Limitations of Markov Chain Monte Carlo Algorithms for Bayesian Inference of Phylogeny

Elchanan Mossel* Eric Vigoda †

July 5, 2005

Abstract

Markov Chain Monte Carlo algorithms play a key role in the Bayesian approach to phylogenetic inference. In this paper, we present the first theoretical work analyzing the rate of convergence of several Markov Chains widely used in phylogenetic inference. We analyze simple, realistic examples where these Markov chains fail to converge quickly. In particular, the studied data is generated from a pair of trees, under a standard evolutionary model. We prove that many of the popular Markov chains take exponentially long to reach their stationary distribution. Our construction is pertinent since it is well known that phylogenetic trees for genes may differ within a single organism. Our results are a cautionary light for phylogenetic analysis using Bayesian inference, and highlight future directions for potential theoretical work.

1 Introduction

Bayesian inference of phylogeny has had a significant impact in Evolutionary Biology, see e.g. [17]. Some of the popular algorithms for Bayesian inference include MrBayes [16], BAMBE [25], and PAML [22, 28]. These algorithms are cited in more than 2000 scientific publications according to scholar.google.com. All of these algorithms rely on Markov Chain Monte Carlo Methods to sample from the posterior probability of a tree given the data. In particular, they design a Markov chain whose stationary distribution is the desired posterior distribution, computed using the likelihood and the priors. Hence, the running time of the algorithm depends on the convergence rate of the Markov chain to its stationary distribution.

Therefore, reliable phylogenetic estimates depend on the Markov chains reaching their stationary distribution before the phylogeny is inferred. A variety of schemes (such as multiple starting points [15]), and increasingly sophisticated algorithms (such as Metropolis Coupled

*Department of Statistics, University of California at Berkeley, Berkeley, CA 94720. mossel@stat.berkeley.edu.
Supported by a Miller fellowship in Computer Science and Statistics and by a Sloan Fellowship in Mathematics
†College of Computing, Georgia Institute of Technology, Atlanta, GA 30332. vigoda@cc.gatech.edu. Supported by NSF Grant CCR-0237834.

Markov Chain Monte Carlo in MrBayes [16]) are heuristically used to ensure the chains converge quickly to their stationary distribution. However, prior to this work there was no theoretical understanding of when the Markov chains converge quickly or slowly. Thus, here we answer a crucial need for theoretical work to guide the multitude of phylogenetic studies using Bayesian inference.

We consider a setting where the data is generated at random, under a standard evolutionary model, from the mixture of two tree topologies. Such a setting is extremely relevant to real-life data sets. A simple example is molecular data consisting of DNA sequences for more than one gene. It is well known that phylogenetic trees can vary between genes (see [14] for an introduction).

A poignant example of varying gene trees is from the study of the phylogeny of humans, chimpanzees, and gorillas. It is now widely believed that humans and chimpanzees are each others closest extant relative (thus, humans and chimpanzees form what is known as a clade), but there are genes (such as the involucrin gene [9]) which favor the chimpanzee-gorilla clade, and other genes (such as the Y-linked RPS4Y locus [24]) which favor the human-gorilla clade. Even for molecular data from just one gene, it is conceivable that different regions within this single gene have different phylogenies (perhaps via intragenic recombination).

We prove that in the above setting, many of the popular Markov chains take extremely long to reach the stationary distribution. In particular, the convergence time is exponentially long in the number of characters of the data set (a character is a sample from the distribution on the pair of trees). This appears to be the first theoretical work analyzing the convergence rates of Markov chains for Bayesian inference. Previously, Diaconis and Holmes [7] analyzed a Markov chain whose stationary distribution is uniformly distributed over all trees, which corresponds to the case with no data.

Our work lays a cautionary tale for Bayesian inference of phylogenies, and suggests that if the data contains more than one phylogeny than great caution should be used before reporting the results from Bayesian inference of the phylogeny. Our results clearly identify further theoretical work that would be of great interest. We discuss possible directions in Section 3.

The complicated geometry of "tree space" poses highly non-trivial difficulties in analyzing maximum likelihood methods on phylogenetic trees, even for constant size trees.

Initial attempts in studying tree-space includes work by Chor et al [3], which constructs several examples where multiple local maxima for likelihood occur. Their examples use non-random data sets (i.e., not generated from any model) on a four species taxa, and the multiple optima occur on a specific tree topology, differing only in the branch lengths.

A different line of work beginning at Yang [27] analytically determines the maximum likelihood over rooted trees on three species and binary characters. Since then, some sophisticated tools from Algebraic Geometry have been used to study the likelihood function and other polynomials on tree space, see e.g. [8, 26]. It seems like the main result on tree spaces needed in this paper does not follow directly from the Algebraic Geometry methodology.

The algebraic approach seems very powerful at first glance. However, even analyzing the likelihood functions on trees on 5 leaves, requires the solution of an optimization problem of a

rational function with 7 variables on a simplex with a large number of low dimensional facets. Thus, this seems like a computationally intractable problem.

Our approach uses an asymptotic expansion where different terms of the expansion correspond to simple combinatorial quantities in terms of the underlying trees. This allows the analysis of the complicated likelihood function on tree space.

1.1 Definitions

We present the formal definitions of the various notions, and then precisely state our results.

Let Ω denote the set of all phylogenetic trees for n species. Combinatorially, Ω is the set of (unrooted) trees $T = (V, E)$ of internal degree 3 and n leaves.

The likelihood of a data set for a tree is defined as the probability the tree generates the data set, under a chosen evolutionary model. For simplicity we discuss our results for one of the simplest evolutionary models, known as the Cavender-Farris-Neyman (CFN) model [2, 11, 21], which uses a binary alphabet. Our results extends to the Jukes-Cantor model with a 4 state alphabet and many other mutation models.

For a tree $T \in \Omega$, let V_{ext} denote the leaves, V_{int} denote the internal vertices, E denote the edge set, and $\overrightarrow{p} : E \to [0, 1/2]$ denote the edge probabilities. The data is a collection of binary assignments to the leaves. Under the CFN model, the probability of an assignment $D : V_{ext} \to \{0, 1\}$ is

$$\mathbf{Pr}\left(D \mid T, p\right) = \sum_{\substack{D' \in \{0,1\}^V : \\ D'(V_{ext}) = D(V_{ext})}} \prod_{\substack{e=(u,v) \in E(T): \\ D'(u) = D'(v)}} (1 - \overrightarrow{p}(e)) \prod_{\substack{e=(u,v) \in E(T): \\ D'(u) \neq D'(v)}} \overrightarrow{p}(e).$$

We will further assume below that the distribution at any nodes of the tree is given by the uniform distribution on $\{0, 1\}$.

Note, that when $\overrightarrow{p}(e)$ close to zero the endpoints are likely to receive the same assignment, whereas, when $\overrightarrow{p}(e)$ is close to $1/2$ the endpoints are likely to receive independently random assignments. Under the "molecular clock" assumption, edge e has length proportional to $-\log_2(1 - 2\overrightarrow{p}(e))$.

An algorithmic way of generating a character D for a tree T with weights $\overrightarrow{p}$, is to first generate a uniformly random assignment for an arbitrary vertex v. Then, beginning at v, for each edge $e = (v, w)$, given the assignment to one of the endpoints, the other endpoint receives the same assignment with probability $1 - \overrightarrow{p}(e)$ and a different assignment with probability $\overrightarrow{p}(e)$.

Finally, for a collection of data $\overrightarrow{D} = (D_1, \ldots, D_N)$:

$$\begin{aligned} \mathbf{Pr}\left(\overrightarrow{D} \mid T, \overrightarrow{p}\right) &= \prod_{D \in \overrightarrow{D}} \mathbf{Pr}\left(D \mid T, \overrightarrow{p}\right) \\ &= \exp\left(\sum_{D \in \overrightarrow{D}} \log\left(\mathbf{Pr}\left(D \mid T, \overrightarrow{p}\right)\right)\right) \end{aligned}$$

Now, applying Bayes law, we can write the posterior probability of a tree given the data:

$$\begin{aligned}\mathbf{Pr}\left(T \mid \overrightarrow{D}\right) &= \frac{\int_p \mathbf{Pr}\left(\overrightarrow{D} \mid T,p\right)\Psi(T,p)dp}{\mathbf{Pr}\left(\overrightarrow{D}\right)} \\ &= \frac{\int_p \mathbf{Pr}\left(\overrightarrow{D} \mid T,p\right)\Psi(T,p)dp}{\sum_{T'}\int_p \mathbf{Pr}\left(\overrightarrow{D} \mid T,p\right)\Psi(T',p)dp}\end{aligned}$$

where $\Psi(T,p)$ is the prior density on the space of trees so that

$$\sum_T \int_{\overrightarrow{p}} \Psi(T,\overrightarrow{p})dp = 1.$$

Since the denominator is difficult to compute, Markov chain Monte Carlo is used to sample from the above distribution. For an introduction to Markov chains in phylogeny see Felsenstein [12].

The algorithms for Bayesian inference differ in their choice of a Markov chain to sample from the distribution, and in their choice of a prior. In practice, the choice of an appropriate prior is an important concern. Felsenstein [12] gives an introduction to many of the possible priors. Rannala and Yang [22] introduce a prior based on a birth-death process, whereas Huelsenbeck and Ronquist's program MrBayes [16] allows the user to input a prior (using either uniform or exponential distributions). Our results holds for all these popular priors, and only require that the priors are so-called ϵ-regular for some $\epsilon > 0$, in the sense that

$$\text{for all } T,p, \;\; \Psi(T,\overrightarrow{p}) \geq \epsilon.$$

Each tree $T \in \Omega$ is given a weight

$$w(T) = \int_{\overrightarrow{p}} \mathbf{Pr}\left(\overrightarrow{D} \mid T,\overrightarrow{p}\right)\Psi(T,\overrightarrow{p})d\overrightarrow{p}.$$

Computing the weight of a tree can be done efficiently via dynamic programming in cases where Ψ admits a simple formula. In other cases, numerical integration is needed. See Felsenstein [12] for background.

The transitions of the Markov chain (T_t) are defined as follows. From a tree $T_t \in \Omega$ at time t,

1. Choose a neighboring tree T'. See below for design choices for this step.
2. Set $T_{t+1} = T'$ with probability $\min\{1, w(T')/w(T)\}$, otherwise set $T_{t+1} = T_t$.

Two natural approaches for connecting the tree space Ω are nearest-neighbor interchanges (NNI), and subtree pruning and regrafting (SPR). In NNI, one of the $n-3$ internal edges is

chosen at random and the four subtrees are reconnected randomly in one of the three ways, see Figure 1 on page 6 for an illustration. In SPR, a random edge is chosen, one of the two subtrees attached to it is removed at random and reinserted along a random edge in the remaining subtree, see Figure 2 on page 6. We refer to the above chains as Markov chains with discrete state space and NNI and/or SPR transitions.

Some Markov chains instead walk on the continuous state space where a state consists of a tree with an assignment of edge probabilities. Our results extend to chains with continuous state space where transitions only modify the tree topology by an NNI or SPR transition, and edge probabilities are always in $(0, 1/2)$. Some examples of continuous state space chains are Li, Pearl and Doss [19], and Larget and Simon [18].

See Felsenstein [12] for an introduction to the various Markov chains, and also Durbin et al [5] for a description of the Markov chain of Larget and Simon.

The mixing time of the Markov chain T_{mix} is defined as the number of transitions until the chain is within total variation distance $1/4$ from the stationary distribution.

1.2 Formal Statement of Results

We consider data coming from a mixture of two trees $T_1(a, a^2)$ and $T_2(a, a^2)$. T_1 is given by $((12), 3), (45)$ while T_2 is given by $((14), 3), (25)$, see Figure 3 on page 6. On the trees $T_1(a, a^2)$ and $T_2(a, a^2)$ we have two edge probabilities, one for those edges incident to the leaves, and a different edge probabilities for internal edges. We let the probabilities of edges going to the leaves be a^2 and the internal edges have probability a where a will be chosen as a sufficiently small constant. The trees $T_1(a, a^2), T_2(a, a^2)$ will have small edge probabilities, as commonly occurs in practice.

We let $\mathcal{D}_1$ be the distribution of the data according to $T_1(a, a^2)$ and $\mathcal{D}_2$ according to $T_2(a, a^2)$. We let $\mathcal{D} = 0.5(\mathcal{D}_1 + \mathcal{D}_2)$, and consider a data set consisting of N characters.

We prove the following theorem.

Theorem 1. *There exist a constant $c > 0$ such that for all $\epsilon > 0$ the following holds. Consider a data set with N characters, i.e., $\overrightarrow{D} = (D_1, \ldots, D_N)$, chosen independently from the distribution $\mathcal{D}$. Consider the Markov chains on tree topologies defined by nearest-neighbor interchanges or subtree pruning and regrafting. Then with probability $1 - \exp(-cN)$ over the data generated, the mixing time of the Markov chains, with priors which are ϵ-regular, satisfies*

$$T_{mix} \geq c\epsilon \exp(cN).$$

Note that ϵ only has a small effect on the mixing time lower bound. Naturally, there is very little interest in studying Markov chains on trees with 5 leaves (there are 15 such trees). However, the result above immediately implies slow mixing for Markov chains on larger trees, assuming the pair of trees generating the data contain copies of T_1, T_2.

Corollary 2. *Let $n \geq 5$. Let $\mathcal{D} = 0.5(\mathcal{D}_1 + \mathcal{D}_2)$ be a distribution on n leaves, where $\mathcal{D}_1$ is generated according to a phylogenetic tree that has $T_1(a, a^2)$ as an induced subtree on some set*

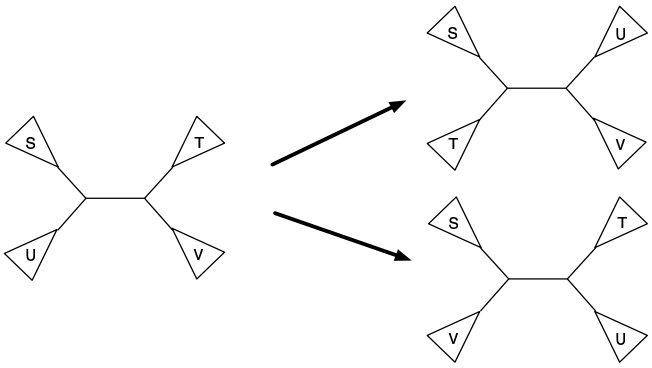

Figure 1: Illustration of NNI transition. An internal edge has 4 subtrees attached. The transition reattaches the subtrees randomly. Since the trees are unrooted, there are three ways of attaching the subtrees, one of which is the same as the original tree.

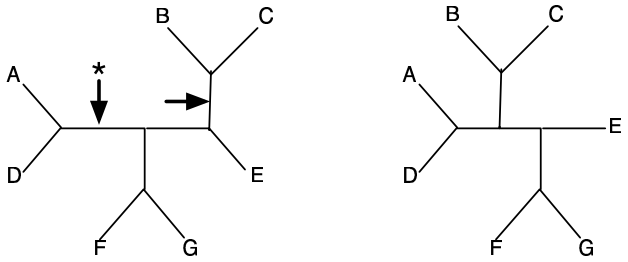

Figure 2: Illustration of SPR transition. The randomly chosen edge is marked by an arrow. The subtree containing B,C is removed and reattached at the random edge marked by a starred arrow. The resulting tree is illustrated.

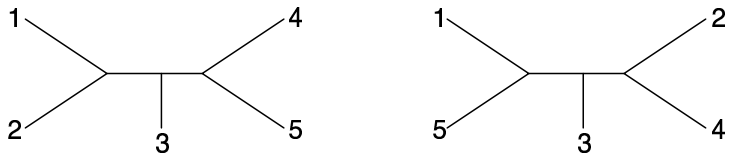

Figure 3: The trees T_1 and T_2

$S \subset [n]$ of size $|S| = 5$ and $\mathcal{D}_2$ is generated according to a phylogenetic tree that has $T_2(a, a^2)$ as an induced subtree on the same set S.

Consider the Markov chains on tree topologies defined by nearest-neighbor interchanges or subtree pruning and regrafting. Then there exist a constant $c > 0$ such that for all $\epsilon > 0$, with probability $1 - \exp(-cN)$ over the data generated, the mixing time of the Markov chains, with priors which are ϵ-regular, satisfies

$$T_{mix} \geq c\epsilon \exp(cN).$$

Remark 1. *One may still try to overcome the slow mixing above by using random starts. However, note that if the trees generating $\mathcal{D}_1$ and $\mathcal{D}_2$ have r disjoint sets S_i for in which the induced subtree in one is $T_1(a, a^2)$ and in the other is $T_2(a, a^2)$ one needs to average on at least 2^r starting points, and these starting points need to be cleverly chosen.*

1.3 General Mutation models

As mentioned above, our theorem is valid for many of the mutation models discussed in the literature. We now define these models and derive some elementary features of them that will be used below. In the general case, it is easier to define the evolution model on rooted trees. However, since we will only discuss reversible models, the trees may be rooted arbitrarily. For general models we consider rooted trees with edge lengths, as opposed to unrooted trees with edge probabilities. As it is well known for the CFN model, the edge probability $\overrightarrow{p}(e)$ is related to the edge length $\overrightarrow{\ell}(e)$ define below by $\overrightarrow{p}(e) = (1 - \exp(-\overrightarrow{\ell}(e)))/2$.

The mutation models are defined on a finite character set $\mathcal{A}$ of size q. We will denote the letters of the alphabet by α, β etc. The mutation model is given by an $q \times q$ mutation rate matrix Q that is common to all edges of the tree along with edge length $\overrightarrow{\ell}(e)$ for all edges of the tree. The mutation along edge e is given by

$$\exp(\overrightarrow{\ell}(e)Q) = I + \overrightarrow{\ell}(e)Q + \frac{\overrightarrow{\ell}^2(e)Q^2}{2!} + \frac{\overrightarrow{\ell}^3(e)Q^3}{3!} + \cdots$$

Thus, the probability of an assignment $D : V_{ext} \to \mathcal{A}$ is

$$\mathbf{Pr}\left(D \mid T, \overrightarrow{\ell}\right) = \sum_{\substack{D' \in \mathcal{A}^V : \\ D'(V_{ext}) = D(V_{ext})}} \pi_{D'(r)} \prod_{e=(u,v) \in E(T)} \left[\exp(\overrightarrow{\ell}(e)Q)\right]_{D'(u), D'(v)}$$

where all the edges (u, v) are assumed to be directed away from the root r.

We will further make the following assumptions below:

Assumption 1. *1. The Markov semi-group $(\exp(\overrightarrow{\ell}Q))_{\ell \geq 0}$ has a unique stationary distribution π, such that $\pi_\alpha > 0$ for all α. Moreover, the semi-group is* reversible *with respect to π, i.e., $Q\pi = \pi Q$.*

2. *The character at the root has marginal distribution π. This implies that the marginal distribution at every node is π.*

3. *The rate of transitions from a state is the same for all states. More formally, there exists a number q such that for all α:*

$$\sum_{\beta \neq \alpha} q_{\alpha,\beta} = -q_{\alpha,\alpha} = q. \tag{1}$$

In fact, by rescaling the edge-length of all edges we assume WLOG that $q = 1$.

4. *There exists constant $c > 0$ such that $q_{\alpha,\beta} > c$ for all $\alpha \neq \beta$.*

Remark 2. *Parts 1 and 2 of the assumption imply that we obtain the same model for all possible rootings of any specific tree. Thus, the model is in fact defined on unrooted trees.*

Remark 3. *It is straightforward to check that our assumptions include as special cases the CFN model, the Jukes-Cantor model, Kimura's two parameter model and many other models. See [20] for an introduction to the various evolutionary models.*

1.4 Statement of the General Theorem

Definition 3. *Let $\mathcal{T}$ be the space of all trees and edge lengths on 5 leaves. We say that a prior density Ψ on $\mathcal{T}$ is* (ϵ, a)-regular, if for every T and $\overrightarrow{\ell}$ where $\overrightarrow{\ell}(e) \leq 2a$ for all e, it holds that $\Psi(T, \overrightarrow{\ell}) \geq \epsilon$.

Remark 4. *All of the priors used in the literature are (a, ϵ)-regular for an appropriate value of $\epsilon = \epsilon(a)$.*

Theorem 4. *Let Q be a mutation rate matrix that satisfies assumption 1. For trees $(T_1, \overrightarrow{\ell}_1), (T_2, \overrightarrow{\ell}_2)$ on $n \geq 5$ leaves, let the distribution $\mathcal{D}_1$ be generated at the leaves of $(T_1, \overrightarrow{\ell}_1)$ and the distribution $\mathcal{D}_2$ be generated at the leaves of $(T_2, \overrightarrow{\ell}_2)$. Now, let $\mathcal{D} = 0.5(\mathcal{D}_1 + \mathcal{D}_2)$ and let $\overrightarrow{D} = (D_1, \ldots, D_N)$, chosen independently from the distribution $\mathcal{D}$.*

There exists an $a > 0$, a constant $c > 0$, two trees T_1^*, T_2^* on 5 leaves and open sets $L_1^* \subset (0, \infty)^{E(T_1^*)}, L_2^* \subset (0, \infty)^{E(T_2^*)}$ such that if for some $S \subset [n]$ of size $|S| = 5$ the induced subgraph of T_1 on S is T_1^* and has edge weights in the set L_1^*, and the induced subgraph of T_2 on S is T_2^* with edge weights in L_2^*. Then, the following holds for all $\epsilon > 0$.

Consider a Markov chain on discrete or continuous tree space where the only moves that change the topology of the tree are NNI and SPR transitions. Then with probability $1 - \exp(-cN)$ over the data generated, the mixing time of the Markov chain, for priors which are (a, ϵ)-regular, satisfies

$$T_{mix} \geq c\epsilon \exp(cN).$$

Remark 5. *It is straightforward to check that Theorem 1 is a special case of Theorem 4. This follows by the standard translation between edge-lengths and edge-probabilities. As mentioned above, the CFN model (as well as Jukes-Cantor and many other models) satisfy the properties we assumed in Assumption 1.*

2 Proof of the General Theorem

We first expand the distribution $\mathcal{D}$. It is easy to do so in terms of C^*, where C^* is the set of cherries in $T_1 \cup T_2$. We will use the following definition of a cherry.

Definition 5. *Let T be a tree. We say that that a pair of leaves i, j is a* cherry *of T if there exists a single edge e of T such that removing e disconnects i, j from the other leaves of T. For a tree T we let $C(T)$ denote the set of cherries of T.*

Note that according to this definition the "star"-tree has no cherries. We clearly have $C^* = C(T_1) \cup C(T_2) = \{(12), (14), (45), (25)\}$.

Our theorem holds for a sufficiently small. Hence, the asymptotic notation in our proofs is in terms of $1/a \to \infty$. Thus, $a = o(a \log a)$ and $a^2 = o(a \log a)$ since $-\log a$ grows as $a \to 0$.

Part 3 of Assumption 1 implies that for an edge of length a, given a character assignment α for one endpoint, the other endpoint has a different assignment with a probability $a + O(a^2)$ independently of α. This is used implicitly throughout the following proof. Dropping this assumption would complicate many of our calculations depending on the probability of a set of mutations.

It is easy to estimate $\mathcal{D}$ for small a. This follows from the following lemma.

Lemma 6. *For an edge e of length b, conditioned on the character at the end point of the edge, the probability that the other end-point has the same label is $1 - b + O(b^2)$. The probability that it obtains a different label is $b + O(b^2)$.*

Proof. Part 4 of Assumption 1 along with the expansion of $\exp(bQ)$ implies

$$\exp(bQ) = I + bQ + O(b^2).$$

□

We will use the following notation for characters. By α we denote the character that is constant α. We let $F_\emptyset$ denote the set of all constant characters. By (α, i, β) we denote the character that is α on all leaves except i where it is β. The set of all such characters is denote by F_i. By (α, i, j, β) we denote the character that is β on i, j and α on all other leaves. The set of all such characters is denoted by $F_{i,j}$. We denote the set of all other characters by G.

We begin with some easy bounds on the dominant terms for the probabilities of various characters. Consider the probability $D \notin F_\emptyset$. There are two ways this can occur, either: there is a mutation on exactly one of the internal edges which occurs with probability $2a(1-a)+O(a^2) = 2a+O(a^2)$; or there is a mutation on a terminal branch (i.e., an edge connected to a leaf) and/or both internal edges, these occur with probability $O(a^2)$ by our choice of edge lengths on T_1 and T_2. Hence,

$$\mathcal{D}[F_\emptyset] = 1 - 2a + O(a^2) \tag{2}$$

For $D \in F_i$ there needs to be a mutation on a terminal branch, hence

$$\mathcal{D}[F_i] = O(a^2)$$

Consider a cherry $(i,j) \in C^*$, say $(i,j) = (1,2) \in C(T_1)$. To generate $D \in F_{i,j}$ we need to be generating from T_1, and need a mutation on the internal edge to the parent of leaves 1 and 2, or a mutation on more than one terminal branch. Thus, for $(i,j) \in C^*$,

$$\mathcal{D}[F_{i,j}] = a/2 + O(a^2) \tag{3}$$

For $(i,j) \notin C^*$, we instead have

$$\mathcal{D}[F_{i,j}] = O(a^2)$$

Finally, the remaining characters will be lower order terms, i.e.,

$$\mathcal{D}[G] = O(a^2).$$

Remark 6. *The lemma above summarizes what we need to know on the distribution $\mathcal{D}$ for the rest of the theorem. Note that the same estimates would hold if all the internal branch length of T_1, T_2 are in $[a - a^3, a + a^3]$ and the terminal edges are in $[a^2 - a^3, a^2 + a^3]$. Thus we will in fact show that there are two open sets S_1, S_2 of edge-length for which the conclusion of the theorem holds.*

Definition 7. *The expected likelihood of a tree T with edge lengths $\overrightarrow{\ell}$ given the data is defined as*

$$L_{\mathcal{D}}(T, \overrightarrow{\ell}) = \mathbf{E}_{x \in \mathcal{D}} \log \mathbf{Pr}\left(x \mid T, \overrightarrow{\ell}\right)$$

Let $L_{\mathcal{D}}(T, \ell)$ denote the expected likelihood of the tree T with all edge length ℓ. We will show that tree T_1 with all edge length a (including terminal branches) has large likelihood.

Lemma 8. *The tree T_1 satisfies*

$$L_{\mathcal{D}}(T_1, a) \geq H(\pi) + (1 + o(1))3a \log a.$$

and similarly for T_2, where

$$H(\pi) = \sum_{\alpha} \pi_\alpha \log \pi_\alpha.$$

Proof. Consider T_1. We first consider the sequences in $F_\emptyset$. By (2), the $\mathcal{D}$ probability of the sequence α is

$$\pi_\alpha(1 - 2a + O(a^2)) = \pi_\alpha + o(a \log a).$$

while the log-likelihood of α according to (T_1, a) is given by

$$\begin{aligned} \log \mathbf{Pr}\left(\alpha \mid T, a\right) &= \log(\pi_\alpha(1 - 7a + O(a^2))) \\ &= \log(\pi_\alpha) + 7a + O(a^2) \\ &= \log(\pi_\alpha) + o(a \log a). \end{aligned}$$

Thus, the total contribution to $L_{\mathcal{D}}(T_1, a)$ coming from $F_\emptyset$ is

$$H(\pi) + o(a \log a).$$

All sequences in F_i have a contribution of $O(a^2)$ up to log corrections, which is also $o(a \log a)$. Similarly for sequences in $F_{i,j}$ such that $(i,j) \notin C^*$, and for sequences in G. If (i,j) belongs to C^* there are two possibilities. First, if (i,j) is a cherry of T_1 and $(\alpha, i, j, \beta) \in F_{i,j}$ then

$$\log \mathbf{Pr}\left((\alpha, i, j, \beta) \mid T, a\right) = (1 + o(1)) \log a.$$

(This follows by considering a single mutation along the edge the separates the cherry (i,j) from the rest of the tree). Thus, using (3), for $(i,j) \in C(T_1)$ we have a total contribution of

$$(0.5 + o(1))a \log a.$$

For $(i,j) \in C^* \setminus C(T_1)$ (i.e., $(i,j) \in C(T_2)$), then for all $(\alpha, i, j, \beta) \in F_{i,j}$ this character occurs if the only mutations are on the pair of terminal edges connected to i and j, otherwise it requires at least 3 mutations. Hence,

$$\log \mathbf{Pr}\left((\alpha, i, j, \beta) \mid T, a\right) = (2 + o(1)) \log a,$$

Thus, using (3), we get a total contribution of

$$(1 + o(1))a \log a.$$

Since C^* contains two cherries from T_1 and two from T_2, the total contribution of $F_{i,j}$ is $(1 + o(1))3a \log a$ as needed. □

Remark 7. *Repeating the proof above shows that*

$$L_{\mathcal{D}}(T_1, \overrightarrow{\ell}) \geq H(\pi) + (1 + o(1))3a \log a$$

if all the edge lengths $\overrightarrow{\ell}$ *are in* $[a/2, 2a]$.

Next we show that for any tree that is close to the optimum, all of its edge lengths are at most $O(a \log(1/a))$.

Lemma 9. *Let* (T, p) *be a tree such that at least one of the edge lengths is greater than* $4a \log(1/a)$. *Then,*

$$L_{\mathcal{D}}(T, \overrightarrow{\ell}) \leq H(\pi) + 4a \log a + o(a \log a).$$

Proof. Let T be any tree for which the sum of the edge lengths is more than $4a \log(1/a)$. It is easy to see that the probability that this tree generates the sequence α is at most $\pi_\alpha(1 + 4a \log a + o(a \log a))$. Since all the terms appearing in the likelihood are negative, we obtain

$$\begin{aligned} L_{\mathcal{D}}(T, \overrightarrow{\ell}) &\leq \sum_{\alpha} (\pi_\alpha + o(a \log a)) \log \left(\pi_\alpha (1 + 4a \log a + o(a \log a))\right) \\ &\leq H(\pi) + 4a \log a \sum_{\alpha} \pi_\alpha + o(a \log a) \\ &= H(\pi) + 4a \log a + o(a \log a), \end{aligned}$$

which proves the claim. □

Once we restrict to trees all of whose edges lengths are at most $4a\log(1/a)$ it is easier to see that the optimal tree must have the correct topology.

Lemma 10. *Let $(T, \overrightarrow{\ell})$ be a tree all of whose edges length are at most $4a\log(1/a)$ and suppose further that T has a topology different than T_1 or T_2. Then*

$$L_{\mathcal{D}}(T,p) \leq H(\pi) + (1+o(1))3.5a\log a.$$

Before proving the above lemma we state the following combinatorial observation.

Observation 11. *Let $T \neq T_1, T_2$ then*

$$|C(T) \cap C^*| \leq 1$$

To see the observation, consider a tree T that contains at least one of the cherries of C^*, say $(1,2)$. Clearly, T can not also contain the cherries $(1,4)$ or $(2,5)$. And if it contains the cherry $(4,5)$ then $T = T_1$.

Proof of Lemma 10. Many of the calculations in this proof are identical to the proof of Lemma 8. We first observe that as in Lemma 8 the contribution to $L_{\mathcal{D}}(T, \overrightarrow{\ell})$ coming from $F_\emptyset$ is at most $H(\pi) + o(a\log a)$. We will now show that the contribution to $L_{\mathcal{D}}(T, \overrightarrow{\ell})$ coming from the cherries in C^* is smaller than $3.5(1+o(1))a\log a$ (recall that all terms are negative). By Observation 11 the number of cherries of T that belong to C^* is at most 1. Note that if $(i,j) \in C^*$ and $(i,j) \notin C(T)$ then for all α and β,

$$\log \mathbf{Pr}\left((\alpha,i,j,\beta) \mid T, \overrightarrow{\ell}\right) \leq \log(O(a^2\log^2(1/a))) = 2(1+o(1))\log a.$$

If $(i,j) \in C(T) \cap C^*$ then (as in the proof of Lemma 8) we have

$$\begin{aligned} \log \mathbf{Pr}\left((\alpha,i,j,\beta) \mid T, \overrightarrow{\ell}\right) &\leq \log(a\log(1/a) + o(a\log(1/a)) \\ &= (1+o(1))\log a. \end{aligned}$$

Thus,

$$\begin{aligned} L_{\mathcal{D}}(T, \overrightarrow{\ell}) &\leq H(\pi) + o(a\log a) + (1+o(1))a(3*2\log a + \log a)/2 \\ &= H(\pi) + (1+o(1))3.5a\log a. \end{aligned}$$

□

Based on Lemmas 8, 9 and 10, we can now make the following assumption.

Assumption 2. *From now on we fix $a > 0$ sufficiently small so that if $\mathcal{D}$ is generated as a mixture from the two trees $(T_1, \overrightarrow{\ell}_1)$ and $(T_2, \overrightarrow{\ell}_2)$ where, for all $i = 1,2$, all internal edges e satisfy $\overrightarrow{\ell}_i(e) \in [a-a^3, a+a^3]$ and all terminal edges (i.e., connected to a leaf) satisfy $\overrightarrow{\ell}_i(e) \in [a^2-a^3, a^2+a^3]$, then*

- *If $T = T_1$ or $T = T_2$ and the edge lengths $\overrightarrow{\ell}$ are in $[a/2, 2a]$ then*

$$L_{\mathcal{D}}(T, \overrightarrow{\ell}) \geq H(\pi) + 3.1a \log a$$

- *If $T \neq T_1, T_2$, then for any edge lengths $\overrightarrow{\ell}$,*

$$L_{\mathcal{D}}(T, \overrightarrow{\ell}) \leq H(\pi) + 3.4a \log a$$

Definition 12. *Let $\overrightarrow{D} = (D_1, \ldots, D_N)$ be N characters. We let*

$$L_{\overrightarrow{D}}(T, p) = \sum_{D \in \overrightarrow{D}} \log \mathbf{Pr}\left(D \mid T, p\right).$$

Using Chernoff bound, we get the following lemma.

Lemma 13. *Suppose $\overrightarrow{D}$ is drawn according to N independent samples from the distribution $\mathcal{D}$. Then, with probability $1 - e^{-\Omega(N)}$, for all trees $(T, \overrightarrow{\ell})$ with the topology $T_1(T_2)$ and edge length $\overrightarrow{\ell}(e)$ in $[a/2, 2a]$ for all e, it holds that*

$$L_{\overrightarrow{D}}(T, \overrightarrow{\ell}) \geq \left(H(\pi) + (3.2a \log a)\right) N. \tag{4}$$

and for all trees $(T, \overrightarrow{\ell})$ with topologies different than T_1, T_2 it holds that

$$L_{\overrightarrow{D}}(T, \overrightarrow{\ell}) \leq \left(H(\pi) + (3.3a \log a)\right) N. \tag{5}$$

Proof. Use Chernoff for each of the q^5 sequences in $\mathcal{A}^5$. □

Lemma 14. *Let $\epsilon > 0$ and let Ψ be an (ϵ, a)-regular prior on $\mathcal{T}$. Then with probability $1 - e^{-\Omega(N)}$ it holds that if $T \neq T_1, T_2$ then*

$$\frac{w(T)}{w(T_1)} \leq \frac{1}{\epsilon} \exp\left(-0.1a \log(1/a) N\right).$$

Proof. With probability $1 - e^{-\Omega(N)}$ we have that (5) and (4) hold. We show that this implies

$$\frac{w(T)}{w(T_1)} \leq \frac{1}{\epsilon} \exp\left(-0.1a \log(1/a) N\right).$$

Since Ψ is (ϵ, a)-regular we see that:

$$\begin{aligned} w(T_1) &= \int_{\overrightarrow{\ell}} \exp(L_{\overrightarrow{D}}(T_1, \overrightarrow{\ell})) \Psi(T_1, \overrightarrow{\ell}) dp \\ &\geq \epsilon a^7 \exp(H(\pi) N) \exp((3.2a \log a) N). \end{aligned}$$

On the other hand,

$$\begin{aligned} w(T) &= \int_{\overrightarrow{\ell}} \exp(L_{\overrightarrow{D}}(T, \overrightarrow{\ell}))\Psi(T, \overrightarrow{\ell})dp \\ &\leq \exp(H(\pi)N)\exp((3.3a\log a)N). \end{aligned}$$

The claim follows. □

To finish off the proof of Theorem 1 we need to show that bad conductance implies slow mixing. Since we also work in the continuous setting, we prove the following claim below.

Lemma 15. *Consider a discrete time Markov chain P on a discrete or continuous state space with a unique stationary measure μ. Assume furthermore that there exists a partition of the state space into 3 sets A_1, A_2, B such that the probability of a move from A_2 to A_1 is 0 (in the sense that $\int d\mu(x)1(x \in A_2)\int dP(x,y)1(y \in A_1) = 0$) and $\mu(A_1) \geq \mu(A_2), \mu(B)/\mu(A_i) \leq \epsilon$ for $i = 1, 2$.*

Let μ^t denote the distribution of the chain after t steps, where the initial distribution μ^0 is given by μ conditioned to A_2. Then the total variation distance between $\mu^{1/4\epsilon}$ and μ is at least $1/4$.

Proof. Let $t = 1/4\epsilon$ and consider sequences $(x_1, \ldots, x_t)$ of trajectories of the chain where x_1 is chosen according to the stationary distribution. Since each x_i is distributed according to the stationary distribution, the fraction of sequences that contain an element of B is by the union bound at most $t\epsilon\mu(A_2) = \mu(A_2)/4$. The fraction of sequences that have their first element in A_2 is $\mu(A_2)$. Thus, conditioned on having $x_1 \in A_2$, the probability that $x_t \in B \cup A_1$ is at most $1/4$. Since the stationary measure of $B \cup A_1$ is at least $1/2$, the claim follows. □

Proof of Theorem 1. The proof now follows from Lemmas 14 and 15 – we take the two sets corresponding to T_1 and T_2 with all edge lengths strictly between 0 and ∞. The proof follows by the observation that T_1 and T_2 are not connected by a transition by either NNI or SPR transitions. □

3 Future Directions

A popular program is MrBayes [16] which additionally uses what is known as Metropolis Coupled Markov Chain Monte Carlo, referred to as $(MC)^3$ [13]. Analysis of this approach requires more detailed results, and it is unclear whether our techniques can be extended to this extent. Some theoretical work analyzing MC^3 in a different context was done by Bhatnagar and Randall [1].

An interesting future direction is to prove a positive result. In particular, is there a class of trees where we can prove fast convergence to the stationary distribution when the data is generated by a tree in this class. More generally, if the data is generated by a single tree, do the Markov chains always converge quickly to the stationary distribution?

Acknowledgments

We thank Bernd Sturmfels and Josephine Yu for interesting discussions on the Algebraic Geometry of tree space.

References

[1] N. Bhatnagar and D. Randall, Torpid mixing of simulated tempering on the Potts model, In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA), 478-487, 2004.

[2] J. A. Cavender, Taxonomy with confidence, *Math. Biosci.*, **40**: 271–280, 1978.

[3] B. Chor, M. D. Hendy, B. R. Holland, and D. Penny, Multiple Maxima of Likelihood in Phylogenetic Trees: An Analytic Approach, *Mol. Biol. Evol.*, 17(10:1529-1541, 2000.

[4] B. Chor and T. Tuller, Maximum Likelihood of Evolutionary Trees is Hard, In Proceedings of the Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB), 2005.

[5] R. Durbin, S. Eddy, A. Krogn, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[6] M. Dyer, A. Frieze, and M. Jerrum, On Counting Independent Sets in Sparse Graphs, *SIAM J. Comput.*, 31(5):1527-1541, 2002.

[7] P. Diaconis and S. P. Holmes, Random walks on trees and matchings, *Electron. J. Probab.*, 7, article 6, 2002.

[8] M. Develin, and B. Sturmfels, Tropical convexity, *Doc. Math.* 9:1–27, 2004.

[9] P. Djian and H. Green, Vectorial expansion of the involucrin gene and the relatedness of the hominoids, *Proc. Natl. Acad. Sci. USA*, 86(21):8447-8451, 1989.

[10] P. L. Erdös, M. A. Steel, L. A. Székely, T. J. Warnow, A Few Logs Suffice to Build (Almost) All Trees (I), *Rand. Struct. Alg.*, 14(2):153-184, 1999.

[11] J. S. Farris, A probability model for inferring evolutionary trees, *Syst. Zool.*, 22: 250–256, 1973.

[12] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Inc., Sunderland, MA, 2004.

[13] C. J. Geyer, Markov chain Monte Carlo maximum likelihood, *Computing Science and Statistics: Proc. 23rd Symp. Interface*, 156-163, 1991.

[14] D. Graur and W.-H. Li, *Fundamentals of Molecular Evolution*, Second Edition, Sinauer Associates, Inc., Sunderland, MA, 1999.

[15] J. P. Huelsenbeck, B. Larget, R. E. Miller, and F. Ronquist, Potential Applications and Pitfalls of Bayesian Inference of Phylogeny, *Syst Biol.*, 51(5):673-88, 2002.

[16] J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees, *Bioinformatics*, 17(8):754-5, 2001.

[17] J. P. Huelsenbeck, F. Ronquist, R. Nielsen, J. P. Bollback, Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology, *Science*, 294:2310-2314, 2001.

[18] B. Larget and D. L. Simon, Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees, *Mol. Biol. Evol.*, 16(6):750-759, 1999.

[19] S. Li, D. K. Pearl, and H. Doss, Phylogenetic tree construction using Markov chain Monte Carlo, *J. Am. Stat. Assoc.*, 95:493-508, 2000.

[20] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, 2000.

[21] J. Neyman, Molecular studies of evolution: a source of novel statistical problems. In *Statistical decision theory and related topics*, S.S Gupta and J. Yackel (eds), 1–27, 1971.

[22] B. Rannala and Z. Yang, Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference, *J. Mol. Evol.*, 43: 304-311, 1996.

[23] S. Roch, A Short Proof that Phylogenetic Tree Reconstruction by Maximum Likelihood is Hard, Preprint appears on arXiv at http://arxiv.org/abs/math.PR/0504378.

[24] P. B. Samollow, L. M. Cherry, S. M. Witte, and J. Rogers, Interspecific variation at the Y-linked RPS4Y locus in hominoids: Implications for phylogeny, *Am. J. Phys. Anthropol.*, 101(3):333-343, 1996.

[25] D. L. Simon and B. Larget. Bayesian analysis in molecular biology and evolution (BAMBE), Version 2.03 beta, Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, PA.

[26] D. Speyer, and B. Sturmfels, The tropical Grassmannian, *Adv. Geom.*, 4(3):389–411, 2004.

[27] Z. Yang, Complexity of the simplest phylogenetic estimation problem, *Proc. R. Soc. Lond. B Biol. Sci.*, 267(1439):109-116, 2000.

[28] Z. Yang and B. Rannala, Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method, *Mol. Biol. Evol.*, 14(7):717-724, 1997.