Phylogenetic Inference for Binary Data on Dendograms Using Markov Chain Monte Carlo

Bob Mau; Michael A. Newton

Journal of Computational and Graphical Statistics, Vol. 6, No. 1 (Mar., 1997), 122-131.

Phylogenetic Inference for Binary Data on Dendograms Using Markov Chain Monte Carlo

Bob MAU and Michael A. NEWTON

Using a stochastic model for the evolution of discrete characters among a group of organisms, we derive a Markov chain that simulates a Bayesian posterior distribution on the space of dendograms. A transformation of the tree into a canonical cophenetic matrix form, with distinct entries along its superdiagonal, suggests a simple proposal distribution for selecting candidate trees "close" to the current tree in the chain. We apply the consequent Metropolis algorithm to published restriction site data on nine species of plants. The Markov chain mixes well from random starting trees, generating reproducible estimates and confidence sets for the path of evolution.

Key Words: Labeled history; Metropolis–Hastings algorithms; Phylogeny reconstruction.

1. INTRODUCTION

"A phylogeny is a branching tree diagram showing the course of evolution in a group of organisms"—Felsenstein (1983, p. 246).

Inference about phylogenies is often based on comparable molecular data, aligned nucleotide sequences being the canonical example. Given such data, one goal of life scientists is to reconstruct the phylogeny whence these organisms arose. Botanists and zoologists called systematists use the phylogeny to aggregate organisms into monophyletic groups for taxonomic purposes (hence, the generic term taxa).

Existing reconstruction techniques in the biological sciences each purport to find the best phylogeny compatible with the data, yet seldom in practice are those solutions the same. Among the more popular methods, one can categorize those using distance matrices in numerical taxonomy as clustering algorithms, while maximum parsimony and maximum likelihood each optimize an objective function over the space of trees. Felsenstein (1988) provided a comprehensive review of current methods.

Phylogenetic inference can also be dichotomized functionally. Maximum likelihood, maximum parsimony, and distance-matrix methods are practical for data sets relating

---

Bob Mau is Research Associate, Department of Genetics, University of Wisconsin–Madison, Madison, WI 53706; e-mail: robertm@stat.wisc.edu. Michael A. Newton is Associate Professor, Departments of Statistics and Biostatistics, University of Wisconsin–Madison, Madison, WI 53706; e-mail: newton@stat.wisc.edu.

Journal of Computational and Graphical Statistics, Volume 6, Number 1, Pages 122–131

many taxa, but beyond point estimates, these methods do not produce exact inferences. Measures of uncertainty rely exclusively on computer-intensive and approximate bootstrap analyses (Felsenstein 1985). More recently developed techniques, such as the phylogenetic invariants (see Evans and Speed 1993) and a Bayesian approach such as Sinsheimer, Lake, and Little (1995), allow exact inference, but mathematical and computational complexity have limited these methods to extremely small problems (no more than four taxa).

We elect a Bayesian approach and use Markov Chain Monte Carlo (MCMC) methods to develop a computationally feasible technique that meets practitioners' demands for methods capable of handling more taxa while keeping statistical inference on a sound footing. Our method assesses uncertainty in the estimate without bootstrapping by constructing a probability distribution on the space of trees. Measurement of uncertainty in our optimal tree accompanies tree construction. By contrast, other methods capable of handling as many taxa must first find the optimal tree, generate bootstrap samples from the data, and rerun the original estimation procedure on each bootstrap sample.

1.1 TREE TERMINOLOGY

Mathematically, a tree Ψ is a connected graph (V, E), with vertex set V and edge set E, characterized by the absence of loops. Vertices are classified as leaves if they are attached to a single edge, and internal nodes otherwise. In rooted binary trees, each internal node has exactly three edges, with the exception of the root node ρ, whose position relative to the leaves determines the direction of time. For each $v \in V\backslash\{\rho\}$, there is a unique parent node $\sigma(v)$, closer to ρ and connected to v by an edge in E. Identifying the leaves with present-day taxa, internal nodes become the common ancestral populations whence the current organisms descended. At each such node, two nodes have coalesced into their most recent common ancestor. The ordering in which these merges occur defines coalescent levels, and the corresponding temporal intervals between consecutive merges comprise the coalescent times.

Figure 1 depicts a seven-taxa example. The shape of the tree, determined by which pairs of nodes coalesce, is called the topology. Different orderings of coalescent levels within a particular topology generate distinct labeled histories, alternative characterizations of shape. We restrict attention to trees having contemporaneous leaves, called dendograms, where a simple additive relationship exists between the time along edges, called branch lengths, and the coalescent times. Such a tree Ψ can be specified either by its topology and branch lengths or its labeled history and coalescent times. The numbers of topologies and labeled histories grow rapidly as the number of taxa n increases, at rates $(2n-3) \times (2n-5) \ldots \times 1$ and $n! \times (n-1)!/2^{n-1}$ respectively (e.g., Felsenstein 1978).

2. A STOCHASTIC MODEL FOR DISCRETE CHARACTERS

Evolution has two stochastic components that can be modeled as Markovian processes: the branching processes of speciation and extinction that generated the organisms

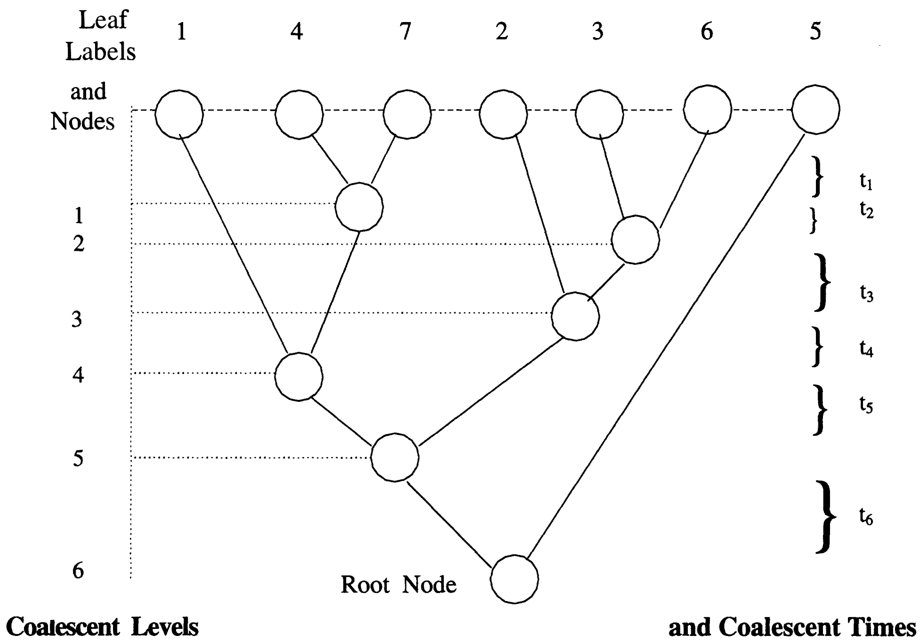

Figure 1. Graphical Depiction of a Sample Phylogeny With Seven Leaves Representing Present-Day Taxa.

of interest, and the changes of molecular characters along each lineage so created that ultimately result in the observed data. We do not model the branching process, choosing instead to treat the phylogeny as a fixed parameter in a model for the evolution of that data.

Generic discrete data take on values from a set $\mathcal{D}$ of size d. Typical data sets on n taxa consist of $n \times N$ matrices, where N is the common number of sites (also called locations) at which data are observed. We adopt a standard Markovian model for the evolution of the data along a known tree (see Goldman 1990). Let $y = \{y_{vk},\ v \in V,\ k = 1,\dots,N\}$ denote a realization in $\mathcal{D}^{2n-1}$ at each site k. Along each branch, transition probabilities are $P_{i,j}(t,\beta)$ for jumps from state i to state j in time t, where β contains the infinitesimal change parameters of the process. Indexing by the vertex set suggests an alternate notation, $p(y_{vk}|y_{\sigma(v)k}t_v,\beta)$, where $y_{\sigma(v)k}$ is the state of the parent node of v at the kth site, y_{vk} is the corresponding state at v, and t_v is the intervening branch length. If the sites are iid, the probability of a particular realization y can be written as

$$P_{\Psi,\beta}(Y = y) \quad = \quad \prod_{\text{all sites } k} \left(\pi_0(y_{\rho k}) \prod_{v \in V \setminus \rho} p(y_{vk}|y_{\sigma(v)k}, t_v, \beta) \right),$$

where $\pi_0(y_{\rho k})$ is an initial distribution on the ancestral states.

Statistical analysis requires an inversion of the probability model. What had been known, the tree and ancestral root state, become candidates for estimation, and what had been the end product of the probability model is now the observed data. Because only data on the leaves are observed, we marginalize over all possible combinations of $\mathcal{D}$-

values at the internal nodes, denoting this collection of matrices by $\mathcal{M}(\mathcal{D})$. Partitioning y into leaf and internal node components, the likelihood of Ψ is written as

$$L(\Psi, \beta \mid y^{\mathcal{L}}) = \sum_{u \in \mathcal{M}(\mathcal{D})} P_{\Psi,\beta}(Y = (y^{\mathcal{L}}, u)). \qquad (2.1)$$

Bayesian analysis requires a prior distribution on the parameter space of the model. The prior should reflect the researcher's prior beliefs on how likely particular parameter values are before the data have been observed. The posterior is a probability measure on the parameter space representing the scientist's beliefs in light of new evidence revealed by the data. Once a prior is specified, we denote the corresponding posterior distribution of interest by $\pi(\Psi)$. Though one cannot easily produce a random sample from the posterior, Metropolis–Hastings algorithms can be constructed to sample dependently from it in an efficient manner. The general algorithm is quite simple: take an arbitrary transition mechanism that proposes trees Ψ^* with density $Q(\Psi, \Psi^*)$, conditional on being at Ψ. We subject a draw from $Q(\Psi, \cdot)$ to a randomized test, accepting the draw Ψ^* with probability

$$\alpha(\Psi, \Psi^*) = \min\left(1, \frac{\pi(\Psi^*) Q(\Psi^*, \Psi)}{\pi(\Psi)\; Q(\Psi, \Psi^*)}\right)$$

otherwise staying put at Ψ.

Standard MCMC theory states that such chains converge to the $\pi(\Psi)$, the posterior distribution of interest, when Q is irreducible (see Tierney 1994). The practical consequence of theory is that

$$\frac{1}{K}\sum_{i=1}^{K} f(\Psi_i) \rightarrow E_\pi[f(\psi)] \quad \text{a.s.},$$

where the Ψ_i are the elements in the chain, and f is some function whose expectation is desired. For example, when f is the indicator function of a particular labeled history, the previous ergodic theorem states that the relative frequency of that labeled history in the Markov chain converges to its marginal posterior probability. Labeled histories are collapsed into their corresponding topologies, and the topology with the highest relative frequency is our reconstruction. Credible sets, the Bayesian counterparts of confidence regions, are comprised of those topologies (labeled histories) occurring most frequently in the simulation.

In the next section, we introduce an algorithm for the efficient exploration of posterior distributions over the space of trees with contemporaneous leaves.

3. APPLICATION TO PHYLOGENY

Standard tree representations may permit efficient computation of the likelihood, but they fail to suggest simple and workable proposal distributions Q on the state space of trees. To discover a proposal distribution that nominates candidate trees near the incumbent, yet is irreducible, we must consider a representation based on a metric.

Table 1. Cophenetic Matrix With the Labels in Standard Numerical Order

Labels	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1	0	3.2*	3.2	2.3	4.7	3.2	2.3
2		0	1.8*	3.2	4.7	1.8	3.2
3			0	3.2*	4.7	1.1	3.2
4				0	4.7*	3.2	.8
5					0	4.7*	4.7
6						0	3.2*
7							0

3.1 THE COPHENETIC MATRIX REPRESENTATION

Following Lapointe and Legendre (1991), binary trees with n leaves can be transformed into $n \times n$ cophenetic matrices, whose array elements $d_{i,j}$ represent the time to coalescence for each pair of leaf nodes i and j. These matrices are symmetric, with zeroes along the diagonal, and the space of matrices is metrizable under the standard Euclidean norm. The following numerical example for the coalescent times, $\mathbf{T} = (.8,.3,.7,.5,.9,1.5)'$, yielding cumulative coalescent times, $\mathbf{T}^{\text{cum}} = (.8,1.1,1.8,2.3,3.2,4.7)'$, will be exercised throughout this section. Our illustrative tree has a cophenetic matrix representation in Table 1. Observe that all off-diagonal elements are elements of $\mathbf{T}^{\text{cum}}$. In particular, the starred entries, $(3.2, 1.8, 3.2, 4.7, 4.7, 3.2)$, form the superdiagonal, defined as the diagonal of the submatrix formed by deleting the first column and the nth row of the cophenetic matrix. Redundant lower triangular entries are suppressed.

The cophenetic matrix form is not unique, but depends on the ordering of the labels. Labels can always be rearranged so that the superdiagonal elements, the $\{d_{i,i+1}, i = 1, \ldots, n-1\}$ of the corresponding cophenetic matrix, are distinct. The truth of this statement is evident from the depiction of a tree in Figure 1, wherein no branches intersect. The order of those labels: $(1, 4, 7, 2, 3, 6, 5)$ is one such permutation leading to a superdiagonal with distinct entries. Each internal node in Figure 1 provides two possible left-right orientations for the descendent branches, resulting in 2^{n-1} distinct arrangements of the labels that preserve both the coalescent structure and the coalescent times, and hence the tree itself. We call all such permutations the set of canonical orderings. Another such ordering, (5,7,4,1,2,6,3), generates the canonical cophenetic matrix in Table 2, where elements above the superdiagonal are easily determined from the ultrametric property of cophenetic matrices (see Lapointe and Legendre 1991). The crucial observation is

Table 2. Canonical Cophenetic Matrix With Labels Permuted in One of the Canonical Orderings

Standard labeling ↓ *permuted*	*1* ↓ *5*	*2* ↓ *7*	*3* ↓ *4*	*4* ↓ *1*	*5* ↓ *2*	*6* ↓ *6*	*7* ↓ *3*
5	0	4.7	4.7	4.7	4.7	4.7	4.7
7		0	0.8	2.3	3.2	3.2	3.2
4			0	2.3	3.2	3.2	3.2
1				0	3.2	3.2	3.2
2					0	1.8	1.8
6						0	1.1
3							0

that a canonical ordering and the corresponding superdiagonal completely determine a dendogram.

3.2 A PROPOSAL DISTRIBUTION ON DENDOGRAMS

We construct a two-stage proposal distribution that acts on the family of canonical cophenetic matrices in the following manner. Starting at a given tree, Q_1 produces a random draw from among the 2^{n-1} canonical orderings of that tree by independently flipping a fair coin to determine the orientation of the descendent branches at each internal node, generating a superdiagonal denoted by $\{a_i = d_{i,i+1}, i = 1, \ldots, n-1\}$. In the second stage, Q_2 acts on the a_i componentwise as follows: for a fixed $\delta > 0$, each a_i^* is drawn independently from

$$Q_{2,i}(a_i) \sim \begin{cases} U(a_i - \delta, a_i + \delta) & \text{if } a_i > \delta \\ \frac{\delta - a_i}{\delta} U(0, \delta - a_i) + \frac{a_i}{\delta} U(\delta - a_i, a_i + \delta) & \text{if } a_i < \delta. \end{cases}$$

The first case is a uniform density with positive support; the second case is a mixture of two uniform densities with positive support, derived from a single uniform density centered near zero, with the probability mass on negative values folded over onto its positive reflection. We can write this compactly as $a_i^* = \mid U_i \mid$, for U_i uniformly distributed on $(a_i - \delta, a_i + \delta)$. The complete second stage is

$$Q_2(a_1, \ldots, a_{n-1}) = \prod_i^{n-1} Q_{2,i}(a_i).$$

Note that δ is a tuning parameter which determines how "far" one can jump from the current tree, and hence can be used to modulate the overall acceptance rate of the chain.

The action of the second stage is illustrated in Figure 2 on a six leaved tree with selected leaf ordering (4,6,3,2,1,5). With leaf nodes equally spaced along a horizontal axis, perpendiculars whose lengths equal the cumulative coalescent times between adjacent leaves are dropped from the interleaf midpoints. Internal nodes are appended at the base of each perpendicular. The vector of these $n-1$ vertical lengths is precisely the superdiagonal under the chosen ordering of leaf labels. The underlying phylogeny (dotted lines in Figure 2), can be recovered from this graphic by the following inductive recipe. The internal node nearest to the top initiates the induction by defining the first coalescence—the two leaf nodes on either side of it. As each successive level is reached, the corresponding internal node is the coalescence of the two nearest unattached nodes, one from each side. Candidate trees, accessible from the current tree under this label ordering, can be characterized by intervals of size 2δ, centered at the current internal nodes (demarcated by bold lines in Fig. 2), that constrain the repositioning of the internal nodes. Observe that Q_2 can propose candidates with different labeled histories, and even different topologies, than the current tree. For example, when the perpendicular between $\{3, 2\}$ is maximally increased while that of $\{4, 6\}$ is simultaneously decreased, a new labeled history is nominated. Likewise, a similar adjustment to the perpendiculars between $\{1, 5\}$ and $\{2, 3\}$, respectively, proposes a different topology where the leaf $\{1\}$ coalesces with the common ancestor of $\{2, 3\}$ before attaching to $\{5\}$.

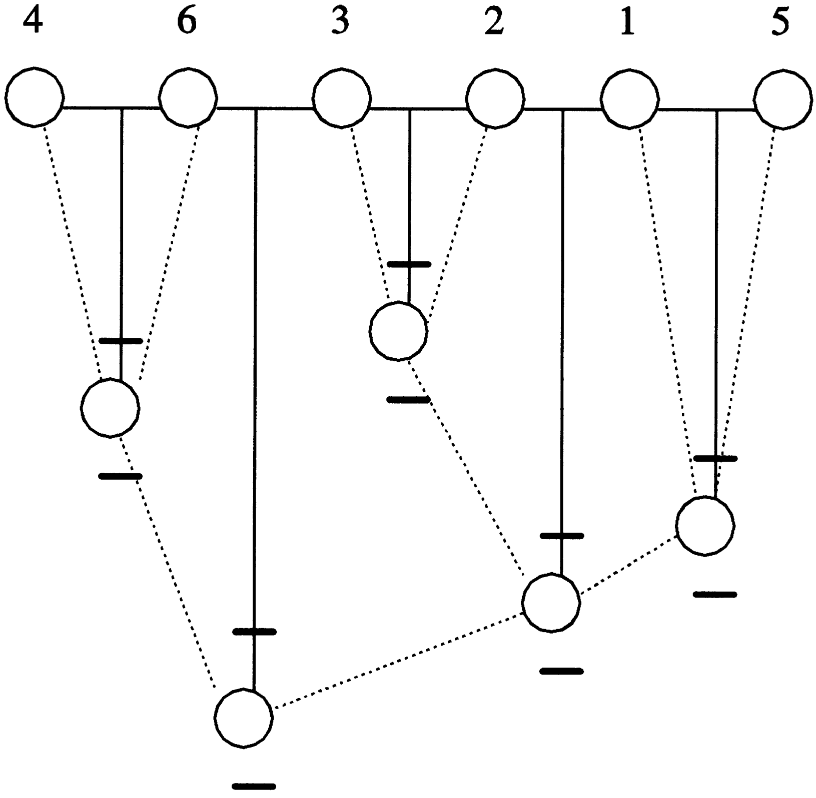

Figure 2. How the Second Stage of the Proposal Distribution Generates Candidates That May Change Shape.

Randomness in the selection of a label ordering is crucial to ensure irreducibility. For example, it is easy to show that any fixed canonical ordering for four taxa restricts Q_2 to move among six of the eighteen possible labeled histories. A proof of irreducibility of the composite driver $Q = Q_1 \circ Q_2$ is given in Mau, Newton, and Larget (1996).

4. A PROBABILITY MECHANISM FOR BINARY DATA

Sytsma and Gottlieb (1986) studied the evolutionary relationships among nine species of the genus Clarkia, weed-like plants indigenous to California. They exposed chloroplast DNA from each specie to 29 restriction enzymes. A restriction enzyme acts on DNA by physically cutting the string in two wherever it recognizes a particular base pattern. At each location where a cut occurs, a restriction site is said to be present. Six hundred nine such positions were identified, of which 490 were noninformative (sites at every species). The informative data were translated into a 9×119 matrix of zeroes and ones. We entertain a simple stochastic model for the evolution of restriction sites; a two-state Markov process with infinitesimal rates λ and μ, representing the intensity of the instantaneous transition from 0 to 1 and 1 to 0, respectively. The matrix of transition probabilities, $P(t)$, satisfies the Chapman–Kolmogorov equation $P'(t) = P(t) \cdot A$, where

A is the generator matrix

$$A = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

Reparameterization of the model by its mutation rate $\theta = \lambda + \mu$ and the ratio of its infinitesimal rates, $r = \frac{\mu}{\lambda}$, yields the following compact representation of the transition probabilities:

$$P_{i,j}(t, \beta = (\theta, r)) = \frac{1}{1+r}\left(r^{1-j} + (-1)^{I\{i \neq j\}} r^i e^{-\theta t}\right)$$

Ignorance priors are placed on the space of labeled histories, coalescent times, and r, suitably restricted to compact subsets. Thus r is updated (via a uniform proposal distribution centered at its current value) as part of the MCMC cycle. Fortuitously, θ is confounded with time, so the branch lengths are proportional to amount of evolution, with proportionality constant θ. Hence, we fix θ throughout the simulation. We derive the initial distribution from biological principles. Finally, the likelihood in Equation 1 must be conditioned on seeing at least one restriction site at each location.

$$L(\Psi, \beta \mid y^{\mathcal{L}}) / \left[1 - p_0(\Psi, \beta)\right]^N,$$

where $p_0(\Psi, \beta)$ is the probability that $\{y_v = 0,\ v \in \mathcal{L}\}$ at one site.

4.1 COMPUTER IMPLEMENTATION AND ANALYSIS

Our computer code produces realizations from the Metropolis algorithm on the space of trees described above. Chains of length 250,000 were subsampled to produce 2,500 stored realizations of labeled histories, total cumulative coalescent times, and infinitesimal rates ratios. The main program generates diagnostics for monitoring the convergence of the chain: the log posterior and cumulative coalescent times of the current tree print to the screen after every 1,000th iteration, as do the intervening average acceptance rates for candidate trees and generator ratios. For this data set, the combination of fixed mutation rate $\theta = .005$ and tuning parameter $\delta = .35$ yields cumulative acceptance rates within the the recommended range (30% to 70%) , determined empirically on the basis of several short preliminary runs. By monitoring the log posterior, burn-in appears to be on the order of 200 iterations. Burn-in is that initial portion of the chain representing draws made before the chain has significantly converged to the posterior. A separate program summarizes simulation output by generating frequency counts of labeled histories, collapsing them into their corresponding topologies, and sorting the topologies by decreasing relative frequency.

Only three topologies, representing approximately 100 labeled histories, are required to construct a 99% credible region for the branching pattern of the phylogeny. In dozens of repeated runs from random starts, the observed relative frequencies deviate by at most ±3% from the reported frequencies for samples of this size. The four most frequently visited topologies, with relative frequencies of 64.9%, 17.9%, 16.8%, and .02%, are shown in Figure 3. Topology III is the reconstruction favored by Systsma and Gottlieb

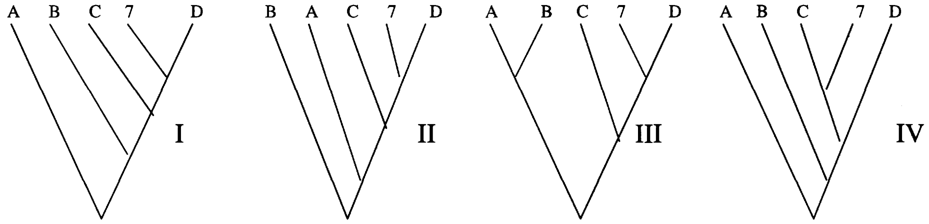

Figure 3. Topologies with Highest Posterior Probability using Markov Chain Monte Carlo, where $A = \{1,2\}$, $B = \{3,4\}$, $C = \{5,6\}$, *and* $D = \{8,9\}$, *Denote Coalescent Pairs Common in the Above Topologies.*

(1986), using maximum parsimony with an outgroup to determine the location of the root. Note that Topologies I–III are the same shape except for the root. Only Topology IV is a different unrooted tree, which minimally reflects the weakness (61%) reported in Sytsma and Gottlieb's (1986) bootstrap analysis of the branch that attaches $\{7\}$. Subsequent computer runs that include the outgroup as a tenth taxa support the rooting in our original estimate (Topology I).

5. DISCUSSION

Our method simulates a Markov chain on the space of dendograms, visiting each labeled history with a long run relative frequency proportional to its marginal posterior probability. For moderate number of taxa and a reasonably discriminating set of data, the vast majority of histories have a near zero posterior probability, and as such will not be hit even in relatively long chains. Although these run lengths are considerably less than the total number of labeled histories, representative samples from the posterior are nonetheless efficiently generated. A single 20-minute simulation on a Sparc 10 workstation generates not only an estimate of the phylogeny for the Clarkia, but a measure of uncertainty in that estimate as well. This result is consistent with the published parsimony solution, but places most of the topological variability in the root, which was fixed under parsimony.

Because much of the genetic data available today are DNA or RNA nucleotide sequences, the logical extension is to four nucleotide bases. The HKY85 model (Hasegawa, Kishino, and Yano 1985), which generalizes the models of Jukes–Cantor, Kimura, and Felsenstein, has recently been implemented by Mau et al. (1996).

Several assumptions have been made that result in a relatively simple expression for the likelihood. In particular, the hypothesis of a molecular clock posits that θ is constant across all branches of the tree. In Mau (1996), an extension to additive trees, in which leaves are no longer constrained to be contemporaneous and branch lengths are measured in units of evolution, is implemented. Additive trees decompose nonuniquely into dendograms plus star components (see Lapointe and Legendre 1992). The Metropolis algorithm presented herein can be applied directly to the dendogram, while a modified copy of the second stage, Q_2, independently updates the coordinates of the star component.

Relaxation of the assumptions of iid sites effects only the likelihood calculation, and do not impact our proposal distribution. As long as the likelihood is computable, one substitutes it into the acceptance ratio of the Metropolis algorithm, and the chain advances. This generality, which extends to tree structures unrelated to phylogenetic analysis, may be the ultimate power of our method.

ACKNOWLEDGMENTS

The authors thank Carter Denniston, David Baum, Tom Kurtz, Peter Donnelly, Simon Tavaré, Joe Felsenstein, François Lapointe, and Bret Larget for their comments and criticisms.

REFERENCES

Evans, S., and Speed, T. (1993), "Invariants of Some Probability Models Used in Phylogenetic Inference," *The Annals of Statistics,* 21, 355–377.

Felsenstein, J. (1978), "The Number of Evolutionary Trees," *Systematic Zoology,* 27, 27–33.

——— (1983), "Statistical Inference of Phylogenies," *Journal of the Royal Statistical Society,* Series A, 146, 246–272.

——— (1985), "Confidence Limits on Phylogenies: An Approach Using the Bootstrap," *Evolution,* 39, 783–791.

——— (1988), "Phylogenies From Molecular Sequences: Inference and Reliability," *Annual Review of Genetics,* 22, 1–65.

Goldman, N. (1990), "Maximum Likelihood Inference of Phylogenetic Trees, With Special Reference to a Poisson Process Model of DNA Substitution and to Parsimony Analyses," *Systematic Zoology,* 39, 345–351.

Hasegawa, M., Kishino, H., and Yano, T (1985), "Dating the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA," *Journal of Molecular Evolution,* 22, 160–174.

Lapointe, F.- J., and Legendre, P. (1991), "The Generation of Random Ultrametric Matrices Representing Dendograms," *Journal of Classification,* 8, 177–200.

——— (1992), "A Statistical Framework to Test the Consensus Among Additive Trees (cladograms)," *Systematic Biology,* 41, 158–171.

Mau, B. (1996), "Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods," unpublished Ph.D. dissertation, University of Wisconsin–Madison, Dept. of Statistics.

Mau, B., Newton, M. A., Larget, B. (1996), "Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods," *technical report 961,* University of Wisconsin–Madison, Dept. of Statistics.

Sinsheimer, J. S., Lake, J. A., and Little, R. J. A. (1995), "Bayesian Hypothesis Testing of Four-Taxon Topologies Using Molecular Sequence Data," *Biometrics,* 52, 193–210.

Sytsma, K. J., and Gottlieb, L. D. (1986), "Chloroplast DNA Evolution and Phylogenetic Relationships in Clarkia Sect *peripetasma* (Onagraceae)," *Evolution,* 40, 1248–1261.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics,* 22, 1701–1762.