Bull Math Biol (2011) 73: 1627–1644
DOI 10.1007/s11538-010-9584-6

ORIGINAL ARTICLE

On the Inference of Large Phylogenies with Long Branches: How Long Is Too Long?

Elchanan Mossel · Sébastien Roch · Allan Sly

Received: 20 January 2010 / Accepted: 7 September 2010 / Published online: 8 October 2010

Abstract The accurate reconstruction of phylogenies from short molecular sequences is an important problem in computational biology. Recent work has highlighted deep connections between sequence-length requirements for high-probability phylogeny reconstruction and the related problem of the estimation of ancestral sequences. In Daskalakis et al. (in Probab. Theory Relat. Fields 2010), building on the work of Mossel (Trans. Am. Math. Soc. 356(6):2379–2404, 2004), a tight sequence-length requirement was obtained for the simple CFN model of substitution, that is, the case of a two-state symmetric rate matrix Q. In particular the required sequence length for high-probability reconstruction was shown to undergo a sharp transition (from $O(\log n)$ to poly(n), where n is the number of leaves) at the "critical" branch length $g_{\mathrm{ML}}(Q)$ (if it exists) of the ancestral reconstruction problem defined roughly as follows: below $g_{\mathrm{ML}}(Q)$ the sequence at the root can be accurately estimated from sequences at the leaves on deep trees, whereas above $g_{\mathrm{ML}}(Q)$ information decays exponentially quickly down the tree.

Here, we consider a more general evolutionary model, the GTR model, where the $q \times q$ rate matrix Q is reversible with $q \geq 2$. For this model, recent results of Roch

E. Mossel supported by NSF Career Award (DMS 054829), by ONR award N00014-07-1-0506, by ISF grant 1300/08 and by Marie Curie grant PIRG04-GA-2008-239317.
S. Roch supported by NSF grant DMS-1007144.

E. Mossel
Weizmann Institute, Rehovot, Israel

E. Mossel
U.C. Berkeley, Berkeley, CA, USA

S. Roch (✉)
Department of Mathematics and Bioinformatics Program, UCLA, Los Angeles, CA, USA
e-mail: roch@math.ucla.edu

A. Sly
Microsoft Research, Redmond, WA, USA

(Preprint, 2009) show that the tree can be accurately reconstructed with sequences of length $O(\log(n))$ when the branch lengths are below $g_{\mathrm{Lin}}(Q)$, known as the Kesten–Stigum (KS) bound, up to which ancestral sequences can be accurately estimated using simple linear estimators. Although for the CFN model $g_{\mathrm{ML}}(Q) = g_{\mathrm{Lin}}(Q)$ (in other words, linear ancestral estimators are in some sense best possible), it is known that for the more general GTR models one has $g_{\mathrm{ML}}(Q) \geq g_{\mathrm{Lin}}(Q)$ with a *strict* inequality in many cases. Here, we show that this phenomenon also holds for phylogenetic reconstruction by exhibiting a family of symmetric models Q and a phylogenetic reconstruction algorithm which recovers the tree from $O(\log n)$-length sequences for some branch lengths in the range $(g_{\mathrm{Lin}}(Q), g_{\mathrm{ML}}(Q))$. Second, we prove that phylogenetic reconstruction under GTR models requires a polynomial sequence-length for branch lengths above $g_{\mathrm{ML}}(Q)$.

1 Introduction

Background Recent years have witnessed a convergence of models and problems from evolutionary biology, statistical physics, and computer science (Mossel and Steel 2005). Standard stochastic models of molecular evolution, such as the Cavender–Farris–Neyman (CFN) model (a.k.a. the Ising model or Binary Symmetric Channel (BSC)) or the Jukes–Cantor (JC) model (a.k.a. the Potts model), have been extensively studied from all these different perspectives and fruitful insights have emerged, notably in the area of computational phylogenetics.

Phylogenetics (Semple and Steel 2003; Felsenstein 2004) is centered around the reconstruction of evolutionary histories from molecular data extracted from modern species. The assumption is that molecular data consists of aligned sequences and that each position in the sequences evolves independently according to a Markov model on a tree, where the key parameters are (see Sect. 2 for formal definitions):

- *Rate matrix.* A $q \times q$ mutation rate matrix Q, where q is the alphabet size. A typical alphabet is the set of nucleotides {A, C, G, T}, but here we allow more general state spaces. Without loss of generality, we denote the alphabet by $[q] = \{1, \ldots, q\}$. The (i, j)'th entry of Q encodes the rate at which state i mutates into state j.
- *Binary tree.* An evolutionary tree T, where the leaves are the modern species and each branching represents a past speciation event. We denote the leaves by $[n] = \{1, \ldots, n\}$.
- *Branch lengths.* For each edge e, we have a scalar branch length $\tau(e)$ which measures the expected total number of substitutions per site along edge e. Roughly speaking, $\tau(e)$ is the time duration between the end points of e multiplied by the mutation rate.

We consider the following two closely related problems:

1. *Phylogenetic Tree Reconstruction (PTR).* Given n molecular sequences of length k (one for each leaf)

$$\left\{s_a = \left(s_a^i\right)_{i=1}^k\right\}_{a\in[n]}$$

with $s_a^i \in [q]$, which have evolved according to the process above with independent sites, reconstruct the topology of the evolutionary tree.

2. *Ancestral State Reconstruction (ASR).* Given a fully specified rooted tree (that is, with known topology and edge lengths) and a single state s_a^1 at each leaf a of the tree, estimate—better than "random"—the state at the root of the tree, independently of the depth of the tree.

In both cases, longer edge lengths correspond to more mutations—and hence more noise—making both reconstruction problems more challenging (Steel and Székely 2002). Our overriding goal is to extend efficient phylogenetic reconstruction to trees with as large branch lengths as possible.

Reconstruction Thresholds Alternatively, the second problem can be interpreted in terms of correlation decay along the tree or as a broadcasting problem on a tree-network. It has thus been extensively studied in statistical physics, probability theory, and computer science. See, e.g., Evans et al. (2000) and references therein. A crucial parameter in the ASR problem is $\tau^+(T) = \max_e \tau(e)$, the maximal branch length in the tree.

One class of ancestral estimators is particularly well understood, the so-called linear estimators. See Sect. 2 for a formal definition. In essence, linear estimators are simply a form of weighted majority. In Mossel and Peres (2003), it was shown that there exists a critical parameter $g_{\mathrm{Lin}}(Q) = \lambda_Q^{-1} \ln \sqrt{2}$, where $-\lambda_Q$ is the largest negative eigenvalue of the rate matrix Q, such that:

- if $\tau^+ < g_{\mathrm{Lin}}(Q)$, for all trees T with $\tau^+(T) = \tau^+$ a *well-chosen* linear estimator provides a good solution to the ASR,
- if $\tau^+ > g_{\mathrm{Lin}}(Q)$, there exist trees T with $\tau^+(T) = \tau^+$ for which ASR is impossible for *any* linear estimator, that is, the correlation between the best linear root estimate and the true root value decays exponentially in the depth of the tree.

For formal definitions, see Mossel and Peres (2003). The threshold $g_{\mathrm{Lin}}(Q) = \lambda_Q^{-1} \ln \sqrt{2}$ is also known to be the critical threshold for *robust (ancestral) reconstruction*, see Janson and Mossel (2004) for details.

For more general ancestral estimators, only partial results are known. For the two-state symmetric Q (the CFN model), impossibility of reconstruction as above holds, when $\tau^+(T) > g_{\mathrm{Lin}}(Q)$, not only for linear estimators but also for *any* estimator, including for instance maximum likelihood. In other words, for the CFN model linear estimators are in some sense best possible. This phenomenon also holds for symmetric models (i.e., where all nondiagonal entries of Q are identical) with $q = 3$ states (Sly 2009) (at least, for high degree trees). However, for symmetric models on $q \geq 5$ states, it is known that ASR is possible beyond $g_{\mathrm{Lin}}(Q)$, up to a critical branch length $g_{\mathrm{ML}}(Q) > g_{\mathrm{Lin}}(Q)$ which is not known explicitly (Mossel 2001; Sly 2009). Larger values of q here correspond for instance to models of protein evolution. ASR beyond $g_{\mathrm{Lin}}(Q)$ can be achieved with a maximum likelihood estimator although in some cases special estimators have been devised (for instance, symmetric models with large q) (Mossel 2001). In this context, $g_{\mathrm{Lin}}(Q)$ is referred to as the *Kesten–Stigum bound* (Kesten and Stigum 1967). We sometimes call the condition $\tau^+(T) < g_{\mathrm{Lin}}(Q)$ the "KS phase" and the condition $\tau^+(T) < g_{\mathrm{ML}}(Q)$ the "reconstruction phase."

For general reversible rate matrices, it is not even known whether there is a *unique* reconstruction threshold $g_{\mathrm{ML}}(Q)$ such that ASR is possible for $\tau^+(T) < g_{\mathrm{ML}}(Q)$ and impossible for $\tau^+(T) > g_{\mathrm{ML}}(Q)$. The general question of finding the threshold $g_{\mathrm{ML}}(Q)$ for ASR is extremely challenging and has been answered for only a very small number of channels.

Steel's Conjecture A striking conjecture of Steel (2001) postulates a deep connection between PTR and ASR. More specifically, the conjecture states that for CFN models if $\tau^+(T) < g_{\mathrm{Lin}}(Q)$ then PTR can be achieved with sequence length $k = O(\log n)$. This says that when we can accurately estimate the states of vertices deep inside a *known* tree, then it is also possible to accurately reconstruct the topology of an *unknown* tree with very short sequence lengths.

In fact, since the number of trees on n labeled leaves is $2^{\Theta(n \log n)}$, this is an optimal sequence length up to constant factors—that is, we cannot hope to distinguish so many trees with fewer potential datasets. The proof of Steel's conjecture was established in Mossel (2004) for balanced trees and in Daskalakis et al. (2010) for general (under the additional assumption that branch lengths are discretized). Furthermore, results of Mossel (2003, 2004) show that for $\tau^+(T) > g_{\mathrm{Lin}}(Q)$ a polynomial sequence length is needed for correct phylogenetic reconstruction. For symmetric models, the results of Mossel (2004), Daskalakis et al. (2010) imply that it is possible to reconstruct phylogenetic trees from sequences of length $O(\log n)$ when $\tau^+(T) < g_{\mathrm{Lin}}(Q)$. These results cover classical models such as the JC model ($q = 4$). Recent results of Roch (2009), building on Roch (2008), Peres and Roch (2009), show that for any reversible mutation matrix Q, it is possible to reconstruct phylogenetic trees from $O(\log(n))$-length sequences again when $\tau^+(T) < g_{\mathrm{Lin}}(Q)$.

However, these results leave the following important problem open:

- As we mentioned before, for symmetric models on $q \geq 5$ states, it is known that ASR is possible for $\tau^+(T) < g_{\mathrm{ML}}(Q)$, where $g_{\mathrm{ML}}(Q) > g_{\mathrm{Lin}}(Q)$. A natural question is to ask if the "threshold" for PTR is $g_{\mathrm{ML}}(Q)$ (i.e., the threshold for ASR) or $g_{\mathrm{Lin}}(Q)$ or perhaps another value. (Note that for the CFN model, the threshold for PTR has been shown to be $g_{\mathrm{Lin}}(Q)$ but in that case it so happens that $g_{\mathrm{Lin}}(Q) = g_{\mathrm{ML}}(Q)$.)

Our Contributions Our main results are the following:

- We show that for symmetric models Q with large q, it is possible to reconstruct phylogenetic trees with $O(\log n)$-length sequences whenever $\tau^+(T) < g_q^+$ where $g_{\mathrm{Lin}}(Q) < g_q^+ < g_{\mathrm{ML}}(Q)$. We thus show that PTR from logarithmic sequences is sometimes possible for branch lengths *above* the KS bound.
- We also show how to generalize the arguments of Mossel (2003, 2004) to show that for any Q and $\tau^+(T) > g_{\mathrm{ML}}(Q)$ it holds that correct phylogenetic reconstruction requires polynomial-length sequences in general. The same idea is used in Mossel (2003, 2004) and the argument presented here. The main difference is that in the arguments in Mossel (2003, 2004) used mutual information together with coupling while the more elegant argument presented here uses coupling only. The results of Mossel (2003) apply for general models but are not tight even for

the CFN model. The argument in Mossel (2004) gives tight results for the CFN model. It is possible to extend that argument to more general models, but we prefer the simpler proof given in the current paper.

Organization We begin with preliminaries and the formal statements of our results in Sect. 2. The proof of our upper bound can be found in Sect. 3. The proof of our lower bound can be found in Sect. 4.

2 Definitions and Results

2.1 Basic Definitions

Phylogenies We define phylogenies and evolutionary distances more formally.

Definition 1 (Phylogeny) A *phylogeny* is a rooted, edge-weighted, leaf-labeled tree $\mathcal{T} = (V, E, [n], \rho; \tau)$ where: V is the set of vertices; E is the set of edges; $L = [n] = \{1, \ldots, n\}$ is the set of leaves; ρ is the root; $\tau : E \to (0, +\infty)$ is a positive edge weight function. We further assume that all internal nodes in $\mathcal{T}$ have degree 3 except for the root ρ which has degree 2. We let $\mathbb{Y}_n$ be the set of all such phylogenies on n leaves and we denote $\mathbb{Y} = \{\mathbb{Y}_n\}_{n \geq 1}$.

Definition 2 (Tree Metric) For two leaves $a, b \in [n]$, we denote by $\mathrm{Path}(a, b)$ the set of edges on the unique path between a and b. A *tree metric* on a set $[n]$ is a positive function $\mathrm{d} : [n] \times [n] \to (0, +\infty)$ such that there exists a tree $T = (V, E)$ with leaf set $[n]$ and an edge weight function $w : E \to (0, +\infty)$ satisfying the following: for all leaves $a, b \in [n]$

$$\mathrm{d}(a, b) = \sum_{e \in \mathrm{Path}(a,b)} w_e.$$

For convenience, we denote by $(\tau(a, b))_{a,b \in [n]}$ the tree metric corresponding to phylogeny $\mathcal{T} = (V, E, [n], \rho; \tau)$. We extend $\tau(u, v)$ to all vertices $u, v \in V$ in the obvious way.

Example 1 (Homogeneous Tree) For an integer $h \geq 0$, we denote by $\mathcal{T}^{(h)} = (V^{(h)}, E^{(h)}, L^{(h)}, \rho^{(h)}; \tau)$ a rooted phylogeny where $T^{(h)}$ is the h-level complete binary tree with arbitrary edge weight function τ and $L^{(h)} = [2^h]$. For $0 \leq h' \leq h$, we let $L^{(h)}_{h'}$ be the vertices on level $h - h'$ (from the root). In particular, $L^{(h)}_0 = L^{(h)}$ and $L^{(h)}_h = \{\rho^{(h)}\}$. We let $\mathbb{HY} = \{\mathbb{HY}_n\}_{n \geq 1}$ be the set of all phylogenies with homogeneous underlying trees.

Model of Molecular Sequence Evolution Phylogenies are reconstructed from molecular sequences extracted from the observed species. The standard model of evolution for such sequences is a Markov model on a tree (MMT).

Definition 3 (Markov Model on a Tree) Let $q \geq 2$. Let $n \geq 1$ and let $T = (V, E, [n], \rho)$ be a rooted tree with leaves labeled in $[n]$. For each edge $e \in E$, we are given a $q \times q$ stochastic matrix $M^e = (M^e_{ij})_{i,j\in[q]}$, with fixed stationary distribution $\pi = (\pi_i)_{i\in[q]}$. An MMT $(\{M^e\}_{e\in E}, T)$ associates a state s_v in $[q]$ to each vertex v in V as follows: pick a state for the root ρ according to π; moving away from the root, choose a state for each vertex v independently according to the distribution $(M^e_{s_u,j})_{j\in[q]}$, with $e = (u, v)$ where u is the parent of v.

The most common MMT used in phylogenetics is the so-called general time-reversible (GTR) model.

Definition 4 (GTR Model) Let $[q]$ be a set of character states with $q = |[q]|$ and π be a distribution on $[q]$ satisfying $\pi_i > 0$ for all $i \in [q]$. For $n \geq 1$, let $\mathcal{T} = (V, E, [n], \rho; \tau)$ be a phylogeny. Let Q be a $q \times q$ rate matrix, that is, $Q_{ij} > 0$ for all $i \neq j$ and $\sum_{j\in[q]} Q_{ij} = 0$, for all $i \in [q]$. Assume Q is reversible with respect to π, that is, $\pi_i Q_{ij} = \pi_j Q_{ji}$, for all $i, j \in [q]$. The GTR model on $\mathcal{T}$ with rate matrix Q is an MMT on $T = (V, E, [n], \rho)$ with transition matrices $M^e = e^{\tau_e Q}$, for all $e \in E$. By the reversibility assumption, Q has q real eigenvalues $0 = \Lambda_1 > \Lambda_2 \geq \cdots \geq \Lambda_q$. We normalize Q by fixing $\Lambda_2 = -1$. We denote by $\mathbb{Q}_q$ the set of all such rate matrices. We let $\mathbb{G}_{n,q} = \mathbb{Y}_n \otimes \mathbb{Q}_q$ be the set of all q-state GTR models on n-leaf trees. We denote $\mathbb{G}_q = \{\mathbb{G}_{n,q}\}_{n\geq 1}$. We denote by s_W the vector of states on the vertices $W \subseteq V$. In particular, $s_{[n]}$ are the states at the leaves. We denote by $\mathcal{L}_{\mathcal{T},Q}$ the distribution of $s_{[n]}$.

GTR models are often used in their full generality in the biology literature, but they also encompass several popular special cases such as the CFN model and the JC model.

Example 2 (q-State Symmetric Model) The *q-state symmetric model* (also called q-state Potts model) is the GTR model with $q \geq 2$ states, $\pi = (1/q, \ldots, 1/q)$, and $Q = Q^{(q)}$ where

$$Q^{(q)}_{ij} = \begin{cases} -\frac{q-1}{q} & \text{if } i = j, \\ \frac{1}{q} & \text{o.w.} \end{cases}$$

It is easy to check that $\Lambda_2(Q) = -1$. The special cases $q = 2$ and $q = 4$ are called respectively the CFN and JC models in the biology literature. We denote their rate matrices by Q^{CFN}, Q^{JC}. For an edge e of length $\tau_e > 0$, let

$$\delta_e = \frac{1}{q}\left(1 - e^{-\tau_e}\right).$$

Then we have

$$(M_e)_{ij} = \left(e^{\tau_e Q}\right)_{ij} = \begin{cases} 1 - (q-1)\delta_e & \text{if } i = j, \\ \delta_e & \text{o.w.} \end{cases}$$

Phylogenetic Reconstruction A standard assumption in molecular evolution is that each site in a sequence (DNA, protein, etc.) evolves *independently* according to a Markov model on a tree, such as the GTR model above. Because of the reversibility assumption, the root of the phylogeny cannot be identified and we reconstruct phylogenies up to their root.

Definition 5 (Phylogenetic Reconstruction Problem) Let $\widetilde{\mathbb{Y}} = \{\widetilde{\mathbb{Y}}_n\}_{n\geq 1}$ be a subset of phylogenies and $\widetilde{\mathbb{Q}}_q$ be a subset of rate matrices on q states. Let $\mathcal{T} = (V, E, [n], \rho; \tau) \in \widetilde{\mathbb{Y}}$. If $T = (V, E, [n], \rho)$ is the rooted tree underlying $\mathcal{T}$, we denote by $T_-[\mathcal{T}]$ the tree T where the root is removed: that is, we replace the two edges adjacent to the root by a single edge. We denote by $\mathbb{T}_n$ the set of all leaf-labeled trees on n leaves with internal degrees 3 and we let $\mathbb{T} = \{\mathbb{T}_n\}_{n\geq 1}$. A *phylogenetic reconstruction algorithm* is a collection of maps $\mathcal{A} = \{\mathcal{A}_{n,k}\}_{n,k\geq 1}$ from sequences $(s^i_{[n]})_{i=1}^k \in ([q]^{[n]})^k$ to leaf-labeled trees $T \in \mathbb{T}_n$. We only consider algorithms $\mathcal{A}$ computable in time polynomial in n and k. Let $k(n)$ be an increasing function of n. We say that $\mathcal{A}$ solves the *phylogenetic reconstruction problem* on $\widetilde{\mathbb{Y}} \otimes \widetilde{\mathbb{Q}}_q$ with sequence length $k = k(n)$ if for all $\delta > 0$, there is $n_0 \geq 1$ such that for all $n \geq n_0$, $\mathcal{T} \in \widetilde{\mathbb{Y}}_n$, $Q \in \widetilde{\mathbb{Q}}_q$,

$$\mathbb{P}\big[\mathcal{A}_{n,k(n)}\big(\big(s^i_{[n]}\big)_{i=1}^{k(n)}\big) = T_-[\mathcal{T}]\big] \geq 1 - \delta,$$

where $(s^i_{[n]})_{i=1}^{k(n)}$ are i.i.d. samples from $\mathcal{L}_{\mathcal{T},Q}$.

An important result of this kind was given by Erdös et al. (1999). Let $\alpha \geq 1$ and $q \geq 2$. The set of rate matrices $Q \in \mathbb{Q}_q$ such that $\mathrm{tr}(Q) \geq -\alpha$ is denoted $\mathbb{Q}_{q,\alpha}$. Let $0 < f < g < +\infty$ and denote by $\mathbb{Y}^{f,g}$ the set of all phylogenies $\mathcal{T} = (V, E, [n], \rho; \tau)$ satisfying $f < \tau_e < g$, $\forall e \in E$. Then Erdos et al. showed (as rephrased in our setup) that, for all $\alpha \geq q - 1$, $q \geq 2$, and all $0 < f < g < +\infty$, the phylogenetic reconstruction problem on $\mathbb{Y}^{f,g} \otimes \mathbb{Q}_{q,\alpha}$ can be solved with $k = \mathrm{poly}(n)$. (In fact, they proved a more general result allowing rate matrices to vary across different edges.) In the case of the Potts model, this result was improved by Daskalakis et al. (2010) (building on (Mossel 2004)) in the Kesten–Stigum (KS) reconstruction phase, that is, when $g < g_{\mathrm{Lin}}(Q) = g^\star_{\mathrm{Lin}} \equiv \ln\sqrt{2}$. They showed that, for all $0 < f < g < g^\star_{\mathrm{Lin}}$, the phylogenetic reconstruction problem on $\mathbb{Y}^{f,g} \otimes \{Q^{(q)}\}$ can be solved with $k = O(\log(n))$. More recently, the latter result was extended to GTR models by Roch (2009), building on Roch (2008), Peres and Roch (2009). But prior to our work, no PTR algorithm had been shown to extend beyond $g^\star_{\mathrm{Lin}}$.

2.2 Our Results

Positive Result In our first result, we extend logarithmic reconstruction results for q-state symmetric models to $\ln\sqrt{2} < g < \ln 2$ for large enough q. This is the first result of this type beyond the KS bound.

Theorem 1 (Logarithmic Reconstruction beyond the KS Transition) *Let* $0 < f < g < +\infty$ *and denote by* $\mathbb{HY}^{f,g}$ *the set of all homogeneous phylogenies* $\mathcal{T} =$

$(V, E, [n], \rho; \tau)$ *satisfying* $f < \tau_e < g$, $\forall e \in E$. *Let* $g^\star_{\text{Perc}} = \ln 2$. *Then, for all* $0 < f < g < g^\star_{\text{Perc}}$, *there is* $R \geq 2$ *such that for all* $q > R$ *the phylogenetic reconstruction problem on* $\mathbb{HY}^{f,g} \otimes \{Q^{(q)}\}$ *can be solved with* $k = O(\log(n))$.

Theorem 1 can be extended to general phylogenies using the techniques of Daskalakis et al. (2010), although then one requires discretized branch lengths. See Daskalakis et al. (2010) for details.

Negative Result In our second result, we show that for $g > g_{\text{ML}}(Q)$ the number of samples k must grow polynomially in n. In particular, this is true for the q-state symmetric model for all $q \geq 2$ and $g > \ln 2$ by the results of Mossel (2001).

Theorem 2 (Polynomial Lower Bound Above $g_{\text{ML}}(Q)$ (see also Mossel 2003, 2004)) *Let* $Q \in \mathbb{Q}_q$ *and* $f = g > g_{\text{ML}}(Q)$. *Then the phylogenetic reconstruction problem on* $\mathbb{HY}^{f,g} \otimes \{Q\}$ *requires* $k = \Omega(n^\alpha)$ *for some* $\alpha > 0$ (*even assuming* Q *and* g *are known exactly beforehand*).

Remark 1 (Biological Convention) Our normalization of Q differs from standard biological convention where it is assumed that the total rate of change per unit time at stationarity is 1, that is,

$$\sum_i \pi_i Q_{ii} = -1.$$

See, e.g., Felsenstein (2004). Let $-\lambda_Q$ denote the largest negative eigenvalue under this convention. Then the Kesten–Stigum bound is given by the solution to

$$2e^{-2\lambda_Q g_{\text{Lin}}(Q)} = 1.$$

In the case of symmetric models with q states, one can check that

$$\lambda_Q = \frac{q}{q-1},$$

and hence,

$$g_{\text{Lin}}(Q) = \frac{q-1}{2q} \ln 2.$$

Here are a few typical values:

$$q = 2 \text{ (CFN model)}: \quad g_{\text{Lin}}(Q) = \frac{1}{4} \ln 2 \approx 0.17,$$

$$q = 4 \text{ (JC model)}: \quad g_{\text{Lin}}(Q) = \frac{3}{8} \ln 2 \approx 0.26,$$

$$q = 16: \quad g_{\text{Lin}}(Q) = \frac{15}{32} \ln 2 \approx 0.32,$$

$$q \to +\infty: \quad g_{\text{Lin}}(Q) \to \frac{1}{2} \ln 2 \approx 0.35.$$

Values for $g_{\mathrm{ML}}(Q)$ are not known in general—except when they coincide with $g_{\mathrm{Lin}}(Q)$. This is known to happen in the symmetric case with $q = 2, 3$ (Bleher et al. 1995; Ioffe 1996; Sly 2009). See also Borgs et al. (2006).

3 Upper Bound for Large q

3.1 Root Estimator

The basic ingredient behind logarithmic reconstruction results is an accurate estimator of the root state. In the KS phase, this can be achieved by majority-type procedures. See Mossel (1998, 2004), Evans et al. (2000). In the reconstruction phase beyond the KS phase, however, a more sophisticated estimator is needed. In this subsection, we define an accurate root estimator which does not depend on the edge lengths.

Random Cluster Methods We use a convenient percolation representation of the ferromagnetic Potts model on trees. Let $q \geq 2$ and $\mathcal{T} = (V, E, [n], \rho; \tau) \in \mathbb{HY}_n$ with corresponding $(\delta_e)_{e\in E}$. Run a percolation process on $T = (V, E)$ where edge e is open with probability $1 - q\delta_e$. Then associate to each open cluster a state according to the uniform distribution on $[q]$. The state so obtained $(s_v)_{v\in V}$ has the same distribution as the GTR model $(\mathcal{T}, Q^{(q)})$.

We will use the following definition. Let T' be a subtree of T which is rooted at ρ. We say that T' is an *l-diluted binary tree* if, for all s, all the vertices of T' at level sl have exactly 2 descendants at level $(s+1)l$. (Assume for now that $\log_2 n$ is a multiple of l.) For a state $i \in [q]$ and assignment $s_{[n]}$ at the leaves, we say that the event $\mathcal{B}_{i,l}$ holds if there is a l-diluted binary tree with state i at all its leaves according to $s_{[n]}$. Let B_l be the set of all i such that $\mathcal{B}_{i,l}$ holds. Consider the following estimator: pick a state X uniformly at random in $[q]$ and let

$$\bar{s}_\rho^l = \begin{cases} X, & \text{if } X \in B_l, \\ \text{pick uniformly in } [q] - \{X\}, & \text{o.w.} \end{cases}$$

We use the following convention. If $\log_2 n$ is not a multiple of l, we add levels of 0-length edges to $\mathcal{T}$ so as to make the total number of levels be a multiple of l and we copy the states at the leaves of $\mathcal{T}$ to all their descendants in the new tree. We then apply the estimator as above.

Error Channel We show next that $\bar{s}_\rho = \bar{s}_\rho^l$ is a good estimator of the root state under the conditions of Theorem 1. Let

$$\overline{M}^{\rho,l} = \left(\mathbb{P}[\bar{s}_\rho = j \mid s_\rho = i]\right)_{i,j\in[q]}.$$

Proposition 1 shows that this "error channel" is of the Potts type with bounded length, no matter how deep the tree. The key behind our reconstruction algorithm in the next section will be to think of this error channel as an "extra edge" in the Markov model.

Proposition 1 (Root Estimator from Diluted Trees) *Let* $g^\star_{\mathrm{Perc}} = \ln 2$. *Then, for all* $0 < g < g^\star_{\mathrm{Perc}}$, *we can find* $l > 0$, $R \geq 2$ *and* $0 < \bar{\mathrm{b}} < +\infty$ *such that*

$$\overline{M}^{\rho,l} = e^{\mathrm{b}_\rho Q},$$

where $\mathrm{b}_\rho \leq \bar{\mathrm{b}}$ *and* $Q = Q^{(q)}$, *for all* $q > R$ *and all* $\mathcal{T} \in \mathbb{HY}^{0,g}$.

Proof The proof is based on a random cluster argument of Mossel (2001). Fix $0 < f < g < g^\star_{\mathrm{Perc}}$. In Mossel (2001), it is shown that one can choose $\varepsilon > 0$ small enough and l, R large enough such that

$$\mathbb{P}[\mathcal{B}_{i,l} \,|\, s_\rho = i] \geq \varepsilon, \tag{1}$$

and

$$\mathbb{P}[\mathcal{B}_{i,l} \,|\, s_\rho \neq i] \leq \varepsilon/2, \tag{2}$$

for all $q > R$ and all $\mathcal{T} = (V, E, [n], \rho; \tau) \in \mathbb{HY}^{0,g}$. The proof in Mossel (2001) actually assumes that all τ_e's are equal to g. However, the argument still holds when $\tau_e \leq g$ for all e since smaller τ's imply smaller δ's which can only strengthen inequalities (1) and (2) by a standard domination argument. (For (2), see the original argument in Mossel (2001).)

Therefore, we have

$$\begin{aligned}
\overline{M}^{\rho,l}_{ii} &= \mathbb{P}[i \in B_l \,|\, s_\rho = i]\mathbb{P}[X = i] + \frac{1}{q-1}\mathbb{P}[X \notin B_l \,|\, s_\rho = i, X \neq i]\mathbb{P}[X \neq i] \\
&\geq \varepsilon\left(\frac{1}{q}\right) + \frac{1}{q-1}(1 - \varepsilon/2)\left(\frac{q-1}{q}\right) \\
&= \frac{1}{q} + \frac{\varepsilon}{2q}.
\end{aligned}$$

Also, by symmetry, we have for $i \neq j$

$$\begin{aligned}
\overline{M}^{\rho,l}_{ij} &= \frac{1}{q-1}\left(1 - \overline{M}^{\rho,l}_{ii}\right) \\
&\leq \frac{1}{q} - \frac{\varepsilon}{2q(q-1)}.
\end{aligned}$$

Hence, the channel $\overline{M}^{\rho,l}$ is of the form $e^{\mathrm{b}_\rho Q}$ with $\mathrm{b}_\rho \leq \bar{\mathrm{b}}$ where, by the relation between δ and τ given in Example 2, we can take

$$\begin{aligned}
\bar{\mathrm{b}} &= -\ln\left(1 - q\left(\frac{1}{q} - \frac{\varepsilon}{2q(q-1)}\right)\right) \\
&= -\ln\left(\frac{\varepsilon}{2(q-1)}\right).
\end{aligned}$$

This concludes the proof. □

3.2 Reconstruction Algorithm

Our reconstruction algorithm is based on standard distance-based quartet techniques. Let $\mathcal{T} = (V, E, [n], \rho; \tau) \in \mathbb{HY}^{f,g}$ be a homogeneous phylogeny that we seek to reconstruct from k samples of the corresponding Potts model at the leaves $(s^i_{[n]})_{i=1}^k \in ([q]^{[n]})^k$.

Distances For two nodes $u, v \in V$, we may relate their distance to the probability that their states agree

$$\tau(u, v) = \sum_{e \in \mathrm{Path}(u,v)} \tau_e = -\ln\left(1 - \left(\frac{q}{q-1}\right)\mathbb{P}[s_u \neq s_v]\right),$$

and so a natural way to estimate $\tau(u, v)$ is to consider the estimator

$$\hat{\tau}(u, v) = -\ln\left(1 - \left(\frac{q}{q-1}\right)\frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\{s^i_u \neq s^i_v\}\right).$$

Of course, given samples at the leaves, this estimator can only be used for $u, v \in [n]$. Instead, when u, v are internal nodes we first reconstruct their sequence using Proposition 1. We will then over-estimate the true distance by an amount not exceeding $2\bar{\mathrm{b}}$ on average. For $u, v \in V - [n]$, let

$$\tau_{\mathrm{b}}(u, v) = \tau(u, v) + \mathrm{b}_u + \mathrm{b}_v,$$

using the notation of Proposition 1. We also let $\{\bar{s}^i_u\}_{i=1}^k, \{\bar{s}^i_v\}_{i=1}^k$ be the reconstructed states at u, v. By convention, we let

$$\tau_{\mathrm{b}}(a, b) = \tau(a, b),$$

and

$$\bar{s}^i_a = s^i_a, \quad \forall i = 1, \ldots, k,$$

for $a, b \in [n]$. Note that, at the beginning of the algorithm, the phylogeny is not known, making it impossible to compute $\{\bar{s}^i_u\}_{i=1}^k$ for internal nodes. However, as we reconstruct parts of the tree, we will progressively compute the estimated sequences of uncovered internal nodes.

By standard concentration inequalities, $\tau_{\mathrm{b}}(u, v)$ can be well approximated with $k = O(\log n)$ as long as $\tau_{\mathrm{b}}(u, v) = O(1)$. For $u, v \in V$ let

$$\hat{\tau}(u, v) = -\ln\left(1 - \left(\frac{q}{q-1}\right)\frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\{\bar{s}^i_u \neq \bar{s}^i_v\}\right).$$

Recall the notation of Example 1.

Lemma 1 (Distorted Metric: Short Distances (Erdös et al. 1999)) *Let $0 \leq h' < h$ and let $u, v \in L^{(h)}_{h'}$ be distinct leaves. For all $D > 0$, $\delta > 0$, $\gamma > 0$, there exists $c = c(D, \delta, \gamma) > 0$, such that if the following conditions hold*:

- [Small Diameter] $\tau_{\mathrm{b}}(u,v) < D$,
- [Sequence Length] $k = c' \log n$ *for* $c' > c$,

then

$$\left|\tau_{\mathrm{b}}(u,v) - \hat{\tau}(u,v)\right| < \delta,$$

with probability at least $1 - n^{-\gamma}$.

Lemma 2 (Distorted Metric: Diameter Test (Erdös et al. 1999)) *Let* $0 \le h' < h$ *and* $u, v \in L_{h'}^{(h)}$. *For all* $D > 0$, $W > 5$, $\gamma > 0$, *there exists* $c = c(D, W, \gamma) > 0$, *such that if the following conditions hold*:

- [Large Diameter] $\tau_{\mathrm{b}}(u,v) > D + \ln W$,
- [Sequence Length] $k = c' \log n$ *for* $c' > c$,

then

$$\hat{\tau}(u,v) > D + \ln \frac{W}{2},$$

with probability at least $1 - n^{-\gamma}$. *On the other hand, if the first condition above is replaced by*

- [Small Diameter] $\tau_{\mathrm{b}}(u,v) < D + \ln \frac{W}{5}$,

then

$$\hat{\tau}(u,v) \le D + \ln \frac{W}{4},$$

with probability at least $1 - n^{-\gamma}$.

Quartet Tests Let $0 \le h' < h$ and $\mathcal{Q}_0 = \{a_0, b_0, c_0, d_0\} \subseteq L_{h'}^{(h)}$. The topology of $T^{(h)}$ restricted to $\mathcal{Q}_0$ is completely characterized by a bipartition or *quartet split* q_0 of the form: $a_0b_0|c_0d_0$, $a_0c_0|b_0d_0$ or $a_0d_0|b_0c_0$. In words $a_0b_0|c_0d_0$ indicates that it is possible, by removing an appropriate edge, to split the tree into two subtrees with a_0 and b_0 on one side and c_0 and d_0 on the other. The most basic operation in quartet-based reconstruction algorithms is the inference of such quartet splits. In distance-based methods in particular, this is usually done by performing the so-called *four-point test*: letting

$$\mathcal{F}(a_0b_0|c_0d_0) = \frac{1}{2}\big[\tau(a_0,c_0) + \tau(b_0,d_0) - \tau(a_0,b_0) - \tau(c_0,d_0)\big],$$

we have

$$q_0 = \begin{cases} a_0b_0|c_0d_0 & \text{if } \mathcal{F}(a_0,b_0|c_0,d_0) > 0, \\ a_0c_0|b_0d_0 & \text{if } \mathcal{F}(a_0,b_0|c_0,d_0) < 0, \\ a_0d_0|b_0c_0 & \text{o.w.} \end{cases}$$

Note that adding "extra edges" at the nodes a_0, b_0, c_0, d_0 as implied in Proposition 1 does not affect the topology of the quartet.

Since Lemma 1 applies only to short distances, we also perform a diameter test. We let $\widehat{\mathcal{F}}(a_0b_0|c_0d_0) = -\infty$ if $\max_{u,v\in\mathcal{Q}_0} \hat{\tau}(u,v) > D + \ln\frac{W}{4}$ and otherwise

$$\widehat{\mathcal{F}}(a_0b_0|c_0d_0) = \frac{1}{2}\big[\hat{\tau}(a_0,c_0) + \hat{\tau}(b_0,d_0) - \hat{\tau}(a_0,b_0) - \hat{\tau}(c_0,d_0)\big].$$

Finally, we let

$$\overline{\mathbb{FP}}(a_0,b_0|c_0,d_0) = \mathbb{1}\big\{\widehat{\mathcal{F}}(a_0b_0|c_0d_0) > f/2\big\}.$$

Proof of Theorem 1 The algorithm is summarized in Fig. 1. The basis of the algorithm was also used in Mossel (2004) in the setting of the two-state symmetric channel where a simpler ancestral reconstruction algorithm (that is, recursive majority) could be used. The key idea is to recursively reconstruct the tree *one layer at a time* from the leaves up to the root. In the first step, this involves pairing leaves according to cherries, that is, pairs of leaves with a common ancestor. A similar procedure is applied to each layer of the tree consecutively.

Given g, the upper bound on the edge lengths, choose l, R, and $\bar{\mathrm{b}}$ such that Proposition 1 holds. Now take c' large enough so that, using $k = c'\log n$ samples and setting $D = 10g + 2\bar{\mathrm{b}}$, $W = 10$, $\delta = \frac{f}{20}$, and $\gamma = 10$, we have that Lemmas 1 and 2 hold.

We begin with the first level of the tree. We apply the quartet test to every 4-tuple of leaves. By a union bound over all such 4-tuples it follows that with high probability

- If $\max_{u,v\in\mathcal{Q}_0} \tau(u,v) \le D$ and the quartet splits as $a_0, b_0|c_0, d_0$ then $\overline{\mathbb{FP}}(a_0, b_0|c_0, d_0) = 1$.
- If $a_0, b_0|c_0, d_0$ is not displayed by the tree then $\overline{\mathbb{FP}}(a_0, b_0|c_0, d_0) = 0$.

In other words, the algorithm correctly identifies the splits of all quartets with small enough diameter and does not identify any false splits. Following the procedure in Fig. 1, we may then pair up vertices which never appear on opposite sides of a split, thus identifying all cherries. We thus accurately reconstruct the first level of the tree. Having correctly identified the bottom level ($h' = 0$) of the tree the algorithm now repeats the procedure to iteratively reconstruct the remainder of the tree from layer $h' = 1$ to $h' = h - 1$. More precisely, since the first h' levels are known correctly by induction, we may treat the internal vertices at level h' as being leaves of a shortened tree of depth $h - h'$. The key difference of course is that *we are not given the sequences for these internal vertices* but instead have to estimate them. Here lies the importance of our ancestral sequence reconstruction algorithm from Proposition 1, which we apply to the reconstructed subtrees below each vertex on level h'. The errors of these estimators are independent and depend only on the subtrees below the vertex. It follows that they can be treated as coming from an adjusted Markov process on the shortened tree, where the edges from the vertices in level $h' + 1$ to level h' are extended according to the error channels which by Proposition 1 are at most $\bar{\mathrm{b}}$. With the estimated sequences on the vertices lying on level h', we may proceed as we did with the leaves by estimating the quartets and determining which vertices form cherries, thus reconstructing the next level. A global union bound (i.e., over the success of all events described above) ensures the success of the algorithm with high probability. This concludes the proof of Theorem 1.

Algorithm
Input: Sequences $(s^i_{[n]})_{i=1}^k \in ([q]^{[n]})^k$;
Output: Estimated tree $\hat{T}$;

- Initialize the output $\hat{T}$ to the set of isolated leaves.
- Let $\mathcal{Z}_0$ be the set of leaves.
- For $h' = 0, \ldots, h-1$,
  1. **Four-Point Test.** Let
  $$\mathcal{R}_{h'} = \{q = ab|cd \ : \ \forall a, b, c, d \in \mathcal{Z}_{h'} \text{ distinct such that } \overline{\mathbb{FP}}(q) = 1\}.$$
  2. **Cherries.** Identify the cherries in $\mathcal{R}_{h'}$, that is, those pairs of vertices that only appear on the same side of the quartet splits in $\mathcal{R}_{h'}$. Let
  $$\mathcal{Z}_{h'+1} = \{a_1^{(h'+1)}, \ldots, a_{2^{h-(h'+1)}}^{(h'+1)}\},$$
  be the parents of the cherries in $\mathcal{Z}_{h'}$.
  3. **Growing the Tree.** Add the cherries identified in the previous step to $\hat{T}$.
  4. **Reconstructed Sequences.** For all $u \in \mathcal{Z}_{h'+1}$, compute $(\bar{s}^i_u)_{i=1}^k$.
- Output $\hat{T}$.

Fig. 1 Algorithm

4 General Lower Bound

Here, we prove the following statement which implies Theorem 2:

Theorem 3 (Polynomial Lower Bound on PTR) *Consider the phylogenetic reconstruction problem for homogeneous trees with fixed edge length $\tau(e) = \tau > 0$ for all edges $e \in E$. Assume further that the ASR problem for edge length τ and matrix Q is not solvable and that moreover $\tau > g^\star_{\text{Lin}}$. Then there exists $\alpha = \alpha(\tau) > 0$ such that the probability of correctly reconstructing the tree is at most $O(n^{-\alpha})$ assuming $k \leq n^\alpha$.*

For general mutation rates Q, it is not known if there is a *unique* reconstruction threshold $g_{\text{ML}}(Q)$ such that ASR is possible for $\tau < g_{\text{ML}}(Q)$ and impossible for $\tau > g_{\text{ML}}(Q)$. For models for which such a threshold exists, Theorem 3 above shows the impossibility of phylogenetic reconstruction for $\tau > g_{\text{ML}}(Q)$. The existence of the threshold $g_{\text{ML}}(Q)$ has been established for a few models, e.g., for so-called random cluster models, which include the binary asymmetric channel and the Potts model (Mossel 2001).

The proof of Theorem 3 is based on the following two lemmas. It is useful to write $n = 2^\ell$ for the number of leaves of a homogeneous tree with ℓ levels.

Lemma 3 (Reconstructing a Deep Subtree) *Consider the PTR problem for homogeneous trees with fixed edge length τ. Let $\mu_Q^{\ell,i}$ denote the distribution at the leaves on a homogeneous ℓ-level tree with fixed edge length τ, root value i, and rate matrix Q. Suppose there exists a number $0 < \alpha < 1$ such that for every ℓ and all i one*

can write $\mu_Q^{\ell,i} = (1-\varepsilon)\bar{\mu} + \varepsilon\mu'^i$ *for some probability measures* $\mu'^i, i \in [q]$, $\bar{\mu}$, *and* $\varepsilon = O(2^{-\alpha\ell})$. *Then the probability of correctly reconstructing homogeneous phylogenetic trees with edge length* τ *assuming* $k \leq n^{\alpha/10}$ *is at most* $O(n^{-\alpha/2})$.

Lemma 4 (Leaf Distribution Decomposition) *Consider the ASR problem for homogeneous trees with fixed edge length* τ. *Assume further that the ASR problem for* Q *with edge length* τ *is not solvable and further* $\tau > g^\star_{\mathrm{Lin}}$. *Then there is an* $\alpha = \alpha(\tau) > 0$ *for which the following holds. There exists a sequence* $\varepsilon_\ell = O(2^{-\alpha\ell})$ *such that for all* $i \in [q]$ *one can write* $\mu_Q^{\ell,i} = (1-\varepsilon)\bar{\mu} + \varepsilon\mu'^i$ *for some probability measures* $\mu'^i, i \in [q]$ *and* $\bar{\mu}$.

Proof of Lemma 3 Let r be chosen so that $2^{r-1} < n^{\alpha/20} \leq 2^r$. (Note that $r < \ell$.) Consider the following distribution: first, pick a homogeneous tree T on ℓ levels, where the first r levels are chosen uniformly at random among r-level homogeneous trees and the remaining levels are fixed (i.e., deterministic); second, pick k samples of a Markov model with rate matrix Q and fixed edge length τ on the resulting tree.

Let $\mathcal{A}$ be a phylogenetic reconstruction algorithm. Our goal is to bound the success probability of $\mathcal{A}$ on the random model above. We may assume that the bottom $\ell - r$ levels are given to $\mathcal{A}$ (as it may ignore this information) and that $\mathcal{A}$ is deterministic (as a simple convexity argument shows that deterministic algorithms achieve the highest success probability).

Note that the assumption of the lemma implies that, for a single sample, we can *simultaneously* couple the distribution at the leaves of all the given subtrees of $\ell - r$ levels—except with probability $O(2^r 2^{-\alpha(\ell-r)}) = O(n^{-9\alpha/10})$. This can be achieved by starting the coupling at level r (from the root) of the tree. Repeating this for the $n^{\alpha/10}$ samples we obtain the following. Let μ_T denote the measure on the $n^{\alpha/10}$ samples at leaves of T. Then there exists measures μ, μ'_T and $\varepsilon = O(n^{-8\alpha/10})$ such that $\mu_T = (1-\varepsilon)\mu + \varepsilon\mu'_T$.

Write N_r for the number of leaf-labelled complete binary trees on r levels. Write $\mathcal{E}(s, \mathcal{A}, T)$ for the indicator of the event that the k samples are given by s and that $\mathcal{A}$ recovers T. The success probability of $\mathcal{A}$ is then given by

$$\sum_T N_r^{-1}\Big(\sum_s \mu_T\big(\mathcal{E}(s,\mathcal{A},T)\big)\Big) = (1-\varepsilon)N_r^{-1}\sum_s\sum_T \mu\big(\mathcal{E}(s,\mathcal{A},T)\big) + \varepsilon N_r^{-1}\sum_T\sum_s \mu'_T\big(\mathcal{E}(s,\mathcal{A},T)\big). \quad (3)$$

For the second term note that

$$\sum_s \mu'_T\big(\mathcal{E}(s,\mathcal{A},T)\big) \leq \sum_s \mu'_T(s) = 1,$$

and, therefore, the second term in (3) is bounded by ε. Furthermore, for each s, $\sum_T \mu(\mathcal{E}(s,\mathcal{A},T)) = \mu(s)$ by definition and $\sum_s \mu(s) = 1$ so the first term in (3) is bounded by $(1-\varepsilon)N_r^{-1}$.

Thus overall, the bound on the probability of correct reconstruction is $\varepsilon + (1-\varepsilon)N_r^{-1}$. Using the facts that $N_r = \Omega(2^{2^r}) = \Omega(2^{n^{0.1\alpha}}) = \Omega(n^{\alpha/2})$ and $\varepsilon = O(n^{-8\alpha/10})$ concludes the proof. □

Proof of Lemma 4 For $\delta > 0$ and $r' > 0$, let $\mu_Q^{\ell-r',i}(\delta)$ be the same measure as $\mu_Q^{\ell-r',i}$, except that, for each leaf, independently with probability $1-\delta$, the state at the leaf is replaced by $*$ (which does not belong to the original alphabet). The key to the proof is the main result of Janson and Mossel (2004) where it is shown that if $\tau > g_{\text{Lin}}^{\star}$ then the following holds: There exist fixed $\delta > 0, \alpha > 0$ such that

$$\mu_Q^{\ell-r',i}(\delta) = (1-\varepsilon)\bar{\mu}(\delta) + \varepsilon\mu'^{i}(\delta), \tag{4}$$

where $\varepsilon = O(2^{-\alpha(\ell-r')})$ for some probability measures $\mu'^{i}(\delta)$ and $\bar{\mu}(\delta)$.

The fact that there is no reconstruction (ASR) at edge length τ implies that there exists a fixed r' and measures $\bar{\nu}$ and ν'^{i} such that

$$\mu_Q^{r',i} = (1-\delta)\bar{\nu} + \delta\nu'^{i}.$$

This implies in particular that we can simulate the mutation process on an ℓ-level tree by first using the measure $\mu_Q^{\ell-r',i}(\delta)$ and then applying the following rule: for each node v at level $\ell - r'$ independently

- If the label at v is $*$ then generate the leaf states on the subtree rooted at v according to the measure $\bar{\nu}$.
- Else if it is labeled by i, sample leaf states on the subtree below v from the measure ν'^{i}.

The desired property of the measures $\mu_Q^{\ell,i}$ now follows from the fact that the measures $\mu_Q^{\ell-r',i}(\delta)$ have the desired property by (4). □

Proof of Theorem 3 Lemma 4 guarantees the existence of measures $\mu'^{i}, i \in [q]$, and $\bar{\mu}$ with the desired property. By Lemma 3, the existence of such measures immediately implies the required bound on the probability of reconstruction. This concludes the proof. □

5 Concluding Remarks

The ultimate aim of the line of work discussed in this paper is to establish the following central conjecture—which we call Steel's Program.

Conjecture 1 (Steel's Program) *Whenever the Ancestral Sequence Reconstruction problem is solvable, Phylogenetic Tree Reconstruction can be achieved with sequences of logarithmic length.*

A key feature of our reconstruction algorithm is that the ancestral reconstruction procedure *does not depend on edge lengths*. It is thus robust to uncertainty in the

knowledge of the edge lengths which, in practice, can only be estimated to some precision. Achieving this robustness is a major challenge. In particular, a significant simplification arises when the edge lengths are assumed to be discretized. In this setting, it is possible to recursively estimate the edge lengths *exactly*, as was shown by Daskalakis et al. in the case of the CFN model (Daskalakis et al. 2010). We conjecture that the results of Daskalakis et al. can be extended to more general error channels by using a maximum likelihood ancestral estimator, although the analysis of such an estimator may be somewhat complex. This would establish Steel's Program in the discretized setting. An important direction for future research is to remove the above assumption of discretized edge lengths. To this end, a better understanding is needed of how likelihood-based estimators are affected by uncertainty in the edge lengths.

Acknowledgements We thank the referees for useful comments.

References

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

Borgs, C., Chayes, J. T., Mossel, E., & Roch, S. (2006). The Kesten–Stigum reconstruction bound is tight for roughly symmetric binary channels. In *FOCS* (pp. 518–530).

Daskalakis, C., Mossel, E., & Roch, S. (2010). Evolutionary trees and the Ising model on the Bethe lattice: a proof of Steel's conjecture. Probab. Theory Relat. Fields. doi:10.1007/s00440-009-0246-2

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

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

Felsenstein, J. (2004). *Inferring phylogenies*. Sunderland: Sinauer.

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

Janson, S., & Mossel, E. (2004). Robust reconstruction on trees is determined by the second eigenvalue. *Ann. Probab.*, *32*, 2630–2649.

Kesten, H., & Stigum, B. P. (1967). Limit theorems for decomposable multi-dimensional Galton–Watson processes. *J. Math. Anal. Appl.*, *17*, 309–338.

Mossel, E. (1998). Recursive reconstruction on periodic trees. *Random Struct. Algorithms*, *13*(1), 81–97.

Mossel, E. (2001). Reconstruction on trees: beating the second eigenvalue. *Ann. Appl. Probab.*, *11*(1), 285–300.

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

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

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

Mossel, E., & Steel, M. (2005). How much can evolved characters tell us about the tree that generated them? In O. Gascuel (Ed.), *Mathematics of evolution and phylogeny* (pp. 384–412). Oxford: Oxford University Press.

Peres, Y., & Roch, S. (2009). *Reconstruction on trees: Exponential moment bounds for linear estimators*. Preprint.

Roch, S. (2008). Sequence-length requirement for distance-based phylogeny reconstruction: Breaking the polynomial barrier. In *FOCS* (pp. 729–738).

Roch, S. (2009). *Phase transition in distance-based phylogeny reconstruction*. doi:10.1126/science.1182300.

Semple, C., & Steel, M. (2003). *Mathematics and its applications series: Vol. 22*. *Phylogenetics*. Oxford: Oxford University Press.

Sly, A. (2009). Reconstruction for the Potts model. In M. Mitzenmacher (Ed.), *STOC* (pp. 581–590). New York: ACM.

Steel, M. (2001). *My favourite conjecture*. Preprint.

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