Data Requirement for Phylogenetic Inference from Multiple Loci: A New Distance Method

Gautam Dasarathy[†], Robert Nowak[†], and Sebastien Roch[#]

[†] Wisconsin Institutes for Discovery

[#] Department of Mathematics

University of Wisconsin - Madison

Abstract

We consider the problem of estimating the evolutionary history of a set of species (phylogeny or species tree) from several genes. It is known 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 reconstruction algorithm that provably improves over all previous methods in a regime of interest.

Index Terms

phylogenetic inference, incomplete lineage sorting, multispecies coalescent, distance methods, sample complexity, molecular clock

1 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 [1]. The subject of this paper is an important source of such gene tree incongruence, known as *incomplete lineage sorting* (ILS), where two lineages fail to coalesce in their most recent common ancestral population. That failure may lead one of the lineages to first coalesce with a more distantly related population thereby producing a gene tree whose topology differs from the species tree that we are trying to estimate. Several species tree reconstruction methods have recently been developed that address ILS. See for instance [2], [3] and references therein. Many such methods rely on a statistical model known as the *multispecies coalescent* which, roughly speaking, generates gene trees by performing independent coalescent processes in each ancestral population and then assembling these together. This process is illustrated in Figure 1 below and explained in a little more detail in Section 2.2. For more background on phylogenetic inference and coalescent theory see, e.g., [4], [5], [6].

The accuracy of multiloci reconstruction methods has been evaluated empirically, for instance, in [7], [8]. The focus of this paper 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., [8], [9], [10], [11]. 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. [12] 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 [10], which is 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 [8], 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., [13], [14], [15]). Intuitively, for instance, the "minimum" used in GLASS may be significantly 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 reconstruct a single gene tree correctly with high probability, it is both necessary [16] and sufficient [17] for the sequence length k to scale as f^{-2}. Therefore, in light of this and the results in [12], one might expect that the total amount of data required, mk, 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 [6], where the mutation rates and population sizes are constant across the populations represented by the branches of 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).

2 PRELIMINARIES AND NOTATION

We will begin with a description of our modeling assumptions and introduce some notation that will be used throughout the paper.

2.1 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 isolated populations are represented by the size n leaf set L of this tree. The goal is to learn the structure of S. We assume that each branch $e \in E$ of the species tree corresponds to t_e generations of evolution and we assume that each generation in this branch has a population of size N_e. As is standard in coalescent theory, we will assign each branch $e \in E$, a length $\tau_e > 0$ in coalescent time units defined as $\tau_e \triangleq t_e/N_e$. The smallest branch length, $f \triangleq \min_e \tau_e$, will play an important role in our analysis and in particular, we will be interested in the case where f is very small. For a pair of vertices $X, Y \in V$, we will use $\pi^S_{XY} \subset E$ to denote the unique path connecting X and Y in S and τ_{XY} will denote the length of this path. Notice that $\{\tau_{AB}\}_{A,B\in L}$ forms a metric on the set L and such a metric that can be written as a sum of path lengths on a tree is called an *additive metric* (see e.g., [6]) with respect to that tree. If we additionally assume that the population sizes in each branch are equal to some constant N, then $\{\tau_{AB}\}_{A,B\in L}$ forms an ultrametric with respect to S, i.e., for any three leaves A, B, C such that S restricted to A, B, C has the topology $((A, B), C)$*, we have that

$$\tau_{AB} \leq \tau_{AC} = \tau_{BC}.$$

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

∗. We will sometimes find it useful to represent trees in the so called Newick Format. For instance, the Newick representations of the trees labelled Gene 1 and Gene 2 in Figure 1 are $((A, B), C)$ and $(A, (B, C))$, respectively.

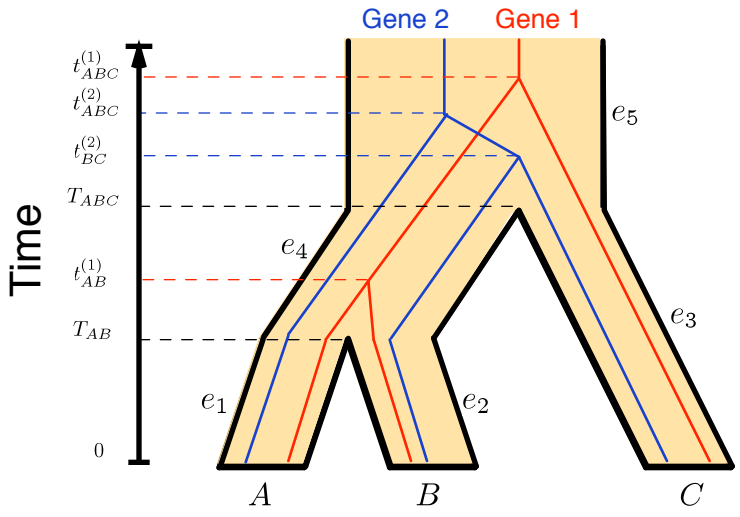

Fig. 1: A species tree (the thick, shaded tree) and two samples from the multispecies coalescent. Notice that while the topology of Gene 1 agrees with the species tree, the topology of Gene 2 does not.

$\mu_L \triangleq \min_{e\in E} \mu_e$ and $\mu_U \triangleq \max_{e\in E} \mu_e$ denote the smallest and largest mutation rates, respectively.

2.2 The Multispecies Coalescent and the Gene Trees

Following [18], we assume that a *multispecies coalescent* (MSC) process produces m (independent) random genealogies $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \ldots, \mathcal{G}^{(m)}$ 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 and in the case where the population size N_e in each branch $e \in E$ is a constant N. Consider the 3 species example of Figure 1, where the thick, shaded tree is the species tree S with edges $\{e_i\}_{i=1}^5$. As is standard in coalescent theory, we will think of time as running backwards, that is, 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) branch (or speciate). Let us first consider one random

draw from the MSC, i.e., the case of one particular gene, Gene 1. $A, B,$ and C each have a copy (or 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 (represented by the branch e_4), they have a chance to coalesce. According to the MSC, the coalescence happens after a random time drawn according to the $\text{Exp}(1)$ distribution, that is,

$$\mathbb{P}\left[t_{AB}^{(1)} - T_{AB} \geq x\right] = 1 - e^{-x}, \qquad x \geq 0. \tag{1}$$

Now, the coalesced A-B lineage and the lineage corresponding to C do not interact until time T_{ABC}, which is when they find themselves in a common population. They then coalesce at a random time $t_{ABC}^{(1)}$ which is again such that $t_{ABC}^{(1)} - T_{ABC} \sim \text{Exp}(1)$. This gives us a random gene tree with the topology $((A, B), C)$. To contrast with this, consider the case of Gene 2. Here, the lineages corresponding to the alleles in A and B do not coalesce in e_4 (since the randomly drawn coalescence time was more than the length of e_4). So, at time T_{ABC}, there are three lineages present in the branch e_5. When there are multiple lineages in the same population, according to the MSC, each pair independently coalesces again after a random time period drawn according to the $\text{Exp}(1)$ distribution. In this case, the genealogies of B and C alleles coalesce (at time $t_{BC}^{(2)}$) before A and B, thus giving us a second random tree with topology $(A, (B, C))$. Notice that while the genealogy (evolutionary history) of Gene 1 agrees with that of the species, the genealogy of Gene 2 does not. This is an example of incomplete lineage sorting which, as mentioned earlier, is a fundamental road block for learning the tree of life.

We refer the reader to [18] for more details on the multispecies coalescent but, we will state the model here for the sake of completeness. Before we proceed, we will record a simple fact about the exponential distribution: If $X_1, \ldots, X_p \overset{\text{iid}}{\sim}$

$\text{Exp}(1)$, then $\min_{i\in\{1,\ldots,p\}} X_i \sim \text{Exp}(p)$. This follows since

$$\mathbb{P}\left(\min_{i\in\{1,\ldots,p\}} X_i \geq t\right) = \prod_{i=1}^{p} P(X_i \geq t) = e^{-pt}. \tag{2}$$

The density of the likelihood of a gene tree $\mathcal{G}^{(i)} = \left(\mathcal{V}^{(i)}, \mathcal{E}^{(i)}\right)$ can now be written down as follows. We will focus our attention on the branch $e \in E$ of the species tree and for the gene tree $\mathcal{G}^{(i)}$, let $I_e^{(i)}$ and $O_e^{(i)}$ be the number of lineages entering and leaving the branch e respectively. For instance, consider Gene 1 in Figure 1. Here, two lineages enter the branch e_4 and one lineage leaves it. On the other hand, in the case of Gene 2 in Figure 1, two lineages enter the branch e_4 and two lineages leave it. Let $t_{e,s}^{(i)}, s = \left\{1, 2, \ldots, I_e^{(i)} - O_e^{(i)} + 1\right\}$ be the $s-$th coalescent time corresponding to $\mathcal{G}^{(i)}$ in the branch e. Recall that each pair of lineages in a population can coalesce at a random time drawn according to the $\text{Exp}(1)$ distribution independently of each other. Therefore, after the $(s-1)$-th coalescent event at time $t_{e,s-1}^{(i)}$, there are $I_e^{(i)} - s + 1$ surviving lineages in branch e and the likelihood that the $s-$th coalescence time in branch e is $t_{e,s}^{(i)}$ corresponds to the event that the minimum of $\binom{I_e^{(i)}-s+1}{2}$ random variables distributed according to $\text{Exp}(1)$ has the value $t_{e,s}^{(i)} - t_{e,s-1}^{(i)}$. Therefore using (2), the density of the likelihood of $\mathcal{G}^{(i)}$ can be written as

$$\prod_{e\in E} \prod_{s=1}^{I_e^{(i)}-O_e^{(i)}+1} \exp\left\{-\binom{I_e^{(i)} - s + 1}{2}\left[t_{e,s}^{(i)} - t_{e,s-1}^{(i)}\right]\right\}, \tag{3}$$

where, for convenience, we let $t_{e,0}^{(i)}$ and $t_{e,I_e^{(i)}-O_e^{(i)}+1}^{(i)}$ be respectively the divergence times of the population in e and of its parent population.

2.3 Observation Model and The Inference Problem

Much of the prior work on understanding the theoretical complexity of learning species trees from multiple loci (or gene trees) has focused on the case where 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 understanding the effect of errors in estimating the gene trees (see e.g.,

[13], [14], [15]). Our approach will be to take this error into account explicitly and in fact bypass the reconstruction of gene trees altogether.

We model the sample generation process according to the standard Jukes-Cantor (JC) model (see e.g., [6]). That is, 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 dependence on the length of $\tilde{e}$ we will make explicit below). Then, 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.

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

The goal then, is to learn the structure of S given the data $\{\chi^{ij}\}_{i \in [m], j \in [k]}$ which is an $n \times m \times k$ array composed of the characters $\{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$, where $\{\chi^{ij}\}_{j \in [k]}$ is the data generated from the random gene tree $\mathcal{G}^{(i)}$ according to the Jukes-Cantor model.

The Jukes-Cantor model was chosen because it lends itself to easy presentation. Since the techniques developed here are *distance-based*, all our results can be generalized to the more realistic Generalized Time-Reversible (GTR) model [19] using spectral techniques as in [20], [21].

3 Main Results

We now state the main results of the paper. First, we will deal with the case where the strong molecular clock [6] assumption holds. We will then turn our attention to the more general case that does away with this assumption.

3.1 The Molecular Clock Assumption Holds

Assuming that the molecular clock hypothesis holds is often unrealistic; it is equivalent to believing that all extant and ancestral populations have the same population size and that the mutations happen at the same rate through time and across populations. It has however proven to be a useful abstraction for developing powerful methods. In our setting, this is equivalent to assuming that for all $e \in E$, $\mu_e = \mu > 0$, and $N_e = N$, both constants independent of 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

$$\widehat{p}_{AB} = \frac{1}{mk} \sum_{i \in [m], j \in [k]} \mathbb{1}\{\chi_A^{ij} \neq \chi_B^{ij}\}, \tag{4}$$

which can be thought of as the normalized hamming distance between the concatenated molecular sequences corresponding to species A and B. Our first result, which is proved in Section 5.1, is that, in expectation, $\{\widehat{p}_{AB}\}_{A,B \in L}$ is not only a metric on L, but is in fact an ultrametric with respect to S.

Theorem 1. *$\{\mathbb{E}[\widehat{p}_{AB}]\}_{A,B \in L}$ 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*

$$\mathbb{E}[\widehat{p}_{AC}] = \mathbb{E}[\widehat{p}_{BC}] > \mathbb{E}[\widehat{p}_{AB}] + \frac{3e^{-\frac{4}{3}\mu\tau_{AC}}\mu}{8\mu + 3} f. \tag{5}$$

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], [6] for background on distance based methods). For the sake of simplicity, we may assume that we use the UPGMA algorithm [22], 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*

$$m \geq C_1(\mu, \Delta, n, \epsilon) \times f^{-2} \quad \textit{and} \quad k \geq 1, \tag{6}$$

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, which is proved in Section 5.2, 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 algorithm [8] except instead of using the average coalescent time as the distance measure, we use (4), which can be considered as the normalized hamming distance. It turns out that this modification is crucial to obtaining our improved sample complexity result.

3.2 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 and population size N_e.

First, we observe that $\{\mathbb{E}[\widehat{p}_{AB}]\}_{A,B\in L}$ 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., [6]). However, it is not immediately clear how to define a

metric that satisfies the 4-point condition in our setting. Our next result, which is arguably the most important contribution of this paper, shows that this can be done. As before, we will first consider an idealized measure of dissimilarity as follows:

$$d_{AB} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\mathbb{E}\left[\widehat{p}_{AB}\right]\right), A, B \in L,$$

where $\widehat{p}_{AB}$ is as defined in (4). Our next result, which parallels Theorem 1, shows that this "idealized" dissimilarity measure is actually an *additive metric* with respect to S. 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 the topology of S restricted to these 4 leaves is $((A, B), (C, D))$ or $(((A, B), C), D)$, the above distances satisfy

$$d_{AB} + d_{CD} \leq d_{AC} + d_{BD} = d_{AD} + d_{BC}.$$

See [6], for instance, for more information about tree metrics.

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*

$$d_{AC} + d_{BD} = d_{AD} + d_{BC} > d_{AB} + d_{CD} + \alpha_{\text{add}}, \tag{7}$$

where $\alpha_{\text{add}} = \frac{3}{4}\log\left(\frac{8}{3}\mu_L(1 - e^{-f}) + 1\right) > 0$ and $\mu_L \triangleq \min_{e\in E}\mu_e$ is the smallest mutations rate, as defined in Section 2.1.

It is somewhat surprising that this result is true. It tells us that if one ignores the fact that there are multiple loci and pretends as though all samples came from a single gene tree, then the gene tree estimated from this "concatenated molecular sequence" has the same topology as S. Furthermore, this result is also interesting since phylogenetic mixtures are known to cause problems for distance-based methods [23]. We prove Theorem 3 in Section 5.3.

In light of this, we propose the following algorithm to reconstruct S. First, we define the following sample-based *corrected* measure of dissimilarity (with

$\widehat{p}_{AB}$ as defined in (4))

$$\widehat{d}_{AB} \triangleq -\frac{3}{4}\log\left(1-\frac{4}{3}\widehat{p}_{AB}\right). \tag{8}$$

Now, use any quartet-test based algorithm (like Neighbor Joining [24]) which returns an additive tree using $\{\widehat{d}_{AB}\}_{A,B\in L}$ defined as in (8) as the input dissimilarity measure. We call this algorithm METAL (for Metric algorithm for Estimation of Trees based on Aggregation of Loci).

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 2.1). We then have the following result.

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

$$k \geq 1 \text{ and } m \geq \frac{e^{\frac{8\mu_U\Delta}{3}}(8\mu_U+3)^2(24+8\alpha_{\text{add}})^2}{162\alpha_{\text{add}}^2}\log\left(\frac{16\binom{n}{4}}{\epsilon}\right) \tag{9}$$

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

$$C_2(\mu_U,\mu_L,\Delta,n,\epsilon)\times f^{-2}, \text{ where } C_2(\mu_U,\mu_L,\Delta,n,\epsilon) = \frac{8e^{\frac{8\mu_U\Delta}{3}}(8\mu_U+3)^2}{9\mu_L^2}\log\left(\frac{16\binom{n}{3}}{\epsilon}\right).$$

Remark. Following [17], the diameter Δ can be replaced by the (often much smaller) *depth*[1] of the tree by employing a distance method that uses only those distances that are "small enough".

We prove Theorem 4 using arguments that are similar in spirit to those in the proof of Theorem 2. We refer the reader to Section 5.4 for the exact details.

Theorem 4 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

1. 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.

rate μ_e is constant across gene trees; generalizing this analysis to the case where the mutation rates are allowed to change is an interesting avenue for future work.

4 Discussion

Irrespective of the sequence length k of each gene, the number of genes m required needs to satisfy $m \in \Omega(f^{-1})$ for consistent species tree estimation. To see this, consider the species tree in Figure 1. Given m gene trees drawn according to the MSC based on this species tree, the probability that none of them have a coalescent event in branch e_4 is given by $e^{-m\tau_{e_4}}$ (this is the probability that m independent exponentials are bigger than τ_{e_4}). Therefore, if $m < \tau_{e_4}^{-1}$, then with probability greater than e^{-1}, none of the m the gene trees have a coalescence event in e_4, that is, there is no evidence for the existence of this branch from the sample. This argument can also be formalized by observing that any algorithm that is able to estimate S reliably should be able to perform a reliable hypothesis test between two shifted exponential distributions. Therefore, this result follows from the fact that $D_{\mathrm{KL}}\left(p(x;\tau_{AB}+f)\|p(x;\tau_{AB})\right) = f$, where $p(x;a) = e^{-(x-a)}\mathbb{1}\left\{x \geq a\right\}$ and $D_{KL}\left(\cdot\|\cdot\right)$ is the Kullback-Liebler divergence [25].

On the other hand, we know from [16] that even without the confounding effect of the multispecies coalescent, a total sequence length ($m \times k$) 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

$$m \geq C \max\left\{f^{-1}, \frac{f^{-2}}{k}\right\}. \tag{10}$$

As mentioned earlier, the results in this paper show that $m \in \mathcal{O}(f^{-2})$ is achievable irrespective of the value of k, i.e., 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 [10] (as shown in

[12]) 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 and in particular, is it possible to devise an algorithm that recovers S given $m \in o(f^{-2})$ when the sequence length, k, is moderate, say, $\mathcal{O}(f^{-1})$? (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 METAL meets the lower bound for small k?

5 Proofs of the Main Results

In this section we will prove the main results of the paper.

5.1 Proof of Theorem 1

Recall that for any pair of leaves $A, B \in L$, we define

$$\widehat{p}_{AB} = \frac{1}{mk} \sum_{i \in [m], j \in [k]} \mathbb{1}\{\chi_A^{ij} \neq \chi_B^{ij}\}. \tag{11}$$

Theorem 1. $\{\mathbb{E}[\widehat{p}_{AB}]\}_{A,B \in L}$† 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

$$\mathbb{E}[\widehat{p}_{AC}] = \mathbb{E}[\widehat{p}_{BC}] > \mathbb{E}[\widehat{p}_{AB}] + \frac{3e^{-\frac{4}{3}\mu\tau_{AC}}\mu f}{8\mu + 3}. \tag{12}$$

Proof: Suppose that $A, B, C \in L$ are three arbitrary leaves of the species tree with the topology $((A, B), C)$. By definition, we have that

$$\mathbb{E}[\widehat{p}_{AC}] = \mathbb{E}\left[\frac{3}{4}\left(1 - e^{-\frac{4}{3}\delta_{AC}}\right)\right],$$

where δ_{AC} is the distance between A and C on a random gene tree drawn according to the MSC. Notice that it satisfies $\delta_{AC} = \mu\tau_{AC} + 2\mu Z$ with $Z \sim \text{Exp}(1)$. Therefore, we have

†. Unless otherwise noted, expectations will be with respect all the randomness present.

$$\begin{aligned}\mathbb{E}\left[\widehat{p}_{AC}\right] - \mathbb{E}\left[\widehat{p}_{AB}\right] &= -\frac{3}{4}e^{-\frac{4}{3}\mu\tau_{AC}}\mathbb{E}\left[e^{-\frac{8}{3}\mu Z}\right] + -\frac{3}{4}e^{-\frac{4}{3}\mu\tau_{AB}}\mathbb{E}\left[e^{-\frac{8}{3}\mu Z}\right] \\ &\overset{(a)}{=} \frac{3\left(e^{-\frac{4}{3}\mu\tau_{AB}} - e^{-\frac{4}{3}\mu\tau_{AC}}\right)}{4(\frac{8}{3}\mu+1)} \\ &\overset{(b)}{\geq} \frac{3e^{-\frac{4}{3}\mu\tau_{AC}}\mu f}{(8\mu+3)},\end{aligned}$$

where (a) follows from the fact that if $X \sim \text{Exp}(1)$, for any $\alpha > 0$, $\mathbb{E}[e^{-\alpha X}] = (\alpha+1)^{-1}$ and (b) follows from observing that for any $\alpha > 0$ and $x < y$, we have

$$\frac{e^{-\alpha x}}{\alpha} - \frac{e^{-\alpha y}}{\alpha} = \int_x^y e^{-\alpha t}\, dt \geq (y-x)e^{-\alpha y}$$

Proceeding similarly, It can be seen that $\mathbb{E}\left[\widehat{p}_{AC}\right] = \mathbb{E}\left[\widehat{p}_{BC}\right]$. This concludes the proof. □

5.2 Proof of Theorem 2

We now prove Theorem 2 which guarantees that S can be reliably recovered by using a standard distance-based algorithm like UPGMA or bottom-up agglomerative clustering with $\{\widehat{p}_{AB}\}_{A,B\in L}$ as a dissimilarity measure for L.

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

$$m \geq C_1(\mu, \Delta, n, \epsilon) \times f^{-2} \quad \text{and} \quad k \geq 1, \tag{13}$$

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

Proof: Recall that the algorithm we propose to recover the tree uses $\{\hat{p}_{AB}\}_{A,B\in L}$ as a dissimilarity measure and uses an agglomerative clustering algorithm. Therefore, this procedure errs if for any triple of leaves A, B, C which have the topology $((A,B),C)$ with respect to S, either $\widehat{p}_{AB} > \widehat{p}_{AC}$ or $\widehat{p}_{AB} > \widehat{p}_{BC}$.

Letting $\binom{L}{3}$ denote the set of all unordered triples in L, we can use the union bound to over-estimate the probability of error with the following:

$$\mathbb{P}\left[\bigcup_{((A,B),C)\in\binom{L}{3}}\left\{\text{The triple } ((A,B),C) \text{ is such that } \widehat{p}_{AB} > \widehat{p}_{AC} \text{ or } \widehat{p}_{AB} > \widehat{p}_{BC}\right\}\right]$$
$$\leq \sum_{((A,B),C)\in\binom{L}{3}} \mathbb{P}\left[\widehat{p}_{AB} > \widehat{p}_{AC}\right] + \mathbb{P}\left[\widehat{p}_{AB} > \widehat{p}_{BC}\right]. \tag{14}$$

We will now upper bound the term $\mathbb{P}\left[\widehat{p}_{AB} > \widehat{p}_{AC}\right]$, the other term will satisfy the same upper bound. Defining $\alpha_{\text{um}} = \frac{3e^{-\frac{4}{3}\Delta}\mu f}{(8\mu+3)}$, for an arbitrary triple $((A,B),C)$ we have

$$\begin{aligned}\mathbb{P}\left[\widehat{p}_{AB} - \widehat{p}_{AC} > 0\right] &= \mathbb{P}\left[\widehat{p}_{AB} - \mathbb{E}\left[\widehat{p}_{AB}\right] - \widehat{p}_{AC} + \mathbb{E}\left[\widehat{p}_{AC}\right] > \mathbb{E}\left[\widehat{p}_{AC}\right] - \mathbb{E}\left[\widehat{p}_{AB}\right]\right] \\ &\overset{(a)}{\leq} \mathbb{P}\left[\widehat{p}_{AB} - \mathbb{E}\left[\widehat{p}_{AB}\right] - \widehat{p}_{AC} + \mathbb{E}\left[\widehat{p}_{AC}\right] > \alpha_{\text{um}}\right] \\ &\leq \mathbb{P}\left[\widehat{p}_{AB} - \mathbb{E}\left[\widehat{p}_{AB}\right] > \frac{\alpha_{\text{um}}}{2}\right] + \mathbb{P}\left[\mathbb{E}\left[\widehat{p}_{AC}\right] - \widehat{p}_{AC} > \frac{\alpha_{\text{um}}}{2}\right], \end{aligned} \tag{15}$$

where (a) follows from Theorem 1. Let us first look at the first term in (15). The second one will follow similarly.

$$\begin{aligned}&\mathbb{P}\left[\widehat{p}_{AB} - \mathbb{E}[p_{AB}] > \alpha_{\text{um}}/2\right] \\ &\overset{(a)}{=} \mathbb{E}\left[\mathbb{P}\left(\widehat{p}_{AB} - \frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} + \frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} - \mathbb{E}\left[\widehat{p}_{AB}\right] > \frac{\alpha_{\text{um}}}{2}\middle|\{\delta_{AB}^{(i)}\}_{i\in[m]}\right)\right] \\ &\leq \mathbb{E}\left[\mathbb{P}\left(\widehat{p}_{AB} - \frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} > \frac{\alpha_{\text{um}}}{4}\middle|\{\delta_{AB}^{(i)}\}_{i\in[m]}\right)\right] + \mathbb{P}\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} - \mathbb{E}\left[\widehat{p}_{AB}\right] > \frac{\alpha_{\text{um}}}{4}\right). \end{aligned} \tag{16}$$

In (a), we condition on $\{\delta_{AB}^{(i)}\}_{i\in[m]}$, where $\delta_{AB}^{(i)}$, as before, is the random distance between the leaves A and B on the gene tree $\mathcal{G}^{(i)}$. We then add and subtract $\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)}$, where $p_{AB}^{(i)} \triangleq \frac{3}{4}\left(1 - e^{-\frac{4}{3}\delta_{AB}^{(i)}}\right)$. The next inequality follows from a union bound. The two terms in the above inequality can now be upper bounded

using Hoeffding's inequality:

$$\mathbb{E}\left[\mathbb{P}\left[\frac{1}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k}X_{AB}^{ij}-\frac{1}{m}\sum_{i=1}^{m}p_{AB}^{(i)}>\frac{\alpha_{\text{um}}}{4}\middle|\left\{d_{AB}^{(i)}\right\}\right]\right]\leq e^{-mk\alpha_{\text{um}}^2/16}. \tag{17}$$

$$\mathbb{P}\left(\frac{1}{m}\sum_{i\in[m]}p_{AB}^{(i)}-\mathbb{E}\left[\widehat{p}_{AB}\right]>\frac{\alpha_{\text{um}}}{4}\right)\leq e^{-m\alpha_{\text{um}}^2/16}. \tag{18}$$

These inequalities follow since $\mathbb{E}\left[X_{AB}^{ij}\middle|\delta_{AB}^{(i)}\right]=p_{AB}^{(i)}$ and $\mathbb{E}\left[p_{AB}^{(i)}\right]=\mathbb{E}\left[\widehat{p}_{AB}\right]$.

Substituting these in (14), we have

$$\begin{aligned}\mathbb{P}\left[\text{Error}\right]&\leq\sum_{((AB)C)\in\binom{L}{3}}\mathbb{P}\left[\widehat{p}_{AB}>\widehat{p}_{AC}\right]+\mathbb{P}\left[\widehat{p}_{AB}>\widehat{p}_{BC}\right]\\&\leq\sum_{((AB)C)\in\binom{L}{3}}4\left(e^{-mk\alpha_{\text{um}}^2/16}+e^{-m\alpha_{\text{um}}^2/16}\right)\\&\leq\binom{n}{3}4\left(e^{-mk\alpha_{\text{um}}^2/16}+e^{-m\alpha_{\text{um}}^2/16}\right)\end{aligned}$$

Therefore, the probability of error can be made less than ϵ if we pick m and k as shown in (6) or (13). □

5.3 Proof of Theorem 3

Recall that we define $d_{AB}=-\frac{3}{4}\log\left(1-\frac{4}{3}\mathbb{E}\left[\widehat{p}_{AB}\right]\right)$ and Theorem 3, which we will prove now, tells us that these distances form an additive metric with respect to S.

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

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

where $\alpha_{\text{add}}=\frac{3}{4}\log\left(\frac{8}{3}\mu_L(1-e^{-f})+1\right)>0$.

Proof: We will first show that for any 4 leaves $A,B,C,D\in L$ that are such that either $((A,B),(C,D))$ or $(((A,B),C,D))$ holds with respect to S, then $d_{AC}+d_{BD}>d_{AB}+d_{CD}+\alpha_{\text{add}}$. Using similar techniques, we will next establish that $d_{AC}+d_{BD}=d_{AB}+d_{CD}$.

We begin by observing that by definition,

$$d_{AC} + d_{BD} - d_{AB} - d_{CD} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\mathbb{E}\left[\widehat{p}_{AC}\right]\right) - \frac{3}{4}\log\left(1 - \frac{4}{3}\mathbb{E}\left[\widehat{p}_{BD}\right]\right) + \frac{3}{4}\log\left(1 - \frac{4}{3}\mathbb{E}\left[\widehat{p}_{AB}\right]\right) + \frac{3}{4}\log\left(1 - \frac{4}{3}\mathbb{E}\left[\widehat{p}_{AB}\right]\right) \quad (19)$$

$$= \frac{3}{4}\log\left(\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]}\right), \quad (20)$$

where the expectations in the last equation are with respect to the multispecies coalescent and the δ's are the random gene tree distances as defined in Section 2.3.

We will prove this theorem by lower bounding the quantity $\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]}$ appropriately. Towards this end, we 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 each case separately and bound the above quantity in what follows.

Case (a): $((A, B), (C, D))$ In order to tackle the first case, we will use the notation from Figure 2a below, which shows the species tree S restricted to the leaves A, B, C, D. 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 2a 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 $\frac{1}{2}(\delta_{XY} - \mu_{XY})$, i.e., it is the effective (mutation rate adjusted) coalescent time after the lineages corresponding to X and Y find themselves in a common population.

By the memoryless property of the exponential distribution, it is easy to check that $Z_{AB} - \mu_{o_1o_3}$ conditioned on $\overline{\mathcal{E}_{AB}}$ has the same distribution as $Z_{CD} - \mu_{o_2o_3}$ conditioned on $\overline{\mathcal{E}_{CD}}$. Let Z denote be a random variable with this common distribution. Also observe that Z_{AC} and Z_{BD} have the same distribution as Z. This is depicted diagrammatically in Figure 2a.

Now, using the fact that by definition, $\delta_{AB} = \mu_{AB} + 2Z_{AB}$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right] &= e^{-\frac{4}{3}\mu_{AB}}\mathbb{E}\left[e^{-\frac{8}{3}Z_{AB}}\right] \\
&= e^{-\frac{4}{3}\mu_{AB}}\left\{\mathbb{E}\left[e^{-\frac{8}{3}Z_{AB}}\middle|\mathcal{E}_{AB}\right]\mathbb{P}\left(\mathcal{E}_{AB}\right) + \mathbb{E}\left[e^{-\frac{8}{3}Z_{AB}}\middle|\overline{\mathcal{E}_{AB}}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right\} \\
&\overset{(a)}{\geq} e^{-\frac{4}{3}\mu_{AB}}\left\{e^{-\frac{8}{3}\mu_{o_1o_3}}\mathbb{P}\left(\mathcal{E}_{AB}\right) + e^{-\frac{8}{3}\mu_{o_1o_3}}\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right\} \\
&= e^{-\frac{4}{3}\left(\mu_{AB}+2\mu_{o_1o_3}\right)}\left\{\mathbb{P}\left(\mathcal{E}_{AB}\right) + \mathbb{E}\left[e^{-\frac{8}{3}Z}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right\}, \qquad (21)
\end{aligned}$$

where (a) follows from the fact that conditioned on $\mathcal{E}_{AB}$, $Z_{AB} \leq \mu_{o_1o_3}$ and that conditioned on $\overline{\mathcal{E}_{AB}}$, $Z_{AB} \overset{d}{=} Z + \mu_{o_1o_3}$. Similarly, we get the following lower bound corresponding to the leaves C, D.

$$\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right] \geq e^{-\frac{4}{3}\left(\mu_{CD}+2\mu_{o_2o_3}\right)}\left\{\mathbb{P}\left(\mathcal{E}_{CD}\right) + \mathbb{E}\left[e^{-\frac{8}{3}Z}\right]\mathbb{P}\left(\overline{\mathcal{E}_{CD}}\right)\right\} \qquad (22)$$

On the other hand, notice that $\delta_{AC} = \mu_{AC} + 2Z_{AC} \overset{d}{=} \mu_{AC} + 2Z$ and $\delta_{BD} = \mu_{BD} + 2Z_{BD} \overset{d}{=} \mu_{BD} + 2Z$. Therefore, we have

$$\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right] = e^{-\frac{4}{3}\mu_{AC}}\mathbb{E}\left[e^{-\frac{8}{3}Z}\right], \qquad \text{and} \qquad \mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right] = e^{-\frac{4}{3}\mu_{BD}}\mathbb{E}\left[e^{-\frac{8}{3}Z}\right], \quad (23)$$

From equations (21) - (23), we have

$$\begin{aligned}
&\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]} \times \frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]} \\
&\qquad \geq \frac{e^{-\frac{4}{3}\left(\mu_{AB}+2\mu_{o_1o_3}\right)}\left\{\mathbb{P}\left(\mathcal{E}_{AB}\right) + \mathbb{E}\left[e^{-\frac{8}{3}Z}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right\}}{e^{-\frac{4}{3}\mu_{AC}}\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]} \\
&\qquad\qquad \times \frac{e^{-\frac{4}{3}\left(\mu_{CD}+2\mu_{o_2o_3}\right)}\left\{\mathbb{P}\left(\mathcal{E}_{CD}\right) + \mathbb{E}\left[e^{-\frac{8}{3}Z}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right\}}{e^{-\frac{4}{3}\mu_{BD}}\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]} \qquad (24) \\
&\qquad \overset{(a)}{=} \frac{\left\{\mathbb{P}\left(\mathcal{E}_{AB}\right) + \mathbb{E}\left[e^{-\frac{8}{3}Z}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right\}\left\{\mathbb{P}\left(\mathcal{E}_{CD}\right) + \mathbb{E}\left[e^{-\frac{8}{3}Z}\right]\mathbb{P}\left(\overline{\mathcal{E}_{CD}}\right)\right\}}{\left(\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]\right)^2} \\
&\qquad = \left[\frac{\mathbb{P}\left(\mathcal{E}_{AB}\right)}{\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]} + \mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right] \times \left[\frac{\mathbb{P}\left(\mathcal{E}_{CD}\right)}{\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]} + \mathbb{P}\left(\overline{\mathcal{E}_{CD}}\right)\right] \qquad (25)
\end{aligned}$$

where in (a), we have used the fact that $\mu_{AB} + \mu_{CD} + 2\mu_{o_1o_3} + 2\mu_{o_2o_3} = \mu_{AC} + \mu_{BD}$ and in the last step we divide each term in the numerator by $\mathbb{E}\left[e^{-\frac{8}{3}Z}\right]$.

(a) Case (a) (b) Case (b)

Fig. 2: Pictures showing the random variables and internal nodes used in Proof of Theorem 3

Next, observe that Z stochastically dominates the random variable $\mu_L \tilde{Z}$, where $\tilde{Z} \sim \text{Exp}(1)$. Therefore, we have

$$\mathbb{E}\left[e^{-\frac{8}{3}Z}\right] \leq \mathbb{E}\left[e^{-\frac{8}{3}\mu_L \tilde{Z}}\right] = \frac{1}{\frac{8}{3}\mu_L + 1}. \tag{26}$$

Substituting this in (25) gives us

$$\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]} \geq \left[\frac{8}{3}\mu_L \mathbb{P}\left(\mathcal{E}_{AB}\right) + 1\right] \times \left[\frac{8}{3}\mu_L \mathbb{P}\left(\mathcal{E}_{CD}\right) + 1\right] \tag{27}$$

Finally, we observe that the probability that the event $\mathcal{E}_{AB}$ occurs is given by $1 - e^{-\tau_{o_1 o_3}}$, where $\tau_{o_1 o_3}$ is the length of the path (o_1, o_3) in the species tree; this follows from the memoryless property of the exponential distribution. Since $\tau_{o_1 o_3} \geq f$, we have that $\mathbb{P}\left(\mathcal{E}_{AB}\right) \geq 1 - e^{-f}$, and similarly $\mathbb{P}\left(\mathcal{E}_{CD}\right) \geq 1 - e^{-f}$. Substituting this in (27), we get the following lower bound

$$\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]} \geq \left[\frac{8}{3}\mu_L\left(1 - e^{-f}\right) + 1\right]^2 \tag{28}$$

Next, we consider Case (b).

Case (b) : $(((A, B), C), D)$ Here, we will write o_1, o_2, o_3 to denote the most recent common ancestors of (A, B), (A, C) and (A, D) respectively. Again we will use notation from the previous case for random variables of the form $Z_{XY}, X, Y \in L$.

In this case, we let $\mathcal{E}_{AB}$ denote the event that the lineages corresponding to A and B coalesce in the branch (o_1, o_2) in Figure 2b. Again, from the memoryless property, it can be seen that the random variable $Z_{AB} - \mu_{o_1o_2}$ conditioned on $\overline{\mathcal{E}_{AB}}$ and the random variable Z_{AC} have the same distribution; we let Z_1 denote a random variable with this common distribution. Similarly Z_{CD} and Z_{BD} have the same distribution and we let Z_2 denote a random variable with this distribution.

Reasoning as before, we see that since $\delta_{AB} = \mu_{AB} + 2Z_{AB}$,

$$\begin{aligned}\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right] &= e^{-\frac{4}{3}\mu_{AB}}\left\{\mathbb{E}\left[e^{-\frac{8}{3}Z_{AB}}\middle|\mathcal{E}_{AB}\right]\mathbb{P}\left(\mathcal{E}_{AB}\right) + \mathbb{E}\left[e^{-\frac{8}{3}Z_{AB}}\middle|\overline{\mathcal{E}_{AB}}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right\} \\ &\overset{(a)}{\geq} e^{-\frac{4}{3}\mu_{AB}}\left\{e^{-\frac{8}{3}\mu_{o_1o_2}}\mathbb{P}\left(\mathcal{E}_{AB}\right) + e^{-\frac{8}{3}\mu_{o_1o_2}}\mathbb{E}\left[e^{-\frac{8}{3}Z_1}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right\} \\ &= e^{-\frac{4}{3}\left(\mu_{AB}+2\mu_{o_1o_2}\right)}\left\{\mathbb{P}\left(\mathcal{E}_{AB}\right) + \mathbb{E}\left[e^{-\frac{8}{3}Z_1}\right]\mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right\}, \end{aligned} \tag{29}$$

where, as before, (a) follows from the fact that conditioned on $\mathcal{E}_{AB}$, $Z_{AB} \leq \mu_{o_1o_2}$ and that conditioned on $\overline{\mathcal{E}_{AB}}$, $Z_{AB} \overset{d}{=} Z_1 + \mu_{o_1o_2}$. On the other hand, we have

$$\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right] = e^{-\frac{4}{3}\mu_{CD}}\mathbb{E}\left[e^{-\frac{8}{3}Z_2}\right] \tag{30}$$

$$\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right] = e^{-\frac{4}{3}\mu_{AC}}\mathbb{E}\left[e^{-\frac{8}{3}Z_1}\right] \tag{31}$$

$$\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right] = e^{-\frac{4}{3}\mu_{BD}}\mathbb{E}\left[e^{-\frac{8}{3}Z_2}\right]. \tag{32}$$

Therefore, from (29)-(32), we have that

$$\begin{aligned}\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]} &\geq e^{-\frac{4}{3}\left(\mu_{AB}+\mu_{CD}+2\mu_{o_1o_2}-\mu_{AC}-\mu_{BD}\right)}\left(\frac{1}{\mathbb{E}\left[e^{-\frac{8}{3}Z_1}\right]}\mathbb{P}\left(\mathcal{E}_{AB}\right) + \mathbb{P}\left(\overline{\mathcal{E}_{AB}}\right)\right) \\ &= \frac{\mathbb{P}\left[\mathcal{E}_{AB}\right]}{\mathbb{E}\left[e^{-\frac{8}{3}Z_1}\right]} + \mathbb{P}\left[\overline{\mathcal{E}_{AB}}\right] \end{aligned} \tag{33}$$

where the second step follows from the fact that $\mu_{AB}+\mu_{CD}+2\mu_{o_1o_2} = \mu_{AC}+\mu_{BD}$. Finally, as in case (a), we use the bounds $\mathbb{E}\left[e^{-\frac{8}{3}Z_1}\right] \leq \frac{1}{\frac{8}{3}\mu_L+1}$ and that $\mathbb{P}\left[\mathcal{E}_{AB}\right] \geq 1 - e^{-f}$ to get the following lower bound.

$$\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]} \geq \frac{8}{3}\mu_L(1 - e^{-f}) + 1 \tag{34}$$

Since $\left(\frac{8}{3}\mu_L(1-e^{-f})+1\right) \geq 1$, from (28) and (34), we have that for any 4 leaves A, B, C, D such that the species tree S restricted to these four leaves satisfies either $((A,B),(C,D))$ or $(((A,B),C),D)$, then

$$\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AB}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{CD}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]} \geq \frac{8}{3}\mu_L(1-e^{-f})+1 \tag{35}$$

Substituting this lower bound in (20), we get the result that for any 4 leaves $A, B, C, D \in L$ that are such that $((A,B),(C,D))$ or $(((A,B),C,D))$ holds with respect to S, we have that $d_{AC} + d_{BD} > d_{AB} + d_{CD} + \alpha_{\text{add}}$, where $\alpha_{\text{add}} = \frac{3}{4}\log\left(\frac{8}{3}\mu_L(1-e^{-f})+1\right)$.

To conclude the proof, we will next establish the "equality part" of the theorem. As in (20), notice that the following holds.

$$d_{AC} + d_{BD} - d_{AD} - d_{BC} = \frac{3}{4}\log\left(\frac{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AD}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BC}}\right]}{\mathbb{E}\left[e^{-\frac{4}{3}\delta_{AC}}\right]\mathbb{E}\left[e^{-\frac{4}{3}\delta_{BD}}\right]}\right). \tag{36}$$

Again, we will divide this proof into two cases as above.

Case (a): $((A,B),(C,D))$ Observe that the following hold with μ_{XY} and Z as defined before (cf. Fig 2a)

$$\delta_{AD} = \mu_{AD} + 2Z, \quad \delta_{BC} = \mu_{BC} + 2Z,$$
$$\delta_{AC} = \mu_{AC} + 2Z, \quad \delta_{BD} = \mu_{BD} + 2Z$$

Substituting these in (36) and observing that $\mu_{AD} + \mu_{BC} = \mu_{AC} + \mu_{BD}$ tells us that $d_{AC} + d_{BD} = d_{AD} + d_{BC}$ in case (a).

Case (b): $(((A,B),C),D)$ In this case, observe that the following hold again with μ_{XY} and Z_1 and Z_2 as defined earlier (cf. Fig 2(b)):

$$\delta_{AD} = \mu_{AD} + 2Z_2, \quad \delta_{BC} = \mu_{BC} + 2Z_1$$
$$\delta_{AC} = \mu_{AC} + 2Z_1, \quad \delta_{BD} = \mu_{BD} + 2Z_2$$

Again, substituting these in (36) and observing that $\mu_{AD} + \mu_{BC} = \mu_{AC} + \mu_{BD}$ tells us that $d_{AC} + d_{BD} = d_{AD} + d_{BC}$ in case (b) as well. This concludes the proof.

□

5.4 Proof of Theorem 4

We will now prove the last main result in our paper that shows that Theorem 3 can be used to design a tree reconstruction algorithm when one only has access to molecular data and also provides sample complexity results for this algorithm. Recall that we propose the following measure of dissimilarity from the samples

$$\widehat{d}_{AB} \triangleq -\frac{3}{4}\log\left(1-\frac{4}{3}\widehat{p}_{AB}\right). \tag{37}$$

where $\widehat{p}_{AB}$ is as defined in (4).

In light of Theorem 3, we proposed the following tree reconstruction procedure, which we call METAL: use any distance algorithm (like Neighbor Joining [24]) which returns an additive tree using $\{\widehat{d}_{AB}\}_{A,B\in L}$ as the dissimilarity measure. 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

$$k \geq 1 \text{ and } m \geq \frac{e^{\frac{8\mu_U\Delta}{3}}(8\mu_U+3)^2(24+8\alpha_{\text{add}})^2}{162\alpha_{\text{add}}^2}\log\left(\frac{16\binom{n}{4}}{\epsilon}\right) \tag{38}$$

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

$$C_2(\mu_U,\mu_L,\Delta,n,\epsilon)\times f^{-2}, \text{ where } C_2(\mu_U,\mu_L,\Delta,n,\epsilon) = \frac{8e^{\frac{8\mu_U\Delta}{3}}(8\mu_U+3)^2}{9\mu_L^2}\log\left(\frac{16\binom{n}{3}}{\epsilon}\right).$$

Proof: Notice that the above algorithm makes an error only if there exists a set of four leaves A,B,C,D such that $\tau_{AB}+\tau_{CD} \leq \tau_{AC}+\tau_{BD} = \tau_{AD}+\tau_{BC}$, but the 4-point condition is not satisfied by $\widehat{d}$, that is:

$$\widehat{d}_{AB}+\widehat{d}_{CD}-\widehat{d}_{AC}-\widehat{d}_{BD} > 0 \quad \text{or} \quad \widehat{d}_{AB}+\widehat{d}_{CD}-\widehat{d}_{AD}-\widehat{d}_{BC} > 0$$

Therefore, using the union bound, the probability of error can be upper bounded

as follows:

$$\mathbb{P}\left(\text{Error}\right) \leq \sum_{\substack{A,B,C,D\in L:\\ \tau_{AB}+\tau_{CD}\leq\tau_{AC}+\tau_{BD}=\tau_{AD}+\tau_{BC}}} \mathbb{P}\left[\hat{d}_{AB}+\hat{d}_{CD}-\hat{d}_{AC}-\hat{d}_{BD}>0\right] + \mathbb{P}\left[\hat{d}_{AB}+\hat{d}_{CD}-\hat{d}_{AD}-\hat{d}_{BC}>0\right] \tag{39}$$

We will bound the first term inside the summation of (39) and the second one will follow similarly. Setting $\alpha_{\text{add}} \triangleq \frac{3}{4}\log\left(\frac{8}{3}\mu_L(1-e^{-f})+1\right)$, observe that for a quadruple of leaves A, B, C, D such that $\tau_{AB}+\tau_{CD} \leq \tau_{AC}+\tau_{BD} = \tau_{AD}+\tau_{BC}$, we have

$$\begin{aligned}
&\mathbb{P}\left[\widehat{d}_{AB}+\widehat{d}_{CD}-\widehat{d}_{AC}-\widehat{d}_{BD}>0\right] \\
&\qquad = \mathbb{P}\Big[\widehat{d}_{AB}-d_{AB}+\widehat{d}_{CD}-d_{CD}-\widehat{d}_{AC}+d_{AC}-\widehat{d}_{BD}+d_{BD} \\
&\qquad\qquad > d_{AC}+d_{BD}-d_{AB}-d_{CD}\Big] \\
&\qquad \leq \mathbb{P}\Big[\widehat{d}_{AB}-d_{AB}+\widehat{d}_{CD}-d_{CD}-\widehat{d}_{AC}+d_{AC}-\widehat{d}_{BD}+d_{BD}>\alpha_{\text{add}}\Big],
\end{aligned}$$

where the second inequality follows from Theorem 3 which says that $d_{AC}+d_{BD}-d_{AB}-d_{CD}>\alpha_{\text{add}}$. We will again use the union bound to get

$$\begin{aligned}
&\mathbb{P}\left[\widehat{d}_{AB}+\widehat{d}_{CD}-\widehat{d}_{AC}-\widehat{d}_{BD}>0\right] \\
&\qquad \leq \mathbb{P}\Big[\widehat{d}_{AB}-d_{AB}+\widehat{d}_{CD}-d_{CD}-\widehat{d}_{AC}+d_{AC}-\widehat{d}_{BD}+d_{BD}>\alpha_{\text{add}}\Big] \\
&\qquad \leq \mathbb{P}\left[\widehat{d}_{AB}-d_{AB}>\frac{\alpha_{\text{add}}}{4}\right]+\mathbb{P}\left[\widehat{d}_{CD}-d_{CD}>\frac{\alpha_{\text{add}}}{4}\right] \\
&\qquad\qquad +\mathbb{P}\left[d_{AC}-\widehat{d}_{AC}>\frac{\alpha_{\text{add}}}{4}\right]+\mathbb{P}\left[d_{BD}-\widehat{d}_{BD}>\frac{\alpha_{\text{add}}}{4}\right].
\end{aligned} \tag{40}$$

To proceed, we will focus our attention on the first term in (40). The remaining terms will follow similarly. For notational clarity, let us define the function $\ell(x) \triangleq -\frac{3}{4}\log\left(1-\frac{4}{3}x\right)$ and let $p_{AB}^{(i)}$ denote the random quantity $\frac{3}{4}\left(1-e^{-\frac{4}{3}\delta_{AB}^{(i)}}\right) = \ell^{-1}\left(\delta_{AB}^{(i)}\right)$, where, as usual, $\delta_{AB}^{(i)}$ is the distances between A and B on the random gene tree $\mathcal{G}^{(i)}$ drawn according to the MSC. Now, observe that, by definition, $\widehat{d}_{AB}$ and d_{AB} are equal to $\ell(\widehat{p}_{AB})$ and $\ell(\mathbb{E}\left[\widehat{p}_{AB}\right])$ respectively.

Our strategy will be to first show that with high probability $\widehat{p}_{AB}$ is close to $\frac{1}{m}\sum_{i=1}^{m} p_{AB}^{(i)}$ which is in turn close to $\mathbb{E}\left[\widehat{p}_{AB}\right]$. We will then use the fact that $\ell(x)$ is a well-behaved function to obtain an upper bound on the the first term of (40).

Conditioned on a particular realization of the MSC process $\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}$, let $\mathcal{E}_1(\xi)$ and $\mathcal{E}_2(\xi)$ denote the events that $\left|\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} - \mathbb{E}\widehat{p}_{AB}\right| > \xi$ and $\left|\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} - \widehat{p}_{AB}\right| > \xi$, respectively. Now, notice we can bound the first term in (40) as follows.

$$\begin{aligned}
\mathbb{P}\left[\widehat{d}_{AB} - d_{AB} > \frac{\alpha_{\text{add}}}{4}\right] &= \mathbb{P}\left[\ell(\widehat{p}_{AB}) - \ell(\mathbb{E}\widehat{p}_{AB}) > \frac{\alpha_{\text{add}}}{4}\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\mathbb{P}\left[\ell(\widehat{p}_{AB}) - \ell(\mathbb{E}\widehat{p}_{AB}) > \frac{\alpha_{\text{add}}}{4}\middle|\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}\right]\right] \\
&\overset{(b)}{\leq} \mathbb{E}\left[\mathbb{P}\left[\ell(\widehat{p}_{AB}) - \ell(\mathbb{E}\widehat{p}_{AB}) > \frac{\alpha_{\text{add}}}{4}\middle|\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}, \mathcal{E}_1(\xi)^c, \mathcal{E}_2(\xi)^c\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{P}\left(\mathcal{E}_1(\xi)\middle|\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}\right)\right] + \mathbb{E}\left[\mathbb{P}\left(\mathcal{E}_2(\xi)\middle|\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}\right)\right],
\end{aligned} \tag{41}$$

where in (a) we condition on $\left\{\delta_{AB}^{(i)}\right\}$, a particular realization of the MSC. In (b) we use the following fact: for any three events $\mathcal{E}_a, \mathcal{E}_b, \mathcal{E}_c$, the following inequality holds

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_a) &= \mathbb{P}(\mathcal{E}_a|\mathcal{E}_b \cup \mathcal{E}_c)\mathbb{P}\left(\mathcal{E}_b \cup \mathcal{E}_c\right) + \mathbb{P}(\mathcal{E}_a|\mathcal{E}_b^c \cap \mathcal{E}_c^c)\mathbb{P}\left(\mathcal{E}_b^c \cap \mathcal{E}_c^c\right) \\
&\leq \mathbb{P}(\mathcal{E}_b \cup \mathcal{E}_c) + \mathbb{P}(\mathcal{E}_a|\mathcal{E}_b^c \cap \mathcal{E}_c^c) \\
&\leq \mathbb{P}\left(\mathcal{E}_b\right) + \mathbb{P}\left(\mathcal{E}_c\right) + \mathbb{P}(\mathcal{E}_a|\mathcal{E}_b^c \cap \mathcal{E}_c^c),
\end{aligned}$$

where we identify $\mathcal{E}_a, \mathcal{E}_b$, and $\mathcal{E}_c$ with the events $\widehat{d}_{AB} - d_{AB} > \frac{\alpha_{\text{add}}}{4}, \mathcal{E}_1(\xi)$, and $\mathcal{E}_2(\xi)$ respectively. Our goal now is to pick a value of ξ so that the first term in (41) is 0. Towards this end, we will use the following result that we prove in Section 5.5.

Claim 1. *For any $\xi > 0$, conditioned on a particular realization $\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}$ of the*

MSC process, and the events $\mathcal{E}_1(\xi)^c$ *and* $\mathcal{E}_2(\xi)^c$*, the following inequality holds*

$$|\ell(\widehat{p}_{AB}) - \ell(\mathbb{E}\widehat{p}_{AB})| \leq \frac{2\xi}{\frac{e^{-4\mu_U \Delta/3}}{\frac{8}{3}\mu_U+1} - \frac{8\xi}{3}}. \tag{42}$$

Now, Claim 1 tells us that if we make the following choice for ξ

$$\xi = \xi_0 \triangleq \frac{9\alpha_{\text{add}} e^{-\frac{4}{3}\mu_U \Delta}}{(24+\alpha_{\text{add}})(8\mu_U+3)}, \tag{43}$$

then conditioned on the events $\mathcal{E}_1(\xi_0)^c$ and $\mathcal{E}_2(\xi_0)^c$, we have that

$$\ell(\widehat{p}_{AB}) - \ell(\mathbb{E}\widehat{p}_{AB}) \leq \frac{\alpha_{\text{add}}}{4}$$

Therefore, we have

$$\mathbb{P}\left[\ell(\widehat{p}_{AB}) - \ell(\mathbb{E}\widehat{p}_{AB}) > \frac{\alpha_{\text{add}}}{4} \middle| \left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}, \mathcal{E}_1(\xi_0)^c, \mathcal{E}_2(\xi_0)^c\right] = 0. \tag{44}$$

Using this in (41), we have

$$\begin{aligned}\mathbb{P}\left[\widehat{d}_{AB} - d_{AB} > \frac{\alpha_{\text{add}}}{4}\right] &\leq \mathbb{E}\left[\mathbb{P}\left(\mathcal{E}_1(\xi_0)\middle|\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}\right)\right] + \mathbb{E}\left[\mathbb{P}\left(\mathcal{E}_2(\xi_0)\middle|\left\{\delta_{AB}^{(i)}\right\}_{i\in[m]}\right)\right] \\ &\leq e^{-2m\xi_0^2} + e^{-2mk\xi_0^2},\end{aligned} \tag{45}$$

where the second inequality comes from applying Hoeffding's inequality to each term, as in (17) and (18). Since this upper bound is independent of the choice of the pair of leaves, we can use (45) and (40) in (39) to get

$$\begin{aligned}\mathbb{P}\left[\text{Error}\right] &\leq \sum_{\substack{A,B,C,D\in L: \\ \tau_{AB}+\tau_{CD}\leq\tau_{AC}+\tau_{BD}=\tau_{AD}+\tau_{BC}}} 8\left(e^{-2m\xi_0^2} + e^{-2mk\xi_0^2}\right) \\ &\leq 8\binom{n}{4}\left(e^{-2m\xi_0^2} + e^{-2mk\xi_0^2}\right).\end{aligned} \tag{46}$$

Now, if we pick m and k as in (38) (also (9)), we see that the right side above is less than ϵ, which concludes the proof. The limit as $f \to 0$ can also be readily computed by observing that $\alpha_{\text{add}} \to 2\mu_L f$ as $f \to 0$. □

5.5 Proof of Claim 1

We will begin by using the fact that $\ell(x)$ satisfies the following Lipschitz property: for any $0 \leq x \leq y \leq B$, we have

$$\begin{aligned}\ell(y) - \ell(x) &= -\frac{3}{4}\log\left(1 - \frac{4}{3}y\right) + \frac{3}{4}\log\left(1 - \frac{4}{3}x\right) \\ &= \int_x^y \frac{1}{1 - \frac{4}{3}t}\, dt \\ &\leq \frac{(y-x)}{1 - \frac{4}{3}B}. \end{aligned} \tag{47}$$

From this, we have that

$$\left|\ell\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)}\right) - \ell\left(\mathbb{E}[p_{AB}]\right)\right| \leq \frac{\xi}{1 - \frac{4}{3}\left(\mathbb{E}\left[\widehat{p}_{AB}\right] + \xi\right)}, \quad \text{conditioned on } \mathcal{E}_1(\xi), \tag{48}$$

where we have chosen the B (of (47)) to be $\mathbb{E}\left[\widehat{p}_{AB}\right] + \xi$, since conditioned on $\mathcal{E}_1(\xi)$, we have that

$$\frac{1}{m}\sum_{i=1}^{m} p_{AB}^{(i)} \leq \mathbb{E}\left[\widehat{p}_{AB}\right] + \xi. \tag{49}$$

Similarly, conditioned on $\mathcal{E}_2(\xi)^c$ and $\mathcal{E}_1(\xi)^c$, we have

$$\begin{aligned}\left|\ell\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)}\right) - \ell\left(\widehat{p}_{AB}\right)\right| &\leq \frac{\xi}{1 - \frac{4}{3}\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} + \xi\right)} \\ &\leq \frac{\xi}{1 - \frac{4}{3}\left(\mathbb{E}\left[\widehat{p}_{AB}\right] + 2\xi\right)}, \end{aligned} \tag{50}$$

where in the first inequality we have chosen B (of (47)) to be $\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} + \xi$, since conditioned on $\mathcal{E}_2(\xi)^c$, we have that $\widehat{p}_{AB} \leq \frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)} + \xi$, and the second inequality follows from (49). Therefore from (48) and (50), we have that the following inequality holds conditioned on $\mathcal{E}_1(\xi)^c$ and $\mathcal{E}_2(\xi)^c$:

$$\begin{aligned}\left|\ell(\widehat{p}_{AB}) - \ell(\mathbb{E}\left[\widehat{p}_{AB}\right])\right| &\leq \left|\ell\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)}\right) - \ell\left(\mathbb{E}[p_{AB}]\right)\right| + \left|\ell\left(\frac{1}{m}\sum_{i\in[m]} p_{AB}^{(i)}\right) - \ell\left(\widehat{p}_{AB}\right)\right| \\ &\leq \frac{2\xi}{1 - \frac{4}{3}\left(\mathbb{E}\left[\widehat{p}_{AB}\right] + 2\xi\right)} \end{aligned} \tag{51}$$

Finally, to conclude the proof of the claim, we bound $\mathbb{E}[p_{AB}]$ using the properties of the multispecies coalescent. Notice that, by definition, the random

distance δ_{AB} is equal to $\mu_{AB} + 2Z_{AB}$, where μ_{AB} and Z_{AB} are as defined in Section 5.3. Therefore,

$$\begin{aligned}\mathbb{E}\left[\widehat{p}_{AB}\right] &= \mathbb{E}\left[\frac{3}{4}(1 - e^{-\frac{4}{3}\delta_{AB}})\right] \\ &= \frac{3}{4}\left(1 - e^{-\frac{4}{3}\mu_{AB}}\mathbb{E}\left[e^{-\frac{8}{3}Z_{AB}}\right]\right) \end{aligned} \tag{52}$$

Next, we observe that the random variable Z_{AB} is stochastically dominated by the random variable $\mu_U Z$, where $Z \sim \text{Exp}(1)$. This implies that

$$\begin{aligned}\mathbb{E}\left[e^{-\frac{8}{3}Z_{AB}}\right] &\geq \mathbb{E}\left[e^{-\frac{8}{3}\mu_U Z}\right] \\ &= \frac{1}{\frac{8}{3}\mu_U + 1}.\end{aligned}$$

Using this and the fact that $\mu_{AB} \leq \mu_U \Delta$ in (52), we have

$$\mathbb{E}\left[\widehat{p}_{AB}\right] \leq \frac{3}{4}\left(1 - \frac{e^{-\frac{4}{3}\mu_U\Delta}}{\frac{8}{3}\mu_U + 1}\right).$$

Substituting this in (51) concludes the proof.

REFERENCES

[1] W. P. Maddison, "Gene trees in species trees," *Systematic biology*, vol. 46, no. 3, pp. 523–536, 1997.

[2] R. Nichols, "Gene trees and species trees are not the same," *Trends in Ecology & Evolution*, vol. 16, no. 7, pp. 358–364, 2001.

[3] L. Liu, L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards, "Coalescent methods for estimating phylogenetic trees," *Molecular Phylogenetics and Evolution*, vol. 53, no. 1, pp. 320–328, 2009.

[4] J. Felsenstein, *Inferring phylogenies*, vol. 2. Sinauer Associates Sunderland, 2004.

[5] R. Griffiths and S. Tavaré, "Ancestral inference in population genetics," *Statistical Science*, pp. 307–319, 1994.

[6] C. Semple and M. A. Steel, *Phylogenetics*, vol. 24. Oxford University Press, 2003.

[7] A. D. Leaché and B. Rannala, "The accuracy of species tree estimation under simulation: a comparison of methods," *Systematic Biology*, vol. 60, no. 2, pp. 126–137, 2011.

[8] L. Liu, L. Yu, D. K. Pearl, and S. V. Edwards, "Estimating species phylogenies using coalescence times among sequences," *Systematic Biology*, vol. 58, no. 5, pp. 468–477, 2009.

[9] J. H. Degnan, M. DeGiorgio, D. Bryant, and N. A. Rosenberg, "Properties of consensus methods for inferring species trees from gene trees," *Systematic Biology*, vol. 58, no. 1, pp. 35–54, 2009.

[10] E. Mossel and S. Roch, "Incomplete lineage sorting: consistent phylogeny estimation from multiple loci," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 7, no. 1, pp. 166–171, 2010.

[11] L. Liu, L. Yu, and D. Pearl, "Maximum tree: a consistent estimator of the species tree," *Journal of Mathematical Biology*, vol. 60, pp. 95–106, 2010. 10.1007/s00285-009-0260-0.

[12] S. Roch, "An analytical comparison of multilocus methods under the multispecies coalescent: the three-taxon case.," in *Pacific Symposium on Biocomputing*, pp. 297–306, World Scientific, 2013.

[13] L. Nakhleh, "Computational approaches to species phylogeny inference and gene tree reconciliation," *Trends in ecology & evolution*, vol. 28, no. 12, pp. 719–728, 2013.

[14] J. Yang and T. Warnow, "Fast and accurate methods for phylogenomic analyses," *BMC bioinformatics*, vol. 12, no. Suppl 9, p. S4, 2011.

[15] M. W. Hahn *et al.*, "Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution," *Genome Biol*, vol. 8, no. 7, p. R141, 2007.

[16] M. A. Steel and L. A. Székely, "Inverting random functions. II. Explicit bounds for discrete maximum likelihood estimation, with applications," *SIAM J. Discrete Math.*, vol. 15, no. 4, pp. 562–575 (electronic), 2002.

[17] P. L. Erdos, M. A. Steel, L. A. Székely, and T. J. Warnow, "A few logs suffice to build (almost) all trees (i)," *Random Structures and Algorithms*, vol. 14, no. 2, pp. 153–184, 1999.

[18] B. Rannala and Z. Yang, "Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci," *Genetics*, vol. 164, no. 4, pp. 1645–1656, 2003.

[19] S. Tavaré, "Some probabilistic and statistical problems in the analysis of dna sequences," *Lectures on mathematics in the life sciences*, vol. 17, pp. 57–86, 1986.

[20] S. Roch, "Toward extracting all phylogenetic information from matrices of evolutionary distances," *Science*, vol. 327, no. 5971, pp. 1376–1379, 2010.

[21] E. Mossel, Y. Peres, *et al.*, "Information flow on trees," *The Annals of Applied Probability*, vol. 13, no. 3, pp. 817–844, 2003.

[22] R. R. Sokal and C. D. Michener, *A statistical method for evaluating systematic relationships*. University of Kansas, 1958.

[23] M. Steel, "A basic limitation on inferring phylogenies by pairwise sequence comparisons," *Journal of Theoretical Biology*, vol. 256, no. 3, pp. 467 – 472, 2009.

[24] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[25] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.