Distance-based species tree estimation: information-theoretic trade-off between number of loci and sequence length under the coalescent

Elchanan Mossel
U Penn and UC Berkeley
mossel@wharton.upenn.edu

Sebastien Roch
UW Madison
roch@math.wisc.edu

April 22, 2015

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 \( m = \Theta(1/f^2\sqrt{k}) \).

*Keywords: phylogenetic reconstruction, multispecies coalescent, sequence length requirement."
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 $P_0^{(m)}$

$$H_0^{(m)} : X_i \sim P_0^{(m)},$$

and the same distribution corrupted by a sparse signal $P_1^{(m)}$

$$H_1^{(m)} : X_i \sim Q^{(m)} := (1 - \sigma_m) P_0^{(m)} + \sigma_m P_1^{(m)}.$$

Typically one takes $\sigma_m = m^{-\beta}$, where $\beta \in (0, 1)$. This problem arises in a number of applications \cite{Dob58, JCL10, CJT05, KHH+05}. The Gaussian case in particular is well-studied \cite{Ing97, DJ04, CJJ11}. For instance it is established in \cite{Ing97, DJ04} that, in the case $P_0^{(m)} \sim N(0, 1)$ and $P_1^{(m)} \sim N(\lambda_m, 1)$ 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 \cite{RY03}. The latter problem is of great practical interest in computational evolutionary biology and is currently the subject of intense study. See e.g. \cite{LYK+09, DR09, ALPE12, Nak13} for surveys. There is in particular a growing body of theoretical results \cite{DR06, DDBR09, DD10, MR10, LYP10, ADR11, Roc13, DNR14, RS15, RW15}, 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. \cite{MFP08, BS14, KMRR15} 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. \cite{SS03} for a general introduction to mathematical phylogenetics.

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 \cite{JC69}, one of the simplest Markovian models of molecular evolution, we associate to each edge $e \in E$ a
mutation probability

\[ p_e = 1 - e^{-\nu_e t_e}, \]  

where \( \nu_e \) 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. [SS03].) The Jukes-Cantor process is defined as follows:

- Associate to the root a sequence \( s_r = (s_{r,1}, \ldots, s_{r,k}) \in \{A, C, G, T\}^k \) of length \( k \) where each site \( s_{r,i} \) is uniform in \( \{A, C, G, T\} \).
- Let \( U = \{r\} \).
- Repeat until \( U = \emptyset \):
  - Pick a \( u \in U \).
  - Let \( u^- \) be the parent of \( u \).
  - Associate a sequence \( s_u \in \{A, C, G, T\}^k \) to \( u \) as follows: \( s_u \) is obtained from \( s_{u^-} \) by mutating each site in \( s_{u^-} \) independently with probability \( p(u^-, u) \); when a mutation occurs at a site \( i \), replace \( s_{u,i} \) with a uniformly chosen state in \( \{A, C, G, T\} \).
  - 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 \( s_1, \ldots, s_n \) under the Jukes-Cantor process. We define the single-locus phylogeny estimation problem as follows:

Given sequences at the leaves \( (s_1, \ldots, s_n) \sim \mathcal{L}[T, (p_e)_e, k] \), recover the (leaf-labelled) unrooted tree \( T^-r \).

(One may also be interested in estimating the \( p_e \)'s, 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. [Fel04]. For a survey of the learning perspective on this problem, see e.g. [MSZ+13]. 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 [SS02, Mos03, Mos04, MRS11] and matching algorithmic upper bounds [ESSW99a, DMR11a, DMR11b, Roc10]. More general models of molecular evolution have also been considered in this context; see e.g. [ESSW99b, CGG02, MR05, DR10, ADHR10].

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. [DBP05, Nak13] 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 [Mad97]. 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 $T_j = (V_j, E_j, L_j)$, and sequence length $k_j$. The triple $(T_j, (p^j_e)_{e \in E_j}, k_j)$ is picked at random according to a given distribution $G[\hat{S}, (\nu_e, t_e)_{e \in E}]$ which depends on the unknown species tree, mutation parameters $\nu_e$ 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, $G$, is given by the so-called multispecies coalescent, which is a canonical model for combining speciation history and population genetic effects [RY03]. 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
   
   $$(T_j, (p^j_e)_{e \in E_j}, k_j) \sim G[\hat{S}, (\nu_e, t_e)_{e \in E}], \quad j = 1, \ldots, m.$$  

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,
   
   $$(s^1_j, \ldots, s^n_j) \sim L[T_j, (p^j_e), k_j], \quad j = 1, \ldots, m,$$

   independently of the other genes.

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

Given sequences at the leaves $(s^1_j, \ldots, s^n_j), j = 1, \ldots, m$, 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 [LYK+09, DR09, ALPE12, Nak13]. In particular, there is a number of theoretical results, including [DR06, DDBR09, DD10, MR10, LYP10, ADR11, Roc13, DNR14, RS15, RW15]. However, many of these results concern the statistical properties (identifiability, consistency, convergence rate) of species tree estimators that (unrealistically) assume perfect knowledge of the $T_j$s. 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 $T_j$s. (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 $\ell_1$-distance,

$$\theta(s_a, s_b) := \|s_a - s_b\|_1, \quad \forall a, b \in [n].$$

(2)

We refer to reconstruction methods relying solely on the $\theta(s_a, s_b)$'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., [HTF09]) applied to the distance matrix $(\theta(s_a, s_b))_{a,b \in [n]}$ 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 $a$, $b$, $\frac{1}{k}\theta(s_a, s_b)$ is close enough to its expectation (e.g. [SS03])

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

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 be extended to general $\nu_e$'s. See e.g. [ESSW99a] 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 $\theta(s^j_a, s^j_b)_{a,b \in [n]}, j = 1, \ldots, m$, estimate $d_{ab}$ as defined above within $\varepsilon$ 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.

**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 $m$.

---

Figure 1: Three species trees.
(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 2 where we let the time to the most recent divergence be \( 1 - f \) (from today) and the time to the earlier divergence 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 \( P_0 \) and \( Q \) be the distributions of \( \theta(s_1^1, s_1^2) \) 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 \( P_0^\otimes m \) and \( Q^\otimes m \). To connect the problem to sparse signal detection we observe below that, under the multispecies coalescent, \( Q \) is in fact a mixture of \( P_0 \) and a sparse signal \( P_1 \), i.e.,

\[
Q = (1 - \sigma_f) P_0 + \sigma_f P_1,
\]

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

When testing between \( P_0^\otimes m \) and \( Q^\otimes m \), the optimal sum of Type-I (false positive) and Type-II (false negative) errors is given by (e.g. [CT91])

\[
\inf_A \{ P_0^\otimes m (A) + Q^\otimes m (A^c) \} = 1 - \| P_0^\otimes m - Q^\otimes m \|_{TV},
\]

where \( \| \cdot \|_{TV} \) denotes the total variation distance. Because \( \sigma_f = O(f) \), for any \( k \), in order to distinguish between \( P_0 \) and \( Q \) one requires that, at the very least, \( m = \Omega(f^{-1}) \). Otherwise the probability of observing a sample originating from \( P_1 \) under \( Q \) is bounded away from 1. In [MR10] 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 [SS02] implies that \( m = \Omega(f^{-2}) \) is needed. An algorithm achieving this bound under the multispecies coalescent was recently given in [DNR14].

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

\[
\| P_0^\otimes m - Q^\otimes m \|_{TV} \leq \delta,
\]

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

\[
\| P_0^\otimes m - Q^\otimes m \|_{TV} \geq 1 - \delta,
\]
whenever
\[ m \geq c' \frac{1}{f^2 \sqrt{k}}. \]

Moreover, there is an efficient test to distinguish between \( P_0^{\otimes m} \) and \( Q^{\otimes m} \) 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.

**Proof sketch** Let \( Z \) be an exponential random variable with mean 1. We first show that, under \( P_0 \) (respectively \( Q \)), \( \theta(s_1, s_2) \) 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 \( \sigma_f \) is the probability that \( Z \leq f \).

A recent result of [CW14] gives a formula for the detection boundary of the sparse signal detection problem for general \( P_0, P_1 \). However, applying this formula here is non-trivial. Instead we bound directly the total variation distance between \( P_0^{\otimes m} \) and \( Q^{\otimes m} \). Similarly to the approach used in [CW14], we work instead with the Hellinger distance \( H^2(P_0^{\otimes m}, Q^{\otimes m}) \) which tensorizes as follows (see e.g. [CT91])

\[
\frac{1}{2} H^2(P_0^{\otimes m}, Q^{\otimes m}) = 1 - \left( 1 - \frac{1}{2} H^2(P_0, Q) \right)^m, \tag{5}
\]

and further satisfies

\[
\|P_0^{\otimes m} - Q^{\otimes m}\|_{TV}^2 \leq H^2(P_0^{\otimes m}, Q^{\otimes m}) \left[ 1 - \frac{1}{4} H^2(P_0^{\otimes m}, Q^{\otimes m}) \right]. \tag{6}
\]

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

\[
H^2(P_0, Q) = O \left( f^2 \sqrt{k} \right).
\]

The details are in Section 3.

The proof of Theorem 2 on the other hand involves the construction of a statistical test that distinguishes between \( P_0^{\otimes m} \) and \( Q^{\otimes m} \). In the regime \( k = O(1) \), an optimal test (up to constants) compares the means of the samples [DNRI14]. In the regime \( k = \omega(f^{-2}) \), an optimal test (up to constants) compares the minima of the samples [MR10]. 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.

**Organization.** The gene tree generating model is defined in Section 2. The proof of Theorem 1 can be found in Section 3. For lack of space, most details are in the appendix. The proof of Theorem 2 can also be found in the appendix.

## 2 Further definitions

In this section, we give more details on the model.
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 [RY03]. 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 $\tau$ 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. [Dur08] for an introduction to coalescent theory.)

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

**Lemma 1 (Distance distribution)** Let $S$ be a two-leaf species tree with $d_{12} = 2\tau$ and $\nu_e = 1$ for all $e$ and let $\theta(s^1_1, s^1_2)$ be as in (2) for some $k$. Then the distribution of $\theta(s^1_1, s^1_2)$ 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 2 (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 $P_0$ and $Q$ be the distributions of $\theta(s^1_1, s^1_2)$ for a single gene under $S$ and $S^+$ respectively. Then, there is $P_1$ such that,

$$Q = (1 - \sigma_f) P_0 + \sigma_f P_1,$$

where $\sigma_f = O(f)$, as $f \to 0$. More specifically, $P_1$ is obtained by conditioning $Q$ on the event that $Z$ is $\leq f$ and $\sigma_f$ 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

$$
\prod_{j=1}^{m} \prod_{e \in E} \exp \left(-\frac{O_{j}^{e}}{2}\left[\sigma_{j}^{e,O_{j}^{e}+1} - \sigma_{j}^{e,O_{j}^{e}}\right]\right) \prod_{\ell=1}^{I_{j}^{e}-O_{j}^{e}} \exp \left(-\frac{\ell}{2}\left[\sigma_{j}^{e,\ell} - \sigma_{j}^{e,\ell-1}\right]\right),
$$

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 $\ell$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 $T_j$ 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.

A more abstract setting Before proving Theorem 1, we re-set the problem in a more generic setting that will make the computations more transparent. We consider two distributions $P_0$ and $P_1$ for a random variable $\theta$ taking values in $\{0, \ldots, k\}$ for some $k$. We assume that the distribution of $\theta$ takes the form

$$
P_0[\theta = \ell] = \binom{k}{\ell} \mathbb{E}_0[X^\ell (1-X)^{k-\ell}],
$$

where $\mathbb{E}_0$ is the expectation operator corresponding to $P_0$, and $X$ is some random variable admitting a density over $[0, 1]$. The distribution is similarly defined under $P_1$. We make the following assumptions, which are satisfied in the setting of the previous section:

A1. Under $P_0$ and $P_1$, $X$ admits a density whose support is $(p_0, p^0)$ under $P_0$ and $(p_0 - \phi_f, p_0)$ under $P_1$, where $0 < p_0 < p^0 < 1$ (independent of $f$) and $\phi_f = O(f)$. (In the setting of Lemma 2, $p_0 = \frac{3}{4}(1-e^{-2})$, $p_0 - \phi_f = \frac{3}{4}(1-e^{-(2-2f)})$, and $p^0 = 3/4$.)

A2. Under $P_0$, 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 2 under $P_0$ the density of $X$ on $(p_0, p^0)$ is $\frac{4e^{1/2}}{3}(1-4x/3)^{-3/4}$.)

As before, we let

$$
\mathbb{Q} = (1 - \sigma_f) P_0 + \sigma_f P_1,
$$

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

9
3 Lower bound

In the abstract setting of Section 2, the Hellinger distance can be written as

\[ H^2(P_0, Q) = \sum_{j=0}^{k} \left[ \sqrt{Q[\theta = j]} - \sqrt{P_0[\theta = j]} \right]^2 \]

\[ = \sum_{j=0}^{k} \left[ \sqrt{1 + \sigma_f \left( \frac{P_1[\theta = j]}{P_0[\theta = j]} - 1 \right)} - 1 \right] P_0[\theta = j] \]

\[ = \sum_{j=0}^{k} \left[ \sqrt{1 + \sigma_f \left( \frac{E_1[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} - 1 \right)} - 1 \right] P_0[\theta = j] \]  \hspace{1cm} (7)

We prove the following proposition, which implies Theorem 1.

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

\[ H^2(P_0, Q) = O \left( f^2 \sqrt{k} \right). \]

From (7), in order to bound the Hellinger distance, we need to control the ratio \( \frac{E_1[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} \) and the probability \( P_0[\theta = j] \). Because the standard deviation of \( \theta/k \) 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 \( E_0 \) (an event of probability \( O(1/\sqrt{k}) \)) and \( \theta/k \) being within \( O(1/\sqrt{k}) \) of \( p_0 \) as well (in which case the ratio \( \frac{E_1[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} \) 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, as we describe next.

3.1 Proof of Proposition 1

Let \( C \) be a large constant (not depending on \( f \)) to be determined later. We divide up the sum in (7) into intervals with distinct behaviors. We consider the following intervals for \( \frac{j}{k} \):

\[ J_0 = \left[ p_0, p_0 + C \sqrt{\frac{\log k}{k}} \right], \quad J_1 = \left[ p_0 + C \sqrt{\frac{\log k}{k}}, 1 \right], \]

and

\[ J_0' = [p_0 - \phi_f, p_0], \quad J_1' = \left[ p_0 - C \sqrt{\frac{\log k}{k}}, p_0 - \phi_f \right], \quad J_2' = \left[ 0, p_0 - C \sqrt{\frac{\log k}{k}} \right]. \]

For a subset of \( \frac{j}{k} \)-values \( J \), we write the contribution of \( J \) to the Hellinger distance as

\[ H^2(P_0, Q)|_J = \sum_{j:j/k \in J} \left[ \sqrt{1 + \sigma_f \left( \frac{E_1[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} - 1 \right)} - 1 \right]^2 P_0[\theta = j]. \]
**High number of substitutions** We first consider $J_1$. Let

$$J_{\leq 1} = \left\{ 0 \leq j \leq k : \frac{E_1[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} \leq 1 \right\}.$$  

We show in the appendix that $J_1 \subseteq J_{\leq 1}$ which, in turn, we show implies

$$H^2(\mathbb{P}_0, \mathbb{Q})|_{J_1} = O(f^2).$$  

(8)

**Low number of substitutions** We further divide up $J'_0 \cup J'_1 \cup J'_2$ into

$$I'_0 = \left[ p_0 - \frac{1}{\sqrt{k}}, p_0 \right],$$

$$I'_\ell = \left[ p_0 - \frac{2^\ell - 1}{\sqrt{k}}, p_0 - \frac{2\ell}{\sqrt{k}} \right], \quad \ell \geq 1.$$  

We prove in the appendix the following bound on the expectation ratio. For all $j/k \in J'_0 \cup J'_1 \cup J'_2$,

$$\frac{E_0[X^j(1-X)^{k-j}]}{E_1[X^j(1-X)^{k-j}]} \geq C_2 \frac{1}{\sqrt{k}} \exp \left( -C_1 \sqrt{k} \left( p_0 - \frac{j}{k} \right) \right),$$

for some constants $C_1, C_2 > 0$. We also show that, for all $\ell \geq 0$,

$$\mathbb{P}_0[\theta/k \in I'_\ell] \leq \frac{C_3 \exp \left( -2^{2\ell - 1} \right)}{\sqrt{k}},$$

for some constant $C_3 > 0$. Combining these bounds we show that for this regime

$$H^2(\mathbb{P}_0, \mathbb{Q})|_{J'_0 \cup J'_1 \cup J'_2} = O(f^2 \sqrt{k}).$$  

(9)

**Border regime** We now consider $J_0$. Define $L$ by $2^L = C \sqrt{\log k}$ (and assume for simplicity that it is integer-valued), and

$$I_0 = \left[ p_0, p_0 + \frac{1}{\sqrt{k}} \right],$$

$$I_\ell = \left[ p_0 + \frac{2^\ell - 1}{\sqrt{k}}, p_0 + \frac{2\ell}{\sqrt{k}} \right], \quad \ell \geq 1.$$  

We prove in the appendix that, for all $j \in I_\ell$, $\ell \leq L$,

$$\frac{E_0[X^j(1-X)^{k-j}]}{E_1[X^j(1-X)^{k-j}]} \geq C_2 \frac{1}{\sqrt{k}} \exp \left( C_1 2^{2\ell} \right),$$

for some constants $C_1, C_2 > 0$ and that, for all $0 \leq \ell < L$,

$$\mathbb{P}_0[\theta/k \in I_\ell] \leq \frac{C_3 2^{2\ell}}{\sqrt{k}},$$

for some constant $C_3 > 0$. Combining these bounds we show that for this regime it also holds that

$$H^2(\mathbb{P}_0, \mathbb{Q})|_{J_0} = O(f^2 \sqrt{k}).$$  

(10)

Adding up Equations (8), (9) and (10) proves Proposition 1.
Acknowledgments

We thank Gautam Dasarathy and Rob Nowak for helpful discussions.

References

[ADHR10] 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.

[ADR11] 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.

[ALPE12] 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.

[BS14] 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.

[CGG02] 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.

[CJJ11] 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.

[CJT05] 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.

[CT91] 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.

[CW14] 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.

[DBP05] 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.

[DD10] 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.
[DDBR09] 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.

[DJ04] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 06 2004.

[DMR11a] 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.

[DMR11b] 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.

[DNR14] 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.

[Dob58] 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.

[DR06] J. H. Degnan and N. A. Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5), May 2006.

[DR09] 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.

[DR10] Constantinos Daskalakis and Sébastien Roch. Alignment-free phylogenetic reconstruction. In *RECOMB*, pages 123–137, 2010.

[Dur96] Richard Durrett. *Probability: theory and examples*. Duxbury Press, Belmont, CA, second edition, 1996.

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.

[ESSW99a] 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.

[ESSW99b] 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.

[Fel04] J. Felsenstein. *Inferring Phylogenies*. Sinauer, New York, New York, 2004.
[HTF09] 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.

[Ing97] Yu. I. Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.*, 6(1):47–69, 1997.

[JC69] T. H. Jukes and C. Cantor. Mammalian protein metabolism. In H. N. Munro, editor, *Evolution of protein molecules*, pages 21–132. Academic Press, 1969.

[JCL10] 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.

[KHH+05] Martin Kulldorff, Richard Heffernan, Jessica Hartman, Renato Assuno, and Farzad Mostashari. A space?time permutation scan statistic for disease outbreak detection. *PLoS Med.*, 2(3):e59, 02 2005.

[KMRR15] Junhyong Kim, Elchanan Mossel, Mikls Z. Rcz, and Nathan Ross. Can one hear the shape of a population history? *Theoretical Population Biology*, 100(0):26 – 38, 2015.

[LYK+09] 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.

[LYP10] 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.

[Mad97] Wayne P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.

[MFP08] Simon Myers, Charles Fefferman, and Nick Patterson. Can one learn history from the allelic spectrum? *Theoretical Population Biology*, 73(3):342 – 348, 2008.

[Mos03] E. Mossel. On the impossibility of reconstructing ancestral data and phylogenies. *J. Comput. Biol.*, 10(5):669–678, 2003.

[Mos04] E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404, 2004.

[MR05] 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.

[MR10] 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.
[MRS11] 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.

[MSZ+13] 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.

[Nak13] 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.

[Roc10] Sebastien Roch. Toward extracting all phylogenetic information from matrices of evolutionary distances. *Science*, 327(5971):1376–1379, 2010.

[Roc13] 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.

[RS15] Sebastien Roch and Mike Steel. Likelihood-based tree reconstruction on a concatenation of alignments can be positively misleading. *Theoretical Population Biology*, 2015. To appear.

[RW15] 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.

[RY03] 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.

[SS02] 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.

[SS03] C. Semple and M. Steel. *Phylogenetics*, volume 22 of *Mathematics and its Applications series*. Oxford University Press, 2003.
A Some useful lemmas

The following is Lemma 4 in [CW14]:

**Lemma 3** For $b > 0$, let $h_b(s) = \left(\sqrt{1 + b(s - 1)} - 1\right)^2$.

1. For any $b > 0$, the function $h_b(s)$ is strictly decreasing on $[0, 1]$ and strictly increasing on $[1, +\infty)$.
2. For any $b > 0$ and $s \geq 1$,

$$h_b(s) \leq [b(s - 1)] \land [b(s - 1)]^2 \leq [bs] \land [bs]^2.$$

The following lemmas follow from straightforward calculus.

**Lemma 4** For $j \in \{0, \ldots, k\}$ and $x \in (0, 1)$, let

$$\Phi_j(x) = \frac{j}{k} \log x + \frac{k - j}{k} \log(1 - x).$$

Then

$$\Phi'_j(x) = \frac{j}{k} \frac{1}{x} - \frac{k - j}{k} \frac{1}{1 - x} = \frac{1}{x(1 - x)} \left(\frac{j}{k} - x\right).$$

As a result $\Phi_j$ is increasing on $[0, \frac{j}{k}]$ and decreasing on $[\frac{j}{k}, 1]$, and $\Phi'_j(\frac{j}{k}) = 0$.

**Lemma 5** For $j \in \{0, \ldots, k\}$, $p \in (0, 1)$, and $x \in [0, p)$, let

$$\Psi_{j,p}(x) = \frac{j}{k} \log \frac{p}{p - x} + \frac{k - j}{k} \log \frac{1 - p}{1 - p + x}.$$ 

Then:

1. The first two derivatives are:

$$\Psi'_{j,p}(x) = \frac{j}{k} \frac{1}{p - x} - \frac{k - j}{k} \frac{1}{1 - p + x} = \frac{1}{(p - x)(1 - p + x)} \left(\frac{j}{k} - (p - x)\right),$$

and

$$\Psi''_{j,p}(x) = \frac{j}{k} \left\{\frac{1}{(p - x)^2}\right\} + \frac{k - j}{k} \left\{\frac{1}{(1 - p + x)^2}\right\} \geq \frac{1}{2},$$

(since the terms in curly brackets are at least 1 and one of $\frac{j}{k}$ or $\frac{k - j}{k}$ is greater or equal than $1/2$).

2. By a Taylor expansion around $x = 0$, we have for $x \in [0, p)$ and some $x^* \in [0, x]$

$$\Psi_{j,p}(x) = \frac{1}{p(1 - p)} \left(\frac{j}{k} - p\right) x + \frac{x^2}{2} \Psi''_{j,p}(x^*) \geq \frac{1}{p(1 - p)} \left(\frac{j}{k} - p\right) x + \frac{1}{4} x^2.$$
B Lower bound: proofs

High number of substitution We consider $J_1$ first. Let

$$J_{\leq 1} = \left\{ 0 \leq j \leq k : \frac{E_1[X^j(1 - X)^{k-j}]}{E_0[X^j(1 - X)^{k-j}]} \leq 1 \right\}.$$  

By Lemma 3 in Section A

$$\sum_{j \in J_{\leq 1}} \left[ \sqrt{1 + \sigma_f \left( \frac{E_1[X^j(1 - X)^{k-j}]}{E_0[X^j(1 - X)^{k-j}]} - 1 \right)} - 1 \right]^2 P_0[J = j]$$

$$\leq \sum_{j \in J_{\leq 1}} \left[ \sqrt{1 - \sigma_f} - 1 \right]^2 P_0[J = j]$$

$$= O(\sigma_f^2) = O(f^2),$$

which of course is $O(f^2 \sqrt{k}).$

Lemma 6 (Ratio less than 1)

$$H^2(P_0, Q)|_{J_{\leq 1}} = O(f^2).$$

Hence it suffices to show that $J_1 \subseteq J_{\leq 1}$. Lemma 4 and Assumption A1 imply that, for $\frac{i}{k} \in J_1$,

$$E_1[X^i(1 - X)^{k-j}] = E_1[\exp(k\Phi_j(X))]$$

$$\leq \exp(k\Phi_j(p_0))$$

$$= p_0^i(1 - p_0)^{k-j}.$$  

(11)

Let $\mathcal{E}$ be the event that

$$\mathcal{E} = \left\{ X \in \left[ p_0 + C \sqrt{\frac{\log k}{k}} - \frac{1}{k}, p_0 + C \sqrt{\frac{\log k}{k}} \right] \right\}$$

By Assumption A2, $P_0[\mathcal{E}] \geq \rho/k$. Hence, using Lemma 4 again, for $\frac{i}{k} \in J_1$

$$E_0[X^i(1 - X)^{k-j}] = E_0[X^i(1 - X)^{k-j} | \mathcal{E}]P_0[\mathcal{E}] + E_0[X^i(1 - X)^{k-j} | \mathcal{E}^c]P_0[\mathcal{E}^c]$$

$$\geq \frac{\rho}{k} p_0^i(1 - p_0)^{k-j}.$$  

(12)

where $p = p_0 + C \sqrt{\frac{\log k}{k}} - \frac{1}{k}$.
Combining (11) and (12), and using Lemma 5 with $x = p - p_0 \geq 0$ (for $k > 1$ and $C$ large enough), we have

$$
\frac{\mathbb{E}_0[X^j(1 - X)^{k-j}]}{\mathbb{E}_1[X^j(1 - X)^{k-j}]} \geq \frac{\rho p^j (1 - p)^{k-j}}{k p_0^j (1 - p_0)^{k-j}} = \frac{\rho}{k} \exp(k \Psi_{j,p}(p - p_0))
$$

$$
\geq \frac{\rho}{k} \exp\left(\frac{k}{4} \left(\frac{1}{p(1-p)} \left(\frac{j}{k} - p\right) x + \frac{1}{4} x^2\right)\right)
$$

$$
\geq \frac{\rho}{k} \exp\left(\frac{k}{4} \left(C \sqrt{\frac{\log k}{k}} - \frac{1}{k}\right)^2\right)
$$

$$
\geq \frac{\rho}{k} \exp\left(\frac{C^2}{5} \log k\right)
$$

$$
\geq 1,
$$

for $C$ large enough (assuming $k$ is large), where on the fourth line we used that $j/k - p \geq 0$ for $j/k \in J_1 = \left[p_0 + C \sqrt{\frac{\log k}{k}}, 1\right]$. That implies $J_1 \subseteq J_{\leq 1}$:

Claim 1 (High substitution: Hellinger distance)

$$H^2(\mathbb{P}_0, \mathbb{Q})|_{J_1} = o(f^2).$$

Low number of substitutions  We further divide up $J_0' \cup J_1' \cup J_2'$ into

$$I_0' = \left[p_0 - \frac{1}{\sqrt{k}}, p_0\right]$$

$$I_\ell' = \left[p_0 - \frac{2^\ell}{\sqrt{k}}, p_0 - \frac{2^{\ell-1}}{\sqrt{k}}\right], \quad \ell \geq 1.$$

Define $L$ by $2^L = C \sqrt{\log k}$ (and assume for simplicity that it is integer-valued).

We first bound $\mathbb{E}_1[X^j(1 - X)^{k-j}]$ using Lemma 4 and Assumption A1:

• On $J_0'$,

$$\mathbb{E}_1[X^j(1 - X)^{k-j}] \leq (j/k)^j (1 - j/k)^{k-j}. \quad (13)$$

• On $J_1' \cup J_2'$,

$$\mathbb{E}_1[X^j(1 - X)^{k-j}] \leq (p_0 - \phi_j)^j (1 - p_0 + \phi_j)^{k-j}. \quad (14)$$

To bound $\mathbb{E}_0[X^j(1 - X)^{k-j}]$, we consider the event

$$\mathcal{E} = \left\{ X \in \left[p_0, p_0 + \sqrt{\frac{1}{k}}\right]\right\}. $$
By Assumption A2 and Lemma [4] on \( J'_0 \cup J'_1 \cup J'_2 \), arguing as in (12)

\[
\mathbb{E}_0[X^j(1 - X)^{k-j}] \geq \frac{\rho}{\sqrt{k}} p^j(1 - p)^{k-j},
\]

(15)

where \( p = p_0 + \sqrt{\frac{1}{k}} \) (assuming \( k \) is large). Combining (13), (14), and (15), and using Lemma 5

- On \( J'_0 \),

\[
\frac{\mathbb{E}_0[X^j(1 - X)^{k-j}]}{\mathbb{E}_1[X^j(1 - X)^{k-j}]} \geq \frac{\rho}{\sqrt{k}} \exp(k \Psi_{j,p}(p - j/k))
\]

\[
\geq \frac{\rho}{\sqrt{k}} \exp \left( k \left( -\frac{1}{p(1 - p)} \left( \frac{j}{k} - p \right)^2 + \frac{1}{4} \left( \frac{j}{k} - p \right)^2 \right) \right)
\]

\[
\geq C'_1 \frac{\rho}{\sqrt{k}},
\]

for some constant \( C'_1 \) (not depending on \( f \)), where we used that \( \phi_f \ll \sqrt{1/k} \) so that \( \left( \frac{j}{k} - p \right)^2 = O(1/k) \) and, further, \( p(1 - p) \in (0, 1/4) \).

- On \( J'_1 \cup J'_2 \),

\[
\frac{\mathbb{E}_0[X^j(1 - X)^{k-j}]}{\mathbb{E}_1[X^j(1 - X)^{k-j}]} \geq \frac{\rho}{\sqrt{k}} \exp(k \Psi_{j,p}(p - p_0 + \phi_f))
\]

\[
\geq \frac{\rho}{\sqrt{k}} \exp \left( k \left( \frac{1}{p(1 - p)} \left( \frac{j}{k} - p \right) (p - p_0 + \phi_f) + \frac{1}{4} (p - p_0 + \phi_f)^2 \right) \right)
\]

\[
\geq \frac{\rho}{\sqrt{k}} \exp \left( -C_1 \sqrt{k} \left( p - \frac{j}{k} \right) \right)
\]

\[
= \frac{\rho}{\sqrt{k}} \exp \left( -C_1 \sqrt{k} \left( p_0 + \sqrt{\frac{1}{k}} - \frac{j}{k} \right) \right)
\]

\[
= C_2 \frac{\rho}{\sqrt{k}} \exp \left( -C_1 \sqrt{k} \left( p_0 - \frac{j}{k} \right) \right),
\]

for some constants \( C_1, C_2 \) (not depending on \( f \)), where again we used that \( \phi_f \ll \sqrt{1/k} \) so that \( (p - p_0 + \phi_f)^2 = O(1/k^2) \).

By decreasing \( C_2 \) appropriately we combine the two bounds into:

**Claim 2 (Low substitution: Likelihood ratio)** For all \( j/k \in J'_0 \cup J'_1 \cup J'_2 \),

\[
\frac{\mathbb{E}_0[X^j(1 - X)^{k-j}]}{\mathbb{E}_1[X^j(1 - X)^{k-j}]} \geq C_2 \frac{1}{\sqrt{k}} \exp \left( -C_1 \sqrt{k} \left( p_0 - \frac{j}{k} \right) \right).
\]

(16)

We now bound the integrand in \( H^2(\mathbb{P}_0, \mathbb{Q}) \) over \( J'_0 \cup J'_1 \cup J'_2 \). Observe first that on \( J_{\leq 1} \)

\[
h_{\sigma_f} \left( \frac{\mathbb{E}_1[X^j(1 - X)^{k-j}]}{\mathbb{E}_0[X^j(1 - X)^{k-j}]} \right) = C_0 \rho^2,
\]

(17)

for some constant \( C_0 > 0 \). Then:
Claim 3 (Low substitution: Integrand) For all \( j/k \), because

\[
\sigma_j \frac{\sqrt{k}}{C_2} \exp \left( C_1 \sqrt{k} \left( p_0 - \frac{j}{k} \right) \right) = O(f^\alpha) \exp \left( O(\sqrt{\log f^{-1}}) \right) = o(1),
\]

we have by Lemma 3 (Part 2) and (17)

\[
h_{\sigma_j} \left( \frac{E_j[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} \right) \leq C_0 f^2 \sqrt{k} \exp \left( 2C_1 \sqrt{k} \left( p_0 - \frac{j}{k} \right) \right).
\]

Changing the constants we re-write this expression as

\[
h_{\sigma_j} \left( \frac{E_j[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} \right) \leq C_2 f^2 k \exp \left( C_1 \sqrt{k} \left( p_0 - \frac{j}{k} \right) \right).
\]

On \( J_2 \), the expectation ratio multiplied by \( \sigma_j \) may (or may not) be larger than 1, therefore by Lemma 3 (Part 2) and (17) again we have (adapting the constants)

\[
h_{\sigma_j} \left( \frac{E_j[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} \right) \leq C_0 f^2 \sqrt{k} \exp \left( C_1 \sqrt{k} \left( p_0 - \frac{j}{k} \right) \right).
\]

Changing the constants we re-write this expression as

\[
h_{\sigma_j} \left( \frac{E_j[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} \right) \leq C_2 f \sqrt{k} \exp \left( C_1 \sqrt{k} \left( p_0 - \frac{j}{k} \right) \right).
\]

where we only squared the exponential and used the fact that \( f \sqrt{k} = o(1) \).

We combine the two bounds into:

**Claim 3 (Low substitution: Integrand)** For all \( j/k \in J_0' \cup J_1' \cup J_2' \),

\[
h_{\sigma_j} \left( \frac{E_j[X^j(1-X)^{k-j}]}{E_0[X^j(1-X)^{k-j}]} \right) \leq C_2 (f^2 k \mathbb{1}_{j/k \in J_0' \cup J_1'} + f \sqrt{k} \mathbb{1}_{j/k \in J_2'}) \exp \left( C_1 \sqrt{k} \left( p_0 - \frac{j}{k} \right) \right). \tag{18}
\]

It remains to bound the integrator. We let

\[
I_0 = \left[ p_0, p_0 + \frac{1}{\sqrt{k}} \right], \quad I_\ell = \left[ p_0 + \frac{2^{\ell-1}}{\sqrt{k}} p_0 + \frac{2^\ell}{\sqrt{k}} \right], \quad \ell \geq 1.
\]

Let \( \Lambda > 0 \) be such that \( 2^\Lambda = (\bar{p} - p_0)\sqrt{k} \). Then by Assumption A2

\[
\mathbb{P}_0[\theta/k \in I_\ell'] = \sum_{\lambda \geq 0} \mathbb{P}_0[\theta/k \in I_\ell | X \in I_{\lambda}] \mathbb{P}_0[X \in I_{\lambda}] \leq \sum_{\lambda \geq 0} \mathbb{P}_0[\theta/k \in I_\ell | X \in I_{\lambda}] \frac{2^{\Lambda-1}}{\sqrt{k}} \rho^{-1} + \sum_{\lambda > \Lambda} \mathbb{P}_0[\theta/k \in I_\ell | X \in I_{\lambda}],
\]

20
By Chernoff’s bound
\[ P_0[\theta/k \in I_\ell | X \in I_\lambda] \leq \exp \left( -2(2^{\ell-1} + 2^{\lambda-1})^2 \right) \leq \exp \left( -2^{2\ell-1} - 2^{2\lambda-1} \right). \]

In particular
\[ \sum_{\lambda > \Lambda} P_0[\theta/k \in I_\ell | X \in I_\lambda] \leq \exp \left( -2^{2\ell-1} \right) \sum_{\lambda > \Lambda} \exp \left( -2^{\lambda-1} \right) \leq \exp \left( -2^{2\ell-1} \right) \exp \left( -C_3' k \right), \]

for some constant \( C_3' > 0 \) (not depending on \( f \)). On the other hand,
\[ \sum_{\Lambda \geq 0} P_0[\theta/k \in I_\ell | X \in I_\lambda] \leq \exp \left( -2^{2\ell-1} \right) \rho \sum_{\Lambda \geq 0} 2^{\lambda-1} \exp \left( -2^{\lambda-1} \right) \leq \frac{C_3 \exp \left( -2^{2\ell-1} \right)}{\sqrt{k}}, \]

for a constant \( C_3 > 0 \) (not depending on \( f \)). Combining the bounds and increasing \( C_3 \) appropriately, we get

**Claim 4 (Low substitution: Integrator)** For all \( \ell \geq 0 \),
\[ P_0[\theta/k \in I_\ell] \leq \frac{C_3 \exp \left( -2^{2\ell-1} \right)}{\sqrt{k}}. \tag{19} \]

We can now compute the contribution of \( J'_0 \cup J'_1 \cup J'_2 \) to the Hellinger distance. Recall that \( L \) is defined by \( 2^L = C^{\sqrt{\log k}} \). From (18) and (19), we get:

- For \( 0 \leq \ell \leq L \),
\[ H^2(P_0, Q)|_{I_\ell} \leq C_4 f^2 \sqrt{k} \exp \left( -C_7 C^2 \log k \right) = o(f^{1+\kappa}) = o(f^2 \sqrt{k}), \]

for some constants \( C_4, C_5 > 0 \). Summing over \( \ell \) we get
\[ \sum_{\ell=0}^{L} H^2(P_0, Q)|_{I_\ell} \leq C_6 f^2 \sqrt{k}, \]

for some constant \( C_6 > 0 \).

- Similarly, for \( \ell > L \),
\[ H^2(P_0, Q)|_{I_\ell} \leq C_5 f \exp \left( -C_4 2^{2\ell} \right), \]

adapting constants \( C_4, C_5 > 0 \). Summing over \( \ell \) we get
\[ \sum_{\ell>L} H^2(P_0, Q)|_{I_\ell} \leq C_8 f \exp \left( -C_7 C^2 \log k \right) = o(f^{1+\kappa}) = o(f^2 \sqrt{k}), \]

by choosing \( C \) large enough.
Combining these bounds we get finally:

Claim 5 (Low substitution: Hellinger distance)

\[ H^2(\mathbb{P}_0, \mathbb{Q})|_{J_0' \cup J_1 \cup J_2} = O(f^2 \sqrt{k}). \]

Border regime. We now consider \( J_0 \). We first bound \( \mathbb{E}_1[X_j^j(1 - X)^{k-j}] \) using Lemma 4 and Assumption A1

\[ \mathbb{E}_1[X_j^j(1 - X)^{k-j}] \leq p_j^j(1 - p_0)^{k-j}. \]  

To bound \( \mathbb{E}_0[X_j^j(1 - X)^{k-j}] \), we consider the event

\[ \mathcal{E} = \left\{ X \in \left[ \frac{j}{k}, \frac{j}{k} + \sqrt{\frac{1}{k}} \right] \right\}. \]

By Assumption A2 and Lemma 4 on \( J_0 \),

\[ \mathbb{E}_0[X_j^j(1 - X)^{k-j}] \geq \frac{\rho}{\sqrt{k}} p_j^j(1 - p)^{k-j}, \]  

where \( p = \frac{j}{k} + \sqrt{\frac{1}{k}} \). Combining (20) and (21), and using Lemma 5 on \( J_0 \),

\[ \frac{\mathbb{E}_0[X_j^j(1 - X)^{k-j}]}{\mathbb{E}_1[X_j^j(1 - X)^{k-j}]} \geq \frac{\rho}{\sqrt{k}} \exp(k \Psi_{j,p}(p - p_0)) \geq \frac{\rho}{\sqrt{k}} \exp \left( k \left( \frac{1}{p(1-p)} \left( \frac{j}{k} - p \right) (p - p_0) + \frac{1}{4} (p - p_0)^2 \right) \right). \]

For \( j/k \in I_\ell, \ell \leq L \),

\[ \frac{\mathbb{E}_0[X_j^j(1 - X)^{k-j}]}{\mathbb{E}_1[X_j^j(1 - X)^{k-j}]} \geq \frac{\rho}{\sqrt{k}} \exp \left( k \left( - \frac{1}{p(1-p)} \sqrt{\frac{1}{k}} \left( \frac{2^\ell + 1}{\sqrt{k}} \right) + \frac{1}{4} \left( \frac{2^{\ell-1} + 1}{\sqrt{k}} \right)^2 \right) \right) \geq C_2 \frac{1}{\sqrt{k}} \exp \left( C_1 2^{2\ell} \right), \]

for constants \( C_1, C_2 > 0 \).

Claim 6 (Border regime: Likelihood ratio) For all \( j/k \in I_\ell, \ell \leq L \),

\[ \frac{\mathbb{E}_0[X_j^j(1 - X)^{k-j}]}{\mathbb{E}_1[X_j^j(1 - X)^{k-j}]} \geq C_2 \frac{1}{\sqrt{k}} \exp \left( C_1 2^{2\ell} \right). \]

We now bound the integrand in \( H^2(\mathbb{P}_0, \mathbb{Q}) \). Because \( \sigma_f \sqrt{k} = o(1) \), by Lemma 3 (Part 2) and (17) we have on \( I_\ell \)

\[ h_{\sigma_f} \left( \frac{\mathbb{E}_1[X_j^j(1 - X)^{k-j}]}{\mathbb{E}_0[X_j^j(1 - X)^{k-j}]} \right) \leq C_0 f^2 \sqrt{\frac{\sigma_f k}{C_2}} \exp \left( -2 C_1 2^{2\ell} \right). \]

Changing the constants we re-write this expression as

\[ h_{\sigma_f} \left( \frac{\mathbb{E}_1[X_j^j(1 - X)^{k-j}]}{\mathbb{E}_0[X_j^j(1 - X)^{k-j}]} \right) \leq C_0 f^2 \sqrt{C_2 f^2 k} \exp \left( -C_1 2^{2\ell} \right). \]
Claim 7 (Border regime: Integrand) For all \( j/k \in I_\ell \), \( 0 \leq \ell \leq L \),
\[
    h_{\sigma f} \left( \frac{\mathbb{E}[X^j (1 - X)^{k-j}]}{\mathbb{E}[X]} \right) \leq C_0 f^2 \vee C_2 f^2 k \exp \left( -C_1 2^\ell \right). \tag{23}
\]

It remains to bound the integrator. We have by Assumption A2 (recall that \( 2^\Lambda = (\bar{p} - p_0) \sqrt{k} \))
\[
    \mathbb{P}_0[\theta/k \in I_\ell] = \sum_{\lambda \geq 0} \mathbb{P}_0[\theta/k \in I_\ell | X \in I_\lambda] \mathbb{P}_0[X \in I_\lambda]
    \leq \sum_{0 \leq \lambda \leq \ell} \mathbb{P}_0[X \in I_\lambda] + \sum_{\ell < \lambda \leq \Lambda} \mathbb{P}_0[\theta/k \in I_\ell | X \in I_\lambda] \mathbb{P}_0[X \in I_\lambda]
    + \sum_{\lambda > \Lambda} \mathbb{P}_0[\theta/k \in I_\ell | X \in I_\lambda].
\]

By Chernoff’s bound, for \( \lambda > \ell \),
\[
    \mathbb{P}_0[\theta/k \in I_\ell | X \in I_\lambda] \leq \exp \left( -2(-2^\ell + 2^{\lambda-1})^2 \right) \leq \exp \left( -2^{2\ell+1}(2^{\lambda-1} - 1)^2 \right).
\]

In particular
\[
    \sum_{\lambda > \Lambda} \mathbb{P}_0[\theta/k \in I_\ell | X \in I_\lambda] \leq \exp \left( -C_3' k \right),
\]
for some constant \( C_3' > 0 \) (not depending on \( f \)). On the other hand,
\[
    \sum_{\ell \leq \lambda \leq \Lambda} \mathbb{P}_0[\theta/k \in I_\ell | X \in I_\lambda] \frac{2^{\lambda-1}}{\sqrt{k}} \rho^{-1} \leq \frac{C_3 2^\ell}{\sqrt{k}},
\]
for a constant \( C_3 > 0 \) (not depending on \( f \)). Combining the bounds and increasing \( C_3 \) appropriately, we get

Claim 8 (Border substitution: Integrator) For all \( 0 \leq \ell < L \),
\[
    \mathbb{P}_0[\theta/k \in I_\ell] \leq \frac{C_3 2^\ell}{\sqrt{k}}. \tag{24}
\]

We can now compute the contribution of \( J_0 \) to the Hellinger distance. From (23) and (24), we get for \( 0 \leq \ell \leq L \)
\[
    H^2(\mathbb{P}_0, \mathbb{Q})|_{I_\ell} \leq \left[ C_0 f^2 \vee C_2 f^2 k \exp \left( -C_1 2^\ell \right) \right] \frac{C_3 2^\ell}{\sqrt{k}}.
\]

Summing over \( \ell \) we get
\[
    \sum_{\ell=0}^L H^2(\mathbb{P}_0, \mathbb{Q})|_{I_\ell} \leq C_4 f^2 \sqrt{k},
\]
for some constant \( C_4 > 0 \).
Claim 9 (Border regime: Hellinger distance)

\[ H^2(P_0, Q)_{\|_0} = O(f^2 \sqrt{k}). \]

C Matching upper bound: proofs

We give two proofs of the upper bound.

C.1 Proof of Theorem

Proof: We use (4) and construct an explicit test \( A \) as follows.

Let \( W \) be the number of genes such that \( \theta/k \leq p_0 \). Let \( w = P_0[\theta/k \leq p_0] \) and \( w' = Q[\theta/k \leq p_0] \).

Then \( W \sim \text{Bin}(m, w) \) under \( P_0 \) and \( W \sim \text{Bin}(m, w') \) under \( Q \). Note that, by Assumption A1, \( X \in [p_0 - \phi_f, p_0] \) under \( P_1 \). By the Berry-Esseen theorem (e.g. [Dur96]),

\[ P_1[\theta/k \leq p_0] \geq E_1[P_1[\theta \leq kX | X]] = \frac{1}{2} - O\left( \frac{1}{\sqrt{k}} \right) \geq \frac{1}{3}, \tag{25} \]

for \( k \) large enough. Hence,

\[ w' = \sigma_f P_1[\theta/k \leq p_0] + (1 - \sigma_f)w \geq \frac{1}{3} \sigma_f + (1 - \sigma_f)w, \tag{26} \]

whereas by the computations in the previous section (more specifically, by summing over \( \ell \) in (19))

\[ w = O\left( \frac{1}{\sqrt{k}} \right), \tag{27} \]

and, similarly, since \( f \sqrt{k} = o(1) \)

\[ w' = O\left( \frac{1}{\sqrt{k}} \right), \tag{28} \]

from (26) and (27). Consequently

\[ w' - w \geq \sigma_f \left( \frac{1}{3} - w \right) = \Omega(f). \tag{29} \]

Let

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

and consider the event

\[ A = \{ W \geq w^* \}. \]

By Chebyshev’s inequality,

\[ P_0^m[A] \leq \frac{4mw(1-w)}{m^2(w'-w)^2} = O\left( \frac{1}{mf^2 \sqrt{k}} \right) \leq \frac{\delta}{2}. \]
for $c'$ large enough, where we used (27) and (29). Similarly,

$$Q^\otimes m[A^c] \leq \frac{4mw'(1-w')}{m^2(w'-w)^2} \leq \frac{\delta}{2}.$$ 

\[\square\]

### C.2 Agnostic version

Although Theorem 2 shows that our bound in Theorem 1 is tight, it relies on a test (i.e., the set $A$) that assumes knowledge of the null and alternative hypotheses. Here we relax this assumption.

**Pairwise distance comparisons** We assume that we have two (independent) collections of genes, $\mathcal{T}_1$ and $\mathcal{T}_2$, one from each model, $\mathbb{P}_0$ and $\mathbb{Q}$ as in the previous section. We split the genes into two equal-sized disjoint sub-collections, $(\mathcal{T}^1_1, \mathcal{T}^1_2)$ and $(\mathcal{T}^2_1, \mathcal{T}^2_2)$. Assume for convenience that the total number of genes is in fact $2m$ for each dataset. Let $C > 0$ be a constant, to be determined later. We proceed in two steps.

1. We first compute $\hat{p}_1$ and $\hat{p}_2$, the $C\sqrt{k}$-quantiles based on $\mathcal{T}^1_1$ and $\mathcal{T}^1_2$ respectively. Let $\hat{p} = \max\{\hat{p}_1, \hat{p}_2\}$.

2. Compute the fraction of genes, $\hat{w}_1$ and $\hat{w}_2$, with $\theta/k \leq \hat{p}$ in $\mathcal{T}^2_1$ and $\mathcal{T}^2_2$ respectively.

We infer that the first dataset comes from $\mathbb{P}_0^\otimes 2m$ if $\hat{w}_1 < \hat{w}_2$, and vice versa.

**Remark 1** Simply comparing the $C\sqrt{k}$-quantiles breaks down when $f \ll \frac{1}{k}$, as then the signal is smaller than the typical bin size. This explains the need for the second phase in our algorithm.

We show that this approach succeeds with probability at least $1 - \delta$ whenever $m \geq c'f^2/\sqrt{k}$, for $c'$ large enough. This proceeds from a series of claims.

**Claim 10** For $c'$ large enough, there is $C_1 > 0$ such that

$$\hat{p} \in \left[p_0, p_0 + \frac{C_1}{\sqrt{k}}\right]$$

with probability $1 - \delta/2$.

**Proof:** By summing over $\ell$ in (19),

$$\mathbb{P}_0[\theta/k \leq p_0] \leq \frac{C'_0}{\sqrt{k}},$$

for some $C'_0 > 0$. For any $C''_1 > C'_1$, there is $C_1 > 0$ such that

$$\mathbb{P}_0\left[\theta/k \leq p_0 + \frac{C_1}{\sqrt{k}}\right] \geq \mathbb{P}_0\left[\theta/k \leq p_0 + \frac{C_1}{\sqrt{k}}\right]X \in \left[p_0, p_0 + \frac{C_1}{\sqrt{k}}\right] \mathbb{P}_0\left[X \in \left[p_0, p_0 + \frac{C_1}{\sqrt{k}}\right]\right] \geq \frac{1}{3} \rho C_1 \geq \frac{C''_1}{\sqrt{k}}$$

25
by the Berry-Essen theorem (as in (25)), for \(C_1\) large enough.

Let
\[
C = \frac{C_1' + C_1''}{2}.
\]

Let \(W\) be the number of genes (among \(m\)) such that \(\theta/k \leq p_0\) and \(w = \mathbb{P}_0[\theta/k \leq p_0]\). Repeating the calculations in the proof of Theorem 2,
\[
\mathbb{P}^\otimes_m \left[ W \geq m \frac{C}{\sqrt{k}} \right] \leq \frac{4mw(1-w)k}{m^2(C-C_1')^2} = \frac{1}{m} O \left( \sqrt{k} \right) \leq \frac{1}{c'} O(f^2k) \leq \frac{\delta}{8},
\]
for \(c'\) large enough. Similarly, let \(\tilde{W}\) be the number of genes such that \(\theta/k \leq p_0 + C_1/\sqrt{k}\) and \(\tilde{w} = \mathbb{P}_0[\theta/k \leq p_0 + C_1/\sqrt{k}]\). Then
\[
\mathbb{P}^\otimes_m \left[ \tilde{W} \leq m \frac{C}{\sqrt{k}} \right] \leq \frac{\delta}{8}.
\]

That implies that with probability \(1 - \delta/4\) the \(C/\sqrt{k}\)-quantile under \(\mathbb{P}^\otimes_m\) lies in the interval \([p_0, p_0 + C_1/\sqrt{k}]\). By monotonicity, \(\mathbb{P}_1[\theta/k \leq p_0 + C_1/\sqrt{k}] \geq \tilde{w}\), and we also have
\[
\mathbb{Q}^\otimes_m \left[ \tilde{W} \leq m \frac{C}{\sqrt{k}} \right] \leq \frac{\delta}{8},
\]
which implies the claim. \(\blacksquare\)

Claim 11 For \(c'\) large enough, if \(T_1\) comes from \(\mathbb{P}^\otimes_{2m}\), \(T_2\) comes from \(\mathbb{Q}^\otimes_{2m}\) and (30) holds, then \(\hat{w}_1 < \hat{w}_2\) with probability \(1 - \delta/2\), and vice versa.

Proof: The proof is identical to that of Theorem 2 with \(W\) now being the number of genes such that \(\theta/k \leq \hat{p}\), \(w = \mathbb{P}_0[\theta/k \leq \hat{p}]\), \(w' = \mathbb{Q}[\theta/k \leq \hat{p}]\), and (27) and (28) now following from Claim 10 together with (19) and (24). \(\blacksquare\)

Triplet reconstruction Consider again the three possible species trees depicted in Figure 2. By comparing the pairs two by two as described in the previous paragraph, we can determine which is the correct species tree topology. Such “triplet” information is in general enough (assuming the molecular clock hypothesis) to reconstruct a species tree on any number of species (e.g. [SS03]). We leave out the details.