2$  in  $T$  and thus can be inserted on  $P_1P_2$ , then we return the message “split” and the edge lengths  $\Delta_{P_1P}^*$ ,  $\Delta_{P_2P}^*$ , and  $\Delta_{NP}^*$ . Otherwise, we return a reason why  $P$  cannot be inserted.

S1 Use Equation (2.7) to compute  $\hat{\Delta}_{XP}$ ,  $\hat{\Delta}_{YP}$ ,  $\hat{\Delta}_{NP}$  for  $NXY$ .  
S2 Let  $X_1 \in \{X, Y\} \cap \text{def}(P_1)$  and  $X_2 \in \{X, Y\} \cap \text{def}(P_2)$ .  
S3 For each  $i = 1$  or  $2$ , **if**  $P_i$  is an internal node of  $T^*$   
S4     **then** use Equation (2.7) to compute  $\hat{\Delta}_{X_iP_i}$  for the triplet formed by  $\text{def}(P_i)$   
S5     **else** set  $\hat{\Delta}_{X_iP_i} \leftarrow 0$ .  
S6 Set  $\Delta_1 \leftarrow \hat{\Delta}_{X_1P} - \hat{\Delta}_{X_1P_1}$  and  $\Delta_2 \leftarrow \hat{\Delta}_{X_2P} - \hat{\Delta}_{X_2P_2}$ .  
S7 **if**  $|\Delta_1| < \Delta_{\min}$  or  $|\Delta_2| < \Delta_{\min}$   
S8     **then return** “too close”  
S9     **else begin**  
S10       **if**  $P_2$  (respectively,  $P_1$ ) is on the path between  $P_1$  and  $X_1$  ( $P_2$  and  $X_2$ ) in  $T^*$   
S11         **then** set  $\Delta'_1 \leftarrow -\Delta_1$  ( $\Delta'_2 \leftarrow -\Delta_2$ )  
S12         **else** set  $\Delta'_1 \leftarrow \Delta_1$  ( $\Delta'_2 \leftarrow \Delta_2$ ).  
(Remark. Since  $X_1$  may equal  $X_2$ , the tests for  $P_1$  and  $P_2$  are both needed.)  
S13     Set  $\Delta''_1 \leftarrow (\Delta'_1 + \Delta_{P_1P_2}^* - \Delta'_2)/2$  and  $\Delta''_2 \leftarrow (\Delta'_2 + \Delta_{P_1P_2}^* - \Delta'_1)/2$ .  
(Remark.  $\Delta''_1 + \Delta''_2 = \Delta_{P_1P_2}^*$ ,  $\Delta''_1$  estimates  $\Delta_{P_1P}$ , and  $\Delta''_2$  estimates  $\Delta_{P_2P}$ .)  
S14     **if**  $\Delta''_1 \geq \Delta_{P_1P_2}^*$  or  $\Delta''_2 \geq \Delta_{P_1P_2}^*$   
S15         **then return** “outside this edge”  
S16         **else return** “split”,  $\Delta''_1$ ,  $\Delta''_2$ ,  $\Delta_{NP}$ .  
S17     **end.**

FIG. 3.3. The Split-Edge subroutine.

and  $NXY$  can be used to insert  $N$  and  $P$  into  $T^*$ . We record an insertion by letting  $\text{def}(P) = \{N, X, Y\}$ ; for notational uniformity, let  $\text{def}(X) = \{X\}$  for all leaves  $X$ .

At line F6,  $\mathcal{S}$  is an array indexed by the leaves  $M$  of  $T^*$ . At the beginning of each iteration of the repeat,  $\mathcal{S}[N]$  stores the most suitable  $P_1P_2$  and  $NXY$  for inserting  $N$  into  $T^*$ .  $\mathcal{S}$  is initialized at line F6; it is updated at lines F13 and F14 after a new leaf and a new internal node are inserted into  $T^*$ . The precise content of  $\mathcal{S}$  is described in Lemma 3.6.

To further specify  $\mathcal{S}[N]$ , we call  $NXY$  *relevant* for  $P_1P_2 \in T_k^*$  if it is positive,  $N \notin T_k^*$ ,  $X \in \text{def}(P_1)$ ,  $Y \in \text{def}(P_2)$ , and  $P_1P_2$  is on the path between  $X$  and  $Y$  in  $T_k^*$ . We use Split-Edge to determine whether the center  $P$  of a relevant  $NXY$  is strictly between  $P_1$  and  $P_2$  in  $T$ . We also use Split-Edge to calculate an estimation  $\Delta_{P'P''}^*$  of  $\Delta_{P'P''}$  for each edge  $P'P'' \in T_k^*$ , which is called the *length* of  $P'P''$  in  $T_k^*$ . Split-Edge has three possible outcomes:

1. At line S8,  $P$  is too close to  $P_1$  or  $P_2$  to be a different internal node.
2. At line S15,  $P$  is outside the path between  $P_1$  and  $P_2$  in  $T$  and thus should not be inserted into  $T_k^*$  on  $P_1P_2$ .
3. At line S16,  $P$  is strictly between  $P_1$  and  $P_2$  in  $T$ . Thus,  $P$  can be inserted between  $P_1$  and  $P_2$  in  $T_k^*$ , and the lengths  $\Delta_{P_1P}^*$ ,  $\Delta_{P_2P}^*$ ,  $\Delta_{NP}^*$  of the possible new edges  $P_1P$ ,  $P_2P$ , and  $NP$  are returned.

In the case of the third outcome,  $NXY$  is called a *splitting triplet* for  $P_1P_2$  in  $T_k^*$ , and  $\langle P_1P_2, NXY, P, \Delta_{P_1P}^*, \Delta_{P_2P}^*, \Delta_{NP}^* \rangle$  is a *splitting tuple*. Each  $\mathcal{S}[N]$  is either a single splitting tuple or null. In the latter case, the estimated closeness of the triplet in  $\mathcal{S}[N]$



is regarded as 0 for technical uniformity.

Fast-HGT ensures the accuracy of  $T^*$  in several ways. The algorithm uses only positive triplets to recover internal nodes of  $T$  at lines F1 and F9. These two lines together form the greedy strategy of Fast-HGT. The maximality of the triplet chosen at these two lines favors large triplets over small ones based on Lemmas 2.6 and 2.7. With a relevant triplet as input, Split-Edge compares  $P$  to  $P_1$  and  $P_2$  using the rule of Equation (2.10) and can estimate the distance between  $P$  and  $P_1$  or  $P_2$  from the same leaf to avoid accumulating estimation errors in edge lengths.

The next lemma enables Fast-HGT to grow  $T^*$  by always using relevant triplets.

**LEMMA 3.1.** *For each  $k = 3, \dots, n-1$ , at the start of the  $(k-2)$ -th iteration of the repeat at line F7,  $\text{def}(P_1) \cap \text{def}(P_2) \neq \emptyset$  for every edge  $P_1P_2 \in T_k^*$ .*

*Proof.* The proof is by induction on  $k$ . The base case follows from the fact that the statement holds for  $T_3^*$  at line F3. The induction step follows from the use of a relevant triplet at line F9.  $\square$

*Remark.* A subsequent work [6] shows that Fast-HGT can run with the same time, space, and sample complexities without knowing  $f$  and  $\Delta_{\min}$ ; this is achieved by slightly modifying some parts of Split-Edge.

**3.2. The running time and work space of Fast-HGT.** Before proving the desired time and space complexities of Fast-HGT in Theorem 3.2 below, we note the following three key techniques used by Fast-HGT to save time and space.

1. At line F1,  $ABC$  is selected for a fixed arbitrary  $A$ . This limits the number of triplets considered at line F1 to  $\mathcal{O}(n^2)$ . This technique is supported by the fact that each leaf in  $T$  is contained in a large triplet.
2. At lines F6 and F14,  $\mathcal{S}$  keeps only splitting tuples. This limits the number of triplets considered for each involved edge to  $\mathcal{O}(n)$ . This technique is feasible since by Lemma 3.5,  $\Psi(T)$  can be recovered using only relevant triplets.
3. At line F14,  $\mathcal{S}$  includes no new splitting tuples for the edges  $Q_1Q_2$  that already exist in  $T^*$  before  $N$  is inserted. This technique is feasible because the insertion of  $N$  results in no new relevant triplets for such  $Q_1Q_2$  at all.

**THEOREM 3.2.** *Fast-HGT runs in  $\mathcal{O}(n^2)$  time using  $\mathcal{O}(n)$  work space.*

*Proof.* We analyze the time and space complexities separately as follows.

*Time complexity.* Line F1 takes  $\mathcal{O}(n^2)$  time. Line F6 takes  $\mathcal{O}(n)$  total time to examine  $2(n-3)$  triplets for each  $Q_1Q_2$ . As for the repeat at line F7, lines F8, F9, and F13 take  $\mathcal{O}(n)$  time to search through  $\mathcal{S}$ . For the  $(k-3)$ -th iteration of the repeat where  $k = 4, \dots, n-1$ , line F14 takes  $\mathcal{O}(n)$  total time to examine at most  $9(n-k-1)$  triplets for each of  $P_1P, P_2P$  and  $NP$ . Thus, each iteration of the repeat takes  $\mathcal{O}(n)$  time. Since the repeat iterates at most  $n-3$  times, the time complexity of Fast-HGT is as stated.

*Space complexity.*  $T^*$  and the sets  $\text{def}(G)$  for all nodes  $G$  in  $T^*$  take  $\mathcal{O}(n)$  work space.  $\mathcal{S}$  takes  $\mathcal{O}(n)$  space. Lines F1, F6 and F14 in Fast-HGT and lines U1 and U2 in Update- $\mathcal{S}$  can be implemented to use  $\mathcal{O}(1)$  space. The other variables needed by Fast-HGT take  $\mathcal{O}(1)$  space. Thus, the space complexity of Fast-HGT is as stated.  $\square$

**3.3. Technical lemmas for bounding the sample size.** Let  $L_k$  be the set of the leaves of  $\Psi(T)$  that are in  $T_k^*$ . Let  $\Psi_k$  be the subtree of  $\Psi(T)$  formed by the edges on paths between leaves in  $L_k$ . A *branchless* path in  $\Psi_k$  is one whose internal nodes are all of degree 2 in  $\Psi_k$ . We say that  $T_k^*$  *matches*  $T$  if  $T_k^*$  without the edge lengths can be obtained from  $\Psi_k$  by replacing every maximal branchless path with an edge between its two endpoints.

For  $k = 3, \dots, n$ , we define the following conditions:

- $\mathcal{A}_k$ :  $T_k^*$  matches  $T$ .
- $\mathcal{B}_k$ : For every internal node  $Q \in T_k^*$ , the triplet formed by  $\text{def}(Q)$  is not small.
- $\mathcal{C}_k$ : For every edge  $Q_1Q_2 \in T_k^*$ ,  $|\Delta_{Q_1Q_2}^* - \Delta_{Q_1Q_2}| < 2\Delta_{\min}$ .

In this section, Lemmas 3.3, 3.4, and 3.5 analyze under what conditions Split-Edge can help correctly insert a new leaf and a new internal node to  $T_k^*$ . Later in §3.4, we use these lemmas to show by induction in Lemma 3.7 that the events  $\mathcal{E}_g$  and  $\mathcal{E}_c$ , which are defined before Lemma 2.9, imply that  $\mathcal{A}_k$ ,  $\mathcal{B}_k$ , and  $\mathcal{C}_k$  hold for all  $k$ . This leads to Theorem 3.8, stating that Fast-HGT solves the weighted evolutionary topology problem with a polynomial-sized sample.

Lemmas 3.3, 3.4, and 3.5 make the following assumptions for some  $k < n$ :

- The  $(k-3)$ -th iteration of the repeat at line F7 has been completed.
- $T_k^*$  has been constructed, and  $\mathcal{A}_k$ ,  $\mathcal{B}_k$ , and  $\mathcal{C}_k$  hold.
- Fast-HGT is currently in the  $(k-2)$ -th iteration of the repeat.

LEMMA 3.3. *Assume that  $\mathcal{E}_c$  holds and the triplet  $NXY$  input to Split-Edge is not small. Then, the test of line S7 fails if and only if  $P \neq P_1$  and  $P \neq P_2$  in  $T$ .*

*Proof.* There are two directions, both using the following equation. By line S6,

$$(3.1) \quad \Delta_1 = (\hat{\Delta}_{X_1P} - \Delta_{X_1P}) - (\hat{\Delta}_{X_1P_1} - \Delta_{X_1P_1}) + (\Delta_{X_1P} - \Delta_{X_1P_1}).$$

( $\implies$ ) To prove by contradiction, assume  $P = P_1$  or  $P = P_2$  in  $T$ . If  $P = P_1$ , then  $\Delta_{X_1P} = \Delta_{X_1P_1}$ , and by  $\mathcal{A}_k$ ,  $P_1$  is an internal node in  $T_k^*$ . By  $\mathcal{B}_k$ , the triplet formed by  $\text{def}(P_1)$  is not small. Thus, by  $\mathcal{E}_c$  and Equation (3.1),  $|\Delta_1| < \Delta_{\min}$ . By symmetry, if  $P = P_2$ , then  $|\Delta_2| < \Delta_{\min}$ . In either case, the test of line S7 passes.

( $\impliedby$ ) Since  $P \neq P_1$ ,  $\Delta_{X_1P} - \Delta_{X_1P_1} \geq -\ln(1-\alpha f) \geq 2\Delta_{\min}$ . If  $P_1$  is a leaf in  $T_k^*$ , then by  $\mathcal{A}_k$ ,  $P_1$  is leaf  $X_1$  in  $T$ , and  $\hat{\Delta}_{X_1P_1} = \Delta_{X_1P_1} = 0$ . By  $\mathcal{E}_c$  and Equation (3.1),  $|\Delta_1| > 1.5\Delta_{\min}$ . If  $P_1$  is an internal node in  $T_k^*$ , then by  $\mathcal{B}_k$ ,  $\mathcal{E}_c$ , and Equation (3.1), we have  $|\Delta_1| > \Delta_{\min}$ . In either case,  $|\Delta_1| > \Delta_{\min}$ . By symmetry, since  $P \neq P_2$ ,  $|\Delta_2| > \Delta_{\min}$ . Thus, the test of line S7 fails.  $\square$

LEMMA 3.4. *In addition to the assumption in Lemma 3.3, also assume that  $P \neq P_1$  and  $P \neq P_2$  in  $T$ , i.e., the test of line S7 has failed. Then, the test of line S14 fails if and only if  $P$  is on the path between  $P_1$  and  $P_2$  in  $T$ .*

*Proof.* There are two directions.

( $\impliedby$ ) From lines S6, S10 and Corollary 2.3(2),

$$\begin{aligned} (\Delta'_1 - \Delta'_2) - (\Delta_{P_1P} - \Delta_{P_2P}) &= \pm \left( (\hat{\Delta}_{X_1P} - \Delta_{X_1P}) - (\hat{\Delta}_{X_1P_1} - \Delta_{X_1P_1}) \right) \\ &\quad \pm \left( (\hat{\Delta}_{X_2P} - \Delta_{X_2P}) - (\hat{\Delta}_{X_2P_2} - \Delta_{X_2P_2}) \right). \end{aligned}$$

Thus, whether  $P_1$  and  $P_2$  are leaves or internal nodes in  $T_k^*$ , by  $\mathcal{A}_k$ ,  $\mathcal{B}_k$ , and  $\mathcal{E}_c$ ,  $|(\Delta'_1 - \Delta'_2) - (\Delta_{P_1P} - \Delta_{P_2P})| < 2\Delta_{\min}$ . By line S13 and Corollary 2.3(2),

$$\begin{aligned} \Delta''_1 &< \frac{2\Delta_{\min} + (\Delta_{P_1P} - \Delta_{P_2P}) + \Delta_{P_1P_2}^*}{2} \\ &= \frac{2(2\Delta_{\min} - \Delta_{P_2P}) + (-2\Delta_{\min} + \Delta_{P_1P_2}) + \Delta_{P_1P_2}^*}{2}. \end{aligned}$$

Then, since  $P \neq P_2$  and thus  $\Delta_{P_2P} \geq 2\Delta_{\min}$ , by  $\mathcal{C}_k$ , we have  $\Delta''_1 < \Delta_{P_1P_2}^*$ . By symmetry,  $\Delta''_2 < \Delta_{P_1P_2}^*$ . Thus, the test of line S14 fails.

( $\implies$ ) To prove by contradiction, assume that  $P$  is not on the path between  $P_1$  and  $P_2$ . By similar arguments, if  $\Delta_{P_1P} > \Delta_{P_1P_2}$  (respectively,  $\Delta_{P_2P} > \Delta_{P_1P_2}$ ), then  $\Delta''_1 > \Delta_{P_1P_2}^*$  (respectively,  $\Delta''_2 > \Delta_{P_1P_2}^*$ ). Thus, the test of line S14 passes.  $\square$

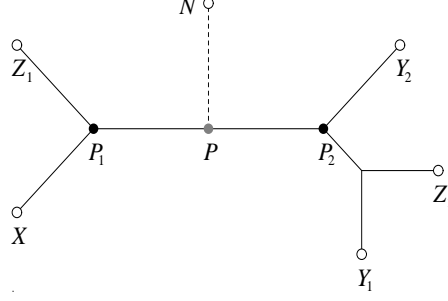


FIG. 3.4. This subgraph of  $T$  fixes some notation used in the proof of Case 1 of Lemma 3.5. The location of  $Y_1$  relative to  $Y_2$  and  $Z_2$  is nonessential; for instance,  $Y_1$  can even be the same as  $Y_2$ . In  $T_k^*$ ,  $\text{def}(P_1) = \{X, Y_1, Z_1\}$  and  $\text{def}(P_2) = \{X, Y_2, Z_2\}$ . Neither  $XY_1Z_1$  nor  $XY_2Z_2$  is small, and  $\Delta_{P_2Y_2} \leq \Delta_{P_2Z_2}$ . We aim to prove that there is a leaf  $N \notin T_k^*$  such that  $NXY_2$  or  $NZ_1Y_2$  is large and defines a node  $P$  strictly between  $P_1$  and  $P_2$  in  $T$ .

LEMMA 3.5. Assume that  $P_1P_2$  is an edge in  $T_k^*$  and some node is strictly between  $P_1$  and  $P_2$  in  $T$ . Then there is a large triplet  $NQ_1Q_2$  with center  $P$  such that  $N \notin T_k^*$ ,  $Q_1 \in \text{def}(P_1)$ ,  $Q_2 \in \text{def}(P_2)$ , and  $P$  is strictly between  $P_1$  and  $P_2$  in  $T$ .

*Proof.* By Lemma 2.6(2), for every node  $P$  strictly between  $P_1$  and  $P_2$  in  $T$ , there exists a leaf  $N \notin T_k^*$  with  $\sigma_{PN} \geq (1 - \alpha g)^{d+1}$ . To choose  $P$ , there are two cases: (1) both  $P_1$  and  $P_2$  are internal nodes in  $T_k^*$ , and (2)  $P_1$  or  $P_2$  is a leaf in  $T_k^*$ .

Case 1. By Lemma 3.1, let  $\text{def}(P_1) = \{X, Y_1, Z_1\}$  and  $\text{def}(P_2) = \{X, Y_2, Z_2\}$ . By  $\mathcal{B}_k$ , neither  $XY_2Z_2$  nor  $XY_1Z_1$  is small. To fix the notation for  $\text{def}(P_1)$  and  $\text{def}(P_2)$  with respect to their topological layout, we assume without loss of generality that Figure 3.4 or equivalently the following statements hold:

- In  $T_k^*$  and thus in  $T$  by  $\mathcal{A}_k$ ,  $P_2$  is on the paths between  $P_1$  and  $Y_2$ , between  $P_1$  and  $Z_2$ , and between  $P_1$  and  $Y_1$ , respectively.
- Similarly,  $P_1$  is on the paths between  $P_2$  and  $Z_1$  and between  $P_2$  and  $X$ .
- $\Delta_{P_2Y_2} \leq \Delta_{P_2Z_2}$ .

Both  $NXY_2$  and  $NZ_1Y_2$  define  $P$ , and the target triplet is one of these two for some suitable  $P$ . To choose  $P$ , we further divide Case 1 into three subcases.

Case 1a:  $\sigma_{XP_2} < \sigma_{Y_2P_2}(1 - \alpha g)$  and  $\sigma_{Y_2P_1} < \sigma_{XP_1}(1 - \alpha g)$ . The target triplet is  $NXY_2$ . Since  $\sigma_{XY_2} \leq \sqrt{\sigma_{XY_2}}$ , by Corollary 2.3(3), let  $P$  be a node on the path between  $X$  and  $Y_2$  in  $T$  with  $\sqrt{\sigma_{XY_2}(1 - \alpha g)} \leq \sigma_{XP} \leq \sqrt{\sigma_{XY_2}(1 - \alpha g)^{-1}}$  and thus by Lemma 2.1  $\sqrt{\sigma_{XY_2}(1 - \alpha g)} \leq \sigma_{Y_2P} \leq \sqrt{\sigma_{XY_2}(1 - \alpha g)^{-1}}$ . By the condition of Case 1a and Lemma 2.1,  $P$  is strictly between  $P_1$  and  $P_2$  in  $T$ . Also, by Corollary 2.4,  $\sigma_{XY_2} \geq \frac{2}{3}\sigma_{XY_2Z_2}$ . Thus, by Lemma 2.1, since  $XY_2Z_2$  is not small,

$$(3.2) \quad \begin{aligned} \sigma_{NXY_2} &= \frac{3}{\frac{1}{\sigma_{XP}\sigma_{PN}} + \frac{1}{\sigma_{Y_2P}\sigma_{PN}} + \frac{1}{\sigma_{XY_2}}} \\ &\geq \frac{1}{\sqrt{\frac{2}{3}\sigma_{XY_2Z_2}^{-1/2}(1 - \alpha g)^{-d-3/2} + \frac{1}{2}\sigma_{XY_2Z_2}^{-1}}} > \sigma_{\text{lg}}. \end{aligned}$$

So  $NXY_2$  is as desired for Case 1a.

Case 1b:  $\sigma_{XP_2} \geq \sigma_{Y_2P_2}(1 - \alpha g)$ . The target triplet is  $NXY_2$ . Let  $P$  be the first node after  $P_2$  on the path from  $P_2$  toward  $P_1$  in  $T$ . Then,  $\sigma_{Y_2P} \geq \sigma_{Y_2P_2}(1 - \alpha g)$ . By Corollary 2.4,  $\sigma_{Y_2P}^2 \geq \sigma_{XY_2Z_2}(1 - \alpha g)^2/3$ . Next, since  $\sigma_{XY_2} \geq \sigma_{XZ_2}$  and  $\sigma_{P_2Y_2} \geq$

$\sigma_{P_2 Z_2}$ ,

$$\sigma_{XY_2 Z_2} \leq \frac{3}{2\sigma_{XY_2}^{-1} + \sigma_{Y_2 P_2}^{-1} \sigma_{P_2 Z_2}^{-1}} \leq \frac{3}{2\sigma_{X P_2}^{-1} \sigma_{Y_2 P_2}^{-1} + \sigma_{Y_2 P_2}^{-2}} \leq \frac{3\sigma_{X P_2}^2}{2(1 - \alpha g) + (1 - \alpha g)^2}.$$

So  $\sigma_{X P}^2 > \sigma_{X P_2}^2 > \sigma_{XY_2 Z_2}(1 - \alpha g)^2$ . Since  $\sigma_{XY_2} \geq \frac{2}{3}\sigma_{XY_2 Z_2}$  and  $XY_2 Z_2$  is not small,

$$(3.3) \quad \sigma_{NXY_2} = \frac{3}{\frac{1}{\sigma_{XP}\sigma_{PN}} + \frac{1}{\sigma_{Y_2 P}\sigma_{PN}} + \frac{1}{\sigma_{XY_2}}} > \frac{1}{\left(\frac{1+\sqrt{3}}{3}\right)\sigma_{XY_2 Z_2}^{-1/2}(1 - \alpha g)^{-d-2} + \frac{1}{2}\sigma_{XY_2 Z_2}^{-1}} > \sigma_{lg}.$$

So  $NXY_2$  is as desired for Case 1b.

*Case 1c:*  $\sigma_{Y_2 P_1} \geq \sigma_{X P_1}(1 - \alpha g)$ . If  $\sigma_{Z_1 P_1} > \sigma_{X P_1}$ , the target triplet is  $NZ_1 Y_2$ ; otherwise, it is  $NXY_2$ . The two cases are symmetric, and we assume  $\sigma_{X P_1} \geq \sigma_{Z_1 P_1}$ . Let  $P$  be the first node after  $P_1$  on the path from  $P_1$  toward  $P_2$  in  $T$ . Then,  $\sigma_{XP} \geq \sigma_{X P_1}(1 - \alpha g)$ . By Corollary 2.4,  $\sigma_{XP}^2 \geq \sigma_{X P_1}^2(1 - \alpha g)^2 \geq \sigma_{XY_1 Z_1}(1 - \alpha g)^2/3$ . Since  $\sigma_{XY_2} \geq \sigma_{X Z_2}$  and  $\sigma_{Y_2 Z_2} > 0$ ,

$$\sigma_{XY_2 Z_2} < \frac{3}{2\sigma_{XY_2}^{-1}} \leq \frac{3}{2\sigma_{Y_2 P_1}^{-1} \sigma_{X P_1}^{-1}} \leq \frac{3\sigma_{Y_2 P_1}^2}{2(1 - \alpha g)}.$$

Hence  $\sigma_{Y_2 P}^2 > \sigma_{Y_2 P_1}^2 > 2\sigma_{XY_2 Z_2}(1 - \alpha g)/3$ . Then, since neither  $XY_2 Z_2$  nor  $XY_1 Z_1$  is small and  $\sigma_{XY_2} \geq \frac{2}{3}\sigma_{XY_2 Z_2}$ ,

$$(3.4) \quad \sigma_{NXY_2} = \frac{3}{\frac{1}{\sigma_{XP}\sigma_{PN}} + \frac{1}{\sigma_{Y_2 P}\sigma_{PN}} + \frac{1}{\sigma_{XY_2}}} > \frac{1}{\frac{1}{\sqrt{3}}\sigma_{XY_1 Z_1}^{-1/2}(1 - \alpha g)^{-d-2} + \frac{1}{\sqrt{6}}\sigma_{XY_2 Z_2}^{-1/2}(1 - \alpha g)^{-d-3/2} + \frac{1}{2}\sigma_{XY_2 Z_2}^{-1}} > \sigma_{lg}.$$

So  $NXY_2$  is as desired for Case 1c with  $\sigma_{X P_1} \geq \sigma_{Z_1 P_1}$ .

*Case 2.* By symmetry, assume that  $P_2 = X$  is a leaf in  $T_k^*$ . Since  $k \geq 3$ ,  $P_1$  is an internal node in  $T_k^*$ . Let  $\text{def}(P_1) = \{X, Y, Z\}$ . By symmetry, further assume  $\sigma_{Y P_1} \geq \sigma_{Z P_1}$ . There are two subcases. If  $\sigma_{X P_1} < \sigma_{Y P_1}(1 - \alpha g)$ , the proof is similar to that of Case 1a and the desired  $P$  is in the middle of the path between  $X$  and  $Y$  in  $T$ . Otherwise, the proof is similar that of Case 1b and  $P$  is the first node after  $P_1$  on the path from  $P_1$  toward  $X$  in  $T$ . In both cases, the desired triplet is  $NXY$ .  $\square$

**3.4. The sample size required by Fast-HGT.** The next lemma analyzes  $\mathcal{S}$ . For  $k = 3, \dots, n - 1$  and each leaf  $M \in T$ , let  $\mathcal{S}_k[M]$  be the version of  $\mathcal{S}[M]$  at the start of the  $(k - 2)$ -th iteration of the repeat at line F7.

LEMMA 3.6. *Assume that for a given  $k \leq n - 1$ ,  $\mathcal{E}_g$ ,  $\mathcal{E}_c$ ,  $\mathcal{A}_{k'}$ ,  $\mathcal{B}_{k'}$ , and  $\mathcal{C}_{k'}$  hold for all  $k' \leq k$ .*

1. *If  $\mathcal{S}_k[M]$  is not null, then it is a splitting tuple for some edge in  $T_k^*$ .*
2. *If an edge  $Q_1 Q_2 \in T_k^*$  and a triplet  $MR_1 R_2$  with  $M \notin T_k^*$  satisfy Lemma 3.5, then  $\mathcal{S}_k[M]$  is a splitting tuple for  $Q_1 Q_2$  in  $T_k^*$  that contains a triplet  $MR_1' R_2'$  with  $\hat{\sigma}_{MR_1' R_2'} \geq \hat{\sigma}_{MR_1 R_2}$ .*

*Proof.* The two statements are proved as follows.

Statement 1. This statement follows directly from the initialization of  $\mathcal{S}$  at line F6, the deletions from  $\mathcal{S}$  at line F13, and the insertions into  $\mathcal{S}$  at lines F6 and F14.

Statement 2. The proof is by induction on  $k$ .

*Base case:*  $k = 3$ . By  $\mathcal{E}_c$ ,  $\mathcal{A}_3$ ,  $\mathcal{B}_3$ ,  $\mathcal{C}_3$ , and Lemmas 3.3 and 3.4,  $MR_1R_2$  is a splitting triplet for  $Q_1Q_2$  in  $T_3^*$ . By the maximization in Update- $\mathcal{S}$  at line F6,  $\mathcal{S}[M]$  is a splitting tuple for some edge  $Q'_1Q'_2 \in T_3^*$  that contains a triplet  $MR'_1R'_2$  with  $\hat{\sigma}_{MR'_1R'_2} \geq \hat{\sigma}_{MR_1R_2}$ . By  $\mathcal{E}_g$ ,  $MR'_1R'_2$  is not small. By Lemmas 3.3 and 3.4,  $Q'_1Q'_2$  is  $Q_1Q_2$ .

*Induction hypothesis:* Statement 2 holds for  $k < n - 1$ .

*Induction step.* We consider how  $\mathcal{S}_{k+1}$  is obtained from  $\mathcal{S}_k$  during the  $(k - 2)$ -th iteration of the repeat at line F7. There are two cases.

*Case 1:*  $Q_1Q_2$  also exists in  $T_k^*$ . By  $\mathcal{A}_k$ ,  $Q_1Q_2$  and  $MR_1R_2$  also satisfy Lemmas 3.3 and 3.4 for  $T_k^*$ . By the induction hypothesis,  $\mathcal{S}_k[M]$  is a splitting tuple for  $Q_1Q_2$  in  $T_k^*$  that contains a triplet  $MR'_1R'_2$  with  $\hat{\sigma}_{MR'_1R'_2} \geq \hat{\sigma}_{MR_1R_2}$ . Then, since  $Q_1Q_2 \neq P_1P_2$  and  $M \neq N$  at line F13,  $\mathcal{S}_k[M]$  is not reset to null. Thus, it can be changed only through replacement at line F14 by a splitting tuple for some edge  $Q'_1Q'_2$  in  $T_{k+1}^*$  that contains a triplet  $MR''_1R''_2$  with  $\hat{\sigma}_{MR''_1R''_2} \geq \hat{\sigma}_{MR'_1R'_2}$ . By  $\mathcal{E}_g$ ,  $MR''_1R''_2$  is not small. Thus, by  $\mathcal{E}_c$ ,  $\mathcal{A}_{k+1}$ ,  $\mathcal{B}_{k+1}$ ,  $\mathcal{C}_{k+1}$ , and Lemmas 3.3 and 3.4,  $Q'_1Q'_2$  is  $Q_1Q_2$ .

*Case 2:*  $Q_1Q_2 \notin T_k^*$ . This case is similar to the base case but uses the maximization in Update- $\mathcal{S}$  at line F14.  $\square$

LEMMA 3.7.  $\mathcal{E}_g$  and  $\mathcal{E}_c$  imply that  $\mathcal{A}_k$ ,  $\mathcal{B}_k$ , and  $\mathcal{C}_k$  hold for all  $k = 3, \dots, n$ .

*Proof.* The proof is by induction on  $k$ .

*Base case:*  $k = 3$ . By Lemma 2.6(2),  $\mathcal{E}_c$ , and the greedy selection of line F1, line F3 constructs  $T_3^*$  without edge lengths. Then,  $\mathcal{A}_3$  holds trivially.  $\mathcal{B}_3$  follows from  $\mathcal{E}_c$ ,  $\mathcal{E}_g$ , and line F1.  $\mathcal{C}_3$  follows from  $\mathcal{B}_3$ ,  $\mathcal{E}_c$  and the use of Equation (2.7) at line F4.

*Induction hypothesis:*  $\mathcal{A}_k$ ,  $\mathcal{B}_k$ , and  $\mathcal{C}_k$  hold for some  $k < n$ .

*Induction step.* The induction step is concerned with the  $(k - 2)$ -th iteration of the repeat at line F7. Right before this iteration, by the induction hypothesis, since  $k < n$ , some  $N'Q_1Q_2$  satisfies Lemma 3.5. Therefore, during this iteration, by  $\mathcal{E}_c$  and Lemmas 3.3, 3.4, and 3.6,  $\mathcal{S}$  at line F8 has a splitting tuple for  $T_k^*$  that contains a triplet  $NXY$  with  $\hat{\sigma}_{NXY} \geq \hat{\sigma}_{N'Q_1Q_2}$ . Furthermore, line F9 finds such a tuple. By  $\mathcal{E}_g$ ,  $NXY$  is not small. Lines F10 and F11 create  $T_{k+1}^*$  using this triplet. Thus,  $\mathcal{B}_{k+1}$  follows from  $\mathcal{B}_k$ . By Lemmas 3.3 and 3.4,  $\mathcal{A}_{k+1}$  follows from  $\mathcal{A}_k$ .  $\mathcal{C}_{k+1}$  follows from  $\mathcal{C}_k$  since the triplets involved at line S13 are not small.  $\square$

THEOREM 3.8. For any  $0 < \delta < 1$ , using sequence length

$$\ell = \mathcal{O} \left( \frac{\log \frac{1}{\delta} + \log n}{(1 - \alpha g)^{4d+8} f^2 c^2} \right),$$

Fast-HGT outputs  $T^*$  with the properties below with probability at least  $1 - \delta$ :

1. Disregarding the edge lengths,  $T^* = \Psi_w(T)$ .
2. For each edge  $Q_1Q_2$  in  $T^*$ ,  $|\Delta_{Q_1Q_2}^* - \Delta_{Q_1Q_2}| < 2\Delta_{\min}$ .

*Proof.* By Lemma 2.9,  $\Pr\{\bar{\mathcal{E}}_g\} \leq \frac{\delta}{2}$  if

$$\ell \geq \ell_g \stackrel{\text{def}}{=} 210\alpha^2 \frac{3 \ln n + \ln \frac{3}{\delta}}{\sigma_{\text{lg}}^2}.$$

Similarly, by Lemma 2.9,  $\Pr\{\bar{\mathcal{E}}_c\} \leq \frac{\delta}{2}$  if

$$\ell \geq \ell_c \stackrel{\text{def}}{=} 81 \frac{3 \ln n + \ln \frac{7}{\delta}}{\sigma_{\text{lg}}^2 f^2 c^2}.$$

We choose  $\ell = \lceil \max\{\ell_g, \ell_c\} \rceil$ . Consequently,  $\Pr\{\mathcal{E}_g \text{ and } \mathcal{E}_c\} \geq 1 - \delta$ . By Lemma 3.7, with probability at least  $1 - \delta$ , Fast-HGT outputs  $T_n^*$ , and  $\mathcal{A}_n$  and  $\mathcal{C}_n$  hold, which correspond to the two statements of the theorem.  $\square$

**4. Further research.** We have shown that theoretically, Fast-HGT has the optimal time and space complexity as well as a polynomial sample complexity. It would be important to determine the practical performance of the algorithm by testing it extensively on empirical and simulated trees and sequences. Furthermore, as conjectured by one of the referees and some other researchers, there might be a trade-off between the time complexity and the practical performance. If this is indeed true empirically, it would be significant to quantify the trade-off analytically.

**Acknowledgments.** We thank Dana Angluin, Kevin Atterson, Joe Chang, Junhyong Kim, Stan Eisenstat, Tandy Warnow, and the anonymous referees for extremely helpful discussions and comments.

#### Appendix A. Proofs of technical lemmas.

**A.1. Proof of Lemma 2.5.** Let  $h_{XY} = \frac{\hat{\sigma}_{XY}}{\sigma_{XY}}$ ;  $h_{XZ} = \frac{\hat{\sigma}_{XZ}}{\sigma_{XZ}}$ ;  $h_{YZ} = \frac{\hat{\sigma}_{YZ}}{\sigma_{YZ}}$ . By Equations (2.6) and (2.7), and by conditioning on the events  $\{h_{XZ} \leq 1 - r\}$  and  $\{h_{YZ} \geq 1 + s\}$  for some  $r, s > 0$ ,

$$\begin{aligned} \Pr\left\{\hat{\Delta}_{XP} - \Delta_{XP} \geq \frac{-\ln(1-\epsilon)}{2}\right\} &= \Pr\{h_{XY}h_{XZ} \leq h_{YZ}(1-\epsilon)\} \\ &\leq \Pr\{h_{XZ} \leq 1-r\} + \Pr\{h_{YZ} \geq 1+s\} + \Pr\left\{h_{XY} \leq (1-\epsilon)\frac{1+s}{1-r}\right\}. \end{aligned}$$

Setting  $\frac{1-r}{1+s} > 1 - \epsilon$ , by Equations (2.3) and (2.4),

$$\begin{aligned} \Pr\left\{\hat{\Delta}_{XP} - \Delta_{XP} \geq \frac{-\ln(1-\epsilon)}{2}\right\} &\leq \\ \exp\left(-\frac{2}{\alpha^2}\ell\sigma_{XZ}^2r^2\right) &+ \exp\left(-\frac{2}{\alpha^2}\ell\sigma_{YZ}^2s^2\right) + \exp\left(-\frac{2}{\alpha^2}\ell\sigma_{XY}^2\left(1 - (1-\epsilon)\frac{1+s}{1-r}\right)^2\right). \end{aligned}$$

Equating these exponential terms yields equations for  $r$  and  $s$ . The solution for  $r$  is

$$r = \frac{t - \sqrt{t^2 - u}}{2\sigma_{XZ}\sigma_{YZ}}; \quad t = \sigma_{XY}\sigma_{YZ} + \sigma_{XZ}\sigma_{YZ} + (1-\epsilon)\sigma_{XY}\sigma_{XZ}; \quad u = 4\sigma_{XY}\sigma_{YZ}^2\sigma_{XZ}\epsilon.$$

Using Taylor's expansion, for  $u > 0$ ,  $(t - \sqrt{t^2 - u})^2 > \frac{u^2}{4t^2}$ . Thus,

$$r^2 > \frac{\epsilon^2}{\left(\frac{1}{\sigma_{XZ}} + \frac{1-\epsilon}{\sigma_{YZ}} + \frac{1}{\sigma_{XY}}\right)^2 \sigma_{XZ}^2} > \frac{\epsilon^2 \sigma_{XY}^2}{9\sigma_{XZ}^2}.$$

So  $\Pr\left\{\hat{\Delta}_{XP} - \Delta_{XP} \geq \frac{-\ln(1-\epsilon)}{2}\right\} \leq 3 \exp\left(-\frac{2}{\alpha^2}\ell\sigma_{XZ}^2r^2\right) < 3 \exp\left(-\frac{2}{9\alpha^2}\ell\sigma_{XY}^2\epsilon^2\right)$ .

**A.2. Proof of Lemma 2.7.** We use the following basic inequalities.

$$(A.1) \quad \min\left\{\frac{\hat{\sigma}_{XY}}{\sigma_{XY}}, \frac{\hat{\sigma}_{XZ}}{\sigma_{XZ}}, \frac{\hat{\sigma}_{YZ}}{\sigma_{YZ}}\right\} \leq \frac{\hat{\sigma}_{XYZ}}{\sigma_{XYZ}} \leq \max\left\{\frac{\hat{\sigma}_{XY}}{\sigma_{XY}}, \frac{\hat{\sigma}_{XZ}}{\sigma_{XZ}}, \frac{\hat{\sigma}_{YZ}}{\sigma_{YZ}}\right\};$$

$$(A.2) \quad \frac{\sigma_{XYZ}}{3} \leq \min\{\sigma_{XY}, \sigma_{XZ}, \sigma_{YZ}\}.$$

The proof of Equation (2.9) is symmetric to that of Equation (2.8). So we only prove the latter. Pick  $\lambda \geq 1$  with  $\sigma_{XYZ} = \sigma_{\text{lg}}\lambda$ . Without loss of generality, we assume  $\min \left\{ \frac{\hat{\sigma}_{XY}}{\sigma_{XY}}, \frac{\hat{\sigma}_{XZ}}{\sigma_{XZ}}, \frac{\hat{\sigma}_{YZ}}{\sigma_{YZ}} \right\} = \frac{\hat{\sigma}_{XY}}{\sigma_{XY}}$ . By Equations (2.3), (A.1), and (A.2),

$$\begin{aligned} \Pr\{\hat{\sigma}_{XYZ} \leq \sigma_{\text{md}}\} &= \Pr\left\{\frac{\hat{\sigma}_{XYZ}}{\sigma_{XYZ}} \leq \frac{\sigma_{\text{md}}}{\sigma_{\text{lg}}\lambda}\right\} \leq \Pr\left\{\frac{\hat{\sigma}_{XY}}{\sigma_{XY}} \leq \frac{\sigma_{\text{md}}}{\sigma_{\text{lg}}\lambda}\right\} \\ &\leq \exp\left(-\frac{2}{\alpha^2}\ell\left(1 - \frac{\sigma_{\text{md}}}{\sigma_{\text{lg}}\lambda}\right)^2\sigma_{XY}^2\right) \leq \exp\left(-\frac{2\left(1 - \frac{\sigma_{\text{md}}}{\sigma_{\text{lg}}}\right)^2}{9\alpha^2}\ell\sigma_{\text{lg}}^2\right). \end{aligned}$$

Then, Equation (2.8) follows from the fact that by the choice of  $\sigma_{\text{md}}$ ,

$$\frac{2\left(1 - \frac{\sigma_{\text{md}}}{\sigma_{\text{lg}}}\right)^2}{9\alpha^2} = \frac{(\sqrt{2} - 1)^2}{36\alpha^2}.$$

**A.3. Proof of Lemma 2.8.** Since Lemma 2.5 can help establish only one half of the desired inequality, we split the probability on the left-hand side of Equation (2.11).

$$\begin{aligned} &\Pr\left\{\left|\hat{\Delta}_{XP} - \Delta_{XP}\right| \geq \frac{\Delta_{\min}}{2}\right\} \\ &\leq \Pr\left\{\hat{\Delta}_{XP} - \Delta_{XP} \geq \frac{\Delta_{\min}}{6}\right\} + \Pr\left\{\hat{\Delta}_{YP} - \Delta_{YP} \geq \frac{\Delta_{\min}}{6}\right\} + \\ &\quad \Pr\left\{\hat{\Delta}_{XP} - \Delta_{XP} \leq -\frac{\Delta_{\min}}{2} \mid \hat{\Delta}_{YP} - \Delta_{YP} < \frac{\Delta_{\min}}{6}\right\}. \end{aligned}$$

Then, since  $\hat{\Delta}_{XY} - \Delta_{XY} = (\hat{\Delta}_{XP} - \Delta_{XP}) + (\hat{\Delta}_{YP} - \Delta_{YP})$ , we have

$$\begin{aligned} &\Pr\left\{\hat{\Delta}_{XP} - \Delta_{XP} \leq -\frac{\Delta_{\min}}{2} \mid \hat{\Delta}_{YP} - \Delta_{YP} < \frac{\Delta_{\min}}{6}\right\} \\ &\leq \Pr\left\{\hat{\Delta}_{XY} - \Delta_{XY} \leq -\frac{\Delta_{\min}}{3}\right\}. \end{aligned}$$

Consequently,

$$\begin{aligned} \text{(A.3)} \quad \Pr\left\{\left|\hat{\Delta}_{XP} - \Delta_{XP}\right| \geq \frac{\Delta_{\min}}{2}\right\} &\leq \Pr\left\{\hat{\Delta}_{XP} - \Delta_{XP} \geq \frac{\Delta_{\min}}{6}\right\} + \\ &\quad \Pr\left\{\hat{\Delta}_{YP} - \Delta_{YP} \geq \frac{\Delta_{\min}}{6}\right\} + \\ &\quad \Pr\left\{\hat{\Delta}_{XY} - \Delta_{XY} \leq -\frac{\Delta_{\min}}{3}\right\}. \end{aligned}$$

By Lemma 2.5,

$$\Pr\left\{\hat{\Delta}_{XP} - \Delta_{XP} \geq \frac{\Delta_{\min}}{6}\right\} \leq 3 \exp\left(-\frac{2}{9\alpha^2}\ell\sigma_{XYZ}^2\left(1 - e^{-\frac{\Delta_{\min}}{3}}\right)^2\right).$$

By Taylor's expansion,  $\left(1 - e^{-\frac{\Delta_{\min}}{3}}\right)^2 \geq \left(1 - (1 - \alpha f)^{\frac{6}{3}}\right)^2 > \frac{c^2}{9}\alpha^2 f^2$ , and thus

$$\text{(A.4)} \quad \Pr\left\{\hat{\Delta}_{XP} - \Delta_{XP} \geq \frac{\Delta_{\min}}{6}\right\} \leq 3 \exp\left(-\frac{c^2}{81}\ell\sigma_{\text{lg}}^2 f^2\right).$$

By symmetry,

$$(A.5) \quad \Pr\left\{\hat{\Delta}_{YP} - \Delta_{YP} \geq \frac{\Delta_{\min}}{6}\right\} \leq 3 \exp\left(-\frac{c^2}{81} \ell \sigma_{\lg}^2 f^2\right).$$

By Equation (2.4),  $\Pr\left\{\hat{\Delta}_{XY} - \Delta_{XY} \leq -\frac{\Delta_{\min}}{3}\right\} \leq \exp\left(-\frac{2}{\alpha^2} \ell \sigma_{XY}^2 \left(e^{\frac{\Delta_{\min}}{3}} - 1\right)^2\right).$

From Equation (A.2),  $\sigma_{XY} \geq \frac{(1-\alpha g)^{2d+2}}{3\sqrt{2}}$ . By Taylor's expansion,  $\left(e^{\frac{\Delta_{\min}}{3}} - 1\right)^2 \geq \left((1-\alpha f)^{-\frac{c}{3}} - 1\right)^2 > \frac{c^2}{9} \alpha^2 f^2$ . Therefore,

$$(A.6) \quad \Pr\left\{\hat{\Delta}_{XY} - \Delta_{XY} \leq -\frac{\Delta_{\min}}{3}\right\} \leq \exp\left(-\frac{c^2}{81} \ell \sigma_{\lg}^2 f^2\right).$$

Lemma 2.8 follows from the fact that putting Equations (A.3) through (A.6) together, we have  $\Pr\left\{\left|\hat{\Delta}_{XY} - \Delta_{XY}\right| \geq \frac{\Delta_{\min}}{2}\right\} \leq 7 \exp\left(-\frac{c^2}{81} \ell \sigma_{\lg}^2 f^2\right).$

#### REFERENCES

- [1] R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson, and M. Thorup. On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM Journal on Computing*, 2000. To appear.
- [2] A. Ambainis, R. Desper, M. Farach, and S. Kannan. Nearly tight bounds on the learnability of evolution. In *Proceedings of the 38th Annual IEEE Symposium on Foundations of Computer Science*, pages 524–533, 1997.
- [3] K. Atteson. The performance of neighbor-joining algorithms of phylogeny reconstruction. *Algorithmica*, 25(2-3):251–278, 1999.
- [4] M. Cryan, L. A. Goldberg, and P. W. Goldberg. Evolutionary trees can be learned in polynomial time in the two-state general Markov-model. In *Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science*, pages 436–445, 1998.
- [5] M. Csűrös and M. Y. Kao. Recovering evolutionary trees through harmonic greedy triplets. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 261–270, 1999.
- [6] M. Csuros. *Reconstructing Phylogenies in Markov Models of Evolution*. PhD thesis, Yale University, 2000. Co-Directors: Dana Angluin and Ming-Yang Kao.
- [7] W. H. E. Day. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49:461–467, 1987.
- [8] W. H. E. Day, D. S. Johnson, and D. Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81:33–42, 1986.
- [9] D.-Z. Du, Y.-J. Zhang, and Q. Feng. On better heuristic for Euclidean Steiner minimum trees (extended abstract). In *Proceedings of the 32nd Annual IEEE Symposium on Foundations of Computer Science*, pages 431–439, 1991.
- [10] P. L. Erdős, M. A. Steel, L. A. Székely, and T. J. Warnow. A few logs suffice to build (almost) all trees. I. *Random Structures & Algorithms*, 14(2):153–184, 1999.
- [11] P. L. Erdős, M. A. Steel, L. A. Székely, and T. J. Warnow. A few logs suffice to build (almost) all trees. II. *Theoretical Computer Science*, 221(1-2):77–118, 1999.
- [12] M. Farach and S. Kannan. Efficient algorithms for inverting evolution. *Journal of the ACM*, 46(4):437–449, 1999.
- [13] J. Felsenstein. Numerical methods for inferring evolutionary trees. *The Quarterly Review of Biology*, 57:379–404, 1982.
- [14] J. Felsenstein. Inferring evolutionary trees from DNA sequences. In B. Weir, editor, *Statistical Analysis of DNA Sequence Data*, pages 133–150. Dekker, 1983.
- [15] J. Felsenstein. Statistical inference of phylogenies. *Journal of the Royal Statistical Society Series A*, 146:246–272, 1983.
- [16] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York, NY, 1997.
- [17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.



- [18] D. Huson, S. Nettles, and T. Warnow. Disk-covering, a fast converging method for phylogenetic tree reconstruction. *Journal of Computational Biology*, 6(3):369–386, 1999.
- [19] M. J. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie. On the learnability of discrete distributions (extended abstract). In *Proceedings of the 26th Annual ACM Symposium on Theory of Computing*, pages 273–282, 1994.
- [20] M. E. Siddall. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. *Cladistics*, 14:209–220, 1998.
- [21] M. Steel. Recovering a tree from the leaf colourations it generates under a Markov model. *Applied Mathematics Letters*, 7(2):19–23, 1994.
- [22] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. K. Mable, editors, *Molecular Systematics*, chapter 11, pages 407–514. Sinauer Associates, Sunderland, Ma, 2nd edition, 1996.