Global Alignment of Molecular Sequences via Ancestral State Reconstruction

Alexandr Andoni[a], Constantinos Daskalakis[b,1], Avinatan Hassidim[c], Sebastien Roch[d,2,*]

[a] *Microsoft Research Silicon Valley*
[b] *EECS, MIT*
[c] *Bar Ilan University and Google Israel*
[d] *Department of Mathematics, UW–Madison*

Abstract

We consider the trace reconstruction problem on a tree (TRPT): a binary sequence is broadcast through a tree channel where we allow substitutions, deletions, and insertions; we seek to reconstruct the original sequence from the sequences received at the leaves. The TRPT is motivated by the multiple sequence alignment problem in computational biology. We give a simple recursive procedure giving strong reconstruction guarantees at low mutation rates. To our knowledge, this is the first rigorous trace reconstruction result on a tree in the presence of indels.

Keywords: Markov models on trees, branching processes, phylogenetic inference

1. Introduction

Trace reconstruction on a star. In the "trace reconstruction problem" (TRP) [1, 2, 3, 4, 5, 6], a random binary string $\mathbf{X}$ of length k generates an i.i.d. collection of traces $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ that are identical to $\mathbf{X}$ except for random *mutations* which consist in *indels*, that is, the deletion of an old site or the insertion of a new site between existing sites, and *substitutions*, that is, the flipping of the state at an existing site. We refer to the positions of a string as *sites.*

[*] Correspondence: roch@math.wisc.edu

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

[2] Supported by NSF grant DMS-1007144.

The goal is to reconstruct efficiently the original string with high probability from as few random traces as possible.

An important motivation for this problem is the reconstruction of ancestral DNA sequences in computation biology [3, 4]. One can think of $\mathbf{X}$ as a gene in an (extinct) ancestor species 0. Through speciation, the ancestor 0 gives rise to a large number of descendants $1, \ldots, n$ and gene $\mathbf{X}$ evolves independently through mutations to sequences $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ respectively. Inferring the sequence $\mathbf{X}$ of an ancient gene from extant descendant copies $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is a standard problem in evolutionary biology [7]. The inference of $\mathbf{X}$ typically requires the solution of an auxiliary problem, the *multiple sequence alignment problem* which is an important problem in its own right in computational biology: site t_i of sequence $\mathbf{Y}_i$ and site t_j of sequence $\mathbf{Y}_j$ are said to be *homologous* (in this simplified TRP setting) if they descend from a common site t of $\mathbf{X}$ *only through substitutions*; in the multiple sequence alignment problem, we seek roughly to uncover the homology relation between $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Once homologous sites have been identified, it is straightforward to estimate the original sequence $\mathbf{X}$ (minus the sites that were deleted in all descendant sequences), for instance, by performing a majority vote.

However, the TRP as defined above is an *idealized* version of the ancestral sequence reconstruction problem in one important aspect. It ignores the actual phylogenetic relationship between species $1, \ldots, n$. A *phylogeny* is a (typically, binary) tree relating a group of species. The leaves of the tree correspond to extant species. Internal nodes can be thought of as extinct ancestors. In particular the root of the tree represents the most recent common ancestor of all species in the tree. Following paths from the root to the leaves, each bifurcation indicates a speciation event whereby a new species is created from a parent. An excellent introduction to phylogenetics is [8].

A standard assumption in computational phylogenetics is that genetic information evolves from the root to the leaves according to a Markov model on the tree. Hence, the stochastic model used in trace reconstruction can be seen as a special case where the phylogeny is *star-shaped*. It may seem that a star is a good first approximation for the evolution of DNA sequences. However extensive work on the so-called reconstruction problem in theoretical computer science and statistical physics has highlighted the importance of taking into account the full tree model in analyzing the reconstruction of ancestral sequences. See below for references. We first discuss the reconstruction problem on a tree without indels. The substitution-only model itself is known in biology as the Cavender-Farris-Neyman (CFN) [9, 10, 11] model.

The reconstruction problem. In the reconstruction problem (RP) on a tree, we have a single site which evolves through substitutions only from the root to the leaves of a tree. In the most basic setup which we will consider here, the tree is a complete d-ary tree and each edge is an independent symmetric indel-free channel where the probability of a substitution is a constant $p_{\mathrm{s}} > 0$. The goal is to reconstruct the state at the root given the vector of states at the leaves. More generally, one can consider a sequence of length k at the root where each site evolves independently according to the Markov process above. Denote by n the number of leaves in the tree. The RP has attracted much attention in the probability theory and theoretical computer science literatures due to its deep connections to computational phylogenetics [12, 13, 14, 15, 16] and statistical physics [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]. See e.g. [30, 31] for background.

Unlike the star case, the RP on a tree exhibits an interesting thresholding effect: on the one hand, information is lost at an exponential rate along each path from the root; on the other hand, the number of paths grows exponentially with the number of levels. When the substitution probability is low, the latter "wins" and vice versa. This "phase transition" has been thoroughly analyzed in the theoretical computer science and mathematical physics literature—although much remains to be understood. More formally, we say that the RP is *solvable* when the correlation between the root and the leaves persists no matter how large the tree is. Note that, unlike the TRP case, we do not require high-probability reconstruction as it is not information-theoretically achievable for d constant. Indeed, consider the information lost on the first level below the root. Moreover the "number of traces" is irrelevant here as it is governed by the depth of the tree and the solvability notion implies nontrivial correlation for *any* depth. When the RP is unsolvable, the correlation decays to 0 for large trees. The results of [32, 18, 33, 23, 21, 24] show that for the CFN (that is, the substitution-only two-state symmetric) model, if $p_{\mathrm{s}} < p^*$, then the RP is solvable, where $d(1 - 2p^*)^2 = 1$. This is the so-called *Kesten-Stigum* bound [34]. If, on the other hand, $p_{\mathrm{s}} > p^*$, then the RP is *unsolvable*. Moreover in this case, the correlation between the root state and any function of the states at the leaves decays as $n^{-\Omega(1)}$. The positive result above is obtained by taking a majority vote over the leaf states.

Results on the RP have been used in previous work to advance the state of the art in rigorous phylogenetic tree reconstruction methods [13, 14, 35, 15]. A central component in these methods is to solve the RP on a partially reconstructed phylogeny to obtain sequence information that is "close" to the evolutionary past; then this sequence information is used to obtain further structural information about the phylogeny. The whole phylogeny is built by alternating these steps.

Our results. However the RP is only an *idealized* version of the ancestral sequence reconstruction problem in that it ignores the presence of indels. In particular, the RP becomes relevant after homologous sites in extant species have been perfectly identified, that is, assuming that the multiple sequence alignment problem has been solved perfectly. This is in fact a long-standing assumption in evolutionary biology where one typically preprocesses sequence data by running it through a multiple sequence alignment heuristic and then one only has to model the substitution process. This simplification has been criticized in the biology literature, where it has been argued that alignment procedures often create systematic biases that affect analysis [36, 37]. Much empirical work has been devoted to the proper joint estimation of alignments and phylogenies [38, 39, 40, 41, 42, 43, 36, 44].

We make progress in this direction by analyzing the RP in the presence of indels which we also refer to as the TRP on a tree (TRPT). We consider a d-ary tree where each edge is an independent channel with substitution probability p_{s}, deletion probability p_{d}, and insertion probability p_{i}. The root sequence has length k and is assumed to be uniform in $\{0,1\}^k$. See Section 1.1 for a precise statement of the model. For the same reasons that applied to the RP problem on a tree, we drop the requirement of high-probability reconstruction and seek instead a reconstructed sequence that exhibits a correlation with the true root sequence bounded away from 0 uniformly in the depth.

We give an efficient recursive procedure which solves the TRPT for $p_{\mathrm{s}} > 0$ a small enough constant (strictly below, albeit close, to the Kesten-Stigum bound) and $p_{\mathrm{d}}, p_{\mathrm{i}} = O(k^{-2/3} \log^{-1} n)$. As a by-product of our analysis we also obtain a partial alignment of the sequences at the leaves. Our method provides a framework for separating the indel process from the substitution process by identifying well-preserved subsequences which then serve as markers for alignment and reconstruction. See Section 1.2 for a high-level description of our techniques. As far as we are aware, our results are the first rigorous results for this problem. Our method also sets up a framework for extending rigorous phylogenetic tree reconstruction techniques beyond substitution-only models.

The results presented here were announced without proof in [45].

Related work. Much work has been devoted to the TRP on a star [1, 2, 3, 4, 5, 6]. In particular, in [5], it was shown that, when there are only deletions, it is possible to tolerate a small constant deletion rate using $n = \mathrm{poly}(k)$ traces. For a different range of parameters, Viswanathan and Swaminathan [6] showed that, under constant substitution probability and $O(1/\log k)$ indel probability, $O(\log k)$ traces suffice. Both results assume that the root sequence $\mathbf{X}$ is uniformly random.

The multiple sequence alignment problem as a combinatorial optimization problem (finding the best alignment under a pairwise scoring function) is known to be NP-hard [46, 47]. Most heuristics used in practice, such as CLUSTAL [48], T-Coffee [49], MAFFT [50], and MUSCLE [51], use the idea of a guide tree, that is, they first construct a very rough phylogenetic tree from the data (using edit distance as a measure of evolutionary distance), and then recursively "align the alignments." Our work can be thought as an attempt to analyze rigorously this type of procedure. Note that the Steiner version of the multiple sequence alignment problem on a fixed phylogeny, the so-called *tree alignment* problem, is known to admit a polynomial-time approximation scheme [52, 53].

Our work is tangentially related to the study of edit distance. Edit distance and pattern matching in random environments has been studied, e.g., by [54, 55, 56].

More recently, following the current work, two of the authors provided a phylogenetic tree reconstruction algorithm using poly-logarithmic sequence lengths under a similar indel model [57].

1.1. Definitions

We now define our basic model of sequence evolution.

Definition 1.1 (Model of sequence evolution). *Let $T_H^{(d)}$ be the d-ary tree with H levels and $n = d^H$ leaves. For simplicity, we assume throughout that d is odd. We consider the following model of evolution on $T_H^{(d)}$. The sequence at the root of $T_H^{(d)}$ has length k and is drawn uniformly at random over $\{0,1\}^k$. Along each edge of the tree, each site (or position) undergoes the following mutations independently of the other sites:*

- ***Substitution.*** *The site state is flipped with probability* $p_{\rm s} > 0$*.*

- ***Deletion.*** *The site is deleted with probability* $p_{\rm d} > 0$*.*

- ***Insertion.*** *A new site is created to the right of the current site with probability* $p_{\rm i} > 0$*. The state of this new site is uniform on* $\{0, 1\}$*.*

These operations occur independently of each other. The last two are called indels*. We let* $p_{\rm id} = p_{\rm i} + p_{\rm d}$ *and* $\theta_{\rm s} = 1 - 2p_{\rm s}$*. The parameters* $p_{\rm s}, p_{\rm d}, p_{\rm i}$ *may depend on* k *and* n*.*

Remark 1.2. *For convenience, our model of mutation is intentionally simplistic. In the biology literature, continuous-time Markov models on the alphabet* $\{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$ *are often used for this type of process [38, 39, 40, 41, 43, 58]. We expect that it should be straightforward to extend our results to such models by proper modifications to the algorithm.*

1.2. Results

Statement of results. Our main result is the following. Denote by $\mathbf{X} = X_1, \ldots, X_k$ a binary uniform sequence of length k. Run the evolutionary process on $T_H^{(d)}$ with root sequence $\mathbf{X}$ and let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be the sequences obtained at the leaves, where $\mathbf{Y}_i = Y_1^i, \ldots, Y_{k_i}^i$.

Theorem 1.3 (Main result). *For all* $\chi > 0$*, there is* $\Phi, \Phi', \Phi'' > 0$ *and* $d'' > 0$ *such that the following holds for* $d \geq d''$ *and* $\beta = d^{-1}$*. There is a polynomial-time algorithm* $\mathbb{A}$ *with access to* $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ *such that for all*

$$(1 - 2p_{\rm s})^2 > \frac{\Phi \log d}{d},$$

$$p_{\rm i} + p_{\rm d} < \frac{\Phi'}{k^{2/3} \log n},$$

$$\Phi'' \log^3 n < k < \mathrm{poly}(n),$$

the algorithm $\mathbb{A}$ *outputs a binary sequence* $\widehat{\mathbf{X}}$ *which satisfies the following with probability at least* $1 - \chi$*:*

1. $\widehat{\mathbf{X}} = \hat{X}_1, \ldots, \hat{X}_k$ *has length* k*.*
2. *For all* $j = 1, \ldots, k$*,* $\mathbb{P}[\hat{X}_j = X_j] > 1 - \beta$*.*

Remark 1.4. *Notice that we assume, for simplicity, that the sequence length of the root is known.*

Remark 1.5. *In fact, we prove a stronger result which shows that the agreement between* $\widehat{\mathbf{X}}$ *and* $\mathbf{X}$ *stochastically dominates an i.i.d. Bernoulli sequence with success probability* $1 - \beta$*.*

Proof sketch. We give a brief proof sketch. The full proof is detailed in Sections 3 to 5. As discussed previously, in the presence of indels the reconstruction of ancestral sequences requires the solution of the *multiple sequence alignment* problem. In addition to being computationally intractable, the standard global alignment approach through the optimization of a pairwise scoring function may create biases and correlations that are hard to quantify. We introduce a more probabilistic approach. From a purely information-theoretic point of view the pairwise alignment of sequences that are far apart in the tree is difficult. A natural solution to this problem is instead to align sequences that are close by in the tree, perform ancestral reconstructions of these sequences, and recurse our way up the tree.

This *recursive* approach raises its own set of issues. Consider a parent node and its d children. It may be easy to align the children's sequences and derive a good approximation to the parent sequence (for example, through site-wise majority). Note however that, to allow a recursion of this procedure all the way to the root, we have to provide strong guarantees about the probabilistic behavior of our level-wise ancestral reconstruction. A careless alignment procedure creates biases and correlations that are hard to control. For instance, it is tempting to treat misaligned sites as independent unbiased noise but this idea presents difficulties:

> Consider a site j of the parent sequence and suppose that for this site we have succeeded in aligning all but two of the children, say 1 and 2. Let $X^i_{j_i}$ denote the site in the i'th child which was used to estimate the j'th site. By the independence assumption on the root sequence and the inserted sites, $X^1_{j_1}$ and $X^2_{j_2}$ are uniform and independent of $(X^i_{j_i})^d_{i=3}$. However, $X^1_{j_1}$ and $X^2_{j_2}$ may originate from the *same* neighboring site of the parent sequence and therefore are themselves correlated.

Quantifying the effect of this type of correlation appears to be nontrivial.

Instead, we use an *adversarial* approach to ancestral reconstruction. That is, we treat the misaligned sites as being controlled by an adversary who seeks to flip the reconstructed value. This comes at a cost: it produces an asymmetry in our ancestral reconstruction. Although the RP is well-studied in the symmetric noise case, much remains to be understood in the asymmetric case. In particular, obtaining tight results in terms of substitution probability here may not be possible as the critical threshold of the RP may be hard to identify. We do however provide a tailored analysis of the particular instance of the RP by recursive majority obtained through this adversarial approach and we obtain results that are close to the known threshold for the symmetric case. Unlike the standard RP, the reconstruction error is not i.i.d. but we show instead that it "dominates" an i.i.d. noise. (See Section 4.2 for a definition.) This turns out to be enough for a well-controlled recursion. We first define a level-wise alignment procedure which has a good success probability (independent of n). However, applying this alignment procedure multiple times in the tree is bound to fail sometimes. We therefore prove that the reconstruction procedure is somewhat robust in the sense that even if one of the d inputs to the reconstruction procedure is faulty, it still has a good probability of success.

As for our level-wise alignment procedure, we adopt an *anchor* approach. Anchors were also used by [4, 5]—although in a quite different way. We imagine a partition of every node's sequence into islands of length $O(k^{1/3})$. (The precise choice of the island length comes from a trade-off between the length and the number of islands in bounding the "bad" events below.) At the beginning of each island we have an anchor of length $O(\log n)$. Through this partition of the sequences in islands and anchors we aim to guarantee the following. Given a specific father node v, with fair probability 1) all the anchors in the children nodes are indel-free; and 2) for all parent islands, almost all of the corresponding children islands have no indel at all and, moreover, at most one child island may have a single indel. The "bad" children islands—those that do not satisfy these properties—are treated as controlled by an adversary. We show that Conditions 1) and 2) are sufficient to guarantee that: the anchors of all islands can be aligned with high probability and single indel events between anchors can be identified. This allows an alignment of all islands with at most one "bad" child per island and is enough to perform a successful adversarial recursive majority vote as described above. The bound on the maximum indel probability sustained by our reconstruction algorithm comes from satisfying Conditions 1) and 2) above.

Notation. For a sequence $\mathbf{X} = X_1, \ldots, X_k$, we let $\mathbf{X}[i:j] = X_i, \ldots, X_j$. We use the expression "with high probability (w.h.p.)" to mean "with probability at least $1 - 1/\text{poly}(n)$" where the polynomial in n can be made of arbitrarily high degree by choosing the appropriate constants large enough. We denote by $\text{Bin}(n,p)$ a random variable with binomial distribution of parameters n, p. For two random variables X, Y we denote by $X \sim Y$ the equality in distribution.

2. Description of the Algorithm

In this section we describe our algorithm for TRPT. Our algorithm is recursive, proceeding from the leaves of the tree to the root. We describe the recursive step applied to a non-leaf node of the tree.

Recursive Setup—Our Goal. For our discussion in this section, let us consider a non-leaf node v with d children, denoted u_i for $i \in [d]$. For notational convenience, we drop the index u and denote its children by $1, \ldots, d$. Our goal for the recursive step of the algorithm is to reconstruct the sequence at the node v given the sequences of the children. Denote the sites of the father by $\mathbf{X}_0 = X_1^0, \ldots, X_{k_0}^0$, and the sites of the i'th child by $\mathbf{X}_i = X_1^i, \ldots, X_{k_i}^i$. During the reconstruction process, we do not have access to the children's sequences, but rather to reconstructed sequences denoted by $\widehat{\mathbf{X}}_i = \hat{X}_1^i, \ldots \hat{X}_{\hat{k}_i}^i$.

Let us consider the following partition of the sequence of v into subsequences, called *islands*. Of course our algorithm does not have access to the sequence at v during the recursive step of the algorithm. We define the partition as a means to describe our algorithm: The sites of v are partitioned into *islands* of length $\ell = k^{1/3}$ (except for the last one which is possibly shorter). Denote by $N_0 = \lceil k_0/\ell \rceil$ the number of islands in v. Each island starts with an *anchor* of a bits. That is, the islands are the bitstrings $\mathbf{X}_0[1:\ell]$, $\mathbf{X}_0[\ell+1:2\ell], \ldots$ and the anchors are the bitstrings $\mathbf{X}_0[1:a]$, $\mathbf{X}_0[\ell+1:\ell+a], \ldots$.

Our algorithm tries to identify, for each island $\mathbf{X}_0[(i-1)\ell+1:i\ell]$, the substrings of each of the d children that correspond to this island (that is, contain the sites of the island), called "child islands" and then performs ancestral reconstruction on the aligned child islands by site-wise majority. This task is not straightforward because of the shifts produced by indels.

We proceed iteratively for $i = 1 \dots N_0$. We use the islands that have been identified as indel-free for ancestral reconstruction.

Some islands do have indels however. This leads to two "modes of failure": one invalidates the entire (parent) node, and the other invalidates only an island of a child. More specifically, a parent node becomes invalidated (that is, useless) when indels are not evenly distributed, that is: when an indel occured in an anchor, or two (or more) indels occured in a specific island over all d children. This is a rare event. Barring this event, each island suffers only at most one indel over all children. The island (of a child) that has exactly one indel is invalidated (second mode of failure), and is thus deemed useless for reconstruction purposes. As long as the parent node is not invalidated, each island will have at least $d-2$ non-invalidated children islands with high probability (one additional island is potentially lost to a child node that may have been invalidated at an earlier stage; see Section 3.2).

Even when the algorithm identifies that a child island has an indel somewhere, the island is not ignored. The algorithm still needs to compute the length of the island in order to know the start of the next island in this child. For this purpose, we use the anchor of the next island and match it to the corresponding anchors of the other (non-invalidated) child islands. In fact the same procedure lets us detect which of the child islands are invalidated.

More formally, we define d functions $f_i : \{1, \dots, k_0\} \to \{1, \dots, k_i\} \cup \{\dagger\}$, where f_i takes a site of v to the corresponding site of the i'th child or to the special symbol $\dagger$ if the site was deleted. Note that for each i, f_i is monotone, when ignoring sites which are mapped to $\dagger$. For $t = \ell r$, let $s_i(r) = f_i(t+1) - (t+1)$ denote the displacement in the i^{th} child of the site corresponding to the $(t+1)^{\text{st}}$ site of the parent, that is, the starting site of the $(r+1)$'th island. (We leave $s_i(r)$ undefined if $f_i(t+1) = \dagger$. Below, we will only be interested in a subtree where this does not happen. See Section 3.2.) By convention, we take $s_i(0) = 0$. If there is no indel between $t = \ell r$ and $t' = \ell r'$ then $s_i(r) = s_i(r')$ (assuming $s_i(r)$ and $s_i(r')$ are well-defined). Note that, in the specific case of one indel operation in the r-th island, we have that $|s_i(r-1) - s_i(r)| = 1$ (assuming $s_i(r-1)$ and $s_i(r)$ are well-defined).

Algorithm. Our algorithm estimates the values of $s_i(r)$ and uses these estimates to match the starting positions of the islands in the children. The full algorithm is given in Figure 1. We use the following additional notation. For $x \in \{0,1\}$, we let $\langle x \rangle = 2x - 1$. Then, for two $\{0,1\}^m$-sequences

$Y = y_1, \ldots, y_m$ and $Z = z_1, \ldots, z_m$, we define their (empirical) correlation as

$$\mathrm{Corr}(Y, Z) = \frac{1}{m} \sum_{j=1}^{m} \langle y_j \rangle \langle z_j \rangle.$$

Note that $y \mapsto \langle y \rangle$ maps 1 to 1 and 0 to -1. One can think of $\mathrm{Corr}(Y, Z)$ as a form of normalized centered Hamming distance between Y and Z. In particular, a large value of $\mathrm{Corr}(Y, Z)$ implies that Y and Z tend to agree. We will use the following threshold (which will be justified in Section 5.1)

$$\gamma = ((1 - \delta)(1 - 2p_{\mathrm{s}})^2 - 4\beta),$$

where δ is chosen so that

$$(1 - \delta)(1 - 2p_{\mathrm{s}})^2 - 8\beta > \delta + 8\beta,$$

where again $\beta = d^{-1}$ and d is large enough.

3. Analyzing the Indel Process

We define $a \geq C \log n$ and $\alpha \leq \varepsilon/d < 1$, for constants C, ε to be determined later. We require $a < k^{1/3} < \mathrm{poly}(n)$. We assume that the indel probability per site satisfies

$$p_{\mathrm{id}} = \frac{\alpha}{4dk^{2/3}a} = O\left(\frac{1}{k^{2/3}\log n}\right).$$

Throughout, we denote the tree by $T = (V, E)$.

3.1. Bound on the Sequence Length

As the indel probability is defined per site, longer sequences suffer more indel operations than shorter ones. We begin by bounding the effect of this process. We claim that with high probability the lengths of all sequences are roughly equal.

Lemma 3.1 (Bound on sequence length). *For all $\zeta > 0$ (small), there exists $C' > 0$ (large) so that for all u in V, we have*

$$k_v \in [\underline{k}, \bar{k}] \equiv [(1 - \zeta)k, (1 + \zeta)k],$$

with high probability given $k \geq C' \log^3 n$. We denote this event by $\mathcal{L}$.

Proof of Lemma 3.1: We prove the upper bound by assuming there is no deletion. The lower bound can be proved similarly. The proof goes by induction. Let v be a node at graph distance i from the root. We show that there is $C'' > 0$ independent of i such that

$$k_v \leq k + i\sqrt{C''k\log n}.$$

Since the depth of T is $O(\log n)$, this implies the main claim as long as

$$\sqrt{C''k\log n}\log n \leq \zeta k,$$

which follows from our assumption for $C' > 0$ large enough.

The base case of the induction is satisfied trivially. Assume the induction claim holds for v, the parent of u. It suffices to show that the number of new insertions is at most $\sqrt{C''k\log n}$. By our induction hypothesis, the number of insertions is bounded above by a binomial Z with parameters $k + (i-1)\sqrt{C''k\log n} \leq (1+\zeta)k$ and p_{id} w.h.p. By Hoeffding's inequality, taking

$$\eta = \sqrt{\frac{C'''\log n}{(1+\zeta)k}},$$

we have

$$\begin{array}{rcl} \mathbb{P}[Z > (1+\zeta)kp_{\mathrm{id}} + (1+\zeta)k\eta] & < & \exp(-2((1+\zeta)k\eta)^2/[(1+\zeta)k]) \\ & = & 1/\mathrm{poly}(n). \end{array}$$

By our assumption on p_{id}, we have

$$(1+\zeta)kp_{\mathrm{id}} = O\left(\frac{\alpha k^{1/3}}{\log n}\right),$$

so that choosing C'' large enough gives

$$(1+\zeta)kp_{\mathrm{id}} + (1+\zeta)k\eta \leq \sqrt{C''k\log n}.$$

This proves the claim. ■

3.2. Existence of a Dense Stable Subtree

We claim that with probability close to 1 there exists a dense subtree of T with a "good indel structure," as defined below. Our algorithm will try to identify this subtree and perform reconstruction on it, as described in Section 4.

Indel structure of a node. Recall that $\ell = k^{1/3}$.

Definition 3.2 (Indel structure). *For a node (parent) v, we say that v is radioactive if one of the following events happen:*

1. *Event $\mathcal{B}_1$: Node v has a child u such that when evolving from v to u an indel operation occurred in at least one of the sites which are located in an anchor.*
2. *Event $\mathcal{B}_2$: There is an island I and two children u, u', such that an indel occurred in I in the transition from v to u and in the transition from v to u'.*
3. *Event $\mathcal{B}_3$: There is an island I and a child u, such that two indel operations (or more) happened in I in the transition from v to u.*

Otherwise the node v is stable. By definition, the leaves of T are stable. A subtree of T is stable if all of its nodes are stable.

Lemma 3.3 (Bound on radioactivity). *For all $0 < \alpha \leq 1$, there exists a choice of $\zeta > 0$ small enough in Lemma 3.1 such that conditioning on the event $\mathcal{L}$ occuring: any vertex v is radioactive with probability at most α.*

Proof of Lemma 3.3: According to Lemma 3.1, the length of the sequence at v is in $[\underline{k}, \bar{k}]$ w.h.p. We denote that event by $\mathcal{L}_v$. We bound the probability of events $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ separately.

Let $\overline{N} = \bar{k}/\ell = (1+\zeta)k^{2/3}$. Conditioned on $\mathcal{L}_v$, there are at most $\overline{N}$ anchors, each of length a. By a union bound, the probability that at least one of the sites in the anchors has an indel operation in any child is upper bounded by

$$\begin{aligned}
\mathbb{P}[\mathcal{B}_1] &= \mathbb{P}[\mathcal{B}_1 \,|\, \mathcal{L}_v]\mathbb{P}[\mathcal{L}_v] + \mathbb{P}[\mathcal{B}_1 \,|\, \mathcal{L}_v^c]\mathbb{P}[\mathcal{L}_v^c] \\
&\leq \overline{N} a d p_{\mathrm{id}} + 1/\mathrm{poly}(n) \\
&= \frac{\alpha a d \overline{N}}{4k^{2/3} a d} + 1/\mathrm{poly}(n) \\
&= \frac{(1+\zeta)k^{2/3}}{k^{2/3}} \cdot \frac{\alpha}{4} + 1/\mathrm{poly}(n) \\
&< \alpha(1/3 - 1/\mathrm{poly}(n)),
\end{aligned}$$

where we choose ζ small enough. The quantity we want to estimate is in fact $\mathbb{P}[\mathcal{B}_1 \,|\, \mathcal{L}]$ (which is not the same as conditioning on $\mathcal{L}_v$ only). But notice that

$$\mathbb{P}[\mathcal{B}_1] = \mathbb{P}[\mathcal{B}_1 \,|\, \mathcal{L}]\mathbb{P}[\mathcal{L}] + \mathbb{P}[\mathcal{B}_1 \,|\, \mathcal{L}^c]\mathbb{P}[\mathcal{L}^c] \geq \mathbb{P}[\mathcal{B}_1 \,|\, \mathcal{L}]\mathbb{P}[\mathcal{L}],$$

which implies

$$\mathbb{P}[\mathcal{B}_1 \,|\, \mathcal{L}] \leq \frac{\alpha(1/3 - 1/\mathrm{poly}(n))}{1 - 1/\mathrm{poly}(n)} < \alpha/3.$$

(This argument shows that it suffices to condition on $\mathcal{L}_v$. We apply the same trick below.)

To bound the probability of the second event, consider an island I and a son u. The probability that there is an indel when evolving from v to u is at most

$$p_{\mathrm{id}}\ell = \frac{\alpha}{4k^{2/3}ad}k^{1/3} = \frac{\alpha}{4k^{1/3}ad}.$$

Thus, the probability that more than one child of v experiences an indel in I is at most

$$\begin{aligned}
\sum_{i=2}^{d} \binom{d}{i} \left(\frac{\alpha}{4k^{1/3}ad}\right)^i &\leq \sum_{i=2}^{d} \frac{d^i}{i!} \left(\frac{\alpha}{4k^{1/3}ad}\right)^i \\
&\leq \sum_{i=2}^{d} \frac{1}{i!} \left(\frac{\alpha}{4k^{1/3}a}\right)^i \\
&\leq e\left(\frac{\alpha}{4k^{1/3}a}\right)^2 \\
&= \frac{e\alpha^2}{16k^{2/3}a^2},
\end{aligned}$$

where we used that the expression in parenthesis on the second line is < 1. Taking a union bound over all islands, the probability that at least two children experience an indel in the same island is at most

$$\begin{aligned}
\mathbb{P}[\mathcal{B}_2 \,|\, \mathcal{L}] &\leq \overline{N} \cdot \frac{e\alpha^2}{16k^{2/3}a^2} \\
&= \frac{(1+\zeta)e\alpha^2}{16a^2} \\
&< \frac{\alpha}{3},
\end{aligned}$$

where we used that $\alpha < 1$.

For the third event, consider again an island I and a child u. The probability that at least two indel operations occur in I when evolving from v to

u is at most

$$\begin{aligned}\sum_{i=2}^{2\ell}\binom{2\ell}{i}\left(\frac{\alpha}{4adk^{2/3}}\right)^i &\leq \sum_{i=2}^{2\ell}\frac{1}{i!}\left(\frac{2\ell\alpha}{4adk^{2/3}}\right)^i\\ &\leq \sum_{i=2}^{2\ell}\frac{1}{i!}\left(\frac{\alpha}{2adk^{1/3}}\right)^i\\ &\leq e\left(\frac{\alpha}{2adk^{1/3}}\right)^2\\ &\leq \frac{e\alpha^2}{4a^2d^2k^{2/3}}.\end{aligned}$$

(We use 2ℓ to account for insertions *and* deletions.) Taking a union bound over all islands and children, the probability that there are two indel operations in the same child in the same island is bounded by

$$\begin{aligned}\mathbb{P}[\mathcal{B}_3 \,|\, \mathcal{L}] &\leq d\overline{N}\frac{e\alpha^2}{4a^2d^2k^{2/3}}\\ &\leq \frac{(1+\zeta)e\alpha^2}{4a^2d}\\ &< \alpha/3.\end{aligned}$$

Taking a union bound over the three ways in which a site can become radioactive proves the lemma. ■

Lemma 3.4 (Existence of a dense stable subtree). *For all $0 < \chi < 1$, there is a choice of $\zeta > 0$ small enough in Lemma 3.1 such that, conditioning on the event $\mathcal{L}$ occuring, with probability at least $1-\chi$, the root of T is the father of a $(d-1)$-ary stable subtree of T. We denote this event by $\mathcal{S}$.*

Proof of Lemma 3.4: We follow a proof of [19]. Let v be a node at distance r from the leaves. We let ν_r be the probability that v is the root of a $(d-1)$-ary stable subtree conditioned on $\mathcal{L}$. Let

$$g(\nu) = \nu^d + d\nu^{d-1}(1-\nu).$$

We argue as in Lemma 3.3. Let ν'_r (respectively ν'_{r-1}) be the probability that v (respectively one of its children) is the root of a $(d-1)$-ary stable subtree *conditioned on* $\mathcal{L}_v$ (defined in Lemma 3.3). Then

$$\nu'_r \geq (1-\alpha)g(\nu'_{r-1}).$$

By the argument in Lemma 3.3, $\nu'_r = \nu_r + 1/\mathrm{poly}(n)$ and $\nu'_{r-1} = \nu_{r-1} + 1/\mathrm{poly}(n)$ so that the previous inequality holds without the primes up to an additive $1/\mathrm{poly}(n)$ term.

Note that

$$g'(\nu) = d(d-2)\nu^{d-2}(1-\nu).$$

In particular, g is monotone, $g(1) = 1$, and $g'(1) = 0$. Hence, for all $0 < \chi < 1$, there is $1 - \chi < \nu^* < 1$ such that

$$g(\nu^*) > \nu^*.$$

Then, taking α small enough that

$$1 - \alpha > \nu^*/g(\nu^*),$$

we have

$$\nu_r \gtrsim (1-\alpha)g(\nu_{r-1}) \gtrsim \frac{\nu^*}{g(\nu^*)} g(\nu_{r-1}) \gtrsim \nu^* > 1 - \chi,$$

by the induction hypothesis that $\nu_{r-1} \geq \nu^*$, where $\gtrsim$ indicates inequality up to an additive $1/\mathrm{poly}(n)$ term. Note in particular that $\nu_0 = 1 \geq \nu^*$. ■

4. A Stylized Reconstruction Process

We describe a hypothetical sequence reconstruction process performed on the stable tree defined from the indels. Assuming that the radioactive nodes and the islands with indels are controlled by an adversary, we argue that the process gives strong reconstruction guarantees. In the next section, we will then argue that the true algorithm performs at least as well as this hypothetical reconstruction process against an adversary. Throughout we suppose that a stable tree exists and is given to us, together with the "orbit" of every site of the sequence at the root of the tree (see function F below). However, we are given no information about the substitution process.

Let $v \in V$ and assume v is the root of a $(d-1)$-ary stable subtree $T^* = (V^*, E^*)$ of T. (We make the stable subtree below v into a $(d-1)$-ary tree by removing nodes from it at random.) Let $u \in V^*$. For each island I in u, at most one child u' of u in T^* contains an indel in which case it contains exactly one indel. We say that such an I is a corrupted island of u'. The basic intuition behind our analysis is that, provided the alignment on T^* is performed correctly (which we defer to Section 5.2), the ancestral reconstruction step of our algorithm is a recursive majority procedure against an adversary which controls the corrupted islands and the radioactive nodes (as well as all their descendants). Below we analyze this adversarial process.

Recursive majority. We begin with a formal definition of recursive majority. Let $\mathrm{Maj} : \{0, 1, \sharp\}^d \to \{0, 1\}$ be the function that returns the majority value over non-$\sharp$ values, and flips an unbiased coin in case of a tie (including the all-$\sharp$ vector). Let $n_0 = d^{H_0}$ be the number of leaves in T below v. Consider the following recursive function of $z = (z_1, z_2, \ldots, z_{n_0}) \in \{0, 1, \sharp\}$: $\mathrm{Maj}^0(z_1) = z_1$, and

$$\begin{aligned}&\mathrm{Maj}^j(z_1, \ldots, z_{d^j})\\ &\quad = \mathrm{Maj}(\mathrm{Maj}^{j-1}(z_1, \ldots, z_{d^{(j-1)}}), \ldots,\\ &\qquad \mathrm{Maj}^{j-1}(z_{d^j - d^{(j-1)}+1}, \ldots, z_{d^j})),\end{aligned}$$

for all $j = 1, \ldots, H_0$. Then, $\mathrm{Maj}^{H_0}(z)$ is the d-wise recursive majority of z.

Let $\mathbf{X}_0 = X^0_1, \ldots, X^0_{k_0}$ be the sequence at v. For $u \in V^*$ and $t = 1, \ldots, k_0$, we denote by $F_u(t)$ the position of site X^0_t in u or $\dagger$ if the site has been deleted on the path to u. We say that $\mathcal{C}_{u,t}$ holds if $F_u(t)$ is in a corrupted island of u. Let $\mathrm{Path}(u, v)$ be the set of nodes on the path between u and v.

Definition 4.1 (Gateway node). *A node u is a* gateway *for site t if:*

1. *$F_u(t) \neq \dagger$; and*
2. *For all $u' \in \mathrm{Path}(u, v) - \{v\}$, $\mathcal{C}_{u',t}$ does not hold.*

We let $T^{**}_t = (V^{**}_t, E^{**}_t)$ be the subtree of T^* containing all gateway nodes for t. By construction, T^{**}_t is at least $(d-2)$-ary and for convenience we remove nodes at random to make it exactly $(d-2)$-ary. Notice that, for $t, t' \in [1 : k_0]$, the subtrees T^{**}_t and $T^{**}_{t'}$ are random and correlated. However, they are independent of the substitution process.

We will argue in Section 5.2 that the reconstructed sequence produced by our method at v "dominates" (see below) the following reconstruction process. Let $L_v = u_1, \ldots, u_{n_0}$ be the leaves below v ordered according to a planar realization of the subtree below v. Denote by $\mathbf{X}_i = X^i_1, \ldots, X^i_{k_i}$ the sequence at u_i. For $t = 1, \ldots, k_0$, let L^{**}_t be the leaves of T^{**}_t. We define the following auxiliary sequences: for $u_i \in L_v$, we let $\Xi_i = \xi^i_1, \ldots, \xi^i_{k_i}$ where for $t = 1, \ldots, k_0$

$$\xi^i_t = \begin{cases} X^i_{F_{u_i}(t)} & \text{if } u_i \in L^{**}_t \\ 1 - X^0_t & \text{o.w.} \end{cases}$$

In words, ξ^i_t is the descendant of X^0_t if u_i is a gateway to t and is the opposite of the value X^0_t otherwise. Because of the monotonicity of recursive majority, the latter choice is in some sense the "worst adversary" (ignoring correlations between sites—we will come back to this point later). We then define a reconstructed sequence at v as $\widehat{\Xi}_0 = \hat{\xi}^0_1, \ldots, \hat{\xi}^0_{k_0}$ where for $t = 1, \ldots, k_0$

$$\hat{\xi}^0_t = \mathrm{Maj}^{H_0}(\xi^1_t, \ldots, \xi^{n_0}_t).$$

We now analyze the accuracy of this (hypothetical) estimator—which we refer to as the *adversarial reconstruction of* $\mathbf{X}_0$.

4.1. Recursive Majority Against an Adversary

To analyze the performance of the adversarial reconstuction $\widehat{\Xi}_0$, we consider the following stylized process.

Definition 4.2 (Adversarial Process). *We consider the following process:*

1. *Run the evolutionary process on $T^{(d-2)}_{H_0}$ at one position only starting with root state 0 without indels, that is, taking $p_{\mathrm{id}} = 0$.*
2. *Then complete $T^{(d-2)}_{H_0}$ into $T^{(d)}_{H_0}$ and associate to each additional node the state 1.*
3. *Let $R^{(d)}_{H_0}$ be the random variable in $\{0, 1\}$ obtained by running recursive majority on the leaf states obtained above.*

We call this process the *recursive majority against an adversary on $T^{(d)}_{H_0}$*.

Lemma 4.3 (Accuracy of recursive majority). *There exists a constant $C'' > 0$ and $d'' > 0$ such that taking*

$$\theta_{\mathrm{s}}^2 > \frac{C'' \log d}{d},$$

and $d \geq d''$, then the probability that the recursive majority against an adversary on $T^{(d)}_{H_0}$ correctly reconstructs root state 0 is at least $1 - \beta$ uniformly in H_0 where $\beta = d^{-1}$. In comparison, note that the Kesten-Stigum bound for binary symmetric channels on d-ary trees is $\theta^2 > d^{-1}$ [34, 59].

Proof of Lemma 4.3: Recall that we assume the root state is 0. Because of the bias towards 1 from Part 2 in Definition 4.2, we cannot apply standard results about recursive majority for symmetric channels [17, 13]. Instead, we perform a tailored analysis of this particular channel.

We take asymptotics as $d \to +\infty$ and we show that the probability of reconstruction can be taken to be

$$1 - \beta = 1 - \frac{1}{d},$$

for C'' large enough. Let v be the root of $T^{(d)}_{H_0}$. We denote by Z_v the number of non-adversarial children of v in state 0 and by Z'_v the number of nodes among them that return 0 upon applying recursive majority to their respective subtree. Let $q^0_{H_0}$ be the probability of incorrect reconstruction at v (given that the state at v is 0). Then

$$\begin{aligned} 1 - q^0_{H_0} &\geq \mathbb{P}\left[Z'_v \geq \frac{d+1}{2}\right] \\ &\geq \sum_{i=0}^{d-2} \mathbb{P}\left[Z'_v \geq \frac{d+1}{2} \mid Z_v = i\right] \mathbb{P}[Z_v = i], \end{aligned} \tag{1}$$

where we simply ignored the contribution of the children who flipped to 1.

We prove $q^0_{H_0} \leq 1/d$ by induction on the height. Let u be a non-adversarial node in $T^{(d)}_{H_0}$ at height h from the leaves to which we associate as above the variables Z_u, Z'_u and the quantity q^0_h. Note that $q^0_0 = 0$. We assume the induction hypothesis holds for $h-1$. Note that conditioned on the state at u being 0 Z_u is $\mathrm{Bin}(d-2, (1-p_{\mathrm{s}}))$ where

$$1 - p_{\mathrm{s}} = \frac{1+\theta_{\mathrm{s}}}{2} = \frac{1}{2} + \Theta\left(\sqrt{\frac{\log d}{d}}\right),$$

as $d \to +\infty$. Similarly, given $Z_u = i$, the variable Z'_u is $\mathrm{Bin}(i, 1 - q^0_{h-1})$. In particular, the quantity

$$\mathbb{P}\left[Z'_u \geq \frac{d+1}{2} \mid Z_u = i\right],$$

is monotone in i. We use Chernoff's bound on Z'_u to truncate the lower bound (1). Indeed, let

$$\mu = (1-p_{\mathrm{s}})(d-2) = \frac{d}{2} + \Upsilon(d),$$

with

$$\Upsilon(d) = \Theta(\sqrt{d \log d}),$$

and

$$\mu(1-\eta) = \frac{d}{2} + \frac{\Upsilon(d)}{2},$$

where in particular

$$\eta = \Theta\left(\sqrt{\frac{\log d}{d}}\right).$$

Then, we have

$$\mathbb{P}[Z_u < \mu(1-\eta)] < \exp\left(-\mu\eta^2/2\right) = d^{-\Omega(1)},$$

for C'' large enough. Applying to (1) leads to the lower bound

$$1 - q_h^0 \geq (1 - d^{-\Omega(1)})\mathbb{P}\left[\mathrm{Bin}\left(\frac{d}{2} + \frac{\Upsilon(d)}{2}, 1 - q_{h-1}^0\right) \geq \frac{d+1}{2}\right],$$

where we used monotonicity. By the induction hypothesis, $q_{h-1}^0 \leq 1/d$. By applying Chernoff's bound again we get

$$\mathbb{P}\left[\mathrm{Bin}\left(\frac{d}{2} + \frac{\Upsilon(d)}{2}, 1 - q_{h-1}^0\right) \geq \frac{d+1}{2}\right] > 1 - d^{-\Omega(1)},$$

and therefore $q_h^0 \leq 1/d$. This proves the claim. ■

Definition 4.4 (Bernoulli sequence). *For $q > 0$ and $m \in \mathbb{N}$, the (q, m)-Bernoulli sequence is the product distribution on $\{0,1\}^m$ such that each position is 1 independently with probability $1 - q$. We denote by $B_{q,m}$ the corresponding random variable.*

Lemma 4.5 (Subsequence reconstruction). *Assume v is the root of a $(d-1)$-ary stable subtree. Choosing $C'' > 0$ and $d'' > 0$ as in Lemma 4.3 is such that the following holds for $d \geq d''$ and $\beta = d^{-1}$. For $t, m \in \{1, \ldots, k_0\}$, let $\Lambda = (\lambda_1, \ldots, \lambda_m)$ be the* agreement vector *between the $\widehat{\Xi}_0[t+1 : t+m]$ and $\mathbf{X}_0[t+1 : t+m]$, that is, $\lambda_i = 1$ if recursive majority correctly reconstructs position i. Then there is $0 \leq \beta' \leq \beta$ such that $\Lambda \sim B_{\beta',m}$. (Here, β' may depend on H_0 but β does not.)*

Proof of Lemma 4.5: As we pointed out earlier, although the subtrees $(T_{t'}^{**})_{t'=t+1}^{t+m}$ are correlated by the construction of the islands, they are independent of the substitution process. By forcing (randomly) the subtrees $(T_{t'}^{**})_{t'=t+1}^{t+m}$ to be $(d-2)$-ary and fixing the adversarial nodes to 1 (as per Part 2 in Definition 4.2), we restore the i.i.d. nature of the reconstruction process on the sites, from which the result follows. ■

4.2. Stochastic Domination and Correlation

In our discussion so far we have assumed that a stable tree exists and is given to us, together with the function F. This allowed us to define the stylized recursive majority process against an adversary for which we claimed strong reconstruction guarantees. In reality, we have no access to the stable tree. We construct it recursively from the leaves to the root. At the same time we align sequences, discover corrupted islands, and reconstruct sequences of internal nodes. The stylized recursive majority process may be used to provide a lower bound on the actual reconstruction process. The notion of lower bound that is of interest to us is captured by *stochastic domination*, which we recall.

Definition 4.6 (Stochastic domination). *Let* $\mathbf{X}$,$\mathbf{Y}$ *be two random variables in* $\{0,1\}^m$. *We say that* $\mathbf{Y}$ *stochastically dominates* $\mathbf{X}$, *denoted* $\mathbf{X} \preceq \mathbf{Y}$, *if there is a joint random variable* $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ *such that the marginals satisfy* $\mathbf{X} \sim \widetilde{\mathbf{X}}$ *and* $\mathbf{Y} \sim \widetilde{\mathbf{Y}}$ *and moreover* $\mathbb{P}[\widetilde{\mathbf{X}} \leq \widetilde{\mathbf{Y}}] = 1$.

Correlation. The analysis of the previous section guarantees that the sequences output by the adversarial reconstruction process are well correlated with the true sequences. Now we establish that, under stochastic domination, the inter-sequence correlation is preserved. We first establish an important property of the adversarial process. Let T_u and T_v be the two disjoint copies of $T_h^{(d)}$ rooted at the nodes u and v respectively, and let $\mathbf{X} = X_1, X_2, \ldots, X_m \in \{0,1\}^m$ and $\mathbf{Y} = Y_1, Y_2, \ldots, Y_m \in \{0,1\}^m$ be sequences at the nodes u and v. Assume that u and v are the roots of $(d-1)$-ary stable subtrees. Let $\widehat{\mathbf{X}}' = \hat{X}'_1, \hat{X}'_2, \ldots, \hat{X}'_m \in \{0,1\}^m$ and $\widehat{\mathbf{Y}}' = \hat{Y}'_1, \hat{Y}'_2, \ldots, \hat{Y}'_m \in \{0,1\}^m$ be the reconstructions of $\mathbf{X}$ and $\mathbf{Y}$ obtained by the adversarial reconstruction process. Let $\Lambda = \lambda_1, \ldots, \lambda_m$ and $\Theta = \theta_1, \ldots, \theta_m$ be the resulting agreement vectors.

Lemma 4.7 (Concentration of bias). *Let* β', β *be as in Lemma 4.5. Then, with probability at least* $1 - e^{-\Omega(m\beta^2)}$ *the following are satisfied*

$$\left| \frac{1}{m} \sum_{i=1}^{m} \langle \lambda_i \rangle \langle \theta_i \rangle - (1-2\beta')^2 \right| \leq \frac{1}{2}\beta;$$

$$\left| \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\langle \lambda_i \rangle = -1} - \beta' \right| \leq \frac{1}{2}\beta;$$

$$\left|\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\langle\theta_i\rangle=-1}-\beta'\right|\leq\frac{1}{2}\beta.$$

Proof of Lemma 4.7: This follows from Lemma 4.5, the independence of Λ and Θ, and three applications of Hoeffding's lemma. ■

Lemma 4.8 (Correlation bound). *Let $\widehat{\mathbf{X}},\widehat{\mathbf{Y}}\in\{0,1\}^m$ be random strings defined on the same probability space as $\widehat{\mathbf{X}}'$ and $\widehat{\mathbf{Y}}'$. Denote by $\mathbf{Z}$ (resp. $\mathbf{W}$) the agreement vectors of $\widehat{\mathbf{X}}$ (resp. $\widehat{\mathbf{Y}}$) with $\mathbf{X}$ (resp. $\mathbf{Y}$). Assume that $\Lambda\leq\mathbf{Z}$ and $\Theta\leq\mathbf{W}$ with probability 1, where Λ and Θ are the agreement vectors of $\widehat{\mathbf{X}}'$ and $\widehat{\mathbf{Y}}'$ with $\mathbf{X}$ and $\mathbf{Y}$ as explained above. Then, conditioned on the conclusions of Lemma 4.7, we have, with probability 1*

$$|\operatorname{Corr}(\mathbf{X},\mathbf{Y})-\operatorname{Corr}(\widehat{\mathbf{X}},\widehat{\mathbf{Y}})|\leq 8\beta.$$

Proof of Lemma 4.8: Note that

$$\operatorname{Corr}(\widehat{\mathbf{X}},\widehat{\mathbf{Y}})=\frac{1}{m}\sum_{i=1}^{m}\langle\hat{X}_i\rangle\langle\hat{Y}_i\rangle=\frac{1}{m}\sum_{i=1}^{m}\langle X_i\rangle\langle Y_i\rangle\langle Z_i\rangle\langle W_i\rangle.$$

Hence,

$$|\operatorname{Corr}(\mathbf{X},\mathbf{Y})-\operatorname{Corr}(\widehat{\mathbf{X}},\widehat{\mathbf{Y}})|\leq\frac{1}{m}\sum_{i=1}^{m}(1-\langle Z_i\rangle\langle W_i\rangle)=1-\frac{1}{m}\sum_{i=1}^{m}\langle Z_i\rangle\langle W_i\rangle.$$

Now notice by case analysis that

$$\langle Z_i\rangle\langle W_i\rangle\geq\langle\lambda_i\rangle\langle\theta_i\rangle-\mathbb{1}_{\langle\lambda_i\rangle=-1}-\mathbb{1}_{\langle\theta_i\rangle=-1}.$$

The claim follows from the bounds in Lemma 4.7 which imply

$$1-\frac{1}{m}\sum_{i=1}^{m}\langle Z_i\rangle\langle W_i\rangle\leq 1-(1-2\beta')^2+2\beta'+\frac{3}{2}\beta\leq 8\beta,$$

where we used $0\leq\beta'\leq\beta$. ■

5. Analyzing the True Reconstruction Process

In Section 5.1 we argue that, if a stable subtree exists, the adversarial reconstructions of aligned children anchors of the same parent node exhibit strong correlation signal between them, while misaligned anchors exhibit weak signal. This holds true for sequences that stochastically dominate the adversarial reconstructions as well. See the "Anchor alignment" step in Figure 1.

Then in Section 5.2 we prove the correctness of our recursive procedure.

5.1. Anchor Alignment

Consider a parent v that is stable. Let i, j be two children with sequences $\mathbf{X}_i = X^i_1, \ldots, X^i_{k_i}$ and $\mathbf{X}_j = X^j_1, \ldots, X^j_{k_j}$. Let $t = \ell r$ and consider the following subsequences (of length a) at i and j

$$\mathscr{A}^i_r = X^i[t + s_i(r) + 1 : t + s_i(r) + a],$$

and

$$\mathscr{A}^j_r = X^j[t + s_j(r) + 1 : t + s_j(r) + a].$$

These are related (but not identical) to the definition of anchors in the algorithm of Section 2. In particular, note that by definition $\mathscr{A}^i_r$ and $\mathscr{A}^j_r$ are always aligned, in the sense that they correspond to the same subsequence of v. Consider also the following subsequences

$$\mathscr{D}^j_r = X^j[t + s_j(r) : t + s_j(r) + a - 1],$$

and

$$\mathscr{I}^j_r = X^j[t + s_j(r) + 2 : t + s_j(r) + a + 1].$$

These are the one-site shifted subsequences for j. We claim that $\mathscr{A}^i_r$ is always significantly more correlated to its aligned brother $\mathscr{A}^j_r$ than to the misaligned ones $\mathscr{D}^j_r$ and $\mathscr{I}^j_r$. This follows from the fact that the misaligned subsequences are sitewise independent. Recall that $\beta = d^{-1}$ and $(1 - 2p_{\mathrm{s}})^2 = \Omega(\frac{\log d}{d})$.

Lemma 5.1 (Anchor correlations). *For all $\delta > 0$ (and d large enough) such that $(1-\delta)(1-2p_{\mathrm{s}})^2 - 8\beta > \delta + 8\beta$, there is $C > 0$ large enough so that with $a = C \log n$, the following hold:*

1. ***Aligned anchors.***

$$\begin{aligned}&\mathbb{P}\left[\mathrm{Corr}(\mathscr{A}^i_r, \mathscr{A}^j_r) > (1-\delta)(1-2p_{\mathrm{s}})^2\right]\\ &\quad > 1 - \exp\left(-\Omega(a)\right) = 1 - 1/\mathrm{poly}(n).\end{aligned}$$

2. ***Misaligned anchors.***

$$\begin{aligned}&\mathbb{P}\left[\mathrm{Corr}(\mathscr{A}^i_r, \mathscr{D}^j_r) < \delta\right]\\ &\quad > 1 - \exp\left(-\Omega(a)\right) = 1 - 1/\mathrm{poly}(n),\end{aligned}$$

and similarly for $\mathscr{I}^j_r$.

We denote by $\mathcal{A}_{i,j,r}$ the above events and their symmetric counterparts under $i \leftrightarrow j$, that is, under the exchange of i and j.

Proof of Lemma 5.1: For the first claim note that, assuming that the parent v is stable, the expectation of $\mathrm{Corr}(\mathscr{A}^i_r, \mathscr{A}^j_r)$ is $\theta_\mathrm{s}^2 = (1-2p_\mathrm{s})^2$ where we used that 1) there is no indel in the sites $[t+1 : t+a]$ between v and i, j; 2) that the sites are perfectly aligned; and 3) that the substitution process is independent of the indel process. We also used the fact that the θ_s's behave multiplicatively along a path under our model of substitution [8]. The result then follows from Hoeffding's inequality.

For the second claim, because the anchors are now misaligned the t'-th term in $\mathrm{Corr}(\mathscr{A}^i_r, \mathscr{D}^j_r)$ for $t' \in [t{+}1 : t{+}a]$ is the variable $\langle X^i_{t'+s_i(r)}\rangle\langle X^j_{t'+s_j(r)-1}\rangle$ which is uniform in $\{-1, +1\}$. In particular, we now have the expectation of $\mathrm{Corr}(\mathscr{A}^i_r, \mathscr{D}^j_r)$ is 0. The result follows from the method of bounded differences applied to the independent vectors

$$\{(X^i_{t'+s_i(r)}, X^j_{t'+s_j(r)})\}_{t'=t}^{t+a}.$$

■

Lemma 5.2 (Reconstructed version). *Let $\widehat{\mathbf{X}}_i = (\hat{X}^i_\iota)_{\iota=1}^{k_i}$ and $\widehat{\mathbf{X}}_j = (\hat{X}^j_\iota)_{\iota=1}^{k_j}$ dominate the adversarial reconstructions $\widehat{\mathbf{X}}'_i$ and $\widehat{\mathbf{X}}'_j$ of $\mathbf{X}_i$ and $\mathbf{X}_j$, as defined in Lemma 4.8. Let $\hat{\mathscr{A}}^i_r = \hat{X}^i[t + s_i(r) + 1 : t + s_i(r) + a]$ and similarly for all other possibilities $\hat{\mathscr{A}} \leftrightarrow \hat{\mathscr{D}}, \hat{\mathscr{I}}$ and/or $i \leftrightarrow j$. Denote by $\mathcal{B}_{i,j,r}$ the event that the conclusions of Lemma 4.7 hold for $\widehat{\mathbf{X}}'_i$ and $\widehat{\mathbf{X}}'_j$ over all pairs of intervals involving $[t + s_i(r) : t + s_i(r) + a - 1]$, $[t + s_i(r) + 1 : t + s_i(r) + a]$, and $[t + s_i(r) + 2 : t + s_i(r) + a + 1]$, with $i \leftrightarrow j$ as necessary. Then, conditioned on $\mathcal{B}_{i,j,r}$ and $\mathcal{A}_{i,j,r}$ we have*

$$\mathrm{Corr}(\hat{\mathscr{A}}^i_r, \hat{\mathscr{A}}^j_r) > (1-\delta)(1-2p_\mathrm{s})^2 - 8\beta,$$

$$\mathrm{Corr}(\hat{\mathscr{A}}^i_r, \hat{\mathscr{D}}^j_r) < \delta + 8\beta,$$

and

$$\mathrm{Corr}(\hat{\mathscr{A}}^i_r, \hat{\mathscr{I}}^j_r) < \delta + 8\beta,$$

as well as their symmetric counterparts under $i \leftrightarrow j$.

Proof of Lemma 5.2: This follows from Lemmas 4.8 and 5.1 and the triangle inequality. ■

5.2. Proof of Correctness

Recall the definitions of the events $\mathcal{L}$, $\mathcal{S}$, $\mathcal{B}_{i,j,r}$, $\mathcal{A}_{i,j,r}$ from Lemmas 3.1, 3.4, 5.1 and 5.2. Conditioning on $\mathcal{L}$ and $\mathcal{S}$, denote by $T^* = (V^*, E^*)$ the stable $(d-1)$-ary subtree of T. Then, for all $v \in V^*$, all pairs of children i, j of v in T^*, and all $r = 1, \ldots, \bar{k}/\ell$, we condition on the events $\mathcal{B}_{i,j,r}$ and $\mathcal{A}_{i,j,r}$. Note that having conditioned on $\mathcal{L}$ there is only a polynomial number of such events, since all sequence lengths are bounded by $\bar{k}$. (If $r\ell$ is larger than a node's sequence length we assume that the corresponding events are vacuously satisfied.) Finally recall that, conditioning on $\mathcal{L}$, the event $\mathcal{S}$ occurs with probability $1 - \chi$ and all other events occur with high probability. We denote the collection of events by $\mathcal{E}$.

Conditioning on $\mathcal{E}$, the proof of correctness of the algorithm follows from a bottom-up induction. Suppose that at a recursive step of the algorithm we have reconstructed sequences for all children of a node v, which are strongly correlated with the true sequences (in the sense of dominating the corresponding adversarial reconstructions). Having conditioned on the events $\mathcal{A}_{i,j,r}$ and $\mathcal{B}_{i,j,r}$, it follows then that the correct alignments of anchors exhibit strong correlation signal while the incorrect alignments weak correlation signal. Hence, our correlation tests between anchors discover the corrupted islands and do the anchor alignments correctly (at least for all nodes lying inside the stable tree). Hence the shift functions $\hat{s}_i$'s are correctly inferred, and the reconstruction of v's sequence can be shown to dominate the corresponding adversarial reconstruction. We proceed with a formal proof.

Proof of Theorem 1.3: Having conditioned on the event $\mathcal{E}$, we justify the correctness of our reconstruction method via the following induction. The top level of the induction establishes Theorem 1.3. Below we use the notation introduced in Figure 1.

Induction hypothesis. Consider a parent v in T^*; in particular, v is stable. We assume that the following conditions, denoted by $(\star)$, are satisfied: For all children $i \in [d]$ of v belonging to T^*

1. **Alignment.** For all children i' of i with $i' \in T^*$ and all $r = 1, \ldots, \bar{k}/\ell - 1$,
$$\hat{s}_{i'}(r) = s_{i'}(r). \tag{2}$$
(This condition is trivially satisfied for values of $r\ell$ that are larger than the sequence length of i'.)

2. **Reconstruction.** Moreover, we have $\hat{k}_i = k_i$ and for all $t = 1, \ldots, k_i$, the following holds:

> Let L_i be the leaves below i with $n_i = |L_i|$. Let H be the level of v. Let L_t^{**} be the gateway leaves for site t. For $u \in L_t^{**}$ let $F_u(t)$ be the position of site t in u. Note that $\hat{X}_t^i$ can be written as $\hat{X}_t^i = \mathrm{Maj}^{H-1}(z_1, \ldots, z_{n_i})$, where z_j is either $\sharp$ or $X_{\flat_j}^j$ for an appropriate function $\flat_j$. Our hypothesis is that
>
> $$\forall u \in L_t^{**},\ \flat_u = F_u(t). \tag{3}$$

In particular, the ancestral reconstruction $\widehat{\mathbf{X}}_i$ dominates the adversarial reconstruction $\widehat{\mathbf{X}}'_i$.

The base case where v is a leaf is trivially satisfied.

Alignment. We begin with the correctness of the alignment.

Lemma 5.3 (Induction: Alignment). *Assuming $\mathcal{E}$ and $(\star)$, the algorithm infers s_i correctly for all children $i \in [d]$ which are also in T^*, that is, (2) holds for v.*

Proof of Lemma 5.3: Let Π denote the set of children of v in T^*. The proof follows by induction on r. The base case $r = 0$ is trivial. Assume correctness for $r - 1$.

If there is no indel in any of the children $i \in \Pi$ between the sites $(r-1)\ell$ and $r\ell$ of v, then under $\mathcal{E}$, $(\star)$ and Lemma 5.2 we have $\Pi \subseteq G_r$. In that case, for all $i \in \Pi$ we have $\hat{s}_i(r) = \hat{s}_i(r-1) = s_i(r-1) = s_i(r)$, where the second equality is from $(\star)$.

If there is an indel operation in island r, then since v is stable only one indel operation occurred in one child. Denote the child with an indel by j. Assume the indel is a deletion. (The case of the insertion is handled similarly.) If j is not in T^* we are back to the previous case. So assume j is in T^*. Again, from $\mathcal{E}$, $(\star)$ and Lemma 5.2 the other children in T^* are added to the set G_r, and the shift value will be computed correctly for them. Moreover by $(\star)$, for every $i \in \Pi - \{j\}$,

$$\begin{aligned} f_i(r\ell + 1) &= r\ell + 1 + s_i(r) \\ &= r\ell + 1 + \hat{s}_i(r) \\ &= r\ell + 1 + \hat{s}_i(r-1), \end{aligned}$$

which is the starting point of $\widehat{A}^i_r$. Also,

$$\begin{aligned} f_j(r\ell+1) &= r\ell+1+s_j(r) \\ &= r\ell+1+s_j(r-1)-1 \\ &= r\ell+1+\hat{s}_j(r-1)-1 \\ &= r\ell+\hat{s}_j(r-1), \end{aligned}$$

which is the starting point of $\widehat{D}^j_r$. Thus according to Lemma 5.2 $\widehat{D}^j_r$ matches $\widehat{A}^i_r$ for all $i \in \Pi \cap G_r$. As there are $d-2$ children in $\Pi \cap G_r$, we get that the algorithm sets

$$\hat{s}_j(r) = \hat{s}_j(r-1) - 1 = s_j(r-1) - 1 = s_j(r),$$

as required. Note also that in this case, according to Lemma 5.2 again, $\widehat{A}^j_r$ does not have high correlation with $\widehat{A}^i_r$ for any $i \in \Pi \cap G_r$, and thus we will consider $\widehat{I}^j_r$ and $\widehat{D}^j_r$. Similarly, $\widehat{I}^j_r$ does not have high correlation with $\widehat{A}^i_r$ for any $i \in \Pi \cap G_r$, and thus we will not try to set $\hat{s}_j(r)$ twice. ■

Ancestral reconstruction. We use Lemma 5.3 to prove that the ancestral reconstruction dominates the adversarial reconstruction. In the algorithm, we perform a sitewise majority vote over the children of v in G_r (these are the aligned children—see the description of the algorithm in Figure 1). For notational convenience, we assume that in fact we perform a majority vote over *all* children but we replace the states of the children outside G_r with $\sharp$.

Lemma 5.4 (Induction: Reconstruction). *Assuming $\mathcal{E}$, $(\star)$ and the conclusion of Lemma 5.3, (3) holds for v. In particular, the ancestral reconstruction $\widehat{\mathbf{X}}_v$ dominates the adversarial reconstruction $\widehat{\mathbf{X}}'_v$.*

Proof of Lemma 5.4: The second claim follows from the first one together with the construction of the adversarial process and the monotonicity of Maj (in the sense that, assuming the root state is 0, flipping the adversary's 1s to 0s or $\sharp$s cannot flip Maj to 1).

As for the first claim, by Lemma 5.3 for each site of v there are $d-2$ uncorrupted children islands containing this site such that the children are also in T^*. In particular, the $d-2$ corresponding sites in the children are correctly aligned. Moreover, by the induction hypothesis, each corresponding site in the children satisfy (3). By taking a majority vote over these sites we get (3) for v as well.

One last detail is handling the case where the last island is shorter than ℓ. In that case, we add the last island to the previous one (and treat the juxtaposition as a regular island in the analysis above). ■

This concludes the proof of Theorem 1.3. ■

6. Discussion

We have provided a novel algorithm for reconstructing ancestral sequences in the presence of indels. The algorithm also provides a partial alignment of the sequences at the leaves.

Several open problems remain. The bounds we obtained on the mutation parameters are likely not tight. In particular, it is not clear whether the bound on the indel probability should depend on k and n. Removing such dependence appears to be a significant challenge.

Also, we have only considered trees with sufficiently high degrees. In the biological context, one is generally interested in binary trees instead. It may be possible to extend our result to that case by dividing the tree into large subtrees. Such an approach was used successfully in the indel-free case [17, 13]

References

[1] V. I. Levenshtein, Efficient reconstruction of sequences, IEEE Transactions on Information Theory 47 (1) (2001) 2–22.

[2] V. I. Levenshtein, Efficient reconstruction of sequences from their subsequences or supersequences, J. Comb. Theory Ser. A 93 (2) (2001) 310–332. doi:http://dx.doi.org/10.1006/jcta.2000.3081.

[3] T. Batu, S. Kannan, S. Khanna, A. McGregor, Reconstructing strings from random traces, in: SODA '04: Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2004, pp. 910–918.

[4] S. Kannan, A. McGregor, More on reconstructing strings from random traces: insertions and deletions, in: Proceedings of ISIT, 2005, pp. 297–301.

[5] T. Holenstein, M. Mitzenmacher, R. Panigrahy, U. Wieder, Trace reconstruction with constant deletion probability and related results, in: SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008, pp. 389–398.

[6] K. Viswanathan, R. Swaminathan, Improved string reconstruction over insertion-deletion channels, in: SODA '08: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008, pp. 399–408.

[7] J. W. Thornton, Resurrecting ancient genes: experimental analysis of extinct molecules, Nat. Rev. Genet. 5 (5) (2004) 366–375.

[8] C. Semple, M. Steel, Phylogenetics, Vol. 22 of Mathematics and its Applications series, Oxford University Press, 2003.

[9] J. A. Cavender, Taxonomy with confidence, Math. Biosci. 40 (3-4).

[10] J. S. Farris, A probability model for inferring evolutionary trees, Syst. Zool. 22 (4) (1973) 250–256.

[11] J. Neyman, Molecular studies of evolution: a source of novel statistical problems, in: S. S. Gupta, J. Yackel (Eds.), Statistical desicion theory and related topics, Academic Press, New York, 1971, pp. 1–27.

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

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

[14] C. Daskalakis, E. Mossel, S. Roch, Optimal phylogenetic reconstruction, in: STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing, ACM, New York, 2006, pp. 159–168.

[15] S. Roch, Sequence-length requirement for distance-based phylogeny reconstruction: Breaking the polynomial barrier, in: FOCS, 2008, pp. 729–738.

[16] S. Roch, Toward extracting all phylogenetic information from matrices of evolutionary distances, Science 327 (5971) (2010) 1376–1379. doi:10.1126/science.1182300.
URL `http://dx.doi.org/10.1126/science.1182300`

[17] E. Mossel, Recursive reconstruction on periodic trees, Random Struct. Algor. 13 (1) (1998) 81–97.
URL `http://www.stat.berkeley.edu/ mossel/publications/recursive.ps`

[18] W. S. Evans, C. Kenyon, Y. Peres, L. J. Schulman, Broadcasting on trees and the Ising model, Ann. Appl. Probab. 10 (2) (2000) 410–433.

[19] E. Mossel, Reconstruction on trees: beating the second eigenvalue, Ann. Appl. Probab. 11 (1) (2001) 285–300.
URL `http://www.stat.berkeley.edu/ mossel/publications/second.ps`

[20] E. Mossel, Y. Peres, Information flow on trees, Ann. Appl. Probab. 13 (3) (2003) 817–844.

[21] F. Martinelli, A. Sinclair, D. Weitz, Glauber dynamics on trees: boundary conditions and mixing time, Comm. Math. Phys. 250 (2) (2004) 301–334.

[22] S. Janson, E. Mossel, Robust reconstruction on trees is determined by the second eigenvalue, Ann. Probab. 32 (2004) 2630–2649.
URL `http://www.stat.berkeley.edu/ mossel/publications/robust.pdf`

[23] N. Berger, C. Kenyon, E. Mossel, Y. Peres, Glauber dynamics on trees and hyperbolic graphs, Probab. Theory Rel. 131 (3) (2005) 311–340, extended abstract by Kenyon, Mossel and Peres appeared in proceedings of 42nd IEEE Symposium on Foundations of Computer Science (FOCS) 2001, 568–578.

[24] C. Borgs, J. T. Chayes, E. Mossel, S. Roch, The Kesten-Stigum reconstruction bound is tight for roughly symmetric binary channels., in: FOCS, 2006, pp. 518–530.

[25] A. Gerschenfeld, A. Montanari, Reconstruction for models on random graphs, in: FOCS '07: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, IEEE

Computer Society, Washington, DC, USA, 2007, pp. 194–204. doi:http://dx.doi.org/10.1109/FOCS.2007.58.

[26] N. Bhatnagar, J. C. Vera, E. Vigoda, D. Weitz, Reconstruction for colorings on trees, SIAM J. Discrete Math. 25 (2) (2011) 809–826.

[27] A. Sly, Reconstruction of random colourings, Communications in Mathematical Physics 288 (3) (2009) 943–961.

[28] A. Sly, Reconstruction for the Potts model, in: STOC '09: Proceedings of the 41st annual ACM symposium on Theory of computing, ACM, New York, NY, USA, 2009, pp. 581–590. doi:http://doi.acm.org/10.1145/1536414.1536493.

[29] Y. Peres, S. Roch, Reconstruction on trees: Exponential moment bounds for linear estimators, Electron. Comm. Probab. 16 (2011) 251–261 (electronic).

[30] S. Roch, Markov models on trees: Reconstruction and applications, Ph.D. thesis, UC Berkeley (2007).

[31] A. Sly, Spatial and temporal mixing of gibbs measures, Ph.D. thesis, UC Berkeley (2009).

[32] P. M. Bleher, J. Ruiz, V. A. Zagrebnov, On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice, J. Statist. Phys. 79 (1-2) (1995) 473–482.

[33] D. Ioffe, On the extremality of the disordered state for the Ising model on the Bethe lattice, Lett. Math. Phys. 37 (2) (1996) 137–143.

[34] H. Kesten, B. P. Stigum, Additional limit theorems for indecomposable multidimensional Galton-Watson processes, Ann. Math. Statist. 37 (1966) 1463–1481.

[35] R. Mihaescu, C. Hill, S. Rao, Fast phylogeny reconstruction through learning of ancestral sequences, CoRR abs/0812.1587.

[36] A. Loytynoja, N. Goldman, Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis, Science 320 (5883) (2008) 1632–1635.

arXiv:http://www.sciencemag.org/cgi/reprint/320/5883/1632.pdf, doi:10.1126/science.1158395.
URL http://www.sciencemag.org/cgi/content/abstract/320/5883/1632

[37] K. M. Wong, M. A. Suchard, J. P. Huelsenbeck, Alignment Uncertainty and Genomic Analysis, Science 319 (5862) (2008) 473–476. arXiv:http://www.sciencemag.org/cgi/reprint/319/5862/473.pdf, doi:10.1126/science.1151532.
URL http://www.sciencemag.org/cgi/content/abstract/319/5862/473

[38] J. L. Thorne, H. Kishino, J. Felsenstein, An evolutionary model for maximum likelihood alignment of dna sequences, Journal of Molecular Evolution 33 (2) (1991) 114–124.

[39] J. L. Thorne, H. Kishino, J. Felsenstein, Inching toward reality: An improved likelihood model of sequence evolution, Journal of Molecular Evolution 34 (1) (1992) 3–16.

[40] D. Metzler, Statistical alignment based on fragment insertion and deletion models, Bioinformatics 19 (4) (2003) 490–499. arXiv:http://bioinformatics.oxfordjournals.org/cgi/reprint/19/4/490.pdf, doi:10.1093/bioinformatics/btg026.
URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/4/490

[41] I. Miklos, G. A. Lunter, I. Holmes, A "Long Indel" Model For Evolutionary Sequence Alignment, Mol Biol Evol 21 (3) (2004) 529–540. arXiv:http://mbe.oxfordjournals.org/cgi/reprint/21/3/529.pdf, doi:10.1093/molbev/msh043.
URL http://mbe.oxfordjournals.org/cgi/content/abstract/21/3/529

[42] M. A. Suchard, B. D. Redelings, BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny, Bioinformatics 22 (16) (2006) 2047–2048. arXiv:http://bioinformatics.oxfordjournals.org/cgi/reprint/22/16/2047.pdf, doi:10.1093/bioinformatics/btl175.
URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/16/2047

[43] E. Rivas, S. R. Eddy, Probabilistic phylogenetic inference with insertions and deletions, PLoS Comput Biol 4 (9) (2008) e1000172. doi:10.1371/journal.pcbi.1000172.

[44] K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, T. Warnow, Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees, Science 324 (5934) (2009) 1561–1564. arXiv:http://www.sciencemag.org/cgi/reprint/324/5934/1561.pdf, doi:10.1126/science.1171243.
URL http://www.sciencemag.org/cgi/content/abstract/324/5934/1561

[45] A. Andoni, C. Daskalakis, A. Hassidim, S. Roch, Global alignment of molecular sequences via ancestral state reconstruction, in: ICS, 2010.

[46] L. Wang, T. Jiang, On the complexity of multiple sequence alignment, Journal of Computational Biology 1 (4) (1994) 337–348.

[47] I. Elias, Settling the intractability of multiple alignment, Journal of Computational Biology 13 (7) (2006) 1323–1339, pMID: 17037961. arXiv:http://www.liebertonline.com/doi/pdf/10.1089/cmb.2006.13.1323, doi:10.1089/cmb.2006.13.1323.
URL http://www.liebertonline.com/doi/abs/10.1089/cmb.2006.13.1323

[48] D. G. Higgins, P. M. Sharp, Clustal: a package for performing multiple sequence alignment on a microcomputer, Gene 73 (1) (1988) 237–244.

[49] C. Notredame, D. Higgins, J. Heringa, T-coffee: A novel method for fast and accurate multiple sequence alignment.

[50] K. Katoh, K. Misawa, K.-i. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucl. Acids Res. 30 (14) (2002) 3059–3066. arXiv:http://nar.oxfordjournals.org/cgi/reprint/30/14/3059.pdf, doi:10.1093/nar/gkf436.
URL http://nar.oxfordjournals.org/cgi/content/abstract/30/14/3059

[51] R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucl. Acids Res. 32 (5) (2004) 1792–1797. arXiv:http://nar.oxfordjournals.org/cgi/reprint/32/5/1792.pdf, doi:10.1093/nar/gkh340.
URL http://nar.oxfordjournals.org/cgi/content/abstract/32/5/1792

[52] D. Sankoff, Minimal mutation trees of sequences, SIAM Journal on Applied Mathematics 28 (1) (1975) 35–42. doi:10.1137/0128004.
URL http://link.aip.org/link/?SMM/28/35/1

[53] L. Wang, T. Jiang, E. L. Lawler, Approximation algorithms for tree alignment with a given phylogeny, Algorithmica 16 (3) (1996) 302–315.

[54] G. Navarro, A guided tour to approximate string matching, ACM Computing Surveys (CSUR) 33 (1) (2001) 31–88.

[55] G. Navarro, R. Baeza-Yates, E. Sutinen, J. Tarhio, Indexing methods for approximate string matching, Bulletin of the Technical Committee on 19.

[56] A. Andoni, R. Krauthgamer, The smoothed complexity of edit distance, Lecture Notes in Computer Science 5125 (2008) 357–369.

[57] A. Andoni, M. Braverman, A. Hassidim, Phylogenetic reconstruction with insertions and deletions, preprint (2010).

[58] C. Daskalakis, S. Roch, Alignment-free phylogenetic reconstruction, in: RECOMB, 2010, pp. 123–137.

[59] Y. Higuchi, Remarks on the limiting Gibbs states on a $(d+1)$-tree, Publ. Res. Inst. Math. Sci. 13 (2) (1977) 335–348.

```
1. Input. Children sequences $\hat{X}^1, \dots, \hat{X}^d$.
2. Initialization. Set $\hat{s}_i(0) := 0$, $\forall i$, $\ell = k^{1/3}$, $r = 1$, and $t = \ell$.
3. Main loop. While $\hat{X}^i[t + \hat{s}_i(r-1) + 1 : t + \hat{s}_i(r-1) + a]$ is non-empty for all $i$,
   (a) Current position. Set $t = \ell r$.
   (b) Anchor definition. For each $i$, set $\widehat{A}_r^i = \hat{X}^i[t + \hat{s}_i(r-1) + 1 : t + \hat{s}_i(r-1) + a]$. We say that $\widehat{A}_r^i$ is the $r$'th anchor of the $i$'th child. (If any of the remaining sequences is shorter than $a$, redo the previous loop with the entire remaining sequences, that is, use the anchor from the previous loop to align the rest of the sequences.)
   (c) Anchor alignment. For each anchor, we define the set of anchors which agree with it. Formally, $G_r^i = \{j \in [d] : \mathrm{Corr}(\widehat{A}_r^i, \widehat{A}_r^j) \geq \gamma\}$.
   (d) Update. Define the set of aligned children $G_r = \{i : |G_r^i| \geq d - 2\}$.
       i. Aligned anchors. For each $i \in G_r$, set $\hat{s}_i(r) = \hat{s}_i(r-1)$.
       ii. Misaligned anchors. For each $i \notin G_r$ define two strings $\widehat{D}_r^i = \hat{X}^i[t + \hat{s}_i(r-1) : t + \hat{s}_i(r-1) + a - 1]$ and $\widehat{I}_r^i = \hat{X}^i[t + \hat{s}_i(r-1) + 2 : t + \hat{s}_i(r-1) + a + 1]$. If
               $|\{j \in [d] - \{i\} \ : \ \mathrm{Corr}(\widehat{D}_r^i, \widehat{A}_r^j) \geq \gamma\}| \geq d - 2,$
           set $\hat{s}_i(r) = \hat{s}_i(r-1) - 1$. If
               $|\{j \in [d] - \{i\} \ : \ \mathrm{Corr}(\widehat{I}_r^i, \widehat{A}_r^j) \geq \gamma\}| \geq d - 2,$
           set $\hat{s}_i(r) = \hat{s}_i(r-1) + 1$.
   (e) Ancestral sequence. Compute $\hat{X}^0_{t-\ell+1}, \dots \hat{X}^0_t$ by performing a sitewise majority on the children in $G_r$.
   (f) Increment. Set $r := r + 1$.
4. Output. Output $\hat{X}^0$ and set $\hat{k}_0$ to its length.
```

Figure 1: This is the basic recursive step of our reconstruction algorithm. It takes as input the d inferred sequences of the children $\hat{X}^1, \dots, \hat{X}^d$ and computes a sequence for the parent $\hat{X}^0$. If any of the steps above cannot be accomplished, we abort the reconstruction of the parent and declare it radioactive.