New Sample Complexity Bounds for Phylogenetic Inference from Multiple Loci

Gautam Dasarathy
Electrical & Computer Engineering
University of Wisconsin - Madison
dasarathy@wisc.edu

Robert Nowak
Electrical & Computer Engineering
University of Wisconsin - Madison
nowak@ece.wisc.edu

Sebastien Roch
Mathematics
University of Wisconsin - Madison
roch@math.wisc.edu

***Abstract*—We consider the problem of estimating the evolutionary history of a set of species (phylogeny or species tree) from several genes. It has been known however that the evolutionary history of individual genes (gene trees) might be topologically distinct from each other and from the underlying species tree, possibly confounding phylogenetic analysis. A further complication in practice is that one has to estimate gene trees from molecular sequences of finite length. We provide the first full data-requirement analysis of a species tree reconstruction method that takes into account estimation errors at the gene level. Under that criterion, we also devise a novel algorithm that provably improves over all previous methods in a regime of interest.**

I. Introduction

We consider the problem of estimating the common evolutionary history, more precisely the *species tree*, of a set of n species using sequence data from multiple genes or loci. It is well known that the estimated genealogical history of a gene (*gene tree*) may be topologically distinct from the species tree that encapsulates it, possibly confounding phylogenetic analysis [11]. Here we consider an important source of such incongruence, known as *incomplete lineage sorting* (ILS), where two lineages fail to coalesce in their most recent common ancestral population. This may lead one of the lineages to first coalesce with a more distantly related population thereby producing incongruence. Several species tree reconstruction methods have recently been developed that address ILS. See for instance [8] and references therein. Several such methods rely on a statistical model known as the *multispecies coalescent* (MSC): independent coalescent processes are performed in each ancestral population and these are assembled to produce a gene tree. This process is illustrated in Figure 1 below. For more on phylogenetic inference and coalescent theory see, e.g., [4], [5], [19].

The accuracy of multiloci reconstruction methods has been evaluated empirically, for instance, in [7], [10]. The focus here is the mathematical characterization of the performance of such methods. Prior theoretical work has focused mainly on statistical consistency under the multispecies coalescent; see e.g., [10], [2], [13], [9]. That is, assuming access to either correct gene trees or correct pairwise distances (or coalescence times) for each gene, a method is *statistically consistent* if it is guaranteed to converge on the correct species tree as the number of genes, m, tends to infinity. [17] studies the rates of convergence (in m) for several such methods. For instance, letting $f > 0$ denote the smallest branch length in the species tree, in the limit $f \to 0$, it was shown that the GLASS algorithm [13], an agglomerative clustering method in which the dissimilarity between each pair of species is taken to be the *minimum* of the coalescent times among the m genes, needs the number of genes m to scale as f^{-1}. On the other hand, m needs to scale as f^{-2} for the STEAC algorithm [10], which is also an agglomerative clustering method which instead uses the *average* of the coalescent times across the m genes as the measure of dissimilarity. In reality, however, one has to estimate gene trees and coalescent times from finite, say, length-k molecular sequences. Taking into account the resulting estimation errors at the gene level is key to mathematically quantify and compare the performance of different methods (see e.g., [14], [24], [6]). Intuitively, for instance, the "minimum" used in GLASS may be more sensitive to estimation errors than the "average" used in STEAC. We make progress towards this goal by performing the first full data requirement analysis of some species tree reconstruction methods.

Our contribution is two-fold. First it is known that, in order to reliably reconstruct a single gene tree, it is both necessary [22] and sufficient [3] for the sequence length k to scale as f^{-2}. Therefore, in light of this and the results in [17], one might expect that the total amount of data required mk (since there are m genes, each of length k) must scale as f^{-3} and f^{-4} for GLASS and STEAC respectively. We show that, by a crucial modification of STEAC, one obtains an algorithm that is guaranteed to reconstruct the species tree exactly with high probability as long as m scales like f^{-2} and $k \geq 1$. In particular, it suffices for the overall sample complexity, mk, to scale like f^{-2} (which is much smaller than f^{-3} and f^{-4} in the regime of interest, where $f \ll 1$). Secondly, unlike GLASS, STEAC only works under the restrictive molecular clock assumption [19], where the mutation rates are constant across the populations in the species tree. We extend the previous data requirement result beyond the molecular clock by devising a novel STEAC-like species tree reconstruction algorithm which we call METAL (Metric algorithm for Estimation of Trees based on Aggregation of Loci). This algorithm is a distance based method where the distances are defined by concatenating the molecular sequences corresponding to all the loci (genes).

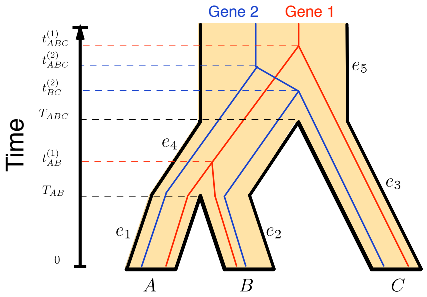

Fig. 1: A species tree (thick, shaded tree) and two samples from the MSC. Notice that while the topology of Gene 1 agrees with the species tree, the topology of Gene 2 does not.

A. Preliminaries and Notation

1. The Species Tree. At the heart of the model is an unknown *species tree* $S = (V, E)$ which represents the evolutionary history of n isolated populations; these are represented by the size n leaf set L of this tree. The goal is to learn S. For the sake of simplicity, we will assume that each branch of the species tree corresponds to a population of size N and as is standard in coalescent theory, we will assign each branch $e \in E$ a length $\tau_e > 0$ in coalescent time units (which is proportional to the number of generations represented by the branch). The smallest branch length, $f \triangleq \min_e \tau_e$, will play an important role in our analysis and results. For a pair of vertices $X, Y \in V$, we will use $\pi^S_{XY} \subset E$ to denote the path connecting X and Y in S and τ_{XY} will denote the length of this path. Notice that $\{\tau_{AB}\}_{A,B \in L}$ is an *ultrametric* with respect to S[§]. We will let $\Delta \triangleq \max_{A,B \in L} \tau_{AB}$ denote the diameter of the species tree. Finally, to each branch $e \in E$, we will also associate a mutation rate, μ_e and we will let $\mu_L \triangleq \min_{e \in E} \mu_e$ and $\mu_U \triangleq \max_{e \in E} \mu_e$ denote the smallest and the largest mutation rates, respectively.

2. The Multispecies Coalescent and Gene Trees. Following [15], we assume that a *multispecies coalescent* (MSC) process produces m (independent) random genealogies $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(m)}$ based on S. These encode, say, the evolutionary history of m different genes or loci on the genome and will be referred to as *gene trees* henceforth. It is easier to understand the MSC constructively. Consider Figure 1, where the thick, shaded tree is the species tree S with edges $\{e_i\}_{i=1}^5$. Time (in coalescent time units) starts at 0 at the leaves and increases towards the root of the tree. By T_{AB} (resp. T_{ABC}), we mean the time when the parent population of A and B (resp. the parent population of A, B, and C) speciates. Let us first consider one random draw from the MSC, Gene 1. A, B, and C each have a copy (allele) of Gene 1 and the MSC describes the evolutionary history of the lineages corresponding to these alleles. From time 0 until T_{AB}, the lineages corresponding to A and B are in isolated populations and hence do not "coalesce". However, once these lineages reach the parent population of A and B (i.e., e_4), they have a chance to coalesce. According to the MSC, the coalescence happens after a random time drawn according to the Exp(1) distribution, i.e., $\mathbb{P}(t^{(1)}_{AB} - T_{AB} \geq x) = e^{-x}, x \geq 0$. Now, this coalesced lineage and the lineage corresponding to C do not interact until time T_{ABC}. They then coalesce at a random time $t^{(1)}_{ABC}$, where $t^{(1)}_{ABC} - T_{ABC} \sim \text{Exp}(1)$. This gives us a random gene tree with the topology $((A, B), C)$. On the other hand, in the case of Gene 2, the lineages corresponding to A and B do not coalesce in e_4 So, at time T_{ABC}, there are three lineages present in e_5. According to the MSC, when there are multiple lineages in the same population, each pair independently coalesces after a random time period drawn according to the Exp(1) distribution. In this case, the genealogies of B and C alleles coalesce (at time $t^{(2)}_{BC}$) before A and B, thus giving us a second random tree with topology $(A, (B, C))$. Notice that while the genealogy of Gene 1 agrees with the topology of S, the genealogy of Gene 2 does not. This is an example of ILS which, as mentioned earlier, is a fundamental road block for learning the tree of life.

We refer the reader to [1] for more on our modeling assumptions and to [15] for more details on the MSC. However, we will state the model here for completeness. The density of the likelihood of a gene tree $\mathcal{G}^{(i)} = \left(\mathcal{V}^{(i)}, \mathcal{E}^{(i)}\right)$ can be written down as follows. For each branch $e \in E$ of the species tree let $I^{(i)}_e$ and $O^{(i)}_e$ be the number of lineages entering and leaving the branch e respectively. Let $t^{(i)}_{e,s}$, be the $s-$th coalescent time in the branch e. Then, the density of the likelihood of $\mathcal{G}^{(i)}$ is given by $\prod_{e \in E} \prod_{s=1}^{I^{(i)}_e - O^{(i)}_e + 1} \exp\left\{-\binom{I^{(i)}_e - s + 1}{2}\left[t^{(i)}_{e,s} - t^{(i)}_{e,s-1}\right]\right\}$, where, for convenience, we let $t^{(i)}_{e,0}$ and $t^{(i)}_{e, I^{(i)}_e - O^{(i)}_e + 1}$ be respectively the divergence times of the population in e and of its parent population.

3. Observation Model and the Inference Problem. Much of the prior work on understanding the theoretical complexity of learning species trees from multiple loci has assumed that exact gene trees are available. However, in reality one needs to estimate these gene trees from molecular sequences and indeed there has been a recent thrust towards investigating the effect of errors in estimating the gene trees (see e.g., [14], [24], [6]). Our approach will be to take this error into account explicitly and in fact bypass the reconstruction of the gene trees altogether.

We model the sample generation process according to the standard Jukes-Cantor (JC) model (see e.g., [19]). For this, given a gene tree $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we will associate to each $\tilde{e} \in \mathcal{E}$, a probability $p_{\tilde{e}}$ (whose form we will make explicit below). The JC model assigns a character from $\{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$ uniformly at random to the root of $\mathcal{G}$. Moving away from the root, with probability $p_{\tilde{e}}$, each edge $\tilde{e}$ changes the state of its ancestor to one of the other three, chosen uniformly at random. The states at the leaves of $\mathcal{G}$ are assembled into a length n vector to get the first sample; this process is repeated k times to generate the data set. Notice that k models the number of sites or the sequence length of each gene.

[§] that is, for any three leaves A, B, C such that S restricted to A, B, C has the topology $((A, B), C)$, we have that $\tau_{AB} \leq \tau_{AC} = \tau_{BC}$

Now, we define $p_{\tilde{e}}$. To each edge $\tilde{e}$ of the random gene tree $\mathcal{G}$ is associated a random length $\sigma_{\tilde{e}}$ according to the MSC. Also, given an edge $e \in E$ of the species tree, we will write $\sigma_{e \cap \tilde{e}}$ to denote the length of the portion of $\tilde{e}$ that overlaps with e. This lets us define the effective (mutation rate adjusted) branch lengths, $\delta_{\tilde{e}} = \sum_{e \in E} \mu_e \sigma_{e \cap \tilde{e}}$. As before, for any two vertices $X, Y \in \mathcal{V}$, $\pi^{\mathcal{G}}_{XY}$ denotes the path joining X and Y in $\mathcal{G}$ and σ_{XY} (resp. δ_{XY}) denotes the length of this path under σ (resp. under δ). Now, for an edge $\tilde{e} \in \mathcal{E}$, we define $p_{\tilde{e}} \triangleq \frac{3}{4}(1 - e^{-\frac{4}{3}\delta_{\tilde{e}}})$[1].

The goal then, is to learn the structure of S given the data $\{\chi^{ij}\}_{i \in [m], j \in [k]}$, where $\{\chi^{ij}\}_{j \in [k]}$ is the data generated from $\mathcal{G}^{(i)}$ according to the JC model.

The JC model was chosen because it lends itself to easy presentation. Since the techniques developed here are *distance-based*, they can be generalized to the more realistic Generalized Time-Reversible (GTR) model [23] using spectral techniques as in [16], [12].

II. Main Results

We now state the main results of the paper. First, we will deal with the case where the strong molecular clock [19] assumption holds. We will then turn to the more general case.

1. The Molecular Clock Assumption Holds. While assuming the molecular clock hypothesis (which is equivalent to believing that all mutations happen at the same rate through time and across populations) is often unrealistic, it has proved to be a useful abstraction for developing powerful methods. In our setting, this is equivalent to assuming that $\mu_e = \mu > 0, \forall e \in E$.

In order to infer the species tree from samples, we will begin by defining a distance measure on the leaves. For each pair of leaves $A, B \in L$, we define

$$\widehat{p}_{AB} = \frac{1}{mk} \sum_{i \in [m], j \in [k]} \mathbb{1}\{\chi^{ij}_A \neq \chi^{ij}_B\}, \tag{1}$$

which can be thought of as the normalized hamming distance between the molecular sequences corresponding to species A and B. Our first result is that, in expectation, this is not only a metric on L, but is in fact an ultrametric with respect to S.

Theorem 1. $\{\mathbb{E}\left[\widehat{p}_{AB}\right]\}_{A,B \in L}$ *forms an ultrametric with respect to the true species tree S. In fact, for any triple $A, B, C \in L$ with the topology $((A, B), C)$ in S, we have*

$$\mathbb{E}\left[\widehat{p}_{AC}\right] = \mathbb{E}\left[\widehat{p}_{BC}\right] > \mathbb{E}\left[\widehat{p}_{AB}\right] + \frac{3e^{-\frac{4}{3}\mu\tau_{AC}}\mu f}{8\mu + 3}. \tag{2}$$

The proof follows from the observation that, by definition, we have $\mathbb{E}\left[\widehat{p}_{AC}\right] = \mathbb{E}\left[\frac{3}{4}\left(1 - e^{-\frac{4}{3}\delta_{AC}}\right)\right]$, where δ_{AC} is the random gene tree distance that satisfies $\delta_{AC} = \mu\tau_{AC} + 2\mu Z$ with $Z \sim \text{Exp}(1)$. We refer the reader to [1] for the exact details. This result inspires the following procedure for reconstructing S: Use $\{\widehat{p}_{AB}\}_{A,B \in L}$ as a dissimilarity measure for L and use a standard algorithm that accepts a dissimilarity measure and returns an ultrametric tree (see e.g., [4], [19] for background on distance based methods). For the sake of simplicity, we may assume that we use the UPGMA algorithm [20], the standard method for bottom-up agglomerative clustering, in order to produce an ultrametric tree. Then, recalling that μ denotes the (common) mutation rate across the populations represented by the species tree, S and Δ denotes diameter of S, we have the following performance guarantee.

Theorem 2. *Given an $\epsilon > 0$, using UPGMA on L with the dissimilarity measure $\{\widehat{p}_{AB}\}_{A,B \in L}$ results in the correct tree S being output with probability no less than $1 - \epsilon$ as long as the number of genes m, and the sequence length k satisfy*

$$m \geq C_1(\mu, \Delta, n, \epsilon) \times f^{-2} \ \ \textit{and} \ \ k \geq 1, \tag{3}$$

where $C_1(\mu, \Delta, n, \epsilon) = \frac{16\, e^{\frac{8}{3}\mu\Delta}(8\mu+3)^2}{9\mu^2} \log\left(\frac{8\binom{n}{3}}{\epsilon}\right)$.

Theorem 2, whose proof is sketched in Section IV-A , tells us that the above procedure succeeds with high probability as long as we get molecular sequences of length at least one from at least $\mathcal{O}(f^{-2})$ genes. That is, a total sequence length of $mk = \mathcal{O}(f^{-2})$ suffices for reliable learning. Notice that the procedure we propose is similar to the STEAC [10] algorithm except, instead of using the average coalescent time as the distance measure, we use (1), which can be considered as the normalized hamming distance. It turns out that this fact is important in obtaining our improved sample complexity result.

2. The Molecular Clock Assumption Does Not Hold. We will now consider the more general case where the strong molecular clock assumption does not hold. That is, we will assume that each branch e of the species tree has a (possibly) distinct mutation rate μ_e.

First, observe that $\mathbb{E}[\widehat{p}_{AB}]$ as defined above is no longer an ultrametric with respect to S and therefore, the above procedure (and for a similar reason, the STEAC algorithm) cannot be used to recover the species tree. In such situations, one usually turns to distance methods that rely on the 4-point condition (see e.g., [19]). However, it is not immediately clear that one can define a metric that satisfies this condition in our setting. Our next result, which is arguably the most important contribution of this paper, shows precisely that this can be done. Towards this end, for $A, B \in L$, we define the following measure of dissimilarity $d_{AB} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\mathbb{E}\left[\widehat{p}_{AB}\right]\right)$, where $\widehat{p}_{AB}$ is as defined in (1). Theorem 3, which parallels Theorem 1, shows that this "idealized" dissimilarity measure is actually an *additive metric* with respect to S[‡]. See, e.g., [19] for more on tree metrics. This result is especially interesting since phylogenetic mixtures are known to cause problems for distance-based methods [21].

[1] Notice that this definition implies that the probability p_{XY} of disagreement between the characters at vertices X and Y satisfies, $p_{XY} = \frac{3}{4}(1 - e^{-\frac{4}{3}\delta_{XY}})$.

[‡] Recall that this means that the four point condition holds, i.e., for a quadruple of leaves A, B, C, D that are such that $((A, B), (C, D))$ or $(((A, B), C), D)$ holds with respect to S, the distances satisfy $d_{AB} + d_{CD} \leq d_{AC} + d_{BD} = d_{AD} + d_{BC}$.

Theorem 3. *The set of dissimilarities $\{d_{AB}\}_{A,B\in L}$ forms an additive metric with respect to S. In fact, suppose the leaves $A, B, C, D \in L$ are such that either $((A,B),(C,D))$ or $(((A,B),C),D)$ holds with respect to S, then*

$$d_{AC} + d_{BD} = d_{AD} + d_{BC} > d_{AB} + d_{CD} + \alpha_{\text{add}}, \qquad (4)$$

where $\alpha_{\text{add}} = \frac{3}{4}\log\left(\frac{8}{3}\mu_L(1-e^{-f})+1\right) > 0$ *and* $\mu_L \triangleq \min_{e\in E}\mu_e$*, as defined in Section I-A.*

It is somewhat surprising that this result is true. This theorem tells us that if one pretends as though all samples came from a single gene tree and estimate it, then this tree suggested by the "concatenated molecular sequence" has the same topology as S. We sketch the proof in Section IV-B

In light of this, we propose the following algorithm, which we call METAL (for Metric algorithm for Estimation of Trees based on Aggregation of Loci), to reconstruct S. First, define the following sample-based *corrected* measure of dissimilarity (with $\widehat{p}_{AB}$ as defined in (1)):

$$\widehat{d}_{AB} \triangleq -\frac{3}{4}\log\left(1-\frac{4}{3}\widehat{p}_{AB}\right). \qquad (5)$$

Now, use any algorithm that returns an additive tree (like Neighbor Joining [18]) using $\{\widehat{d}_{AB}\}_{A,B\in L}$ (from (5)) as the input dissimilarity measure. Recall that μ_U and μ_L are respectively the maximum and minimum mutation rates, and Δ is the diameter of the species tree S (c.f. Section I-A). We then have the following result.

Theorem 4. *For any $\epsilon > 0$, the METAL algorithm succeeds in reconstructing (the unrooted version of) S with probability at least $1-\epsilon$ as long as m and k satisfy*

$$k \geq 1 \textit{ and } m \geq \frac{e^{\frac{8\mu_U\Delta}{3}}(8\mu_U+3)^2(24+8\alpha_{\text{add}})^2}{162\alpha_{\text{add}}^2}\log\left(\frac{16\binom{n}{4}}{\epsilon}\right)$$

where $\alpha_{\text{add}} = \frac{3}{4}\log\left(\frac{8}{3}\mu_L(1-e^{-f})+1\right)$.

In the limit as $f \to 0$, the right side above approaches $C_2(\mu_U,\mu_L,\Delta,n,\epsilon) \times f^{-2}$,, where $C_2(\mu_U,\mu_L,\Delta,n,\epsilon) = \frac{8e^{\frac{8\mu_U\Delta}{3}}(8\mu_U+3)^2}{9\mu_L^2}\log\left(\frac{16\binom{n}{3}}{\epsilon}\right)$.

Remark. Following [3], the diameter Δ can be replaced by the (often much smaller) *depth*[2] by using only those distances that are "small enough".

We refer the reader to [1] for the proof of Theorem 4 which is similar in spirit to the proof of Theorem 2. This result tells us that as long as m scales like $\mathcal{O}(f^{-2})$ and $k \geq 1$, the species tree can be reconstructed (upto the location of the root) reliably. It should be noted here that we assume that for each population/branch $e \in E$, the mutation rate μ_e is constant across gene trees; generalizing our analysis to the case where the mutation rates are allowed to change is an interesting avenue for future work.

III. Discussion

Irrespective of the sequence length k, the number of genes m needs to satisfy $m \in \Omega(f^{-1})$ for consistent species tree estimation. To see this, observe that any algorithm that is able to estimate S reliably should be able to perform a reliable hypothesis test between two shifted exponential distributions. So, the lower bound follows from the fact that $D_{\text{KL}}\left(p(x;\tau_{AB}+f)\|p(x;\tau_{AB})\right) = f$, where $p(x;a) = e^{-(x-a)}\mathbb{1}\{x \geq a\}$ and $D_{KL}(\cdot\|\cdot)$ is the Kullback-Liebler divergence. On the other hand, we know from [22] that even without the confounding effect of the MSC, a total sequence length (mk) of at least $\Omega(f^{-2})$ is needed for consistent estimation. These two together imply that there is a constant $C > 0$ such that m needs to satisfy the following for consistent estimation of the species tree $m \geq C\max\left\{f^{-1}, \frac{f^{-2}}{k}\right\}$.

The results in this paper show that $m \in \mathcal{O}(f^{-2})$ is achievable irrespective of the value of k; in particular, a total data set size of $mk \in \mathcal{O}(f^{-2})$ is achievable. Prior to this, to the best of our knowledge, the best complexity bounds were provably attained by GLASS [13] (as shown in [17]) which requires that $m \geq \mathcal{O}(f^{-1})$ and $k \geq \mathcal{O}(f^{-2})$, i.e., a total data set size of $mk \in \mathcal{O}(f^{-3})$. This raises two very interesting open questions. (A) What is the precise tradeoff between m and k for reliable recovery of S? (B) Is there a procedure that attains all points (values of m and k) in this tradeoff, as opposed to the current situation where it appears as though GLASS meets the lower bounds for large k and our procedure meets the lower bound for small k?

IV. Sketch of the Proofs

In this section, we will sketch the proofs of Theorem 2 and 3. For more details, we refer the interested reader to [1].

A. Sketch of the Proof of Theorem 2. For a pair of leaves $A, B \in L$, observe that for $\alpha_{\text{um}} > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left[\widehat{p}_{AB} - \mathbb{E}[\widehat{p}_{AB}] > \alpha_{\text{um}}/2\right] \\
&\quad = \mathbb{E}\left[\mathbb{P}\left(\widehat{p}_{AB} - \mathbb{E}\left[\widehat{p}_{AB}\right] > \frac{\alpha_{\text{um}}}{2}\middle|\{\delta_{AB}^{(i)}\}_{i\in[m]}\right)\right] \\
&\quad \leq \mathbb{E}\left[\mathbb{P}\left(\widehat{p}_{AB} - \frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} > \frac{\alpha_{\text{um}}}{4}\middle|\{\delta_{AB}^{(i)}\}_{i\in[m]}\right)\right. \\
&\qquad\qquad \left. + \mathbb{P}\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} - \mathbb{E}\left[\widehat{p}_{AB}\right] > \frac{\alpha_{\text{um}}}{4}\middle|\{\delta_{AB}^{(i)}\}_{i\in[m]}\right)\right],
\end{aligned}$$

where in the first equation, $\delta_{AB}^{(i)}$ is the distance between leaves A and B on the random gene tree $\mathcal{G}^{(i)}$. Using Hoeffding's inequality, we get $\mathbb{P}\left[\widehat{p}_{AB} - \mathbb{E}[\widehat{p}_{AB}] > \frac{\alpha_{\text{um}}}{2}\right] \leq e^{-\frac{mk\alpha_{\text{um}}^2}{16}} + e^{-\frac{m\alpha_{\text{um}}^2}{16}}$.

Now, notice that the probability of error is upper bounded by the sum over all triplets $((A,B),C)$ of $\mathbb{P}[\widehat{p}_{AB} > \widehat{p}_{AC}]$. Using Theorem 1, we know that $\mathbb{P}[\widehat{p}_{AB} > \widehat{p}_{AC}]$ is upper bounded by $\mathbb{P}[\widehat{p}_{AB} - \mathbb{E}[\widehat{p}_{AB}] > \frac{\alpha_{\text{um}}}{2}] + \mathbb{P}\left[\mathbb{E}\left[\widehat{p}_{AC}\right] - \widehat{p}_{AC} > \frac{\alpha_{\text{um}}}{2}\right]$. Therefore, from above, we get that $\sum_{((A,B),C)}\mathbb{P}\left[\widehat{p}_{AB} > \widehat{p}_{AC}\right]$

[2] The depth of an edge e is the length (under τ) of the shortest path between two leaves crossing e; the depth of a tree is the maximum edge depth.

is no less than $2\binom{n}{3}\left(e^{-mk\alpha_{\mathrm{um}}^2/16}+e^{-m\alpha_{\mathrm{um}}^2/16}\right)$. Picking m and k as prescribed will guarantee that the probability of error is upper bounded by ϵ.

B. Sketch of the Proof of Theorem 3. Note that for any 4 leaves of the species tree A, B, C, D, there are only 2 possible topologies with respect to S (upto relabeling): (a) $((A, B), (C, D))$ and (b) $(((A, B), C), D)$. We will consider the first case here and refer the reader to [1] for the second case and other details.

In order to tackle case (a), we will use the notation from Figure 2. Let o_1, o_2 and o_3 be the common ancestors of (A, B), (C, D) and (A, C) respectively. Let $\mathcal{E}_{AB}$ be the event that the lineages corresponding to A and B coalesce in the segment (o_1, o_3) of the tree in Figure 2 and let $\overline{\mathcal{E}_{AB}}$ be the event that this does not occur. Similarly, we define the events $\mathcal{E}_{CD}$ and $\overline{\mathcal{E}_{CD}}$. To reduce notational clutter, for $w, v \in S$, we will write μ_{wv} to denote $\sum_{e\in\pi^S_{wv}}\mu_e\tau_e$. Now, for leaves $X, Y \in L$, let Z_{XY} denote the random quantity $\frac{1}{2}(\delta_{XY}-\mu_{XY})$, It is easy to check that the quantities $Z_{AB}-\mu_{o_1o_3} \mid \overline{\mathcal{E}_{AB}}$, $Z_{CD}-\mu_{o_2o_3} \mid \overline{\mathcal{E}_{AB}}$, Z_{AC}, and Z_{BD} have the same distribution. Let Z denote this common random variable; this is shown in Figure 2. Now, since $\delta_{AB}=\mu_{AB}+2Z_{AB}$, and since conditioned on $\mathcal{E}_{AB}$, $Z_{AB}\le\mu_{o_1o_3}$, we have that

$$\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right] \ge e^{-\frac{4}{3}\mu_{AB}}\Bigg\{\mathbb{E}\left[e^{-\frac{8}{3}\mu_{o_1o_3}}\middle|\mathcal{E}_{AB}\right]\mathbb{P}\left(\mathcal{E}_{AB}\right) + e^{-\frac{8}{3}\mu_{o_1o_3}}\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\Bigg\} \quad (6)$$

A similar calculation yields a lower bound for $\mathbb{E}\left[e^{-2\delta_{CD}}\right]$. On the other hand, notice that $\delta_{AC}=\mu_{AC}+2Z$ and $\delta_{BD}=\mu_{BD}+2Z$. Therefore, we have

$$\begin{aligned}&\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]}\\ &\ge\left[\frac{\mathbb{P}\left(\mathcal{E}_{AB}\right)}{\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]}+\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right]\times\left[\frac{\mathbb{P}\left(\mathcal{E}_{CD}\right)}{\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]}+\mathbb{P}\left(\overline{\mathcal{E}_{CD}}\right)\right]\\ &\ge\left[\frac{8}{3}\mu_L\left(1-e^{-f}\right)+1\right]^2, \qquad (7)\end{aligned}$$

where in the last step we have used the fact that the random variable Z stochastically dominates the random variable $\mu_L\tilde{Z}$, where $\tilde{Z}\sim\mathrm{Exp}(1)$ and that that $\mathbb{P}\left[\mathcal{E}_{XY}\right]\ge 1-e^{-f}$ for each pair of leaves $X, Y \in L$. For the second case, $(((A, B), C), D)$, one gets a lower bound $\left[\frac{8}{3}\mu_L(1-e^{-f})+1\right]$ instead of (7). Taking logarithms on either side gives us that $d_{AC}+d_{BD}\ge d_{AB}+d_{CD}+\log\left(\frac{8}{3}\mu_L\left(1-e^{-f}\right)+1\right)$. Using a similar procedure, one can show that $d_{AC}+d_{BD}=d_{AD}+d_{BC}$.

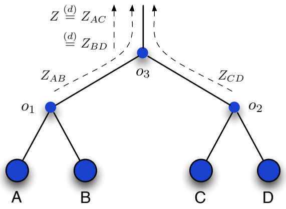

Fig. 2: Species tree restricted to A, B, C, D shown with the random variables used in Proof of Theorem 3

References

[1] Gautam Dasarathy, Robert Nowak, and Sebastien Roch. Data requirement for phylogenetic inference from multiple loci: A new distance method. *arXiv Preprint*, April 2014.

[2] James Degnan, Michael DeGiorgio, David Bryant, and Noah Rosenberg. Properties of consensus methods for inferring species trees from gene trees. *Systematic Biology*, 58(1):35–54, 2009.

[3] Peter Erdos, Michael Steel, László Székely, and Tandy Warnow. A few logs suffice to build (almost) all trees (i). *Random Structures and Algorithms*, 14(2):153–184, 1999.

[4] Joseph Felsenstein. *Inferring phylogenies*. Sinauer Associates, 2004.

[5] R. C. Griffiths and Simon Tavaré. Ancestral inference in population genetics. *Statistical Science*, pages 307–319, 1994.

[6] M Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Gen Biol*, 8(7):R141, 2007.

[7] Adam Leaché and Bruce Rannala. The accuracy of species tree estimation under simulation: a comparison of methods. *Systematic Biology*, 60(2):126–137, 2011.

[8] Liang Liu, Lili Yu, Laura Kubatko, Dennis Pearl, and Scott Edwards. Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, 53(1):320–328, 2009.

[9] Liang Liu, Lili Yu, and Dennis Pearl. Maximum tree: a consistent estimator of the species tree. *Journal of Mathematical Biology*, 60:95–106, 2010. 10.1007/s00285-009-0260-0.

[10] Liang Liu, Lili Yu, Dennis K Pearl, and Scott V Edwards. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5):468–477, 2009.

[11] Wayne Maddison. Gene trees in species trees. *Systematic biology*, 46(3):523–536, 1997.

[12] Elchanan Mossel and Yuval Peres. Information flow on trees. *The Annals of Applied Probability*, 13(3):817–844, 2003.

[13] Elchanan Mossel and Sebastien Roch. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(1):166–171, 2010.

[14] Luay Nakhleh. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in ecology & evolution*, 28(12):719–728, 2013.

[15] Bruce Rannala and Ziheng Yang. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4):1645–1656, 2003.

[16] Sebastien Roch. Toward extracting all phylogenetic information from matrices of evolutionary distances. *Science*, 327(5971):1376–1379, 2010.

[17] Sebastien Roch. An analytical comparison of multilocus methods under the multispecies coalescent: the three-taxon case. In *Pacific Symposium on Biocomputing*, pages 297–306. World Scientific, 2013.

[18] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[19] Charles Semple and Michael Steel. *Phylogenetics*, volume 24. Oxford University Press, 2003.

[20] Robert Sokal and Charles Michener. *A statistical method for evaluating systematic relationships*. University of Kansas, 1958.

[21] Michael Steel. A basic limitation on inferring phylogenies by pairwise sequence comparisons. *Journal of Theoretical Biology*, 256(3):467 – 472, 2009.

[22] Michael Steel and László Székely. Inverting random functions. II. Explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM J. Discrete Math.*, 15(4):562–575 (electronic), 2002.

[23] Simon Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17:57–86, 1986.

[24] Jimmy Yang and Tandy Warnow. Fast and accurate methods for phylogenomic analyses. *BMC bioinformatics*, 12(Suppl 9):S4, 2011.