JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 11, 2011

Pp. 1599–1609
DOI: 10.1089/cmb.2011.0164

An Optimization-Based Sampling Scheme for Phylogenetic Trees

NAVODIT MISRA,[1] GUY BLELLOCH,[2] R. RAVI,[3] and RUSSELL SCHWARTZ[4]

ABSTRACT

Much modern work in phylogenetics depends on statistical sampling approaches to phylogeny construction to estimate probability distributions of possible trees for any given input data set. Our theoretical understanding of sampling approaches to phylogenetics remains far less developed than that for optimization approaches, however, particularly with regard to the number of sampling steps needed to produce accurate samples of tree partition functions. Despite the many advantages in principle of being able to sample trees from sophisticated probabilistic models, we have little theoretical basis for concluding that the prevailing sampling approaches do in fact yield accurate samples from those models within realistic numbers of steps. We propose a novel approach to phylogenetic sampling intended to be both efficient in practice and more amenable to theoretical analysis than the prevailing methods. The method depends on replacing the standard tree rearrangement moves with an alternative Markov model in which one solves a theoretically hard but practically tractable optimization problem on each step of sampling. The resulting method can be applied to a broad range of standard probability models, yielding practical algorithms for efficient sampling and rigorous proofs of accurate sampling for heated versions of some important special cases. We demonstrate the efficiency and versatility of the method by an analysis of uncertainty in tree inference over varying input sizes. In addition to providing a new practical method for phylogenetic sampling, the technique is likely to prove applicable to many similar problems involving sampling over combinatorial objects weighted by a likelihood model.

Key words: algorithms, linear programming, molecular evolution, Monte Carlo likelihood, phylogenetic trees.

1. INTRODUCTION

MUCH OF THE THEORY AND CLASSIC METHODS OF PHYLOGENY RECONSTRUCTION were developed for a variety of optimization formulations of the problem (e.g., parsimony, likelihood, or distance-based). Optimization approaches have fallen into disfavor, however, due to the frequent presence of multiple optima or near-optima and a general desire to quantify uncertainty in the resulting trees. As a result, algorithms based

[1]Max Planck Institute for Molecular Genetics, Berlin, Germany.
[2]Computer Science Department, [3]Tepper School of Business, and [4]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania.

on the idea of sampling over the space of phylogenetic trees for a given data set are now generally preferred to optimization approaches in order to provide better statistical support while answering questions such as whether a given bipartition is more likely than another conflicting bipartition. Popular sampling methods such as MrBayes (Huelsenbeck and Ronquist, 2001) use Markov chains over a class of tree rearrangement moves such as nearest neighbor interchange (NNI), subtree pruning and re-grafting (SPR), and tree bisection and reconnection (TBR) (Felsenstein, 2004) to estimate partition functions of trees under some implied probability distribution on tree topologies, branch lengths, ancestral sequences, and population genetic parameters. Despite their advantages, though, sampling-based approaches suffer from a comparatively poorly developed theoretical literature. In particular, there are few theoretical results regarding the mixing properties of their underlying Markov chains, particularly the number of steps for which one must run a model to sample accurately from its partition function. As a result, we rarely have any sound theoretical basis for concluding that a phylogeny sampling algorithm has been run sufficiently long to generate an accurate sample.

Among the few positive results are the methods of Diaconis and Holmes (2002), for uniform sampling over all phylogenetic trees, and the recent result of Stefankovic and Vigoda (2010), showing rapid mixing of SPR Markov chains when data is generated by phylogenies with sufficiently short branches. Mossel and Vigoda (2005) and Stefankovic and Vigoda (2007) have shown that Markov chains based on standard NNI or SPR moves do not always mix well. Their results are valid for a likelihood-based method on problem instances where input data can be represented by a mixture of two tree topologies. The question of a polynomial bound for mixing time on data generated from a single tree, with arbitrary branch lengths, is still open. There is, therefore, a need for either new theoretical insights into the mixing properties of the prevailing methods or the development of new sampling methods for which we can more readily analyze these properties.

In this article, we pursue the latter approach, developing an alternative Markov chain-based phylogeny sampling algorithm that is more amenable to theoretical mixing time analysis and allows one to prove non-trivial mixing time bounds in important special cases. The key algorithmic insight of our method is that one can convert the hard sampling problem inherent in standard tree sampling into an easier sampling problem that uses, as a subroutine, the solution of a theoretically hard but practically tractable optimization problem (an instance of the minimum spanning tree problem with degree constraints). By repeatedly solving the embedded optimization problem provably to optimality, one can in turn solve the sampling problem with a small number of Monte Carlo steps. Our method can be used to sample from the likelihood distribution of labelled tree topologies, also known as the *ancestral likelihood*. We use this optimization-based method to theoretically bound the mixing time for a heated version of the well known Cavender-Farris-Neyman (CFN) model (Cavender, 1978; Farris, 1973; Neyman, 1971), proving that our optimization-based Markov chain mixes in time polynomial in the number of leaves (taxa) and the number of characters in the input for the CFN model. We then demonstrate the practical effectiveness of the method through a small empirical analysis of how uncertainty in tree topology changes with increasing numbers of taxa under a standard likelihood model. Our method can be readily generalized to sampling from the set of Steiner trees on arbitrary weighted graphs and might have applications to many similar problems involving sampling over combinatorial objects weighted by a likelihood function.

We begin by presenting some basic notation and background on likelihood models, and explain our new approach in their context. We then describe the Integer Linear Programming (ILP) formulation for the optimization subroutine in our sampler, and show how using this powerful procedure in each move, the mixing time of the CFN model of ancestral likelihood can be bounded by a polynomial number of steps in the input size. Finally, we close with some experimental results intended to demonstrate the practical use of our method.

2. NOTATION

We begin by defining some basic terminology and notation used throughout this manuscript. We refer the reader to Felsenstein (2004) for a more thorough introduction to the topic of phylogenetics and the concepts and terminology presented below. Let H be an input matrix that specifies a set χ of N taxa, over a set $C = \{c_1, \ldots c_M\}$ of M characters, such that H_{ij} represents the j^{th} character of the i^{th} taxon. Further, let n_k be the set of admissible states of the k^{th} character c_k. The set of all possible states is the space

$\mathcal{S} \equiv n_1 \times n_2 \ldots \times n_M$. We will represent the i^{th} character of any element $b \in S$, by $(b)_i$. The state space $\mathcal{S}$ can be represented as a graph $\mathcal{G} = (V_\mathcal{G}, E_\mathcal{G})$ with the vertex set $V_\mathcal{G} = \mathcal{S}$ and edge set $E_\mathcal{G} = \{(u, v) | u, v \in \mathcal{S}, d_h[u, v] = 1\}$, where $d_h[u, v]$ is the Hamming distance between u and v.

The set of all possible trees can be conveniently classified using the concept of phylogenetic X-Trees. A phylogenetic X-Tree $T(\chi)$ displaying a set of taxa χ is defined as follows: there is a bijection or labeling between the set of taxa χ and the leaves of $T(\chi)$. Furthermore, all internal nodes are of degree three or more. The latter requirement is equivalent to contracting the edges between any pair of degree-two internal nodes. Clearly, removing an edge in any tree disconnects the tree into two subtrees, each of which has a non-empty intersection with the set of taxa. Thus, each edge corresponds to a bipartition or a *split* of the taxa. The *topology* of a phylogenetic X-Tree is defined as the set of all splits obtained by removing an edge of $T(\chi)$. A popular approach to solving the phylogeny inference problem is to search through the space of all topologically distinct phylogenetic X-Trees for a given set of taxa. This search space is usually defined over the space of binary tree topologies (i.e., where all internal nodes are of degree three). Any instance of a phylogenetic X-Tree with an internal node with degree greater than three (also known as a *polytomy*) can be treated as a special case of a binary tree where two internal nodes represent the same vertex in the graph $\mathcal{G}$. From now on we will refer to such binary phylogenetic X-Trees simply as phylogenies. Each phylogeny on N leaves has $N - 2$ degree 3 internal nodes.

It is well known that there are $\frac{(2N-2)!}{2^{N-1}(N-1)!}$ distinct rooted tree topologies for phylogenies with N leaves. Diaconis and Holmes (2002) have shown that the set of tree topologies can be conveniently visualized using a connection between perfect matchings on $2N - 2$ points and phylogenies with N leaves. Given a phylogeny T, we will use their method to assign a number to each internal node. We arbitrarily assign distinct labels to the leaves (from 1 to N) each of which corresponds to an element in χ. Since we are interested in unrooted phylogenies, we arbitrarily root the tree along any of the $2N - 3$ edges. Initially, all internal nodes are unlabeled. Each internal node is assigned a label between $N + 1$ and $2N - 2$ in the following sequence: (1) At each step, find an unlabeled internal node that has both its descendants labeled; in case there is more than one such internal node, choose the one that has a descendant with the lowest label. (2) Assign the lowest available label to this internal node: (3) Recurse until all nodes are labeled. Diaconis and Holmes showed that this mapping is a bijection by showing how to transform any matching into a binary tree as follows. Assume we have a perfect matching P on $2N - 2$ points, such that the first N points represent the leaves of some binary tree T. Since, more than half of the points are leaves, at least one pair of leaves (say u and v) must be matched in P. We can represent this matched pair by a subtree where u and v are joined to an internal node with the smallest label (namely $N + 1$). If there is more than one pair of matched leaves, we take the pair that contains the leaf with the smallest label and connect them with the internal node $N + 1$. Now if we remove this pair of matched leaves and treat the internal node $N + 1$ as a new leaf, the remaining perfect matching on $2N - 4 = 2(N - 1) - 2$ points has $N - 2 + 1$ leaves along with a subtree associated with the new leaf node $N + 1$. If we iterate this process k times, we reduce the matching to $2(N - k) - 2$ points and obtain a forest with $2N - 2 - 2k$ components on the vertices $\{1, \ldots, 2N - 1\}$, such that the leaf of each subtree has a label between 1 and N. After $N - 2$ steps, we join the final pair of nodes to get a binary tree with nodes 1 to N as leaves.

3. LIKELIHOOD MODEL

We will represent each phylogeny by a 4-tuple $T(\chi, \phi, \alpha, \tau)$. We overload the symbol T to represent both the topology of the phylogeny as well as a bijection between leaves $l(T)$ and input taxa χ and a mapping from internal nodes of T onto a set $\phi \subseteq \mathcal{S}$ such that ϕ_i represents the label for the i^{th} internal node. Next, we assign a likelihood to T assuming that the taxa have evolved via point mutations. Let $\alpha = \{\alpha_k | \alpha_k[j, i] > 0 \ \forall i \neq j, c_k \in C\}$ be a set of rate matrices, such that $\alpha_k[j, i]$ represents the rate for a transition from state i to j for character c_k. We will assume α is reversible with respect to $\pi_k[i]$ (representing the equilibrium frequency of state i at site k), such that $\alpha_k[i, j]\pi_k[j] = \alpha_k[j, i]\pi_k[i]$ and satisfies the conservation equation

$$\alpha_k[i, i] = -\sum_{j \neq i} \alpha_k[j, i] \tag{1}$$

The likelihood of each edge $e = (u, v) \in T$ is given by

$$L(e) = \prod_{c_k \in C} \exp(\tau_e \alpha_k)[(u)_k, (v)_k]$$
$$= \prod_{c_k \in C} \left(I[(u)_k, (v)_k] + \tau_e \alpha_k [(u)_k, (v)_k] + \ldots + \frac{\tau_e^n}{n!} \alpha_k^n [(u)_k, (v)_k] + \ldots \right) \quad (2)$$

where, $\tau_e \in [0, \infty)$ is the branch length representing relative time and I is the identity matrix. We root the tree arbitrarily at internal node $2N - 2$, represented by sequence r, and compute likelihood of edges directed away from the root. Let $\pi[r] = \prod_{c_k \in C} \pi_k[(r)_k]$ be the equilibrium density of the sequence representing the root. The ancestral likelihood of the phylogeny is then given by

$$L(T(\chi, \phi, \alpha, \tau)) = \pi[r] \prod_{e \in T} L(e) \quad (3)$$

The problem we want to solve is that of generating random samples from this likelihood distribution. We can simplify the problem somewhat by integrating out the set of branch lengths τ. Since we know the end points for each edge, this integral is easy to compute using a spectral decomposition of α_k in terms of its eigenvalues $\Lambda_k = \{-\lambda\}$ and corresponding eigenvectors $\{|\lambda\rangle\}$. However, since the smallest eigenvalue is zero (corresponding to the equilibrium distribution), we need to provide a suitable prior over branch lengths, in order to ensure that the integral is convergent. Choosing a suitable prior for an unbounded parameter requires care because a flat prior is not always an uninformative prior (Felsenstein, 2004). In practice, an exponentially decaying prior $Pr(\tau_e) = \eta e^{-\eta \tau_e}$ is usually recommended by popular methods such as MrBayes, and we will follow the same convention in this article. Combining this prior with our likelihood model gives us an expression for the posterior distribution:

$$L(T(\chi, \phi, \alpha)) = \pi[r] \prod_{e \in T} \int_0^\infty L(e) Pr(\tau_e) d\tau_e$$
$$= \pi[r] \prod_{(u,v) \in T} \int_0^\infty \prod_{c_k \in C} \left(\sum_{-\lambda \in \Lambda_k} \langle (u)_k | \lambda \rangle \langle \lambda | (v)_k \rangle e^{-\lambda \tau_e} \right) \eta e^{-\eta \tau_e} d\tau_e \quad (4)$$

Note that α_k is typically of dimension 20 or less, so the spectral decomposition is not a computational bottleneck and the likelihood $L(T(\chi, \phi, \alpha))$ can be computed in $O(NM)$ time. We will not focus on sampling the branch lengths τ_e in this article; however, for completeness we note that τ_e can be sampled exactly and efficiently from the distribution represented by the integrand in the previous equation using rejection sampling (Misra and Schwartz, 2008). In Section 5, we will use a particularly simple closed form expression for the likelihood maxima for the standard CFN model to derive some theoretical results for our method. Note, that in contrast to our approach, methods based on fixing the tree topology followed by sampling branch lengths are known to get trapped in local optima.

Our new model. Assume we are given the labels ϕ for the internal nodes and let $T_*(\chi, \phi, \alpha)$ be a most likely phylogeny. Let $\mathcal{T}(\chi, \alpha) = \{T_*(\chi, \phi, \alpha) : \phi \subset \mathcal{S}\}$. We will restrict ourselves to sampling from the likelihood distribution over the set $\mathcal{T}(\chi, \alpha)$. Note that ancestral maximum likelihood is not statistically consistent in general and pruning away suboptimal trees can significantly alter the marginal distribution (on summing over the ancestral node labels) for a tree topology in certain cases. The intuition behind our approach is that this pruning might allow us to sample from the remaining phylogenies efficiently and reliably, if we can solve the optimization problem efficiently.

Now, consider the following family of distributions $L(v)^\beta$ for $\beta \in [0, \infty)$. Such distributions are usually called ''heated'' distributions and β, the inverse temperature, in analogy with the usual definition of temperature in physical processes. These distributions are consistent with our intuitive understanding that high β or low temperatures accentuate the ''roughness'' of the probability landscape. Such distributions are commonly used for approximately sampling from smoothed versions of distributions from which it is hard to sample (case with $\beta < 1$) or in simulated annealing for approximate optimization ($\beta > 1$). In this article, we will focus on the former scenario.

The rest of this article is organized as follows. In Section 4, we present an ILP for finding the optimal topology T_* given the labels ϕ for the set of internal nodes. In Section 5, we present our Monte Carlo Markov Chain (MCMC) algorithm for sampling phylogenies in $\mathcal{T}(\chi, \alpha)$, and show that, for the CFN model, the proposed Markov chain mixes in a number of optimization steps polynomial in the number of taxa and sequence length at sufficiently high temperatures. In Section 6, we present results from experiments performed on simulated data sets. In Section 7, we summarize the main contributions of the work and discuss future directions.

4. ILP FOR SOLVING THE DEGREE CONSTRAINED SPANNING TREE PROBLEM

Each phylogeny is specified by the set ϕ of $N-2$ internal nodes and the set χ of N taxa labeled according to the Diaconis-Holmes convention. Their convention provides valuable information regarding which nodes are potential descendants for a given internal node. We root the tree at internal node $2N-2$ and initially connect each pair of vertices (both taxa and internal nodes) by a directed edge. Edges between taxa are removed and each edge between an internal node and a taxon is directed towards the taxon. Edges between two internal nodes are directed towards the node with smaller label. The internal node with the largest label $(2N-2)$ has all edges directed away from it, from which we have to choose three, and each taxon has all edges directed towards it, from which we have to choose one. For each remaining internal node, we have to choose one edge directed towards it (from its parent) and two edges directed away from it (towards its children). An edge directed from vertex u to v corresponds to a Boolean variable $s_{u,v}$ with edge cost $w_{uv} = -\ln[\int_0^\infty L(e=(u,v))Pr(\tau_e)d\tau_e]$. The set of all such edges and the vertices form the graph G. Since the taxa are assigned labels in arbitrary order, we will try to find the minimum cost phylogeny over all possible orderings of the taxa. The following ILP finds the minimum cost tree compatible with these in and out degree constraints,

$$
\begin{aligned}
\text{Minimize} \quad & \sum_{(u,v)\in G} w_{uv} s_{u,v} & & \\
\text{subject to} \quad & \sum_{v} s_{v,u} = 1 & & \forall u \in G \setminus \{2N-2\} \\
& \sum_{v} s_{u,v} = 2 & & \forall u \in \phi \setminus \{2N-2\} \\
& \sum_{v} s_{2N-2,v} = 3 & & \\
& s_{u,v} \in \{0,1\} & & \forall (u,v) \in G
\end{aligned} \tag{5}
$$

Lemma 1. *The ILP in equation 5 finds the minimum cost directed spanning tree given the edge costs w_{uv}.*

Proof. To prove the correctness of our method we show all feasible solutions to this ILP are acyclic. The degree constraints will then ensure that any feasible solution corresponds to a connected subgraph with no cycles, implying a tree. Suppose for a contradiction a feasible solution contains a cycle over a subset $A \subseteq G$. Since G is finite and elements of G are ordered (the directionality of the edge representing which vertex is a potential descendant of another vertex in G), A must contain a vertex v such that all vertices in $A \setminus \{v\}$ are ancestors of v. The only way to obtain a cycle over A is for v to be connected to two (or more) ancestors. But this would violate the in-degree constraint in the ILP. ■

5. MIXING TIME FOR THE CAVENDER-FARRIS-NEYMAN MODEL

In this section, we use the CFN model for binary sequences to establish some theoretical results regarding the convergence of the heated Markov chain. While we prove rapid mixing only for the CFN model, a special case of the class of likelihood model described above, we note that the proof will apply

trivially to some generalizations of the CFN (e.g., non-binary data) and that the sampling technique itself applies to the full class of likelihood functions. The CFN model assigns an edge probability p_e to each edge, such that the likelihood for k mutations along e is $p_e^k(1-p_e)^{M-k}$. We will use the Hamming distance between two sets of internal nodes as a distance measure over the space $\mathcal{T}(\chi, \alpha)$. We will identify each set of internal nodes ϕ with the minimum cost tree $T(\phi)$ obtained by the method in Section 4. Furthermore, for the purpose of this section, we will restrict ourselves to using the most likely branch length for each phylogeny. Given a phylogeny $T(\phi)$, we will call the set of all phylogenies at a Hamming distance 1 the neighborhood of $T(\phi)$ (represented by $Nbd(\phi)$). We can think of the space $\mathcal{T}(\chi, \alpha)$ as a graph with each phylogeny as a vertex and edges connecting each pair of neighboring phylogenies. The Markov chain is defined by nearest neighbor moves over $\mathcal{T}(\chi, \alpha)$ (Fig. 1). We have the following bound on the change in the likelihood function at each step of the Markov chain.

Lemma 2. *For any two neighboring phylogenies* $u \in Nbd(v)$, $(eM)^{-3\beta} \leq \frac{L(u)}{L(v)} \leq (eM)^{3\beta}$

Proof. Given any edge $e = (u, v)$ with branch length $l = -\ln(1-2p)$, and $d_h(u, v) = k$, the likelihood is given by $(p^k(1-p)^{M-k})^{\beta}$. The optimal branch length maximizing this likelihood can be solved as $l = -\ln(1-2k/M)$. If we perturb one character for one internal node (say u), the maximum fractional change in edge likelihood is $((1-1/M)^{M-1}/M)^{\beta} > (1/eM)^{\beta}$ when $k = 0$. Since each internal node has three edges, the maximum change in likelihood for any tree topology is $(eM)^{3\beta}$. Also, the likelihood for each tree topology is a lower bound on the optimal likelihood. If we consider the topology that is optimal at u with likelihood $L(u)$, then we get an upper bound $(eM)^{-3\beta} \leq \frac{L(u)}{L(v)} \leq (eM)^{3\beta}$. ■

This previous result is sufficient to ensure rapid mixing at sufficiently high temperatures. We use path coupling method of Bubley and Dyer (1997) to prove this result. To readers unfamiliar with path coupling, we recommend the text by Levin et al. (2008) for numerous examples. We will use the Hamming distance $d_h(X, Y)$ between two phylogenies X and Y as a distance metric. Path coupling arguments are based on establishing a coupling of Markov chain moves between each pair of nearest neighbors such that the distance between them decreases on average at each iteration of the Markov chain. For completeness, we state the main lemma from Bubley and Dyer (1997).

Lemma 3. *Let M be Markov chain over a graph* $G(V, E)$ *and let* ρ *define a set of edge distances over* G *such that* $\rho(e) \geq 1 \ \forall e \in E$. *Furthermore, let* d *be a distance metric over* V *such that given any pair of vertices* $u, v \in V$,

$$d(u, v) = min_{P(u, v)} \sum_{(x, y) \in P(u, v)} \rho(x, y) \tag{6}$$

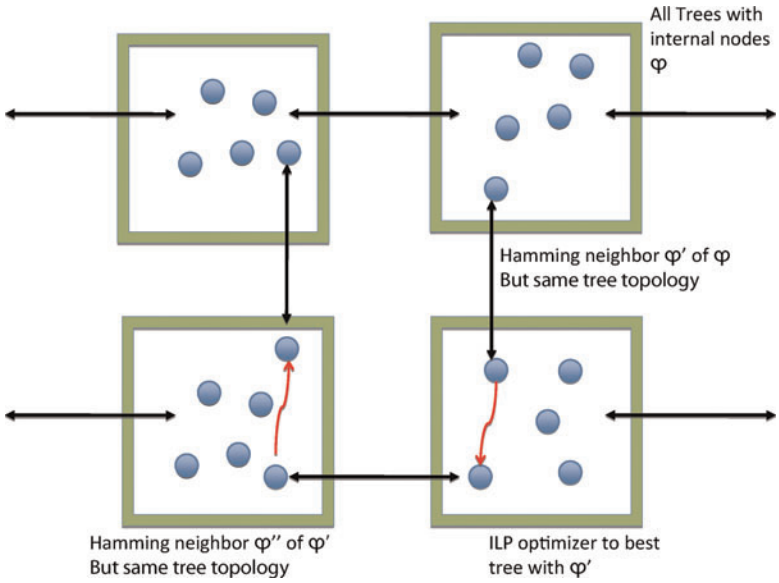

FIG. 1. A typical transition in our modified sampler: In clockwise order, starting from a phylogeny ϕ (**top right**), the black arrow represents a move to a neighboring ancestral sequence ϕ' (**bottom right**), followed by an optimization step (red wiggly line) to reset the topology to be a most likely one. This is followed by another random perturbation of ϕ' to a neighboring ancestral node ϕ'' (**bottom left**).

where the minima is taken over the set of all paths $P(u, v)$ between u and v. Suppose that for all edges $\bar{E} = \{(u, v) | \rho(u, v) = d(u, v)\} \subseteq E$ there exists a coupling of Markov processes $\{u_t = M^t.u\}$ and $\{v_t = M^t.v\}$, such that the following bound holds

$$d(u, v) - E[d(M.u, M.v)] \geq \alpha d(u, v) \ \forall (u, v) \in \bar{E} \tag{7}$$

for some $\alpha > 0$, where $M.v$ represents the state after one step of the Markov chain starting at v. Then the total variation distance $\Delta(t) = max_{u \in V} |M^t.u - \pi| \leq e^{-\alpha t} \ln [D]$, where D is the diameter of G and π is the invariant probability measure on V.

Theorem 1. *For the CFN model, the heated Markov chain mixes in time $O(NM \ln[NM]/(1 - \tanh(3 \ln(M)\beta)))$ for $\beta < -\ln[1 - 1/(NM + 1)]/3 \ln(M)$*

Proof. Suppose we have two random processes X_t and Y_t, evolving according to the Markov chain over phylogenies. We will concatenate the bit strings representing the internal nodes into one string of NM bits for each random variable. Suppose at time $t = 0$, X_0 and Y_0 differ in the k^{th} bit. We define the coupling as follows—Select any bit b (representing a character for one of the internal nodes) uniformly at random. If $b = k$, with probability $1/2$ flip the k^{th} bit of variable X_0 and hold the state for Y_0 or vice versa; if $b \neq k$, with probability $1/2$ select identical proposal states for both random variables X_1 and Y_1 and with probability $1/2$ do nothing. If $b = k$, with probability $> 1/2$ both variables converge in one step. For each of the other choices, the Hamming distance either stays the same or increases by one. In each instance, the probability that Hamming distance decreases is

$$d_h(X_0, Y_0) - E[d_h(X_1, Y_1)] \geq \frac{1/2 - (NM - 1)\frac{1}{2}|Pr[X_1 accepts] - Pr[Y_1 accepts]|}{NM} \tag{8}$$

Now, using lemma 2 we get the bounds

$$Pr[X_1 accepts] = \frac{L(X_1)}{L(X_1) + L(X_0)} \leq \frac{1}{1 + e^{-3 \ln (M)\beta}} \tag{9}$$

and

$$Pr[Y_1 accepts] = \frac{L(Y_1)}{L(Y_1) + L(Y_0)} \geq \frac{e^{-3 \ln (M)\beta}}{1 + e^{-3 \ln (M)\beta}} \tag{10}$$

combining these three equations we get

$$\begin{aligned} d_h(X_0, Y_0) - E[d_h(X_1, Y_1)] &\geq 1/2\left(1/NM - \left(\frac{1}{1 + e^{-3 \ln (M)\beta}} - \frac{e^{-3 \ln (M)\beta}}{1 + e^{-3 \ln (M)\beta}}\right)\right) \\ &= 1/2(1/NM - \tanh (3 \ln (M)\beta)) \end{aligned} \tag{11}$$

This implies for $\beta < -\ln[1 - 1/(NM + 1)]/3 \ln(M)$, the distance between neighboring phylogenies decreases in expectation at a rate greater than $\alpha = 1/2NM(1 - NM \tanh(3 \ln(M)\beta)$. Finally, using the path coupling lemma 3 and the fact that our graph has diameter NM, we get the following condition for the total variation distance $\Delta(t)$ between the distribution of X_t and Y_t for $\alpha > 0$.

$$\Delta(t) = Pr[X_t \neq Y_t] \leq e^{-\alpha t} NM \tag{12}$$

and the mixing time to reach a total variation distance 1/2 is $\tau = \ln[2NM]/\alpha$. ■

While we have provided a proof just for the CFN model, the basic technique can be extended trivially to multi-state models, such as the Jukes-Cantor model.

6. EXPERIMENTS

We implemented our method in C++ and used the gnu linear programming kit GLPK for solving the ILP. The Markov chain was simulated using the *replica sampling* heuristic as described next. In each of the

experiments reported here, three Markov chains were simulated independently, at different values of β, for some user defined time steps N_1. One chain was always maintained at $\beta_0 = 1$ while the i^{th} chain was heated to $\beta_i = (1 + i\delta)^{-1}$, for some user defined value of δ. At each optimization step, the branch lengths were set to the maximum likelihood values, although in principle our method can sample from the full posterior distribution of branch lengths at each step. After N_1 steps of each chain, a pair of chains was picked uniformly at random and an exchange was proposed, followed by the usual Metropolis accept/reject criterion evaluated at the temperature of the colder of the two chains. After every N_{ex} attempts at exchanging states between the Markov chains, a measurement was made. We used data simulated using the CFN model on a user defined tree topology for our experiments, as described in the following section.

Our goals in validation were to verify that the model runs efficiently for moderate sized trees and to demonstrate its ability to ask questions about the ancestral likelihood function. For this purpose, we conducted a small study measuring the accuracy with which the true source tree of each data set can be inferred from the data for varying input sizes. We can assess this uncertainty by examining how often each bipartition in the source tree of a given data set occurs over the sample of trees. We quantify this measure of bipartition mismatch by the mean number of bipartitions that differ between observed tree and source tree across samples.

6.1. Data sets

We report two sets of experiments on simulated data from 10, 25, and 35 taxa trees. Each set was prepared as follows: A tree topology with N leaves was generated by randomly choosing a matching by enumerating $2N - 2$ points in random order and matching successive points. Each edge was assigned a branch length by generating an exponentially distributed random number with user defined mean (mean was fixed by specifying the edge probability) and 100 characters were simulated using the CFN model starting from the root $2N - 2$. We initialized the Markov chain simulator by the simple heuristic of starting with the set of leaves $S = \chi$ and true ancestral nodes. One tenth of the characters in each ancestral node were then randomly perturbed. This process was repeated independently for each chain participating in the replica exchange. All the experiments we report here had edge probabilities no more than 0.1, so these perturbations result in a fairly random initial state.

We first report results on experiments where we varied the number of taxa, keeping all other simulation and sampling parameters constant. In each case, data was simulated on trees with each edge probability fixed at 0.1, so, on average, one in ten characters mutated along each edge of the tree. For the Monte Carlo sampling step, we used three coupled chains maintained at temperatures (or β^{-1}) = 1, 1.01 and 1.02. The temperatures were chosen heuristically (Huelsenbeck and Ronquist, 2001). After every 10 steps, two chains were picked at random and an attempt at swapping their states was made. These experiments were performed to assess the feasibility (both in run time performance and Markov chain convergence) of the proposed method.

For the second set of experiments, we fixed the number of taxa to 25 and simulated data for edge probabilities values of 0.01, 0.05, and 0.1. These experiments were performed to estimate the variation in the uncertainty in inferring the true topology as well as the rate of convergence of the Markov chain.

6.2. Results

Figure 2 shows inferred likelihoods per Monte Carlo step for each tree. The plot reveals that the sampler relaxes to a high likelihood tree in each case. Further, the number of steps until the likelihood plateaus

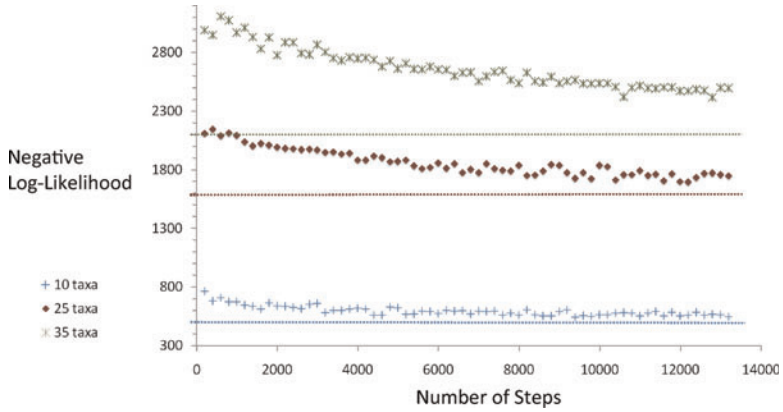

FIG. 2. Inferred negative log-likelihood as a function of Monte Carlo step for trees of 10, 25, and 35 taxa.

TABLE 1. RUN TIME AND MISMATCH BETWEEN TRUE AND SAMPLED TREE TOPOLOGIES FOR THREE INPUT SIZES

No. of taxa	*ILP steps per second*	*Average mismatch*
10	728.21	0.39
25	70.45	2.38
35	29.26	2.58

increases monotonically with the number of taxa, as expected. However, due to the low temperature and our use of the replica exchange heuristic in these empirical tests, we cannot assert with certainty that these chains are well-mixed. The dashed horizontal lines, representing the ancestral likelihood for the source tree, seem to indicate that each chain is quite probably close to equilibrium.

Analysis of run times further shows the method to be very fast in practice despite the need for solving a hard optimization problem at each step. Table 1 shows mean run-times expressed in numbers of Monte Carlo steps solved per second. Run time does increase with numbers of taxa, but is still more than 1500 steps per minute for 35 taxa trees. The method is thus practical for tens of thousands of steps of Markov chain sampling on moderate problem sizes. For instance, each run presented in Figure 2 took less than 23 minutes.

Table 1 also shows the results of the uncertainty analysis. While the most likely ancestral tree is known to be statistically inconsistent in general, we see that the sampler is extremely efficient in identifying true bipartitions for these data sets.

The second set of experiments probes the ancestral likelihood landscape as we vary the mutation probabilities while keeping the number of taxa fixed at 25. Figure 3 shows the negative log-likelihood plots for three experiments with varying edge lengths. Once again we find that the sampled trajectories relax fairly rapidly to a high likelihood tree that is close to the likelihood of the source tree. Data sets with shorter branches (lower mutation probability) seem to relax faster, although we once again cannot assert that with certainty.

Table 2 gives us additional insight into the likely dynamics of the Markov chain. As the mutation probabilities along tree edges increase, so does the accuracy of inferring the source tree. On the other hand, looking at Figure 3 seems to suggest that the sampler relaxes more rapidly for high edge probability data. This set of experiments thus tends to strongly suggest that near equilibrium, the ability of the sampler to estimate the source tree deteriorates as edge probabilities become small. This agrees with our intuition that given two speciation events, it should be relatively easier to infer the order in which they occur if the sequences at the two branch points are ''well separated'' (i.e., the branch length between the internal edges is large). At the same time, in the neighborhood of trees with long branches, the ancestral likelihood is comparably ''flatter'' (for the same reason that the peak of the likelihood curve in Lemma 2 is at the

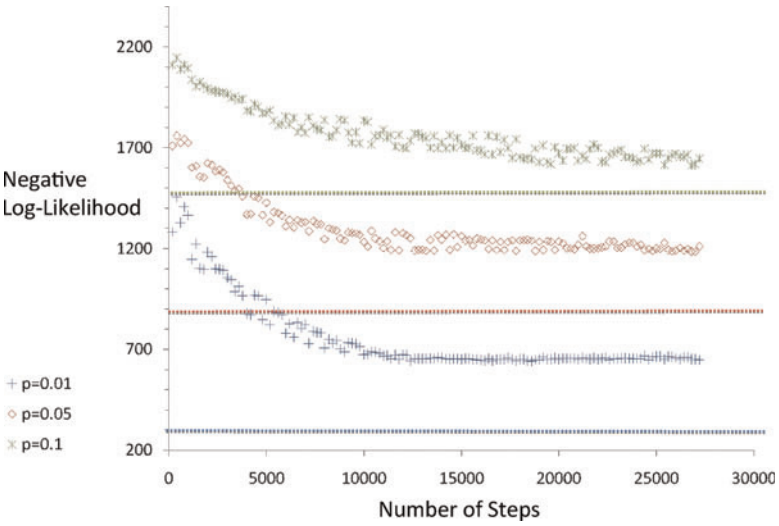

FIG. 3. Inferred negative log-likelihood as a function of Monte Carlo step for trees with 25 taxa and varying edge probabilities.

Table 2. Rejection Ratio and Mismatch between True and Sampled Tree Topologies for Three Different Edge Probabilities for 25 Taxa Data Sets

Average edge probabilities	*Average mismatch*	*Average rejection ratio*
0.01	14.05	0.97
0.05	4.82	0.94
0.10	2.93	0.87

shortest branch). As a result, the Markov chain moves about the state space comparably faster (leading to a lower rejection ratio in Table 2), but at each new node the optimal tree does not differ much from the true tree. On the other hand, for extremely short branches, the most likely ancestral sequences are closely clustered together around a smaller region of the state space and Markov chain rarely ventures out of this central region (leading to a high rejection ratio). However, since most ancestral nodes have largely similar sequences for the low edge probability case, it is relatively more common to swap the order of bifurcation events connecting two ancestral nodes, leading to a higher mismatch between the average estimated topology and the source tree. For the 0.01 edge probability case, the source tree had numerous higher degree internal nodes (or edges with zero branch lengths) and that may be the likely cause for the mismatch. In phylogenetics, it is well known that near polytomies or very closely spaced bifurcation events are generally harder to reconstruct from fixed length sequences (Rokas and Carroll, 2006) and our experiments seem to suggest a similar effect on our ancestral likelihood sampler.

6.3. Discussion

Our experiments show that our method provides an efficient sampler over the ancestral likelihood model for relatively large numbers of taxa. Although the method involves solving a formally intractable problem at each step of Markov chain sampling, in practice it proves extremely fast for even moderately large data sets. Furthermore, our novel formulation allows us to examine aspects of the tree likelihood distribution inaccessible to standard samplers. Maximum ancestral likelihood has been known to be a statistically inconsistent estimator of the tree topology in general. Our experiments seem to indicate that the discrepancy between the topology of the average ancestral reconstruction and the source tree is small and robust to the number of leaves in the tree, except when the source tree has extremely short branch lengths.

7. CONCLUSION

We have developed a novel approach to phylogenetics designed to leverage methods for fast combinatorial optimization to efficiently sample over the space of phylogenetic trees under standard likelihood models. The method depends on an alternative formulation of phylogenetic likelihood to enable sampling over internal node states instead of tree topologies. The work establishes a new approach to performing efficient, accurate sampling over phylogenies and to establishing mixing time bounds for such sampling in practice. To demonstrate its theoretical value, we have established mixing time bounds for the important practical case of the CFN model. These bounds can be extended to some generalizations of that model and provide a new strategy for establishing provable bounds on more general likelihood models. We further demonstrate the practical efficiency and utility of the method through a study of uncertainty in topology inference across samples under a standard likelihood function. The ideas developed here for efficient optimization-based sampling may be applicable to many similar problems involving sampling over likelihoods of combinatorial objects.

ACKNOWLEDGMENTS

This work was supported by U.S. National Science Foundation (grant 0612099 to N.M., G.B., R.R., and R.S.). Additionally, this work was supported by U.S. National Institutes of Health (grants 1R01CA140214 and 1R01AI076318 to R.S.).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Bubley, R., and Dyer, M. 1997. Path coupling: a technique for proving rapid mixing in Markov chains. *Proc. 38th Annu. Symp. Found. Comput. Sci.* 223–231.

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

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

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

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer Publications, Sunderland, MA.

Huelsenbeck, J.P., and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17, 754–755.

Levin, D.A., Peres, Y., and Wilmer, E.A. 2008. *Markov Chains and Mixing Times.* American Mathematical Society, Providence, RI.

Misra, N., and Schwartz, R. 2008. Efficient stochastic sampling of first-passage time with applications to self-assembly simulation. *J. Chem. Phys.* 129, 204109.

Mossel, E., and Vigoda, E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* 309, 2207–2209.

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems, 1–27. *In* Gupta, S.S., and Yackel, J., eds. *Statistical Decision Theory and Related Topics.* Academic Press, New York.

Rokas, A., and Carroll, S.B. 2006. Bushes in the Tree of Life. *PLoS Biol.* 4, e352.

Stefankovic, D., and Vigoda, E. 2007. Pitfalls of heterogeneous processes for phylogenetic reconstruction. *Syst. Biol.* 56, 113–124.

Stefankovic, D., and Vigoda, E. 2010. Fast convergence of MCMC algorithms for phylogenetic reconstruction with homogeneous data on closely related species. Available at arXiv:1003.5964.

Address correspondence to:
Dr. Russell Schwartz
Department of Biological Sciences
Carnegie Mellon University
Pittsburgh, PA 15213

E-mail: russells@andrew.cmu.edu