The Annals of Applied Probability
2013, Vol. 23, No. 2, 693–721
DOI: 10.1214/12-AAP852
© Institute of Mathematical Statistics, 2013

ALIGNMENT-FREE PHYLOGENETIC RECONSTRUCTION: SAMPLE COMPLEXITY VIA A BRANCHING PROCESS ANALYSIS

BY CONSTANTINOS DASKALAKIS[1] AND SEBASTIEN ROCH[2]

Massachusetts Institute of Technology and University of California, Los Angeles

We present an efficient phylogenetic reconstruction algorithm allowing insertions and deletions which provably achieves a sequence-length requirement (or sample complexity) growing polynomially in the number of taxa. Our algorithm is distance-based, that is, it relies on pairwise sequence comparisons. More importantly, our approach largely bypasses the difficult problem of multiple sequence alignment.

1. Introduction. We introduce a new efficient algorithm for the *phylogenetic tree reconstruction* (PTR) problem which rigorously accounts for insertions and deletions.

Phylogenetic background. A *phylogenetic tree* or *phylogeny* is a tree representing the speciation history of a group of organisms. The leaves of the tree are typically existing species. The root corresponds to their *most recent common ancestor* (MRCA). Each branching in the tree indicates a speciation event. It is common to assume that DNA evolves according to a Markovian substitution process on this phylogeny. Under such a model, a *gene* is a sequence in $\{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}^k$. Along each edge of the tree, each site independently mutates according to a Markov rate matrix. The length of a branch is a measure of the amount of substitution along that branch. The precise definition of a branch length depends on the model of evolution. For roughly constant mutation rates, one can think of the branch length as proportional to the amount of time elapsed along a branch. The PTR problem consists of estimating a phylogeny from the genes observed at its leaves. We denote the leaves of a tree by $[n] = \{1, \ldots, n\}$ and their sequences by $\sigma_1, \ldots, \sigma_n$.

The model of sequence evolution above is simplistic: it ignores many mutational events that DNA undergoes through evolution. At the gene level, the most important omissions are insertions and deletions of sites, also called *indels*. Stochastic models taking indels into account have long been known [39, 40], but they are not widely used in practice (or in theory) because of their complexity. Instead, most practical algorithms take a two-phase approach:

Received September 2011; revised February 2012.

[1]Supported by a Sloan Foundation Fellowship and NSF Award CCF-0953960 (CAREER) and CCF-1101491.

[2]Supported by NSF Grant DMS-10-07144.

MSC2010 subject classifications. Primary 60K35; secondary 92D15.

Key words and phrases. Phylogenetic reconstruction, alignment, branching processes.

(1) *Multiple sequence alignment.* Site t_i of sequence σ_i and site t_j of sequence σ_j are said to be *homologous* if they descend from the same site t_0 of a common ancestor u (not necessarily the MRCA) *only through substitutions*. In the *multiple sequence alignment* (MSA) problem, we seek roughly to uncover the homology relation between $\sigma_1, \ldots, \sigma_n$. Typically, the output is represented by a matrix $\mathbf{D}$ of n aligned sequences of equal length with values in $\{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}, -\}$. Each column of the matrix corresponds to homologous sites. The state $-$ is called a *gap* and is used to account for insertions and deletions. For instance, if sequence σ_l does not have a site corresponding to t_0 in u above, then a gap is aligned with positions t_i of σ_i and t_j of σ_j (which belong to the same column).

(2) *Phylogenetic tree reconstruction.* The matrix $\mathbf{D}$ is then cleaned up by removing all columns containing gaps. Let $\mathbf{D}'$ be this new matrix. A standard PTR algorithm is then applied to $\mathbf{D}'$. Note that substitutions alone suffice to explain $\mathbf{D}'$.

Traditionally, most of the research on phylogenetic methods has focused on the second phase.

In fact, current theoretical analyses of PTR assume that the MSA problem has been solved *perfectly*. This has been a long-standing assumption in evolutionary biology. But this simplification is increasingly being questioned in the phylogenetic literature, where it has been argued that alignment heuristics often create systematic biases that affect analysis [26, 42]. Much recent empirical work has been devoted to the proper joint estimation of alignments and phylogenies [25–28, 32, 37, 39, 40]. Here we give the first analysis of an efficient, provably consistent PTR algorithm in the presence of indels. Our new algorithm suggests that a rough alignment suffices for an accurate tree reconstruction (bypassing the computationally difficult multiple alignment problem).

Theoretical properties of PTR. In addition to computational efficiency, an important theoretical criterion in designing a PTR algorithm is the so-called *sequence-length requirement* (SLR). At a minimum, a reconstruction algorithm should be *consistent*, that is, assuming a model of sequence evolution, the output should be guaranteed to converge on the true tree as the sequence length k (the number of *samples*) goes to $+\infty$ [15]. Beyond consistency, the sequence-length requirement (or convergence rate) of a PTR algorithm is the sequence length required for guaranteed high-probability reconstruction. The SLR is typically given as an asymptotic function of n, the number of leaves of the tree. Of course, it also depends on the substitution parameters.

A classical result due to Erdős et al. [13] states that, for general trees under the assumption that all branch lengths are bounded by constants, the so-called short quartet method (SQM) has poly(n)-SLR. The SQM is a particular PTR algorithm based on estimating evolutionary distances between the leaf taxa, that is, the sum of the branch lengths between species. Such algorithms are known as *distance-based methods*. The basic theoretical result behind distance-based methods is the following: the collection of pairwise evolutionary distances between all species forms a

special metric on the leaves known as an additive metric; under mild regularity assumptions, such a metric *characterizes* the underlying phylogeny interpreted as an edge-weighted tree, that is, there is a one-to-one correspondence between additive metrics and phylogenies; moreover, the mapping between them can be computed efficiently [34].

A new approach. In the classical theoretical setting above where the MSA problem is assumed perfectly solved (we refer to this setting below as the ESSW framework), the evolutionary distance between two species is measured using the Hamming distance (or a state-dependent generalization) between their respective sequences. It can be shown that after a proper correction for multiple substitutions (which depends on the model used) the expectation of the quantity obtained does satisfy the additive metric property and can therefore serve as the basis for a distance-based PTR algorithm.

Moving beyond the ESSW framework, it is tempting to account for indels by simply using edit distance instead of the Hamming distance. Recall that the *edit distance* or *Levenshtein distance* between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion or substitution of a single character. However, no analytical expression is known for the expectation of edit distance under standard indel models and computing such an expression appears difficult (if at all possible). An alternative idea is to compute the *maximum likelihood estimator* for the time elapsed between two species given their sequences. But this involves solving a nonconvex optimization problem and the likelihood is only known to be efficiently computable under a rather unrealistic assumption known as reversibility [39] (see below).

We use a different approach. We divide the sequences into quantile blocks (the first $x\%$, the second $x\%$, etc.). We show that by appropriately choosing x above we can make sure that the blocks in different sequences essentially "match" each other, that is, they are made of mostly homologous sites. We then compare the state frequencies in matching blocks and build an additive metric out of this statistic. As we show below, this is in fact a natural generalization of the Hamming estimator of the ESSW framework. However, unlike the Hamming distance which can easily be analyzed through standard concentration inequalities, proving rigorously that our approach works involves several new technical difficulties. Our analysis relies on a branching process analysis of the site displacements. We give a quick proof sketch after the formal statement of our results in Section 1.2.

The results described here were first announced without proof in the special case of ultrametric trees under the CFN model with inverse logarithmic indel rates [10]. Here we give full proofs of stronger results, including extensions to bounded-rate trees under GTR models.

Related work. For more background on models of molecular evolution and phylogenetics, see, for example, [16, 17, 34]. Following the seminal results of [13],

there has been much work on sequence-length requirement, including [4–9, 14, 18, 20, 23, 24, 29, 30, 33, 35, 36].

The multiple sequence alignment problem as a combinatorial optimization problem (finding the best alignment under a given pairwise scoring function) is known to be NP-hard [12, 41]. Most heuristics used in practice, such as CLUSTAL [19], MAFFT [22] and MUSCLE [11], use the idea of a guide tree, that is, they first construct a very rough phylogenetic tree from the data (using, e.g., edit distance as a measure of evolutionary distance), and then recursively construct local alignments produced by "aligning alignments."

To our knowledge, little theoretical work has been dedicated to the joint estimation of alignments and phylogenies, with the exception of Thatte [38] who gave consistency results for the reversible case in the limit where the deletion-to-insertion ratio tends to 1. However, no sequence-length requirement is obtained in [38]. In recent related work, the problem of reconstructing ancestral sequences in the presence of indels was considered [1, 2].

1.1. *Model of sequence evolution.*

Phylogeny. A *phylogeny* is represented by a binary tree $T = (V, E)$, whose leaves $L \subset V$ correspond to extant species, and whose bifurcations denote evolutionary events whereby two new species are generated from an ancestor. The root of the phylogeny, denoted by $r(T)$, represents the common ancestor of all the species in the phylogeny, and we assume that all edges of T are directed away from $r(T)$; so, if $e = (u, v)$ is a branch of the phylogeny, u is the *parent* of v and v is the *child* of u. Moreover, if v' is in the subtree of T rooted at u, we call v' a *descendant* of u and u an *ancestor* of v'.

Along each branch of the phylogeny, the genetic material of the parent species is subject to modifications that produce the genetic material of its child species. A common biological assumption is that the genetic material of each species u can be represented by a binary sequence $\sigma_u = (\sigma_u^1, \ldots, \sigma_u^{K_u})$ of length K_u over a finite alphabet—for ease of presentation, we work with a binary alphabet $\{0, 1\}$ (but see Section 5 for extensions to richer alphabets)—and that the changes to which σ_u is subjected along the branch $e = (u, v)$ are described by a Markov process. In particular, the Markov property implies that, given the sequence σ_u of u, the sequence σ_v is independent of the sequences of the species outside the subtree of T rooted at u.

A simplifying assumption commonly used in phylogenetics is that all species have sequences of the same length and, moreover, that every *site*, that is, every coordinate, in their sequences evolves independently from every other site. In particular, it is assumed that, along each branch $e = (u, v)$ of the phylogeny, every site σ_u^j of the sequence σ_u is flipped with probability p_e to the value $1 - \sigma_u^j$ independently from the other sites. This model is known as the Cavender–Farris–Neyman (CFN) model. A simple generalization to $\{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$ is known as the Jukes–Cantor (JC) model (see, e.g., [16]).

Accouting for indels. In this paper, we consider a more general evolutionary process that accounts for the possibility of insertions and deletions. Our model is similar to the original TKF91 model [39], except that we do not enforce reversibility. In our model, every edge $e = (u, v)$ of the phylogeny is characterized by a quadruple of parameters $(t_e; \eta_e, \mu_e, \lambda_e)$, where t_e is the evolutionary time between the species u and v, and η_e, μ_e and λ_e are, respectively, the *substitution*, *deletion* and *insertion* rates. The Markov process by which the sequence at v is obtained from the sequence at u is defined below (see, e.g., [21] for background on continuous-time Markov processes).

DEFINITION 1.1 (Evolutionary process on a branch). Given an edge $e = (u, v)$, with parameters $(t_e; \eta_e, \mu_e, \lambda_e)$, the sequence σ_v at v is obtained from the sequence σ_u at u according to the following Markov process:

(1) Intialize $\sigma_v := \sigma_u$, $K_v := K_u$ and $t_\ell := t_e$ (where t_ℓ is the remaining time on the edge e).
(2) While $t_\ell > 0$:

- (*Timing of next event*) let $I_0, I_1, \ldots, I_{K_v}$ be exponential random variables with rate λ_e, $D_1, \ldots, D_{K_v}$ exponential random variables with rate μ_e and $M_1, \ldots, M_{K_v}$ exponential random variables with rate η_e; suppose that these random variables are mutually independent and let $\mathcal{T}$ be their minimum;
- if $\mathcal{T} > t_\ell$, the process ends at t_ℓ; otherwise:
  - (*Insertion*) if $I_j = \mathcal{T}$, insert a new site whose value is chosen uniformly at random from $\{0, 1\}$ between the sites σ_v^j and σ_v^{j+1} of σ_v;
  - (*Deletion*) if $D_j = \mathcal{T}$, delete the site σ_v^j from σ_v;
  - (*Substitution*) and if $M_j = \mathcal{T}$, replace σ_v^j by $1 - \sigma_v^j$;

  (If $j = 0$, then σ_v^j is undefined and, if $j = K_v$, then σ_v^{j+1} is undefined.)
- (*Remaining time*) update σ_v according to these changes, and update K_v to reflect the new sequence length; set the remaining time $t_\ell := t_\ell - \mathcal{T}$.

In words, the evolutionary process defined above assumes that every site of the sequence σ_u of the parent species is, independently from the other sites, subjected to a sequence of evolutionary events that flip its value; these events are distributed according to a Poisson point process of intensity η_e in the time interval $[0, t_e]$. However, the site may get deleted and therefore not be inherited by the sequence of the node v; this is determined by whether an exponential random variable of rate μ_e is smaller than t_e. While each site of the parental sequence σ_u is subjected to this process, new sites are introduced in the space between existing sites at rate λ_e, and each of these sites follows a similar process for the remaining time. In essence, insertion and deletion events are governed by an independent branching process for each ancestral site. Note further that the order of the sites, as described above, also plays a role.

REMARK 1.2. Unlike [39], we do not use an "immortal link" and we do not assume that the length process is at stationarity. Our techniques can also be applied to the *TKF*91 model without much modifications. We leave the details to the reader.

Given the evolutionary process on a branch of the phylogeny, the evolutionary process on the whole phylogeny is defined as follows.

DEFINITION 1.3 (Evolutionary process). Suppose that every site of the sequence $\sigma_{r(T)}$ at the root of the phylogeny is chosen to be 0 or 1 uniformly at random. Recursively, if σ_u is the sequence at node u and $e = (u, v)$ is an edge of the phylogeny, the sequence σ_v at node v is obtained from the sequence σ_u by an application of the evolutionary process on a branch described by Definition 1.1.

For ease of exposition, we first present our proof in the special case where the substitution, insertion and deletion rates are the same on all edges of the phylogeny.

DEFINITION 1.4 (Ultrametric assumption). Under the ultrametric assumption, the leaves of the phylogeny are contemporaneous, that is, there exists H such that for each $u \in L$ the sum of evolutionary times t_e on the branches between u and the root is H.

DEFINITION 1.5 (Molecular clock assumption). Under the molecular clock assumption, we assume that the ultrametric assumption holds. Moreover, there exist η, μ and λ such that $\eta_e = \eta$, $\mu_e = \mu$ and $\lambda_e = \lambda$, for all $e \in E$.

We discuss a more general case in Section 5.

Notation. In the sequel, we label the leaves of the phylogeny with the positive integers $1, 2, \ldots, n$, so that $L = \{1, \ldots, n\}$, and the root $r(T)$ of the phylogeny with 0.

1.2. *Main result.*

Statement of results. We begin with a consistency result. Here we consider a completely general phylogeny, that is, neither the ultrametric nor the molecular clock assumptions need hold.

THEOREM 1 (Consistency: finite case). *Assume that* $0 < t_e, \eta_e, \lambda_e, \mu_e < +\infty$, *for all* $e \in E$. *Then there exists a procedure returning the correct tree from the sequences at the leaves, with probability of failure approaching* 0 *as the sequence length at the root of the tree goes to* $+\infty$.

Our main result is the following. For simplicity we first work under the symmetric two-state case and assume that the molecular clock assumption holds.

THEOREM 2 (Main result: two-state, molecular clock case). *Consider the two-state model under the molecular clock assumption. Assume further that there exist constants*

$$0 < f, \qquad g < +\infty,$$

independent of n, *such that*

$$f < t_e < g \qquad \forall e \in E.$$

Moreover, assume that

$$\eta_e = \eta, \qquad \lambda_e = \lambda, \qquad \mu_e = \mu \qquad \forall e \in E,$$

where η, λ *and* μ *are bounded between constants* (*independent of* n) $0 < \underline{\eta} < \bar{\eta} < +\infty$, $0 = \underline{\lambda} < \bar{\lambda} < +\infty$ *and* $0 = \underline{\mu} < \bar{\mu} < +\infty$, *respectively. Under the assumptions above, for all* $\beta' > 0$ *there exists* $\beta'' > 0$ *such that there exists a polynomial-time algorithm solving the phylogenetic reconstruction problem* (*i.e., returning the correct tree*) *with probability of failure* $n^{-\beta'}$, *if the root sequence has length* $k_r \geq n^{\beta''}$.[3]

REMARK 1.6 (Branch lengths). Our assumption that all branch lengths t_e, $e \in E$, satisfy $f < t_e < g$ is standard in the sequence-length requirement literature following the seminal work of [13].

Extensions. In Section 5 we derive the following extension. Let Q be a reversible 4×4 rate matrix with stationary distribution π. (Larger alphabets are also possible.) The GTR sequence evolution process is identical to the one described in Definition 1.1 except that the substitution process is a continuous-time Markov process with rate matrix $\eta_e Q$.

THEOREM 3 (Main result: GTR, bounded-rates case). *Consider the GTR model with rate matrix* Q *under the ultrametric assumption* (*but not necessarily the molecular clock assumption*). *Assume further that there exist constants*

$$0 < f, \qquad g, \underline{\eta}, \bar{\eta}, \underline{\lambda}, \bar{\lambda}, \underline{\mu}, \bar{\mu}, < +\infty,$$

independent of n, *such that*

$$f < t_e < g, \qquad \underline{\eta} < \eta_e < \bar{\eta} \qquad \forall e \in E.$$

[3] In [10], a preliminary version of this result was announced without proof, with the much stronger assumption that $\bar{\lambda}, \bar{\mu} = O(1/\log n)$, that is, that the indel rates are negligible. Here we show that this assumption can be relaxed (at the cost of longer sequences).

Moreover, assume that

$$\lambda_e = \lambda, \qquad \mu_e = \mu \qquad \forall e \in E,$$

where λ *and* μ *are bounded between constants (independent of* n*)* $0 = \underline{\lambda} < \bar{\lambda} < +\infty$ *and* $0 = \underline{\mu} < \bar{\mu} < +\infty$*, respectively. We refer to the conditions above as the bounded-rates assumption. Under the assumptions above, for all* $\beta' > 0$ *there exists* $\beta'' > 0$ *such that there exists a polynomial-time algorithm solving the phylogenetic reconstruction problem (i.e., returning the correct tree) with probability of failure* $n^{-\beta'}$*, if the root sequence has length* $k_r \geq n^{\beta''}$*.*

Proof sketch. Consider the two-state, molecular clock case. As we noted before, unlike the classical setting where the Hamming distance can be analyzed through standard concentration inequalities, proving rigorously that our approach works involves several new technical difficulties. The proof goes through the following steps:

(1) *Expectations.* We first compute expectations of block statistics, which involve analyzing a continuous-time Markov process. We use these calculations to define an appropriate additive metric based on correlations between blocks.

(2) *Sequence length and site displacements.* We give bounds on how much sequence lengths vary across the tree through a moment-generating function argument. Using our bounds on the sequence length process, we bound the worst-case displacements of the sites. Namely, we show that, under our assumptions, all sites move by at most $O(\sqrt{k \log k})$.

(3) *Sequence partitioning.* We divide each sequence in blocks of size roughly k^{ζ} for $\zeta > 1/2$, where k is the sequence length at the root. From our bounds on site displacements, it follows that the blocks roughly match across different sequences. In particular, we bound the number of homologous sites between matching blocks with high probability and show that the expected correlation between these blocks is approximately correct.

(4) *Concentration.* Finally, we show that our estimates are concentrated. The concentration argument proceeds by conditioning on the indel process satisfying the high-probability conditions in the previous points.

The crux of our result is the proper estimation of an additive metric. With such an estimation procedure in hand, we can use a standard distance-based approach to recover the phylogeny.

Organization. The rest of the paper is organized as follows. The evolutionary distance forming the basis of our approach is presented in Section 2. We describe our full distance estimator in Section 3 and prove its concentration in the same section. Extensions are described in Section 5.

2. Evolutionary distances. Consider the two-state, molecular clock case. In this section, we show how to define an appropriate notion of "evolutionary distance" between two species. Although such distances have been widely used in prior phylogenetic work and have been defined for a variety of models [16, 34], to our knowledge our definition is the first that applies to models with indels. We begin by reviewing the standard definition in the indel-free case and then adapt it to the presence of indels. Our estimation procedure is discussed in Section 3.

2.1. *The classical indel-free case.* Suppose first that $\lambda = \mu = 0$, that is, *there is no indel*. In that case, the sequence length remains fixed at k and the alignment problem is trivial. Underlying all distance-based approaches is the following basic definition.

DEFINITION 2.1 (Additive metric). A phylogeny is naturally equipped with a so-called additive metric on the leaves $\mathcal{D}: L \times L \to (0, +\infty)$ defined as

$$\forall a, b \in L \qquad \mathcal{D}(a, b) = \sum_{e \in \mathrm{P}_T(a,b)} \omega_e,$$

where $\mathrm{P}_T(a, b)$ is the set of edges on the path between a and b in T and where ω_e is a nonnegative function of the parameters on e (in our case, t_e, η_e, λ_e and μ_e). For instance, a common choice for ω_e would be $\omega_e = \eta_e t_e$ in which case $\mathcal{D}(a, b)$ is the expected number of substitutions per site between a and b. Often $\mathcal{D}(a, b)$ is referred to as the "evolutionary distance" between species a and b. Additive metrics are characterized by the following four-point condition: for all $a, b, c, d \in L$,

$$\mathcal{D}(a, b) + \mathcal{D}(c, d) \leq \max\{\mathcal{D}(a, c) + \mathcal{D}(b, d), \mathcal{D}(a, d) + \mathcal{D}(b, c)\}.$$

Moreover, assuming $\omega_e > 0$ for all $e \in E$, it is well known that there exists a one-to-one correspondence between $\mathcal{D}$ and T as a weigthed tree with edge weights $\{\omega_e\}_{e \in E}$. We will discuss algorithms for constructing T from $\mathcal{D}$ in Section 4. For more background on tree-based metrics, see [34].

Definition 2.1 implies that phylogenies can be reconstructed by computing $\mathcal{D}(a, b)$ for all pairs of leaves $a, b \in L$. Assume we seek to estimate the evolutionary distance between species a and b using their respective sequences. In a first attempt, one might try the (normalized) Hamming distance between $\sigma_a = (\sigma_a^1, \ldots, \sigma_a^k)$ and $\sigma_b = (\sigma_b^1, \ldots, \sigma_b^k)$. However, the expected Hamming distance (in other words, the probability of disagreement between a site of a and b) does not form an additive metric as defined in Definition 2.1. Instead, it is well known that an appropriate estimator is obtained by "correcting" the Hamming distance for "multiple" substitutions. Denoting by $\widehat{\mathcal{H}}(\sigma_a, \sigma_b)$ the Hamming distance between σ_a and σ_b, a Markov chain calculation shows that $\mathcal{D}(a, b) = -\frac{1}{2}\log(1 -$

$2\mathbb{E}[\widehat{\mathcal{H}}(\sigma_a, \sigma_b)]$), with the choice $\omega_e = \eta_e t_e$ (see, e.g., [16]). In a distance-based reconstruction procedure, one first estimates $\mathcal{D}$ with

$$\widehat{\mathcal{D}}(a, b) = -\tfrac{1}{2}\log\big(1 - 2\widehat{\mathcal{H}}(\sigma_a, \sigma_b)\big) \tag{1}$$

and then applies one of the algorithms discussed in Section 4 below. The sequence-length requirement of such a method can be derived by using concentration results for $\widehat{\mathcal{H}}$ [4, 13].

2.2. *Taking indels into account.* To simplify the presentation, we assume throughout that $\lambda \neq \mu$. The case $\lambda = \mu$ follows from the same argument.

In the presence of indels, the estimator (1) based on the Hamming distance is difficult to apply. One has to first align the sequences, which cannot be done perfectly and causes biases as well as correlations that are hard to analyze. Alternatively, one could try a different string distance such as edit distance. However, computing the expectation of edit distance under indel models appears difficult.

We use a different approach involving correlations between state frequencies. We will eventually apply the estimator to large sub-blocks of the sequences (see Section 3), but we first describe it for the full sequence for clarity. For a node u, let K_u be the (random) length of the sequence at u and Z_u, the number of 0's in the sequence at u. Then, our distance estimator is

$$\widehat{\mathcal{D}}(a, b) = \big(Z_a - \tfrac{1}{2}K_a\big)\big(Z_b - \tfrac{1}{2}K_b\big).$$

We now analyze the expectation of this quantity. For $u \in V$, we let

$$\Delta_u = Z_u - \tfrac{1}{2}K_u$$

be the deviation of Z_u from its expected value (conditioned on the sequence length).

Single channel. Suppose T is made of a single edge from the root r to a leaf a with parameters t, η, λ, μ. Assume first that the original sequence length is $k_r = 1$. Let K_a be the length of the sequence at a. Then K_a is a continuous-time branching process and, by Markov chain calculations ([3], Section III.5), its moment-generating function is

$$F(s, t) \equiv \mathbb{E}[s^{K_a}] = \frac{\mu(s-1) - e^{(\mu-\lambda)t}(\lambda s - \mu)}{\lambda(s-1) - e^{(\mu-\lambda)t}(\lambda s - \mu)}. \tag{2}$$

By differentiating $F(s, t)$ we derive

$$\mathbb{E}[K_a] = e^{-(\mu-\lambda)t} \tag{3}$$

and

$$\mathrm{Var}[K_a] = \frac{\mu + \lambda}{\mu - \lambda}\big[e^{-(\mu-\lambda)t} - e^{-2(\mu-\lambda)t}\big]. \tag{4}$$

Let K_a^* be the number of "new" sites at a, that is, excluding the original site if it survived. (We ignore the substitutions for the time being.) The probability that the original site survives is $e^{-\mu t}$. Then,

$$\mathbb{E}[K_a^*] = \mathbb{E}[K_a - \mathbb{1}\{\text{original site survives}\}] = e^{-(\mu-\lambda)t} - e^{-\mu t}$$

by linearity of expectation.

We now take into account substitutions. Assume that the original sequence length at r is a random variable K_r and that the sequence at r is i.i.d. uniform. Denote by Z_r the number of 0's at r. The probability that a site in r, that is still surviving in a, has flipped its value is

$$\begin{aligned} p &= \mathbb{P}[\text{state flips odd number of times in time } t] \\ &= \sum_{j=0}^{+\infty} e^{-\eta t} \frac{(\eta t)^{2j+1}}{(2j+1)!} \\ &= e^{-\eta t} \sinh \eta t \\ &= \frac{1 - e^{-2\eta t}}{2}. \end{aligned}$$

Also, note that a new site created along the path between r and a has equal chance of being 0 or 1 *at the end of the path*. Then we have the following lemma.

LEMMA 2.2 (Single channel: expected deviation). *The following holds*:

$$\mathbb{E}[\Delta_a | K_r, Z_r] = e^{-(2\eta+\mu)t} \Delta_r.$$

PROOF. We have

$$\begin{aligned} \mathbb{E}[\Delta_a | K_r, Z_r] &= \mathbb{E}\big[(Z_a - \tfrac{1}{2}K_a) | K_r, Z_r\big] \\ &= Z_r e^{-\mu t}(1-p) + (K_r - Z_r)e^{-\mu t} p \\ &\quad + K_r\big(e^{-(\mu-\lambda)t} - e^{-\mu t}\big)\tfrac{1}{2} - K_r e^{-(\mu-\lambda)t}\tfrac{1}{2} \qquad (5) \\ &= Z_r(1-2p)e^{-\mu t} - \tfrac{1}{2}K_r(1-2p)e^{-\mu t} \\ &= e^{-2\eta t} e^{-\mu t} \Delta_r, \end{aligned}$$

where on the first two lines:

(1) the first term is the number of original 0's surviving in state 0;
(2) the second term is the number of original 1's surviving in state 0;
(3) the third term is the number of new sites surviving in state 0 (where recall that new sites are uniformly chosen in $\{0, 1\}$);
(4) the fourth term is half the sequence length at a given the length at r. □

Fork channel. Consider now a "fork" tree, that is, a root r from which emanates a single edge $e_u = (r, u)$ which in turn branches into two edges $e_a = (u, a)$ and $e_b = (u, b)$ (see Figure 1 below). For $x = a, b, u$, we denote the parameters of edge e_x by $t_x, \lambda_x, \mu_x, \eta_x$. Our goal is to compute $\mathbb{E}[\widehat{\mathcal{D}}(a, b)]$ assuming that the sequence length at the root is k_r. We use (5), the Markov property and the fact that Z_u conditioned on K_u is a binomial with parameters $1/2$ and K_u. We get the following lemma.

LEMMA 2.3 (Fork channel: expected distance). *The following holds*:

$$\mathbb{E}[\widehat{\mathcal{D}}(a, b)] = e^{-(2\eta_a+\mu_a)t_a} e^{-(2\eta_b+\mu_b)t_b} e^{-(\mu_u-\lambda_u)t_u} \frac{k_r}{4}.$$

PROOF. We have

$$\begin{aligned}
\mathbb{E}[\widehat{\mathcal{D}}(a, b)] &= \mathbb{E}[\Delta_a \Delta_b] \\
&= \mathbb{E}[\mathbb{E}[\Delta_a \Delta_b | K_u, Z_u]] \\
&= \mathbb{E}[\mathbb{E}[\Delta_a | K_u, Z_u]\mathbb{E}[\Delta_b | K_u, Z_u]] \\
&= e^{-2\eta_a t_a} e^{-\mu_a t_a} e^{-2\eta_b t_b} e^{-\mu_b t_b} \mathbb{E}[\Delta_u^2] \\
&= e^{-2\eta_a t_a} e^{-\mu_a t_a} e^{-2\eta_b t_b} e^{-\mu_b t_b} \mathbb{E}[\mathbb{E}[\Delta_u^2 | K_u]] \\
&= e^{-2\eta_a t_a} e^{-\mu_a t_a} e^{-2\eta_b t_b} e^{-\mu_b t_b} \mathbb{E}\left[\frac{K_u}{4}\right] \\
&= e^{-2\eta_a t_a} e^{-\mu_a t_a} e^{-2\eta_b t_b} e^{-\mu_b t_b} \frac{e^{-(\mu_u-\lambda_u)t_u} k_r}{4},
\end{aligned}$$

where we used (3) and Lemma 2.2. □

Molecular clock. We specialize the previous result to the molecular clock assumption. That is, we assume, for $x = a, b, u$, that $\lambda_x = \lambda$, $\mu_x = \mu$ and $\eta_x = \eta$. Note that by construction $t_a = t_b$ (assuming species a and b are contemporary). We denote $t = t_a$ and $\bar{t} = t_u + t_a$. Denoting $\kappa = \frac{k_r e^{-(\mu-\lambda)\bar{t}}}{4}$, we then get the following lemma.

LEMMA 2.4 (Molecular clock: expected distance). *The following holds*:

$$\mathbb{E}[\widehat{\mathcal{D}}(a, b)] = e^{-(4\eta+\mu+\lambda)t} \kappa.$$

Letting

$$\beta = 4\eta + \mu + \lambda,$$

we get that

$$-2 \log \mathbb{E}[\kappa^{-1} \widehat{\mathcal{D}}(a, b)] = 2\beta t,$$

which is the evolutionary distance between a and b with the choice $\omega_e = \beta t_e$. Therefore, we define the following estimator:

$$\widehat{\mathcal{D}}^*(a,b) = -2\log\kappa^{-1}\widehat{\mathcal{D}}(a,b).$$

3. Distance computation. We now show how to estimate the evolutionary distance between two species by decomposing the sequences into large blocks which serve as roughly independent samples. We use the following notation: $M_t = e^{-(\mu-\lambda)t}$, $D_t = e^{-\mu t}$, $\delta = \mu - \lambda$, $\phi = \mu + \lambda$ and $\Gamma_t = \delta^{-1}\lambda(1 - M_t)$.

We show in Section 4 that the time elapsed between the root and the leaves is bounded by $\frac{g^2}{f}\log_2 n$. Hence, under our assumptions

$$\begin{aligned}\Upsilon_n^{-1} &\equiv e^{-(\bar{\mu}+\bar{\lambda})(g^2/f)\log_2 n} \le e^{-(\bar{\mu}g^2/f)\log_2 n} \\ &\le M_t \le e^{(\bar{\lambda}g^2/f)\log_2 n} \le e^{(\bar{\lambda}+\bar{\mu})(g^2/f)\log_2 n} \equiv \Upsilon_n,\end{aligned} \tag{6}$$

$$\Upsilon_n^{-1} \le e^{-(\bar{\mu}g^2/f)\log_2 n} \le D_t \le 1 \tag{7}$$

and

$$\begin{aligned}0 \le \Gamma_t = \lambda t\frac{1 - e^{-(\mu-\lambda)t}}{(\mu-\lambda)t} &\le \bar{\lambda}\frac{g^2}{f}\log_2 n\frac{e^{(\bar{\lambda}g^2/f)\log_2 n} - 1}{(\bar{\lambda}g^2/f)\log_2 n} \\ &= e^{(\bar{\lambda}g^2/f)\log_2 n} - 1 \le \Upsilon_n,\end{aligned} \tag{8}$$

where we used that the function $x^{-1}(1 - e^{-x})$ is nonnegative and decreasing since its derivative is

$$\frac{xe^{-x} - (1 - e^{-x})}{x^2} = e^{-x}\frac{(1 + x) - e^x}{x^2} \le 0, \qquad x \ne 0.$$

Note that the bounds above are polynomials in n with exponents depending only on f, g, $\bar{\lambda}$ and $\bar{\mu}$. In particular, we will ultimately take sequence lengths k_r of the form $n^{\beta''}$ with β'' chosen much larger than the exponent in Υ_n. We call polynomials in n (such as Υ_n) which have an exponent not depending on β'', *small polynomials*. As a result, the following notation will be useful. For a function $W(k_r)$ of k_r, we use $\mathcal{S}_n(W(k_r))$ to denote a function smaller or equal to $W(k_r)$ *up to a small polynomial factor*. (The latter will be used similarly to the big-O notation.)

Recall the following standard concentration inequalities (see, e.g., [31]).

LEMMA 3.1 (Chernoff bounds). *Let $Z_1, \ldots, Z_m$ be independent $\{0,1\}$-random variables such that, for $1 \le i \le m$, $\mathbb{P}[Z_i = 1] = p_i$ where $0 < p_i < 1$. Then, for $Z = \sum_{i=1}^m Z_i$, $M = \mathbb{E}[Z] = \sum_{i=1}^m p_i$, $0 < \delta_- \le 1$ and $0 < \delta_+ \le U$,*

$$\mathbb{P}[Z < (1 - \delta_-)M] < e^{-M\delta_-^2/2}$$

and

$$\mathbb{P}[Z > (1 + \delta_+)M] < e^{-c(U)M\delta_-^2},$$

where $c(U) = [(1 + U)\ln(1 + U) - U]/U^2$.

3.1. *Concentration of the indel process.*

Sequence length. We first show that the sequence length is concentrated. Let T be single channel consisting of edge $e = (r, a)$. Let k_r be the length at r.

LEMMA 3.2 (Single channel: large deviations of sequence length). *For all $\gamma > 0$ and $\widehat{k}_r \geq k_r = n^{\beta'''}$ with $\beta''' > 0$ large enough, with probability at least $1 - \widehat{k}_r^{-\gamma}$,*

$$K_a = k_r M_t \pm \mathcal{S}_n\big(\sqrt{\widehat{k}_r \log \widehat{k}_r}\big),$$

where the small polynomial factor in $\mathcal{S}_n(\sqrt{\widehat{k}_r \log \widehat{k}_r})$ depends on γ as well.

REMARK 3.3. Although we stated Lemma 3.2 for the full sequence, it will also be needed for "half-sequences" and "blocks." In particular, we use the previous lemma to track the position of sites. In that context, one should think of k_r as the position of a site in r and K_a as its position in a. Then we can use $\widehat{k}_r$ for the full sequence length at r (see Section 3.2).

PROOF OF LEMMA 3.2. We think of K_a as

$$K_a = \sum_{i=1}^{k_r} K_{a,i},$$

where $K_{a,i}$ is the number of sites generated by a single site of the sequence at r. Intuitively, $K_{a,i}$ is the number of sites that were inserted between the sites i and $i+1$ of the sequence at r, plus the site at position i itself, if it survived. Clearly the variables $\{K_{a,i}\}_i$ are mutually independent.

Using (3) we obtain that

$$\mathbb{E}[K_a] = k_r M_t.$$

For $\varepsilon > 0$, by Markov's inequality, we have

$$\mathbb{P}[K_a \geq k_r M_t + k_r \varepsilon] \leq s^{-k_r(M_t+\varepsilon)} \mathbb{E}[s^{K_a}] = \big(s^{-(M_t+\varepsilon)} \mathbb{E}[s^{K_{a,1}}]\big)^{k_r}. \tag{9}$$

We take $s = 1 + C\varepsilon$ for $C > 0$ to be determined.

We have

$$\begin{aligned}
\mathbb{E}[s^{K_{a,1}}] &= \frac{\mu(s-1) - e^{(\mu-\lambda)t}(\lambda s - \mu)}{\lambda(s-1) - e^{(\mu-\lambda)t}(\lambda s - \mu)} = \frac{(\mu - \lambda M_t^{-1})C\varepsilon + \delta M_t^{-1}}{\lambda(1 - M_t^{-1})C\varepsilon + \delta M_t^{-1}} \\
&= \frac{\delta^{-1}(\mu M_t - \lambda)C\varepsilon + 1}{\delta^{-1}\lambda(M_t - 1)C\varepsilon + 1} = \frac{1 - (\lambda^{-1}\mu\Gamma_t - 1)C\varepsilon}{1 - \Gamma_t C\varepsilon} \\
&= [1 - (\lambda^{-1}\mu\Gamma_t - 1)C\varepsilon] \sum_{\iota=0}^{+\infty} [\Gamma_t C\varepsilon]^{\iota},
\end{aligned}$$

whenever $\Gamma_t C\varepsilon < 1$. Hence, if $\Upsilon_n C\varepsilon < 1$ is bounded away from 1 (independently of n), we have, using (8),

$$\begin{aligned}\mathbb{E}[s^{K_{a,1}}] &= [1-(\lambda^{-1}\mu\Gamma_t - 1)C\varepsilon][1+\Gamma_t C\varepsilon + (\Gamma_t C\varepsilon)^2 + O((\Upsilon_n C\varepsilon)^3)] \\ &= 1 + M_t(C\varepsilon) + M_t\Gamma_t(C\varepsilon)^2 + O((\Upsilon_n C\varepsilon)^3).\end{aligned}$$

Moreover, using the binomial series and (6), and assuming $C\varepsilon < 1$

$$\begin{aligned}s^{-(M_t+\varepsilon)} &= \sum_{\iota=0}^{+\infty} \frac{(-M_t-\varepsilon)(-M_t-\varepsilon-1)\cdots(-M_t-\varepsilon-\iota+1)}{\iota!}[C\varepsilon]^\iota \\ &\le 1 - (M_t+\varepsilon)(C\varepsilon) + \frac{(M_t+\varepsilon)(M_t+\varepsilon+1)}{2}(C\varepsilon)^2 \\ &\quad + \sum_{\iota=3}^{+\infty}(M_t+\varepsilon+1)^\iota[C\varepsilon]^\iota \\ &= 1 - (M_t+\varepsilon)(C\varepsilon) + \frac{(M_t+\varepsilon)(M_t+\varepsilon+1)}{2}(C\varepsilon)^2 \\ &\quad + O((\Upsilon_n C\varepsilon)^3),\end{aligned}$$

whenever ε is small and $\Upsilon_n C\varepsilon < 1$ is bounded away from 1 (independently from n). Therefore,

$$\begin{aligned}s^{-(M_t+\varepsilon)}\mathbb{E}[s^{K_{a,1}}] &= 1 - \varepsilon(C\varepsilon) + M_t\Gamma_t(C\varepsilon)^2 + \frac{(M_t+\varepsilon)(M_t+\varepsilon+1)}{2}(C\varepsilon)^2 \\ &\quad - (M_t+\varepsilon)M_t(C\varepsilon)^2 + O((\Upsilon_n C\varepsilon)^3).\end{aligned}$$

Note that the second term on the right-hand side depends on C whereas the remaining terms depend on C^2. Taking $C = \Upsilon_n^{-2}C_0(\gamma)$ with $C_0(\gamma) > 0$ small enough and $c = c_0(\gamma) > 0$ large enough, using (9) with the choice

$$\varepsilon = c\sqrt{\frac{\Upsilon_n^2 \log \widehat{k}_r}{k_r}},$$

we get that

$$\begin{aligned}\mathbb{P}\big[K_a \ge k_r M_t + c\sqrt{\Upsilon_n^2 \widehat{k}_r \log \widehat{k}_r}\big] &\le \mathbb{P}[K_a \ge k_r M_t + k_r\varepsilon] \\ &\le \left(1 - \frac{O(\log \widehat{k}_r)}{k_r}\right)^{k_r} \\ &\le \widehat{k}_r^{-\gamma}.\end{aligned}$$

Note that our choice of ε satisfies $\Upsilon_n C\varepsilon < 1$ for k_r a large enough polynomial of n (compared to the small polynomial Υ_n).

A similar inequality holds for the other direction. □

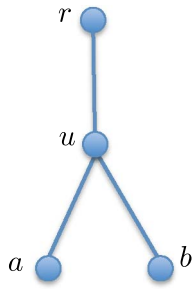

FIG. 1. *The fork channel.*

Correlated sites. Now let T be the fork channel consisting of nodes r, u, a and b as in Figure 1. Assume that a and b are contemporary, call t the time separating them from u and denote by S_{ab} the number of sites in a and b that are jointly surviving from u. These are the sites that produce correlation between the sequences at a and b. All other sites are essentially noise. We bound the large deviations of S_{ab}.

LEMMA 3.4 (Fork channel: large devations of jointly surviving sites). *Condition on the sequence length at u being k_u. Then, for all $\gamma > 0$ and all $\widehat{k}_u \geq k_u = n^{\beta'''}$ with $\beta''' > 0$ large enough, with conditional probability at least $1 - \widehat{k}_u^{-\gamma}$,*

$$S_{ab} = k_u D_t^2 \pm \mathcal{S}_n\big(\sqrt{\widehat{k}_u \log \widehat{k}_u}\big),$$

where the small polynomial factor in $\mathcal{S}_n(\sqrt{\widehat{k}_u \log \widehat{k}_u})$ depends on γ as well.

PROOF. Each site in u survives in a with probability D_t. The same holds for b independently.

The result then follows from Chernoff's bound. We have

$$\begin{aligned}\mathbb{P}\big[S_{ab} < k_u D_t^2 - c\sqrt{\Upsilon_n^2 \widehat{k}_u \log \widehat{k}_u}\big] &\leq \mathbb{P}\left[S_{ab} < k_u D_t^2 - c\sqrt{\frac{\Upsilon_n^2 \log \widehat{k}_u}{k_u}} \cdot k_u D_t^2\right] \\ &\leq \exp(-c^2 \Upsilon_n^2 D_t^2 \log \widehat{k}_u) \\ &\leq \widehat{k}_u^{-\gamma}\end{aligned}$$

for $c = c(\gamma) > 0$ large enough, where we used (7).

The other direction is similar. □

3.2. *Sequence partitioning.* From Lemma 3.2, it follows that the sites of the root sequence (or of an internal sequence) remain fairly close to their expected position at the leaves. We take advantage of this fact by dividing each sequence into blocks of size asymptotically larger than the typical displacement implied by

Lemma 3.2. As a result, matching blocks in different sequences share a significant fraction of sites. Moreover, distinct blocks are roughly independent. We estimate the evolutionary distance between two leaves by comparing the site frequencies in matching blocks. This requires some care as we show next.

Consider the fork channel. We seek to estimate the evolutionary distance $\widehat{\mathcal{D}}(a,b)$ between a and b (normalized by the sequence length at u).

Partitioning the leaf sequences. Let k_0 be some *deterministic* length (to be determined), and consider the first k_0 sites in the sequences σ_a and σ_b at the nodes a and b, respectively. If the sequence at a or b has length smaller than k_0, we declare that our distance estimate $\widetilde{\mathcal{D}}(a,b)$ (see below) is $+\infty$.

We divide the leaf sequences into L blocks of length ℓ where $\ell = \lceil k_0^{\zeta} \rceil$, for some $\frac{1}{2} < \zeta < 1$ to be determined later and $L = \lfloor k_0/\ell \rfloor$. We let $k_0' = \ell L$. For all $i = 1, \ldots, L$, we define the ith *block* $\sigma_{a,i}$ of a to be the subsequence of σ_a ranging from position $(i-1)\ell + 1$ to position $i\ell$. We let $Z_{a,i}$ be the number of zeros inside $\sigma_{a,i}$ and define the *block deviations*

$$\Delta_{a,i} = Z_{a,i} - \frac{\ell}{2}$$

for all $i = 1, \ldots, L$, and similarly for the sequence at b.

Using the above notation we define our distance estimator next. Assume that L is even. Otherwise, we can just drop the last block in the above partition. Our estimator is

$$\widetilde{\mathcal{D}}(a,b) = \frac{2}{L} \sum_{j=0}^{L/2-1} \Delta_{a,2j+1} \Delta_{b,2j+1}.$$

Notice that in our summation above we skipped every other block in our sequence partition to avoid overlapping sites and hence, decrease potential correlations between the terms in the estimator. In the rest of this section, we analyze the properties of $\widetilde{\mathcal{D}}(a,b)$. To do this it is helpful to consider the sequence at u and the events that happened in the channels defined by the edges (u,a) and (u,b).

Partitioning the ancestral sequence. Let us choose ℓ_u to be the largest integer satisfying

$$\ell_u M_t \leq \ell. \tag{10}$$

Suppose that the sequence σ_u at node u is not shorter than $k_u' = (L-1)\ell_u$, and define the ith *ancestral block* $\sigma_{u,i}$ of u to be the subsequence of σ_u ranging from position $(i-1)\ell_u + 1$ to position $i\ell_u$, for all $i \leq L-1$. Given Lemma 3.2, the choice of ℓ_u in (10) is such that the blocks of u and the corresponding blocks at a and b roughly align.

In order to use the expected evolutionary distance as computed in Lemma 2.4, we define an "interior" ancestral block which is guaranteed with high probability to remain entirely "inside" the corresponding leaf block. Let $\delta_u = \lceil L + \frac{1}{M_t}\mathcal{S}_n(\sqrt{k'_u \log k'_u})\rceil$, where the small polynomial factor is the maximum of those in the proofs of Lemma 3.2 and Lemma 3.4 for a given choice of γ. [The $L = o(\sqrt{k_0})$ in δ_u is needed only when (10) is a strict inequality. See the proof of Lemma 3.5 below.] We define the ith (*ancestral*) *interior block* $\sigma'_{u,i}$ of u to be the subsequence of $\sigma_{u,i}$ ranging from position $(i-1)\ell_u + \delta_u$ of σ_u to position $i\ell_u - \delta_u$. Notice that $\delta_u = \mathcal{S}_n(\sqrt{k_0 \log k_0})$, while $\ell_u = \mathcal{S}_n(k_0^\zeta)$. Therefore, for $k_0 > k_0^*$, where k_0^* is sufficiently large, $(i-1)\ell_u + \delta_u < i\ell_u - \delta_u$ so that the sequence $\sigma'_{u,i}$ is well defined.

Also, for all $i = 1, \ldots, L-1$, we define $x'_{a,i}, y'_{a,i}$ to be the position of the left-most (resp., right-most) site in the sequence σ_a descending from the site at position $(i-1)\ell_u + \delta_u$ (resp., $i\ell_u - \delta_u$ of σ_u). Similarly, we define $x'_{b,i}$ and $y'_{b,i}$. Given this notation, we define the following "good" event

$$\begin{aligned}\mathcal{E}'_1 = \{\forall i \le L-1 : (i-1)\ell < x'_{a,i}, x'_{b,i} < (i-1)\ell + 2M_t\delta_u,\\ i\ell - 2M_t\delta_u < y'_{a,i}, y'_{b,i} < i\ell\}.\end{aligned} \tag{11}$$

Intuitively, when the event $\mathcal{E}'_1$ holds, all surviving descendants of the interior block $\sigma'_{u,i}$ are located inside the blocks $\sigma_{a,i}$ and $\sigma_{b,i,}$, respectively (and the blocks remain large enough).

To argue about block independence, we also define the *exterior block* $\sigma''_{u,i}$ of u to be the subsequence of $\sigma_{u,i}$ ranging from position $(i-1)\ell_u - \delta_u$ of σ_u to position $i\ell_u + \delta_u$ with corresponding positions $x''_{a,i}$, $y''_{a,i}$, $x''_{b,i}$ and $y''_{b,i}$ and good event $\mathcal{E}''_1$ defined similarly as above, that is,

$$\begin{aligned}\mathcal{E}''_1 = \{\forall i \le L-1 : (i-1)\ell - 2M_t\delta_u < x''_{a,i}, x''_{b,i} < (i-1)\ell,\\ i\ell < y''_{a,i}, y''_{b,i} < i\ell + 2M_t\delta_u\}.\end{aligned}$$

We define

$$\mathcal{E}_1 = \mathcal{E}'_1 \cup \mathcal{E}''_1.$$

We show that this event holds with high probability, conditioned on the sequence length K_u at u being at least k'_u. Figure 2 shows the structure of the indel process in the case that the event $\mathcal{E}_1$ holds.

LEMMA 3.5 (Interior/exterior block is inside/outside leaf block). *Conditioned on the event* $\{K_u \ge k'_u\}$, *we have*

$$\mathbb{P}[\mathcal{E}_1] \ge 1 - 16L\left(\frac{1}{k'_u}\right)^\gamma.$$

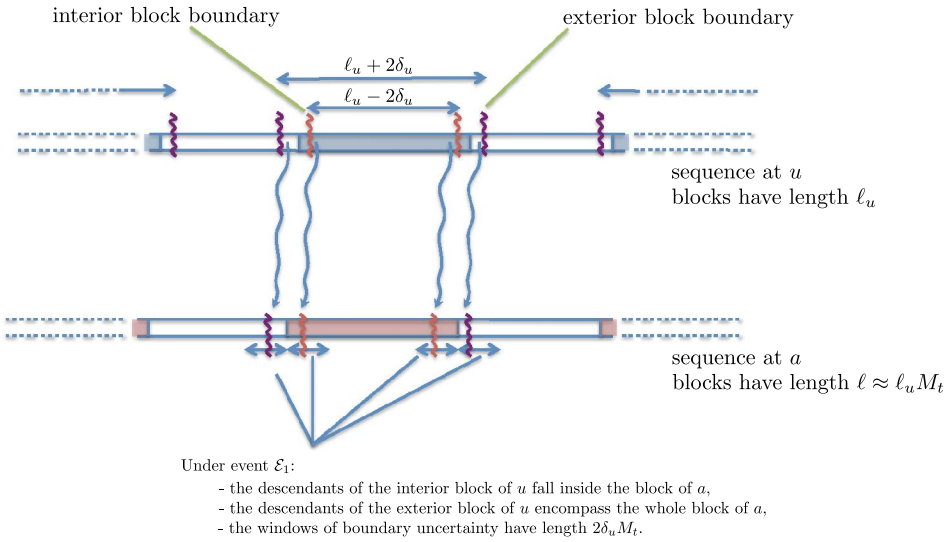

FIG. 2. *Under the event $\mathcal{E}_1$ the descendants of the interior blocks of σ_u fall inside the corresponding blocks of σ_a; the descendants of the exterior blocks of σ_u contain all surviving sites inside the corresponding blocks of σ_a; the windows of uncertainty have length $2M_t\delta_u$.*

PROOF. It follows from Lemma 3.2 that the left-most descendant of the site at position $(i-1)\ell_u + \delta_u$ of σ_u is located inside the sequence of node a at position at least

$$M_t\big((i-1)\ell_u + \delta_u\big) - \mathcal{S}_n\big(\sqrt{k'_u \log k'_u}\big) > M_t\big((i-1)\ell_u + L\big) > (i-1)\ell$$

with probability $\geq 1 - (\frac{1}{k'_u})^\gamma$. The other bounds follow similarly. Taking a union bound over all i's establishes the result. □

Block correlation. Let $S_{ab,i}$ be the number of common sites in the blocks $\sigma_{a,i}$ and $\sigma_{b,i}$ that are jointly surviving from u. Similarly, we define $S'_{ab,i}$ and $S''_{ab,i}$ where, for $\xi = a, b$, $\sigma'_{\xi,i}$ (resp., $\sigma''_{\xi,i}$) denotes the subsequence of σ_ξ ranging from position $x'_{\xi,i}$ (resp., $x''_{\xi,i}$) to position $y'_{\xi,i}$ (resp., $y''_{\xi,i}$). We define a good event for $S_{ab,i}$ as

$$\mathcal{E}_2 = \{\forall i \leq L-1 : \ell_u D_t^2 - 3M_t\delta_u \leq S_{ab,i} \leq \ell_u D_t^2 + 3M_t\delta_u\}.$$

LEMMA 3.6 (Jointly surviving sites in blocks). *Conditioned on the event $\{K_u \geq k'_u\}$, we have*

$$\mathbb{P}[\mathcal{E}_2] \geq 1 - 18L\left(\frac{1}{k'_u}\right)^\gamma.$$

PROOF. We bound

$$\begin{aligned}\mathbb{P}[\mathcal{E}_2^c] &= \mathbb{P}[\mathcal{E}_2^c \cap \mathcal{E}_1] + \mathbb{P}[\mathcal{E}_2^c \cap \mathcal{E}_1^c] \le \mathbb{P}[\mathcal{E}_2^c \cap \mathcal{E}_1] + \mathbb{P}[\mathcal{E}_1^c] \\ &\le \mathbb{P}[\mathcal{E}_2^c \cap \mathcal{E}_1] + 16L\left(\frac{1}{k'_u}\right)^{\gamma}.\end{aligned}$$

By construction, under $\mathcal{E}_1$ we have $S'_{ab,i} \le S_{ab,i} \le S''_{ab,i}$ so that

$$\begin{aligned}\mathbb{P}[\mathcal{E}_2^c \cap \mathcal{E}_1] &\le \mathbb{P}[\exists i,\, S'_{ab,i} \le \ell_u D_t^2 - 3M_t\delta_u] \\ &\quad + \mathbb{P}[\exists i,\, S''_{ab,i} \ge \ell_u D_t^2 + 3M_t\delta_u] \\ &\le \mathbb{P}\big[\exists i,\, S'_{ab,i} \le (\ell_u - 2\delta_u + 1)D_t^2 - \mathcal{S}_n\big(\sqrt{k'_u \log k'_u}\big)\big] \\ &\quad + \mathbb{P}\big[\exists i,\, S''_{ab,i} \ge (\ell_u + 2\delta_u + 1)D_t^2 + \mathcal{S}_n\big(\sqrt{k'_u \log k'_u}\big)\big] \\ &\le 2L\left(\frac{1}{k'_u}\right)^{\gamma}\end{aligned}$$

by Lemma 3.4, where we also used the fact that $D_t^2 \le M_t$. □

3.3. *Estimation guarantees*. We are now ready to analyze the behavior of our estimate $\widetilde{\mathcal{D}}(a,b)$. In this subsection we compute the expectation and variance of $\widetilde{\mathcal{D}}(a,b)$. We denote by $\mathcal{I}$ a realization of the indel process (but not of the substitution process) on the paths between u and a,b. We denote by $\mathcal{E}$ the event that $\{K_u \ge k'_u\}$, $\mathcal{E}_1$ and $\mathcal{E}_2$ are satisfied. Suppose that $k_0 > k_0^*$ (defined in Section 3.2).

LEMMA 3.7 (Block independence). *Conditioning on $\mathcal{I}$ and $\mathcal{E}$, the variables*

$$\{\Delta_{a,2j+1}\Delta_{b,2j+1}\}_{j=1}^{L/2-1}$$

are mutually independent.

PROOF. Observe that when $K_u \ge k'_u$ the ancestral blocks $\sigma_{u,i}$ are well defined. Assuming that $k_0 > k_0^*$, the interior blocks $\sigma'_{u,i}$ are also well defined and disjoint. Hence, for a fixed $\mathcal{I}$ under $\mathcal{E}$, for all $i \le L-1$, both $\Delta_{a,i}$ and $\Delta_{b,i}$ depend on the subsequence of σ_u ranging from position $(i-1)\ell_u - \delta_u + 1$ to position $i\ell_u + \delta_u - 1$. In this case, for $j \in \{1,\dots,L/2-1\}$, different $\Delta_{a,2j+1}\Delta_{b,2j+1}$'s are functions of different subsequences of σ_u. Observe that, since the root sequence is i.i.d. uniform and the insertions above u are also i.i.d. uniform, the state of every site in σ_u is uniform and independent from the other sites. It follows from the above observations that $\{\Delta_{a,2j+1}\Delta_{b,2j+1}\}_{j=1}^{L/2-1}$ are mutually independent. □

LEMMA 3.8 (Expected correlation under good event). *We have*

$$\mathbb{E}[\Delta_{a,i}\Delta_{b,i}|\mathcal{I},\mathcal{E}] = \tfrac{1}{4}e^{-4\eta t}e^{-2\mu t}\ell_u \pm \mathcal{S}_n\big(\sqrt{k_0 \log k_0}\big).$$

PROOF. Let $\Delta^{\mathrm{S}}_{a,i}$ be the contribution to $\Delta_{a,i}$ from those common sites between a and b that are jointly surviving from u. Let $\Delta^{\mathrm{NS}}_{a,i} = \Delta_{a,i} - \Delta^{\mathrm{S}}_{a,i}$, and similarly for b. Then

$$\begin{aligned}\mathbb{E}[\Delta_{a,i}\Delta_{b,i}|\mathcal{I},\mathcal{E}] &= \mathbb{E}[(\Delta^{\mathrm{S}}_{a,i} + \Delta^{\mathrm{NS}}_{a,i})(\Delta^{\mathrm{S}}_{b,i} + \Delta^{\mathrm{NS}}_{b,i})|\mathcal{I},\mathcal{E}] \\ &= \mathbb{E}[\Delta^{\mathrm{S}}_{a,i}\Delta^{\mathrm{S}}_{b,i}|\mathcal{I},\mathcal{E}],\end{aligned}$$

since the contribution from $\Delta^{\mathrm{NS}}_{a,i}$ and $\Delta^{\mathrm{NS}}_{b,i}$ is independent and averages to 0. Write $\Delta^{\mathrm{S}}_{a,i}$ as a sum over the jointly surviving sites, that is,

$$\Delta^{\mathrm{S}}_{a,i} = \sum_{j=1}^{S_{ab,i}} \Big(z^{(j)}_{a,i} - \frac{1}{2}\Big),$$

where $z^{(j)}_{a,i}$ is 1 if the corresponding site of a is 0. Note that the terms in parentheses have zero expectation given $\mathcal{I}$ and $\mathcal{E}$. Then,

$$\mathbb{E}[\Delta^{\mathrm{S}}_{a,i}\Delta^{\mathrm{S}}_{b,i}|\mathcal{I},\mathcal{E}] = \sum_{j=1}^{S_{ab,i}} \mathbb{E}\Big[\Big(z^{(j)}_{a,i} - \frac{1}{2}\Big)\Big(z^{(j)}_{b,i} - \frac{1}{2}\Big)\Big|\mathcal{I},\mathcal{E}\Big]$$

by independence of the sites. We compute the expectation above. We have

$$\begin{aligned}&\mathbb{E}\Big[\Big(z^{(j)}_{a,i} - \frac{1}{2}\Big)\Big(z^{(j)}_{b,i} - \frac{1}{2}\Big)\Big|\mathcal{I},\mathcal{E}\Big] \\ &\quad= \mathbb{E}\Big[\Big(z^{(j)}_{a,i}z^{(j)}_{b,i} - \frac{1}{2}z^{(j)}_{a,i} - \frac{1}{2}z^{(j)}_{b,i} + \frac{1}{4}\Big)\Big|\mathcal{I},\mathcal{E}\Big] \\ &\quad= \mathbb{E}[z^{(j)}_{a,i}z^{(j)}_{b,i}|\mathcal{I},\mathcal{E}] - \frac{1}{4} \\ &\quad= \frac{1}{2}\frac{1+e^{-4\eta t}}{2} - \frac{1}{4} \\ &\quad= \frac{1}{4}e^{-4\eta t}.\end{aligned}$$

Therefore,

$$\mathbb{E}[\Delta^{\mathrm{S}}_{a,i}\Delta^{\mathrm{S}}_{b,i}|\mathcal{I},\mathcal{E}] = \tfrac{1}{4}e^{-4\eta t}S_{ab,i}.$$

The result then follows from the definition of $\mathcal{E}_2$. □

LEMMA 3.9 (Variance under good event). *We have*

$$\mathrm{Var}[\Delta_{a,i}\Delta_{b,i}|\mathcal{I},\mathcal{E}] \le \tfrac{3}{16}\ell^2.$$

PROOF. By Cauchy–Schwarz we have

$$\begin{aligned}\mathbb{E}[\Delta_{a,i}^2\Delta_{b,i}^2|\mathcal{I},\mathcal{E}] &\le (\mathbb{E}[\Delta_{a,i}^4|\mathcal{I},\mathcal{E}]\mathbb{E}[\Delta_{b,i}^4|\mathcal{I},\mathcal{E}])^{1/2}\\ &= \big(\tfrac{1}{16}(3\ell^2-2\ell)\cdot\tfrac{1}{16}(3\ell^2-2\ell)\big)^{1/2}\\ &\le \tfrac{3}{16}\ell^2,\end{aligned}$$

where we used the fact that the length of the sequences $\sigma_{a,i}$ and $\sigma_{b,i}$ is deterministically ℓ, and the number of zeros in $\sigma_{a,i}$ and $\sigma_{b,i}$ follows a binomial distribution with ℓ trials and probability $1/2$. □

LEMMA 3.10 (Distance estimate). *We have*

$$\mathbb{E}[\widetilde{\mathcal{D}}(a,b)|\mathcal{I},\mathcal{E}] = \tfrac{1}{4}e^{-(4\eta+\mu+\lambda)t}\ell \pm \mathcal{S}_n\big(\sqrt{k_0\log k_0}\big)$$

and

$$\mathrm{Var}[\widetilde{\mathcal{D}}(a,b)|\mathcal{I},\mathcal{E}] \le \frac{3}{8}\frac{1}{\lfloor k_0^{1-\zeta}\rfloor}\ell^2.$$

In particular, the standard deviation

$$\mathrm{STD}[\widetilde{\mathcal{D}}(a,b)|\mathcal{I},\mathcal{E}] = O\big(k_0^{(3\zeta-1)/2}\big) = o\big(\sqrt{k_0}\big)$$

for $\zeta > 1/2$ small enough.

PROOF. From Lemma 3.7, the $L/2 = \lfloor k_0/\ell\rfloor/2$ terms in $\widetilde{\mathcal{D}}(a,b)$ are mutually independent. The proof then follows from Lemmas 3.8 and 3.9 and the definition of ℓ_u. □

3.4. *Concentration*. We now show that our distance estimate is concentrated. For notational convenience, we denote by $\mathbb{P}_u^*$ the probability measure induced by conditioning on the event $\{K_u \ge k_u'\}$. Recall that the event $\mathcal{E}$ is contained in $\{K_u \ge k_u'\}$.

LEMMA 3.11 (Concentration of distance estimate). *Let $\alpha > 0$ be such that $\zeta - \alpha > 1/2$, and $\beta = 1-\zeta-2\alpha > 0$ for $\zeta > 1/2$ small enough. Then for k_0 large enough*

$$\mathbb{P}_u^*\left[\left|\frac{4}{\ell}\widetilde{\mathcal{D}}(a,b) - e^{-(4\eta+\mu+\lambda)t}\right| > \frac{1}{k_0^\alpha}\right] \le O\left(\frac{1}{k_0^\beta}\right).$$

PROOF. We use Chebyshev's inequality. We first condition on $\mathcal{I}, \mathcal{E}$. Recalling that $\ell = \lceil k_0^{\zeta} \rceil$, note that

$$\begin{aligned}
&\mathbb{P}_u^*\left[\frac{4}{\ell}\widetilde{\mathcal{D}}(a,b) > e^{-(4\eta+\mu+\lambda)t} + \frac{1}{k_0^{\alpha}}\Big|\mathcal{I},\mathcal{E}\right] \\
&\quad\le \mathbb{P}_u^*\left[\widetilde{\mathcal{D}}(a,b) > \frac{\ell}{4}e^{-(4\eta+\mu+\lambda)t} + \frac{\ell}{4}\frac{1}{k_0^{\alpha}}\Big|\mathcal{I},\mathcal{E}\right] \\
&\quad\le \mathbb{P}_u^*\left[\widetilde{\mathcal{D}}(a,b) > \mathbb{E}[\widetilde{\mathcal{D}}(a,b)|\mathcal{I},\mathcal{E}] - \mathcal{S}_n\big(\sqrt{k_0 \log k_0}\big) + \frac{\ell}{4}\frac{1}{k_0^{\alpha}}\Big|\mathcal{I},\mathcal{E}\right] \\
&\quad\le \frac{3\ell^2/(8\lfloor k_0^{1-\zeta}\rfloor)}{(\ell/(4k_0^{\alpha}) - \mathcal{S}_n(\sqrt{k_0\log k_0}))^2} \\
&\quad= O\left(\frac{1}{k_0^{1-\zeta-2\alpha}}\right).
\end{aligned}$$

The other direction is similar. Taking expectation over $\mathcal{I}$, we have

$$\mathbb{P}_u^*\left[\left|\frac{4}{\ell}\widetilde{\mathcal{D}}(a,b) - e^{-(4\eta+\mu+\lambda)t}\right| > \frac{1}{k_0^{\alpha}}\Big|\mathcal{E}\right] \le O\left(\frac{1}{k_0^{\beta}}\right).$$

Choose $\gamma > 0$ in Lemmas 3.2 and 3.4 large enough so that

$$\gamma - (1-\zeta) > \beta.$$

Then, from Lemmas 3.5 and 3.6, we have

$$\begin{aligned}
&\mathbb{P}_u^*\left[\left|\frac{4}{\ell}\widetilde{\mathcal{D}}(a,b) - e^{-(4\eta+\mu+\lambda)t}\right| > \frac{1}{k_0^{\alpha}}\right] \\
&\quad\le \mathbb{P}_u^*\left[\left|\frac{4}{\ell}\widetilde{\mathcal{D}}(a,b) - e^{-(4\eta+\mu+\lambda)t}\right| > \frac{1}{k_0^{\alpha}}\Big|\mathcal{E}\right]\mathbb{P}_u^*[\mathcal{E}] + \mathbb{P}_u^*[\mathcal{E}^c] \\
&\quad\le O\left(\frac{1}{k_0^{\beta}}\right).
\end{aligned}$$

□

The proofs of Theorems 1 and 2 are given in the next section.

4. Putting it all together.

Large-scale asymptotics. We are ready to prove our main result in the molecular clock case. We postpone the more general case to the next section. A last bit of notation: for a pair of leaves $a, b \in [n]$, we denote by t_{ab} the time between a, b and their most recent common ancestor.

PROOF OF THEOREM 2. We first give a bound on the diameter of the tree. Let h (resp., H) be the length of the shortest (resp., longest) path between the root and a leaf in graph distance. Because the number of leaves is n we must have

$2^h \leq n$ and $2^H \geq n$. Since all leaves are contemporaneous it must be that $Hf \leq hg$. Combining these constraints gives that the diameter Diam satisfies

$$2\frac{f}{g}\log_2 n \leq 2h \leq \text{Diam} \leq 2H \leq 2\frac{g}{f}\log_2 n.$$

Given our bound on the diameter of the tree, it follows that the time from the root r of the tree to any leaf is at most $\frac{g^2}{f}\log_2 n$. Suppose that the length k_r at the root of the tree satisfies $k_r > k_r^* = k_r^*(k_0)$, where k_r^* is the minimum integer satisfying

$$k_r^* \geq e^{(g^2/f)\cdot\mu\log_2 n}\big(k_0 + \mathcal{S}_n\big(\sqrt{k_r^*\log k_r^*}\big)\big),$$

where the small polynomial factor is taken to be the one used in Lemma 3.2.

Lemma 3.2 and a union bound then imply that with probability at least

$$1 - O(n)\cdot(k_r^*)^{-\gamma}$$

for all nodes u

$$K_u \geq k'_u.$$

LEMMA 4.1 (Concentration of distance estimate). *For all $\alpha' > 0$, $\beta' > 0$, there exists $k_0 = n^{\beta'''}$ with $\beta''' > 0$ large enough so that if the sequence length at the root is $k_r > k_r^*(k_0)$, then*

$$\mathbb{P}\left[\forall a, b \in [n], \left|\frac{4}{\ell}\widetilde{\mathcal{D}}(a,b) - e^{-(4\eta+\mu+\lambda)t_{ab}}\right| \leq \frac{1}{n^{\alpha'}}\right] = 1 - O\left(\frac{1}{n^{\beta'}}\right).$$

PROOF. This follows from Lemma 3.11 and our observation above that, if $k_r > k_r^*(k_0)$, with probability at least $1 - O(n)\cdot(k_r^*)^{-\gamma}$, then $K_u \geq k'_u$ for all nodes u. □

Given our bound on the diameter of the tree, it follows that for all pairs of leaves a, b and small $\varepsilon > 0$

$$e^{-(4\eta+\mu+\lambda)t_{ab}\pm\varepsilon} = e^{-(4\eta+\mu+\lambda)t_{ab}}\big(1 \pm O(\varepsilon)\big) \geq \frac{1}{n^{\alpha''}}\big(1 \pm O(\varepsilon)\big).$$

Therefore, choosing α' large enough in Lemma 4.1, we get that all distances can be estimated within a small ε simultaneously with probability going to 1.

Using the standard Buneman algorithm, we can recover the tree efficiently (see, e.g., [34]). □

Constant-size case. The proof of Theorem 1 for the molecular clock case builds on the proof of Theorem 2 by treating n as a constant and letting the sequence length at the root of the tree go to infinity.

PROOF OF THEOREM 1 (Molecular clock case). We can restate Lemma 4.1 in the following form, where the failure probability is expressed more cleanly in terms of the sequence length at the root of the tree. The proof of the lemma is essentially the same.

LEMMA 4.2 (Concentration of distance estimate). *For all $\alpha' > 0$, there exists $k_0^* = n^{\beta'''}$ for $\beta''' > 0$ large enough such that if the sequence length at the root is $k_r > k_r^*(k_0^*)$, then*

$$\mathbb{P}\left[\forall a, b \in [n], \left|\frac{4}{\ell}\widetilde{\mathcal{D}}(a,b) - e^{-(4\eta+\mu+\lambda)t_{ab}}\right| \le \frac{1}{n^{\alpha'}}\right] = 1 - O(n \cdot k_r^{-\gamma}) - O(n^2 \cdot k_r^{-\beta}).$$

Repeating the proof of Theorem 2 above, it follows that the algorithm fails to reconstruct the phylogeny with probability $O(n \cdot k_r^{-\gamma}) + O(n^2 \cdot k_r^{-\beta})$. Letting $k_r \to +\infty$ concludes the proof of Theorem 1. □

5. Extensions.

GTR model. We briefly discuss how our results can be extended to GTR models. For background on GTR models, see, for example, [16]. Let Q be a reversible 4×4 rate matrix with stationary distribution π. Our new sequence evolution process is identical to the one described in Definition 1.1 except that the substitution process is a continuous-time Markov process with rate matrix $\eta_e Q$. The rate matrix Q has 4 nonnegative eigenvalues. For convenience, we assume that the largest negative eigenvalue is -1. We denote by w the corresponding eigenvector which we assume is normalized as

$$\sum_{s \in \{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}} \pi_s w_s^2 = 1.$$

We now perform the following transformation of the state space. For a node u, let $\sigma_u = (\sigma_u^1, \ldots, \sigma_u^{K_u})$ be the transformed sequence at u where $\sigma_u^i = w_{\mathtt{A}}$ (resp., $w_{\mathtt{G}}, w_{\mathtt{C}}, w_{\mathtt{T}}$) if the state at site i is $\mathtt{A}$ (resp., $\mathtt{G}, \mathtt{C}, \mathtt{T}$). Note that, under stationarity, the expectation of the state at site i is 0 by orthogonality of π and w. Then our distance estimator is

$$\widehat{\mathcal{D}}(a,b) = \left(\sum_{i=1}^{K_a} \sigma_a^i\right)\left(\sum_{j=1}^{K_b} \sigma_b^j\right).$$

In particular, in the two-state CFN case, we have $w = (+1, -1)$ and we obtain the same estimate as before, up to a constant. We now analyze the expectation of this quantity. For $u \in V$, we let

$$\Delta_u = \sum_{i=1}^{K_u} \sigma_u^i.$$

LEMMA 5.1. *The following holds*:

$$\mathbb{E}[\Delta_a|\sigma_r] = e^{-(\eta+\mu)t}\Delta_r. \tag{12}$$

REMARK 5.2. Note that this formula is slightly different than that in Lemma 2.2 because of the normalization implied by requiring Q to have second eigenvalue -1.

PROOF OF LEMMA 5.1. The sites created after r contribute 0 in expectation. Of course, so do the deleted sites. The fraction of sites that survive is $e^{-\mu t}$. Suppose site i survives, then note that

$$\mathbb{E}[\sigma_a^i|\sigma_r^i = w_s, i \text{ survives}] = \sum_{s'}(e^{\eta t Q})_{ss'}w_{s'} = e^{-\eta t}w_s.$$

Summing over all sites of r we get

$$\mathbb{E}[\Delta_a|\sigma_r] = e^{-(\eta+\mu)t}\Delta_r$$

as claimed. □

Consider now a "fork" tree, that is, a root r from which emanates a single edge $e_u = (r, u)$ which in turn branches into two edges $e_a = (u, a)$ and $e_b = (u, b)$. For $x = a, b, u$, we denote the parameters of edge e_x by $t_x, \lambda_x, \mu_x, \eta_x$. Our goal is to compute $\mathbb{E}[\widehat{\mathcal{D}}(a, b)]$ assuming that the sequence length at the root is k. The proof is similar to Lemma 2.3.

LEMMA 5.3. *The following holds*:

$$\mathbb{E}[\widehat{\mathcal{D}}(a, b)] = e^{-(\eta_a+\mu_a)t_a}e^{-(\eta_b+\mu_b)t_b}e^{-(\mu_u-\lambda_u)t_u}k.$$

Note that Remark 5.2 also applies here.

PROOF. We have

$$\begin{aligned}
\mathbb{E}[\widehat{\mathcal{D}}(a, b)] &= \mathbb{E}[\Delta_a\Delta_b] \\
&= \mathbb{E}[\mathbb{E}[\Delta_a\Delta_b|\sigma_u]] \\
&= \mathbb{E}[\mathbb{E}[\Delta_a|\sigma_u]\mathbb{E}[\Delta_b|\sigma_u]] \\
&= e^{-\eta_a t_a}e^{-\mu_a t_a}e^{-\eta_b t_b}e^{-\mu_b t_b}\mathbb{E}[\Delta_u^2] \\
&= e^{-\eta_a t_a}e^{-\mu_a t_a}e^{-\eta_b t_b}e^{-\mu_b t_b}\mathbb{E}[\mathbb{E}[\Delta_u^2|K_u]] \\
&= e^{-\eta_a t_a}e^{-\mu_a t_a}e^{-\eta_b t_b}e^{-\mu_b t_b}\mathbb{E}[\mathrm{Var}[\Delta_u|K_u]] \\
&= e^{-\eta_a t_a}e^{-\mu_a t_a}e^{-\eta_b t_b}e^{-\mu_b t_b}\mathbb{E}[K_u\,\mathrm{Var}[\sigma_u^1]] \\
&= e^{-\eta_a t_a}e^{-\mu_a t_a}e^{-\eta_b t_b}e^{-\mu_b t_b}\mathbb{E}[K_u\mathbb{E}[(\sigma_u^1)^2]] \\
&= e^{-\eta_a t_a}e^{-\mu_a t_a}e^{-\eta_b t_b}e^{-\mu_b t_b}e^{-(\mu_u-\lambda_u)t_u}k
\end{aligned}$$

by Lemma 5.1. □

From the previous lemmas, one can adapt the proofs above to the GTR case.

Nonclock case. Using Lemma 5.3, we can get rid of the molecular clock assumption. Consider again the fork tree, but assume that each edge is in fact a path. An adaptation of Lemma 5.3 gives the following lemma.

LEMMA 5.4. *The following holds*:

$$-\ln\left(\frac{\mathbb{E}[\widehat{\mathcal{D}}(a,b)]}{\sqrt{\mathbb{E}[K_a]\mathbb{E}[K_b]}}\right) = \sum_{e\in \mathrm{P}(a,b)} (\eta_e + \mu_e/2 + \lambda_e/2)t_e.$$

Note that Remark 5.2 also applies here.

PROOF. Note that

$$-\ln(k^{-1}\mathbb{E}[K_a]) = \sum_{e\in \mathrm{P}(r,a)} (\mu_e - \lambda_e)t_e$$

and similarly for b. A variant of Lemma 5.3 gives

$$-\ln(k^{-1}\mathbb{E}[\widehat{\mathcal{D}}(a,b)]) = \sum_{e\in \mathrm{P}(a,b)} (\eta_e + \mu_e)t_e + \sum_{e\in \mathrm{P}(r,u)} (\mu_e - \lambda_e)t_e.$$

The result follows by subtracting the previous expressions. □

The expression in Lemma 5.4 provides the additive metric needed to extend our results to nonclock bounded-rates case.

6. Concluding remarks. We have shown how to reconstruct phylogenies under the bounded-rates, GTR model with indels. Our efficient algorithm requires polynomial-length sequences at the root. A natural open problem arises from this work: Can our results be extended to general trees with bounded branch lengths, as opposed to the bounded-rates model? The key difference between the two models is that the former may have a linear diameter whereas the latter has logarithmic diameter. To extend our results, one would need to deal with far away leaves that are almost uncorrelated but for which our block structure does not apply.

REFERENCES

[1] ANDONI, A., BRAVERMAN, M. and HASSIDIM, A. (2011). Phylogenetic reconstruction with insertions and deletions. Preprint.

[2] ANDONI, A., DASKALAKIS, C., HASSIDIM, A. and ROCH, S. (2010). Global alignment of molecular sequences via ancestral state reconstruction. In *ICS* 2010 358–369. Tsinghua University Press, Beijing, China.

[3] ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes*. *Die Grundlehren der Mathematischen Wissenschaften* **196**. Springer, New York. MR0373040

[4] ATTESON, K. (1999). The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica* **25** 251–278. MR1703580

[5] CSURÖS, M. (2002). Fast recovery of evolutionary trees with thousands of nodes. *J. Comput. Biol.* **9** 277–297.

[6] CSURÖS, M. and KAO, M.-Y. (2001). Provably fast and accurate recovery of evolutionary trees through harmonic greedy triplets. *SIAM J. Comput.* **31** 306–322 (electronic). MR1857402

[7] DASKALAKIS, C., HILL, C., JAFFE, A., MIHAESCU, R., MOSSEL, E. and RAO, S. (2006). Maximal accurate forests from distance matrices. In *RECOMB* 2006 281–295. Springer, Berlin.

[8] DASKALAKIS, C., MOSSEL, E. and ROCH, S. (2006). Optimal phylogenetic reconstruction. In *STOC'*06: *Proceedings of the* 38*th Annual ACM Symposium on Theory of Computing* 159–168. ACM, New York. MR2277141

[9] DASKALAKIS, C., MOSSEL, E. and ROCH, S. (2009). Phylogenies without branch bounds: Contracting the short, pruning the deep. In *RECOMB* 2009 451–465. Springer, Berlin.

[10] DASKALAKIS, C. and ROCH, S. (2010). Alignment-free phylogenetic reconstruction. In *RECOMB* 2010 123–137. Springer, Berlin.

[11] EDGAR, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32** 1792–1797.

[12] ELIAS, I. (2006). Settling the intractability of multiple alignment. *J. Comput. Biol.* **13** 1323–1339 (electronic). MR2264995

[13] ERDŐS, P. L., STEEL, M. A., SZÉKELY, L. A. and WARNOW, T. J. (1999). A few logs suffice to build (almost) all trees. I. *Random Structures Algorithms* **14** 153–184. MR1667319

[14] ERDŐS, P. L., STEEL, M. A., SZÉKELY, L. A. and WARNOW, T. J. (1999). A few logs suffice to build (almost) all trees. II. *Theoret. Comput. Sci.* **221** 77–118. MR1700821

[15] FELSENSTEIN, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.* **27** 401–410.

[16] FELSENSTEIN, J. (2004). *Inferring Phylogenies*. Sinauer, New York.

[17] GRAUR, D. and LI, W. H. (1999). *Fundamentals of Molecular Evolution*, 2nd ed. Sinauer, Sunderland, MA.

[18] GRONAU, I., MORAN, S. and SNIR, S. (2008). Fast and reliable reconstruction of phylogenetic trees with very short edges (extended abstract). In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 379–388. ACM, New York. MR2487605

[19] HIGGINS, D. G. and SHARP, P. M. (1988). Clustal: A package for performing multiple sequence alignment on a microcomputer. *Gene* **73** 237–244.

[20] HUSON, D. H., NETTLES, S. H. and WARNOW, T. J. (1999). Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *J. Comput. Biol.* **6** 3–4.

[21] KARLIN, S. and TAYLOR, H. M. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York. MR0611513

[22] KATOH, K., MISAWA, K., KUMA, K.-I. and MIYATA, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30** 3059–3066.

[23] KING, V., ZHANG, L. and ZHOU, Y. (2003). On the complexity of distance-based evolutionary tree reconstruction. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (*Baltimore*, *MD*, 2003) 444–453. ACM, New York. MR1974948

[24] LACEY, M. R. and CHANG, J. T. (2006). A signal-to-noise analysis of phylogeny estimation by neighbor-joining: Insufficiency of polynomial length sequences. *Math. Biosci.* **199** 188–215. MR2211625

[25] LIU, K., RAGHAVAN, S., NELESEN, S., LINDER, C. R. and WARNOW, T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324** 1561–1564.

[26] LÖYTYNOJA, A. and GOLDMAN, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320** 1632–1635.

[27] METZLER, D. (2003). Statistical alignment based on fragment insertion and deletion models. *Bioinformatics* **19** 490–499.

[28] MIKLOS, I., LUNTER, G. A. and HOLMES, I. (2004). A "Long Indel" model for evolutionary sequence alignment. *Mol. Biol. Evol.* **21** 529–540.

[29] MOSSEL, E. (2007). Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Trans. Comput. Bio. Bioinform.* **4** 108–116.

[30] MOSSEL, E. and ROCH, S. (2006). Learning nonsingular phylogenies and hidden Markov models. *Ann. Appl. Probab.* **16** 583–614. MR2244426

[31] MOTWANI, R. and RAGHAVAN, P. (1995). *Randomized Algorithms*. Cambridge Univ. Press, Cambridge. MR1344451

[32] RIVAS, E. and EDDY, S. R. (2008). Probabilistic phylogenetic inference with insertions and deletions. *PLoS Comput. Biol.* **4** e1000172, 20. MR2448496

[33] ROCH, S. (2008). Sequence-length requirement for distance-based phylogeny reconstruction: Breaking the polynomial barrier. In *FOCS* 2008 729–738. IEEE Comput. Soc., Los Alamitos, CA.

[34] SEMPLE, C. and STEEL, M. (2003). *Phylogenetics*. *Oxford Lecture Series in Mathematics and Its Applications* **24**. Oxford Univ. Press, Oxford. MR2060009

[35] STEEL, M. A. and SZÉKELY, L. A. (1999). Inverting random functions. *Ann. Comb.* **3** 103–113. MR1769697

[36] STEEL, M. A. and SZÉKELY, L. A. (2002). Inverting random functions. II. Explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM J. Discrete Math.* **15** 562–575 (electronic). MR1935839

[37] SUCHARD, M. A. and REDELINGS, B. D. (2006). BAli-Phy: Simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* **22** 2047–2048.

[38] THATTE, B. D. (2006). Invertibility of the TKF model of sequence evolution. *Math. Biosci.* **200** 58–75. MR2211928

[39] THORNE, J. L., KISHINO, H. and FELSENSTEIN, J. (1991). An evolutionary model for maximum likelihood alignment of dna sequences. *Journal of Molecular Evolution* **33** 114–124.

[40] THORNE, J. L., KISHINO, H. and FELSENSTEIN, J. (1992). Inching toward reality: An improved likelihood model of sequence evolution. *Journal of Molecular Evolution* **34** 3–16.

[41] WANG, L. and JIANG, T. (1994). On the complexity of multiple sequence alignment. *J. Comput. Biol.* **1** 337–348.

[42] WONG, K. M., SUCHARD, M. A. and HUELSENBECK, J. P. (2008). Alignment uncertainty and genomic analysis. *Science* **319** 473–476. MR2381044

DEPARTMENT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
32 VASSAR STREET
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: costis@csail.mit.edu

DEPARTMENT OF MATHEMATICS
AND BIOINFORMATICS PROGRAM
UNIVERSITY OF CALIFORNIA, LOS ANGELES
520 PORTOLA PLAZA
LOS ANGELES, CALIFORNIA 90095-1555
USA
E-MAIL: roch@math.ucla.edu