Distance-based Species Tree Estimation: Information-Theoretic Trade-off between Number of Loci and Sequence Length under the Coalescent*

Elchanan Mossel[1] **and Sebastien Roch**[2]

1 **U. Penn and UC Berkeley, USA**
`mossel@wharton.upenn.edu`

2 **UW–Madison, USA**
`roch@math.wisc.edu`

Abstract

We consider the reconstruction of a phylogeny from multiple genes under the multispecies coalescent. We establish a connection with the sparse signal detection problem, where one seeks to distinguish between a distribution and a mixture of the distribution and a sparse signal. Using this connection, we derive an information-theoretic trade-off between the number of genes, m, needed for an accurate reconstruction and the sequence length, k, of the genes. Specifically, we show that to detect a branch of length f, one needs $m = \Theta(1/[f^2\sqrt{k}])$.

1998 ACM Subject Classification F.2.0 Analysis of Algorithms and Problem Complexity – General, J.3 Life and Medical Sciences

Keywords and phrases phylogenetic reconstruction, multispecies coalescent, sequence length requirement

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2015.931

1 Introduction

In the **sparse signal detection problem**, one is given m i.i.d. samples $X_1, \ldots, X_m$ and the goal is to distinguish between a distribution $\mathbb{P}_0^{(m)}$

$$H_0^{(m)} : X_i \sim \mathbb{P}_0^{(m)},$$

and the same distribution corrupted by a sparse signal $\mathbb{P}_1^{(m)}$

$$H_1^{(m)} : X_i \sim \mathbb{Q}^{(m)} := (1 - \sigma_m)\,\mathbb{P}_0^{(m)} + \sigma_m\,\mathbb{P}_1^{(m)}.$$

Typically one takes $\sigma_m = m^{-\beta}$, where $\beta \in (0, 1)$. This problem arises in a number of applications [19, 27, 7, 30]. The Gaussian case in particular is well-studied [26, 20, 5]. For instance it is established in [26, 20] that, in the case $\mathbb{P}_0^{(m)} \sim N(0, 1)$ and $\mathbb{P}_1^{(m)} \sim N(\lambda_m, 1)$

* S.R. is supported by NSF grants DMS-1007144, DMS-1149312 (CAREER), and an Alfred P. Sloan Research Fellowship. E.M. is supported by NSF grants DMS-1106999 and CCF 1320105 and DOD ONR grants N000141110140 and N00014-14-1-0823 and grant 328025 from the Simons Foundation. E.M. and S.R. thank the Simons Institute for the Theory of Computing at U.C. Berkeley where this work was done.

18th Int'l Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'15) /
19th Int'l Workshop on Randomization and Computation (RANDOM'15).
Editors: Naveen Garg, Klaus Jansen, Anup Rao, and José D. P. Rolim; pp. 931–942
LIPICS Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

with $\lambda_m = \sqrt{2r\log m}$, a test with vanishing error probability exists if and only if r exceeds an explicitly known *detection boundary* $r^*(\beta)$.

In this paper, we establish a connection between sparse signal detection and the reconstruction of phylogenies from multiple genes or loci under a population-genetic model known as the multispecies coalescent [43]. The latter problem is of great practical interest in computational evolutionary biology and is currently the subject of intense study. See e.g. [31, 17, 2, 42] for surveys. There is in particular a growing body of theoretical results [15, 16, 14, 38, 32, 1, 45, 10, 46, 47], although much remains to be understood. The problem is also closely related to another very active area of research, the reconstruction of demographic history in population genetics. See e.g. [41, 4, 29] for some recent theoretical results.

By taking advantage of the connection to sparse signal detection, we derive a detection boundary for the multilocus phylogeny estimation problem and use it to characterize the trade-off between the number of genes needed to accurately reconstruct a phylogeny and the quality of the signal that can be extracted from each separate gene. Our results apply to an important class of reconstruction methods known as distance-based methods. Before stating our results more formally, we begin with some background. See e.g. [48] for a general introduction to mathematical phylogenetics.

1.1 Species tree estimation

An evolutionary tree, or phylogeny, is a graphical representation of the evolutionary relationships between a group of species. Each leaf in the tree corresponds to a current species while internal vertices indicate past speciation events. In the classical phylogeny estimation problem, one sequences a *single* common gene (or other *locus* such as pseudogenes, introns, etc.) from a representative individual of each species of interest. One then seeks to reconstruct the phylogeny by comparing the genes across species. The basic principle is simple: because mutations accumulate over time during evolution, more distantly related species tend to have more differences between their genes.

Formally, phylogeny estimation boils down to *learning the structure of a latent tree graphical model from i.i.d. samples at the leaves.* Let $T = (V, E, L, r)$ be a rooted leaf-labelled binary tree, with n leaves denoted by $L = \{1, \ldots, n\}$ and a root denoted by r. In the Jukes-Cantor model [28], one of the simplest Markovian models of molecular evolution, we associate to each edge $e \in E$ a mutation probability

$$p_e = 1 - e^{-\nu_e t_e}, \tag{1}$$

where ν_e is the mutation rate and t_e is the time elapsed along the edge e. (The analytical form of (1) derives from a continuous-time Markov process of mutation along the edge. See e.g. [48].) The *Jukes-Cantor process* is defined as follows:

- Associate to the root a sequence $\mathbf{s}_r = (s_{r,1}, \ldots, s_{r,k}) \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^k$ of length k where each site $s_{r,i}$ is uniform in $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.
- Let $U = \{r\}$.
- Repeat until $U = \emptyset$:
  - Pick a $u \in U$.
  - Let u^- be the parent of u.
  - Associate a sequence $\mathbf{s}_u \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^k$ to u as follows: $\mathbf{s}_u$ is obtained from $\mathbf{s}_{u^-}$ by mutating each site in $\mathbf{s}_{u^-}$ independently with probability $p_{(u^-,u)}$; when a mutation occurs at a site i, replace $s_{u,i}$ with a uniformly chosen state in $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.
  - Remove u from U and add the children (if any) of u to U.

Let T^{-r} be the tree T where the root is suppressed, i.e., where the two edges adjacent to the root are combined into a single edge. We let $\mathcal{L}[T, (p_e)_e, k]$ be the distribution of the sequences at the *leaves* $\mathbf{s}_1, \ldots, \mathbf{s}_n$ under the Jukes-Cantor process. We define the **single-locus phylogeny estimation problem** as follows:

> Given sequences at the leaves $(\mathbf{s}_1, \ldots, \mathbf{s}_n) \sim \mathcal{L}[T, (p_e)_e, k]$, recover the (leaf-labelled) unrooted tree T^{-r}.

(One may also be interested in estimating the p_es, but we focus on the tree. The root is in general not identifiable.) This problem has a long history in evolutionary biology. A large number of estimation techniques have been developed. See e.g. [24]. For a survey of the learning perspective on this problem, see e.g. [40]. On the theoretical side, much is known about the sequence length—or, in other words, the number of samples—required for a perfect reconstruction with high probability, including both information-theoretic lower bounds [49, 34, 35, 39] and matching algorithmic upper bounds [22, 11, 12, 44]. More general models of molecular evolution have also been considered in this context; see e.g. [23, 9, 37, 13, 3].

Nowadays, it is common for biologists to have access to *multiple* genes—or even full genomes. This abundance of data, which on the surface may seem like a blessing, in fact comes with significant new challenges. See e.g. [18, 42] for surveys. One important issue is that different genes may have incompatible evolutionary histories—represented by incongruent gene trees. In other words, if one were to solve the phylogeny estimation problem *separately* for several genes, one may in fact obtain *different* trees. Such incongruence can be explained in some cases by estimation error, but it can also result from deeper biological processes such as horizontal gene transfer, gene duplications and losses, and incomplete lineage sorting [33]. The latter phenomenon, which will be explained in Section 2, is the focus of this paper.

Accounting for this type of complication necessitates a *two-level hierarchical model* for the input data. Let $S = (V, E, L, r)$ be a rooted leaf-labelled binary *species tree*, i.e., a tree representing the actual succession of past divergences for a group of organisms. To each gene j shared by all species under consideration, we associate a *gene tree* $T_j = (V_j, E_j, L)$, mutation probabilities $(p_e^j)_{e \in E_j}$, and sequence length k_j. The triple $(T_j, (p_e^j)_{e \in E_j}, k_j)$ is picked at random according to a given distribution $\mathcal{G}[S, (\nu_e, t_e)_{e \in E}]$ which depends on the *unknown* species tree, mutation parameters ν_e and inter-speciation times t_e. It is standard to assume that the gene trees are conditionally independent given the species tree. In the context of incomplete lineage sorting, the distribution of the gene trees, $\mathcal{G}$, is given by the so-called *multispecies coalescent*, which is a canonical model for combining speciation history and population genetic effects [43]. The detailed description of the model is deferred to Section 2, as it is not needed for a high-level overview of our results. For the readers not familiar with population genetics, it is useful to think of T_j as a noisy version of S (which, in particular, may result in T_j having a different (leaf-labelled) topology than S).

Our two-level model of sequence data is then as follows. Given a species tree S, parameters $(\nu_e, t_e)_{e \in E}$ and a number of genes m:

1. **[First level: gene trees]** Pick m independent gene trees and parameters

$$(T_j, (p_e^j)_{e \in E_j}, k_j) \sim \mathcal{G}[S, (\nu_e, t_e)_{e \in E}], \qquad j = 1, \ldots, m.$$

2. **[Second level: leaf sequences]** For each gene $j = 1, \ldots, m$, generate sequence data at the leaves L according to the (single-locus) Jukes-Cantor process, as described above,

$$(\mathbf{s}_1^j, \ldots, \mathbf{s}_n^j) \sim \mathcal{L}[T_j, (p_e^j)_e, k_j], \qquad j = 1, \ldots, m,$$

independently of the other genes.

We define the **multi-locus phylogeny estimation problem** as follows:

> Given sequences at the leaves $(\mathbf{s}_1^j, \ldots, \mathbf{s}_n^j)$, $j = 1, \ldots, m$, generated by the process above, recover the (leaf-labelled) unrooted species tree S^{-r}.

In the context of incomplete lineage sorting, this problem is the focus of very active research in statistical phylogenetics [31, 17, 2, 42]. In particular, there is a number of theoretical results, including [15, 16, 14, 38, 32, 1, 45, 10, 46, 47]. However, many of these results concern the statistical properties (identifiability, consistency, convergence rate) of species tree estimators *that (unrealistically) assume perfect knowledge of the T_js.* We only have a very incomplete picture of the properties of estimators that are based on sequence data, i.e., that do *not* require the knowledge of the T_js. (See below for an overview of prior results.)

Here we consider the data requirement of such estimators based on the sequences. To simplify, we assume that all genes have the same length, i.e., that $k_j = k$ for all $j = 1, \ldots, m$ for some k. (Because our goal is to derive a lower bound, such simplification is largely immaterial.) Our results apply to an important class of methods known as *distance-based methods*, which we briefly describe now. In the single-locus phylogeny estimation problem, a natural way to infer T^{-r} is to use the fraction of substitutions between each pair, i.e., letting $\|\cdot\|_1$ denote the ℓ_1-distance,

$$\theta(\mathbf{s}_a, \mathbf{s}_b) := \|\mathbf{s}_a - \mathbf{s}_b\|_1, \qquad \forall a, b \in [n]. \tag{2}$$

We refer to reconstruction methods relying solely on the $\theta(\mathbf{s}_a, \mathbf{s}_b)$s as distance-based methods. Assume for instance that $\nu_e = \nu$ for all e, i.e., the so-called molecular clock hypothesis. Then it is easily seen that single-linkage clustering (e.g., [25]) applied to the *distance matrix* $(\theta(\mathbf{s}_a, \mathbf{s}_b))_{a,b\in[n]}$ converges to T^{-r} as $k \to +\infty$. (In this special case, the root can be recovered as well.) In fact, T can be reconstructed perfectly as long as, for each a, b, $\frac{1}{k}\theta(\mathbf{s}_a, \mathbf{s}_b)$ is close enough to its expectation (e.g. [48])

$$\theta_{ab} := \frac{3}{4}(1 - e^{-d_{ab}}) \quad \text{with} \quad d_{ab} := \sum_{e \in P(a,b)} \nu_e t_e,$$

where $P(a, b)$ is the edge set on the unique path between a and b in T. Here "close enough" means $O(f)$ where $f := \min_e \nu_e t_e$. This observation can been extended to general ν_es. See e.g. [22] for explicit bounds on the sequence length required for perfect reconstruction with high probability.

Finally, to study distance-based methods in the *multi-locus* case, we restrict ourselves to the following **multi-locus distance estimation problem**:

> Given an accuracy $\varepsilon > 0$ and distance matrices $\theta(\mathbf{s}_a^j, \mathbf{s}_b^j)_{a,b\in[n]}$, $j = 1, \ldots, m$, estimate d_{ab} as defined above within ε for all a, b.

Observe that, once the d_{ab}s are estimated within sufficient accuracy, i.e., within $O(f)$, the species tree can be reconstructed using the techniques referred to in the single-locus case.

1.2 Our results

How is all this related to the sparse signal detection problem? Our main goal here is to provide a lower bound on the amount of data required for perfect reconstruction, in terms of m (the number of genes) and k (the sequence length). Consider the three possible (rooted, leaf-labelled) species trees with three leaves, as depicted in Figure 1, where we let the time to the most recent divergence be $1 - f$ (from today) and the time to the earlier divergence

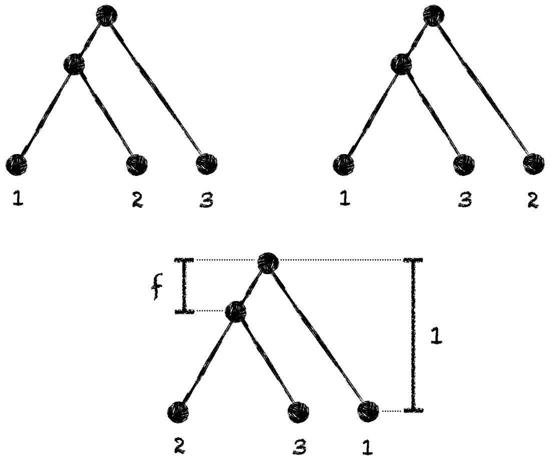

Figure 1 Three species trees.

be 1. In order for a distance-based method to distinguish between these three possibilities, i,e., to determine which pair is closest, we need to estimate the d_{ab}s within $O(f)$ accuracy. Put differently, within the multi-locus distance estimation problem, it suffices to establish a lower bound on the data required to distinguish between a two-leaf species tree S with $d_{12} = 2$ and a two-leaf species tree S^+ with $d_{12} = 2 - 2f$, where in both cases $\nu_e = 1$ for all e. We are interested in the limit $f \to 0$.

Let $\mathbb{P}_0$ and $\mathbb{Q}$ be the distributions of $\theta(\mathbf{s}_1^1, \mathbf{s}_2^1)$ for a single gene under S and S^+ respectively, where for ease of notation the dependence on k is implicit. For m genes, we denote the corresponding distributions by $\mathbb{P}_0^{\otimes m}$ and $\mathbb{Q}^{\otimes m}$. To connect the problem to sparse signal detection we observe below that, under the multispecies coalescent, $\mathbb{Q}$ is in fact a *mixture* of $\mathbb{P}_0$ and a sparse signal $\mathbb{P}_1$, i.e.,

$$\mathbb{Q} = (1 - \sigma_f)\, \mathbb{P}_0 + \sigma_f\, \mathbb{P}_1, \tag{3}$$

where $\sigma_f = O(f)$ as $f \to 0$.

When testing between $\mathbb{P}_0^{\otimes m}$ and $\mathbb{Q}^{\otimes m}$, the optimal sum of Type-I (false positive) and Type-II (false negative) errors is given by (e.g. [8])

$$\inf_A \{\mathbb{P}_0^{\otimes m}(A) + \mathbb{Q}^{\otimes m}(A^c)\} = 1 - \|\mathbb{P}_0^{\otimes m} - \mathbb{Q}^{\otimes m}\|_{\mathrm{TV}}, \tag{4}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance. Because $\sigma_f = O(f)$, for any k, in order to distinguish between $\mathbb{P}_0$ and $\mathbb{Q}$ one requires that, at the very least, $m = \Omega(f^{-1})$. Otherwise the probability of observing a sample originating from $\mathbb{P}_1$ under $\mathbb{Q}$ is bounded away from 1. In [38] it was shown that, provided that $k = \Omega(f^{-2}\log f^{-1})$, $m = \Omega(f^{-1})$ suffices. At the other end of the spectrum, when $k = O(1)$, a lower bound for the single-locus problem obtained by [49] implies that $m = \Omega(f^{-2})$ is needed. An algorithm achieving this bound under the multispecies coalescent was recently given in [10].

We settle the full spectrum between these two regimes. Our results apply when $k = f^{-2+2\kappa}$ and $m = f^{-1-\mu}$ where $0 < \kappa, \mu < 1$ as $f \to 0$.

▶ **Theorem 1** (Lower bound). *For any $\delta > 0$, there is a $c > 0$ such that*

$$\|\mathbb{P}_0^{\otimes m} - \mathbb{Q}^{\otimes m}\|_{\mathrm{TV}} \leq \delta,$$

whenever

$$m \leq c \frac{1}{f^2\sqrt{k}}.$$

Notice that the lower bound on m interpolates between the two extremal regimes discussed above. As k increases, a more accurate estimate of the gene trees can be obtained and one expects that the number of genes required for perfect reconstruction should indeed decrease. The form of that dependence is far from clear however. We in fact prove that our analysis is tight.

▶ **Theorem 2** (Matching upper bound). *For any $\delta > 0$, there is a $c' > 0$ such that*

$$\|\mathbb{P}_0^{\otimes m} - \mathbb{Q}^{\otimes m}\|_{\mathrm{TV}} \geq 1 - \delta,$$

whenever

$$m \geq c' \frac{1}{f^2\sqrt{k}}.$$

Moreover, there is an efficient test to distinguish between $\mathbb{P}_0^{\otimes m}$ and $\mathbb{Q}^{\otimes m}$ in that case.

Our proof of the upper bound actually gives an efficient reconstruction algorithm under the molecular clock hypothesis. We expect that the insights obtained from proving Theorem 1 and 2 will lead to more accurate practical methods as well in the general case.

1.3 Proof sketch

Let Z be an exponential random variable with mean 1. We first show that, under $\mathbb{P}_0$ (respectively $\mathbb{Q}$), $\theta(\mathbf{s}_1^1, \mathbf{s}_2^1)$ is binomial with k trials and success probability $\frac{3}{4}\left(1 - e^{-2(\zeta + Z)}\right)$, where $\zeta = 1$ (respectively $\zeta = 1 - f$). Equation (3) then follows from the memoryless property of the exponential, where σ_f is the probability that $Z \leq f$.

A recent result of [6] gives a formula for the detection boundary of the sparse signal detection problem for general $\mathbb{P}_0$, $\mathbb{P}_1$. However, applying this formula here is non-trivial. Instead we bound directly the total variation distance between $\mathbb{P}_0^{\otimes m}$ and $\mathbb{Q}^{\otimes m}$. Similarly to the approach used in [6], we work instead with the Hellinger distance $H^2(\mathbb{P}_0^{\otimes m}, \mathbb{Q}^{\otimes m})$ which tensorizes as follows (see e.g. [8])

$$\frac{1}{2}H^2(\mathbb{P}_0^{\otimes m}, \mathbb{Q}^{\otimes m}) = 1 - \left(1 - \frac{1}{2}H^2(\mathbb{P}_0, \mathbb{Q})\right)^m, \tag{5}$$

and further satisfies

$$\|\mathbb{P}_0^{\otimes m} - \mathbb{Q}^{\otimes m}\|_{\mathrm{TV}}^2 \leq H^2(\mathbb{P}_0^{\otimes m}, \mathbb{Q}^{\otimes m}) \left[1 - \frac{1}{4}H^2(\mathbb{P}_0^{\otimes m}, \mathbb{Q}^{\otimes m})\right]. \tag{6}$$

All the work is in proving that, as $f \to 0$,

$$H^2(\mathbb{P}_0, \mathbb{Q}) = O\left(f^2\sqrt{k}\right).$$

More details are given in Section 3.1.

The proof of Theorem 2 on the other hand involves the construction of a statistical test that distinguishes between $\mathbb{P}_0^{\otimes m}$ and $\mathbb{Q}^{\otimes m}$. In the regime $k = O(1)$, an optimal test (up to constants) compares the means of the samples [10]. In the regime $k = \omega(f^{-2})$, an optimal test (up to constants) compares the minima of the samples [38]. A natural way to interpolate between these two tests is to consider an appropriate quantile. We show that the $1/\sqrt{k}$-quantile leads to the optimal choice.

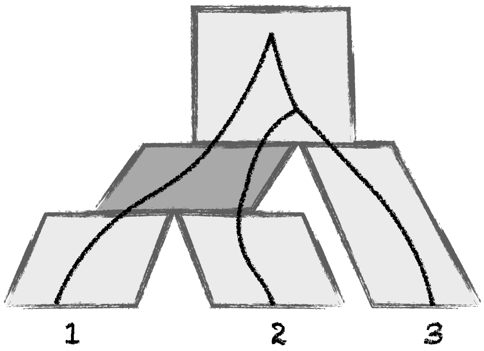

Figure 2 An incomplete lineage sorting event. Although 1 and 2 are more closely related in the species tree (fat tree), 2 and 3 are more closely related in the gene tree (thin tree). This incongruence is caused by the failure of the lineages originating from 1 and 2 to coalesce within the shaded branch.

1.4 Organization.

The gene tree generating model is defined in Section 2. The proofs of the main theorems are omitted from this extended abstract. These can be found at the Arxiv version of the paper [36].

2 Further definitions

In this section, we give more details on the model.

2.1 A little coalescent theory

As we mentioned in the previous section, our gene tree distribution model $\mathcal{G}[S,(\nu_e,t_e)_{e\in E}]$ is the *multispecies coalescent* [43]. We first explain the model in the two-species case. Let 1 and 2 be two species and consider a common gene j. One can trace *back in time* the lineages of gene j from an individual in 1 and from an individual in 2 until the first *common* ancestor. The latter event is called a *coalescence*. Here, because the two lineages originate from different species, coalescence occurs in an ancestral population. Let τ be the time of the divergence between 1 and 2 (back in time). Then, under the multispecies coalescent, the *coalescence time* is $\tau + Z$ where Z is an exponential random variable whose mean depends on the effective population size of the ancestral population. Here we scale time so that the mean is 1. (See e.g. [21] for an introduction to coalescent theory.)

We immediately get for the two-level model of sequence data:

▶ **Lemma 3** (Distance distribution). *Let S be a two-leaf species tree with $d_{12} = 2\tau$ and $\nu_e = 1$ for all e and let $\theta(\mathbf{s}^1_1,\mathbf{s}^1_2)$ be as in (2) for some k. Then the distibution of $\theta(\mathbf{s}^1_1,\mathbf{s}^1_2)$ is binomial with k trials and success probability $\frac{3}{4}\left(1-e^{-2(\tau+Z)}\right)$.*

The memoryless property of the exponential gives:

▶ **Lemma 4** (Mixture). *Let S be a two-leaf species tree with $d_{12} = 2$ and let S^+ be a two-leaf species tree with $d_{12} = 2-2f$, where in both cases $\nu_e = 1$ for all e. Let $\mathbb{P}_0$ and $\mathbb{Q}$ be the distributions of $\theta(\mathbf{s}^1_1,\mathbf{s}^1_2)$ for a single gene under S and S^+ respectively. Then, there is $\mathbb{P}_1$*

such that,

$$\mathbb{Q} = (1-\sigma_f)\,\mathbb{P}_0 + \sigma_f\,\mathbb{P}_1,$$

where $\sigma_f = O(f)$, as $f \to 0$. More specifically, $\mathbb{P}_1$ is obtained by conditioning $\mathbb{Q}$ on the event that Z is $\leq f$ and σ_f is the probability of that event.

More generally (this paragraph may be skipped as it will not play a role below), consider a species tree $S = (V, E; L, r)$ with n leaves. Each gene $j = 1, \ldots, m$ has a genealogical history represented by its gene tree T_j distributed according to the following process: looking backwards in time, on each branch of the species tree, the coalescence of any two lineages is exponentially distributed with rate 1, independently from all other pairs; whenever two branches merge in the species tree, we also merge the lineages of the corresponding populations, that is, the coalescence proceeds on the *union* of the lineages. More specifically, the probability density of a realization of this model for m independent genes is

$$\prod_{j=1}^{m}\prod_{e\in E}\exp\left(-\binom{O_j^e}{2}\left[\sigma_j^{e,O_j^e+1}-\sigma_j^{e,O_j^e}\right]\right)\prod_{\ell=1}^{I_j^e-O_j^e}\exp\left(-\binom{\ell}{2}\left[\sigma_j^{e,\ell}-\sigma_j^{e,\ell-1}\right]\right),$$

where, for gene j and branch e, I_j^e is the number of lineages entering e, O_j^e is the number of lineages exiting e, and $\sigma_j^{e,\ell}$ is the ℓ^{th} coalescence time in e; for convenience, we let $\sigma_j^{e,0}$ and $\sigma_j^{e,I_j^e-O_j^e+1}$ be respectively the divergence times of e and of its parent population. The resulting trees T_js may have topologies that differ from that of the species tree S. This may occur as a result of an incomplete lineage sorting event, i.e., the failure of two lineages to coalesce in a population. See Figure 2 for an illustration.

2.2 A more abstract setting

Before discussing the proofs, we re-set the problem in a more generic setting that will make the computations more transparent. We consider two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ for a random variable θ taking values in $\{0, \ldots, k\}$ for some k. We assume that the distribution of θ takes the form

$$\mathbb{P}_0[\theta = \ell] = \binom{k}{\ell}\mathbb{E}_0[X^\ell(1-X)^{k-\ell}],$$

where $\mathbb{E}_0$ is the expectation operator corresponding to $\mathbb{P}_0$, and X is some random variable admitting a density over $[0,1]$. The distribution is similarly defined under $\mathbb{P}_1$. We make the following assumptions, which are satisfied in the setting of the previous section:

A1. Under $\mathbb{P}_0$ and $\mathbb{P}_1$, X admits a density whose support is (p_0, p^0) under $\mathbb{P}_0$ and $(p_0 - \phi_f, p_0)$ under $\mathbb{P}_1$, where $0 < p_0 < p^0 < 1$ (independent of f) and $\phi_f = O(f)$. (In the setting of Lemma 4, $p_0 = \frac{3}{4}(1-e^{-2})$, $p_0 - \phi_f = \frac{3}{4}(1-e^{-(2-2f)})$, and $p^0 = 3/4$.)

A2. Under $\mathbb{P}_0$, the density of X (on its support) is in $[\rho, \rho^{-1}]$ for some $\rho > 0$ (independent of f) away from p^0, that is, below some $p_0 < \bar{p} < p^0$. (In the setting of Lemma 4, under $\mathbb{P}_0$ the density of X on (p_0, p^0) is $\frac{4e^{1/2}}{3}(1-4x/3)^{-3/4}$.)

As before, we let

$$\mathbb{Q} = (1-\sigma_f)\,\mathbb{P}_0 + \sigma_f\,\mathbb{P}_1,$$

for some $\sigma_f = O(f)$.

3 Main steps of the proof

We give a few more details on the proofs.

3.1 Lower bound

We briefly sketch the main steps of the proof of the lower bound. In the abstract setting of Section 2, the Hellinger distance can be written as

$$\begin{aligned} H^2(\mathbb{P}_0,\mathbb{Q}) &= \sum_{j=0}^{k}\left[\sqrt{\mathbb{Q}[\theta=j]}-\sqrt{\mathbb{P}_0[\theta=j]}\right]^2 \\ &= \sum_{j=0}^{k}\left[\sqrt{1+\sigma_f\left(\frac{\mathbb{P}_1[\theta=j]}{\mathbb{P}_0[\theta=j]}-1\right)}-1\right]^2\mathbb{P}_0[\theta=j] \\ &= \sum_{j=0}^{k}\left[\sqrt{1+\sigma_f\left(\frac{\mathbb{E}_1[X^j(1-X)^{k-j}]}{\mathbb{E}_0[X^j(1-X)^{k-j}]}-1\right)}-1\right]^2\mathbb{P}_0[\theta=j] \qquad (7)\end{aligned}$$

We prove the following proposition, which implies Theorem 1.

▶ **Proposition 5.** *Assume that $k=f^{-2+2\kappa}$ where $0<\kappa<1$ and that Assumptions A1 and A2 hold. As $f\to 0$,*

$$H^2(\mathbb{P}_0,\mathbb{Q}) = O\left(f^2\sqrt{k}\right).$$

From (7), in order to bound the Hellinger distance, we need to control the ratio $\frac{\mathbb{E}_1[X^j(1-X)^{k-j}]}{\mathbb{E}_0[X^j(1-X)^{k-j}]}$ and the probability $\mathbb{P}_0[\theta=j]$. Because the standard deviation of θ/k is $O(1/\sqrt{k})$ and $f\sqrt{k}=o(1)$, the dominant term in the sum (7) turns out to come from X being within $O(1/\sqrt{k})$ of p_0 under $\mathbb{E}_0$ (an event of probability $O(1/\sqrt{k})$) and θ/k being within $O(1/\sqrt{k})$ of p_0 as well (in which case the ratio $\frac{\mathbb{E}_1[X^j(1-X)^{k-j}]}{\mathbb{E}_0[X^j(1-X)^{k-j}]}$ is of order $O(1)$). The contribution of the dominant term is then indeed of order $O(f^2\sqrt{k})$. The full details are somewhat delicate and appear in the Arxiv version of the paper [36].

3.2 Upper bound

To prove the upper bound, we use (4) and construct an explicit test A. Let W be the number of genes such that $\theta/k\le p_0$. Let $w=\mathbb{P}_0[\theta/k\le p_0]$ and $w'=\mathbb{Q}[\theta/k\le p_0]$. Then $W\sim\mathrm{Bin}(m,w)$ under $\mathbb{P}_0$ and $W\sim\mathrm{Bin}(m,w')$ under $\mathbb{Q}$. Let

$$w^* = mw+\frac{m}{2}(w'-w) = mw'-\frac{m}{2}(w'-w),$$

and consider the event

$$A=\{W\ge w^*\}.$$

We show in the Arxiv version of the paper [36] that $\mathbb{P}_0^{\otimes m}[A]\le\frac{\delta}{2}$, and $\mathbb{Q}^{\otimes m}[A^c]\le\frac{\delta}{2}$ when c' is large enough.

Acknowledgments. We thank Gautam Dasarathy and Rob Nowak for helpful discussions.

References

1 Elizabeth S. Allman, James H. Degnan, and John A. Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology*, 62(6):833–862, 2011.

2 Christian N.K. Anderson, Liang Liu, Dennis Pearl, and Scott V. Edwards. Tangled trees: The challenge of inferring species trees from coalescent and noncoalescent genes. In Maria Anisimova, editor, *Evolutionary Genomics*, volume 856 of *Methods in Molecular Biology*, pages 3–28. Humana Press, 2012.

3 Alexandr Andoni, Constantinos Daskalakis, Avinatan Hassidim, and Sébastien Roch. Global alignment of molecular sequences via ancestral state reconstruction (extended abstract). In *ICS*, pages 358–369, 2010.

4 Anand Bhaskar and Yun S. Song. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann. Statist.*, 42(6):2469–2493, 2014.

5 T. Tony Cai, X. Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(5):629–662, 2011.

6 T.T. Cai and Yihong Wu. Optimal detection of sparse mixtures against a given null distribution. *Information Theory, IEEE Transactions on*, 60(4):2217–2232, April 2014.

7 L. Cayon, J. Jin, and A. Treaster. Higher criticism statistic: detecting and identifying non-gaussianity in the wmap first-year data. *Monthly Notices of the Royal Astronomical Society*, 362(3):826–832, 2005.

8 T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley Series in Telecommunications. John Wiley & Sons Inc., New York, 1991. A Wiley-Interscience Publication.

9 M. Cryan, L. A. Goldberg, and P. W. Goldberg. Evolutionary trees can be learned in polynomial time. *SIAM J. Comput.*, 31(2):375–397, 2002. short version, Proceedings of the 39th Annual Symposium on Foundations of Computer Science (FOCS 98), pages 436-445, 1998.

10 Gautam Dasarathy, Robert D. Nowak, and Sébastien Roch. New sample complexity bounds for phylogenetic inference from multiple loci. In *2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, June 29 – July 4, 2014*, pages 2037–2041, 2014.

11 Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Evolutionary trees and the ising model on the bethe lattice: a proof of steel's conjecture. *Probability Theory and Related Fields*, 149:149–189, 2011. 10.1007/s00440-009-0246-2.

12 Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Phylogenies without branch bounds: Contracting the short, pruning the deep. *SIAM J. Discrete Math.*, 25(2):872–893, 2011.

13 Constantinos Daskalakis and Sébastien Roch. Alignment-free phylogenetic reconstruction. In *RECOMB*, pages 123–137, 2010.

14 Michael DeGiorgio and James H Degnan. Fast and consistent estimation of species trees using supermatrix rooted triples. *Molecular Biology and Evolution*, 27(3):552–69, March 2010.

15 J. H. Degnan and N. A. Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5), May 2006.

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

17 James H. Degnan and Noah A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, 24(6):332–340, 2009.

18 Frederic Delsuc, Henner Brinkmann, and Herve Philippe. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, 6(5):361–375, 05 2005.

19 R. L. Dobrusin. A statistical problem arising in the theory of detection of signals in the presence of noise in a multi-channel system and leading to stable distribution laws. *Theory of Probability & Its Applications*, 3(2):161–173, 1958.

20 David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 06 2004.

21 Richard Durrett. *Probability models for DNA sequence evolution.* Probability and its Applications (New York). Springer, New York, second edition, 2008.

22 P. L. Erdös, M. A. Steel;, L. A. Székely, and T. A. Warnow. A few logs suffice to build (almost) all trees (part 1). *Random Struct. Algor.*, 14(2):153–184, 1999.

23 P. L. Erdös, M. A. Steel;, L. A. Székely, and T. A. Warnow. A few logs suffice to build (almost) all trees (part 2). *Theor. Comput. Sci.*, 221:77–118, 1999.

24 J. Felsenstein. *Inferring Phylogenies.* Sinauer, New York, New York, 2004.

25 Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

26 Yu. I. Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.*, 6(1):47–69, 1997.

27 X. Jessie Jeng, T. Tony Cai, and Hongzhe Li. Optimal sparse segment identification with application in copy number variation analysis. *J. Amer. Statist. Assoc.*, 105(491):1156–1166, 2010.

28 T. H. Jukes and C. Cantor. Mammalian protein metabolism. In H. N. Munro, editor, *Evolution of protein molecules*, pages 21–132. Academic Press, 1969.

29 Junhyong Kim, Elchanan Mossel, Miklos Z. Racz, and Nathan Ross. Can one hear the shape of a population history? *Theoretical Population Biology*, 100(0):26–38, 2015.

30 Martin Kulldorff, Richard Heffernan, Jessica Hartman, Renato Assuncao, and Farzad Mostashari. A space time permutation scan statistic for disease outbreak detection. *PLoS Med*, 2(3):e59, 02 2005.

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

32 Liang Liu, Lili Yu, and Dennis K. Pearl. Maximum tree: a consistent estimator of the species tree. *Journal of Mathematical Biology*, 60(1):95–106, 2010.

33 Wayne P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.

34 E. Mossel. On the impossibility of reconstructing ancestral data and phylogenies. *J. Comput. Biol.*, 10(5):669–678, 2003.

35 E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404, 2004.

36 E. Mossel and S. Roch. Distance-based species tree estimation: information-theoretic trade-off between number of loci and sequence length under the coalescent. ArXiv e-print 1504.05289, 2015.

37 Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden Markov models. In *STOC'05: Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 366–375, New York, 2005. ACM.

38 Elchanan Mossel and Sébastien Roch. Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 7(1):166–171, 2010.

39 Elchanan Mossel, Sébastien Roch, and Allan Sly. On the inference of large phylogenies with long branches: How long is too long? *Bulletin of Mathematical Biology*, 73:1627–1644, 2011. 10.1007/s11538-010-9584-6.

40 Raphaël Mourad, Christine Sinoquet, Nevin Lianwen Zhang, Tengfei Liu, and Philippe Leray. A survey on latent tree models and applications. *J. Artif. Intell. Res. (JAIR)*, 47:157–203, 2013.

41 Simon Myers, Charles Fefferman, and Nick Patterson. Can one learn history from the allelic spectrum? *Theoretical Population Biology*, 73(3):342–348, 2008.

42 Luay Nakhleh. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in ecology & evolution*, 28(12):10.1016/j.tree.2013.09.004, 12 2013.

43 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.

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

45 Sebastien Roch. An analytical comparison of multilocus methods under the multispecies coalescent: The three-taxon case. In *Pacific Symposium in Biocomputing 2013*, pages 297–306, 2013.

46 Sebastien Roch and Mike Steel. Likelihood-based tree reconstruction on a concatenation of alignments can be positively misleading. *Theoretical Population Biology*, 2015. To appear.

47 Sebastien Roch and Tandy Warnow. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic Biology*, 2015. In press.

48 C. Semple and M. Steel. *Phylogenetics*, volume 22 of *Mathematics and its Applications series*. Oxford University Press, 2003.

49 M. A. Steel and L. A. Székely. Inverting random functions. II. Explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM J. Discrete Math.*, 15(4):562–575 (electronic), 2002.