Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods

Bob Mau, Michael A. Newton, and Bret Larget

Department of Statistics, University of Wisconsin-Madison,

1210 West Dayton Street, Madison WI 53706-1685 USA

Department of Mathematics and Computer Science, Duquesne University,

College Hall 440, Pittsburgh PA 15282-1704 USA

SUMMARY [1]

We derive a Markov chain to sample from the posterior distribution for a phylogenetic tree given sequence information from the corresponding set of organisms, a stochastic model for these data, and a prior distribution on the space of trees. A transformation of the tree into a canonical cophenetic matrix form suggests a simple and effective proposal distribution for selecting candidate trees close to the current tree in the chain. We illustrate the algorithm with restriction site data on nine plant species, then extend to DNA sequences from thirty-two species of fish. The algorithm mixes well in both examples from random starting trees, generating reproducible estimates and credible sets for the path of evolution.

1. Introduction

"A phylogeny is a branching tree diagram showing the course of evolution in a group of organisms" (Felsenstein 1983, pg. 246). More and more, phylogenetic inference is based on molecular data, such as DNA or protein sequences. Given such data, life scientists wish to reconstruct the phylogeny whence these organisms arose. Their reasons are as diverse as the organisms they study. Systematists use the phylogeny to aggregate organisms into monophyletic groups, or clades, for taxonomic purposes (hence the generic term taxa). For others, the phylogeny may be of peripheral

[1] *Key words:* Cophenetic matrix; Evolution; Metropolis-Hastings algorithm; Phylogeny reconstruction.

importance, yet ignoring it can lead to unwarranted conclusions. For example, given a phylogeny, comparative biologists interested in the correlation between continuous character traits in a group of organisms can correct for dependencies among species (Felsenstein, 1985b). Huelsenbeck and Rannala (1997) discuss a range of inferences that rely on the phylogeny.

Existing reconstruction techniques attempt to find the phylogeny most compatible with the data under consideration. Among the more popular methods, one can categorize those using distance matrices (e.g., Fitch and Margoliash, 1967; Sokal and Sneath, 1963) in numerical taxonomy as clustering algorithms, while maximum parsimony (Camin and Sokal, 1965) and maximum likelihood (Felsenstein, 1981, 1983) each optimize an objective function on the space of trees. Felsenstein (1988) provides a comprehensive review of traditional methods. See also Swofford, Olsen, Waddell, and Hillis (1996).

Phylogenetic inference can also be dichotomized functionally. Maximum likelihood, maximum parsimony, and distance-matrix methods are practical for data sets relating many taxa, but beyond point estimates, these methods do not produce valid inferences. Measures of uncertainty rely exclusively on computer-intensive and approximate bootstrap analyses (Felsenstein, 1985a; Newton 1996). More recently developed techniques, such as the phylogenetic invariants of Cavender, Felsenstein, and Lake (see Evans and Speed, 1993; Navidi, Churchill, and von Haeseler, 1993) and a Bayesian approach (Smouse and Li, 1989; Hasegawa and Kishino, 1989; and Sinsheimer, Lake, and Little, 1996), allow exact inference, but mathematical and computational complexity have constrained these methods to very small problems.

We elect a Bayesian approach and use Markov Chain Monte Carlo (MCMC) methods to provide a computationally feasible technique that meets practitioners' demands for more taxa while keeping statistical inference on a sound footing, provided certain convergence criteria can be adequately demonstrated. Individual components comprising our technique can be found in the phylogenetic literature: Smouse and Li (1989) introduced the Bayesian paradigm, if not the terminology, to phylogeny reconstruction. Goldman (1993) uses non-Bayesian Monte Carlo tests of significance to assess the adequacy of evolutionary models. Griffiths and Tavaré (1994a,b) construct special Markov chains to compute likelihoods for ancestral inference. We apply MCMC to sample trees from the joint posterior distribution. Hence, measurement of uncertainty in our optimal tree accompanies tree construction. By contrast, other practical methods must first find an optimal tree, generate bootstrap samples from the data, and then re-estimate the tree from each bootstrap sample in order to address uncertainty in their reconstruction.

This article is organized as follows: Section 2 opens with an introduction to the requisite terminology for tree representation. We present a general stochastic model for the evolution of discrete molecular data in Section 3. Section 4 articulates the Bayesian perspective and defines Metropolis-Hastings algorithms. In Section 5, we introduce a two-stage proposal distribution that randomly selects a canonical ordering of leaf labels, then acts on the superdiagonal of the corresponding cophenetic matrix. Section 6 describes restriction site data and posits a stochastic model for its evolution. MCMC is applied and key phylogenetic quantities are analyzed. We extend our method to nucleotide sequence data in Section 7. Section 8 describes runs made with computer-simulated DNA sequence data from our model and reconstruction. Finally, we summarize the advantages of our approach and discuss further extensions to more complex models.

2. Tree Terminology and Representation

A phylogeny can be viewed abstractly as a rooted binary weighted tree. Mathematically, a tree is a connected graph (V, E), with vertex set V and edge set E, characterized by the absence of cycles. Vertices are classified as terminal nodes (also called leaves or tips) if they are connected through a single edge, and internal nodes otherwise, with $\mathcal{L}$ and $\mathcal{I}$ denoting the respective subsets. In rooted binary trees, each internal node has exactly three edges, with the exception of the root node ρ, which has only two edges. The placement of the root relative to the leaves determines the direction of time, and hence ancestry. 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. For expository reasons, we prefer to describe the branching pattern of a tree in terms of nodes coalescing or merging backwards in time. Figure 1 illustrates the coalescence structure of a seven taxa example.

[FIGURE 1]

The labeled shape of the tree, determined by which pairs of nodes coalesce, is called the tree topology, and is equivalent to the graph (V, E). The topology can be compactly summarized by using parentheses to indicate coalescences. For example, the topology in Figure 1 is (((1(4 7))(2(3 6)))5).

A weighted tree Ψ is a tree in which each edge has an associated positive weight. The time separating a child from its parent is its edge weight, called its branch length. Branch lengths are the vertical distances between connected nodes in Figure 1. The ordering in which merges occur defines coalescent levels, and the corresponding temporal intervals between consecutive merges constitute coalescent times. Different orderings of coalescent levels within a particular topology generate distinct labeled histories, alternative characterizations of shape. We restrict attention to trees having contemporaneous tips, called dendograms. 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 with n, equal to $(2n-3) \times (2n-5) \ldots \times 1$ and $n! \times (n-1)!/2^{n-1}$ respectively (e.g., Felsenstein, 1978).

For any weighted binary tree with labeled leaf nodes, the tree topology and branch lengths are determined by the within-tree distances between all pairs of leaf nodes (Lapointe and Legendre, 1992). Each permutation of the leaf labels generates a different $n \times n$ symmetric matrix of these distances. In a rooted tree where all leaf nodes are equally distant from the root, such matrices are composed of at most n distinct entries, and are called cophenetic.

When a tree is displayed as in Figure 1, an arbitrary choice is made giving a left-right orientation to the pair of branches emanating from each internal node. This action imparts an order on the leaf nodes. The collection of 2^{n-1} orderings determined in this fashion is called the set of canonical orderings for a given tree. A cophenetic matrix with a canonical ordering has the desirable property that its superdiagonal (the diagonal of the submatrix formed by deleting the first column and nth row) contains each distinct non-zero cophenetic distance. We call such a cophenetic matrix canonical because its superdiagonal completely specifies the tree. One such matrix for the sample phylogeny of Figure 1 is realized in Table 1. A convenient shorthand for the information sufficient to represent the tree is $\{\sigma, \mathbf{a}\}$, where σ is a canonical ordering and $\mathbf{a}$ denotes the times-to-coalescence between adjacent label pairs in σ. Because times-to-coalescence are one half the corresponding cophenetic distances, **a** is simply half the superdiagonal induced by σ. See Lapointe and Legendre (1991) for a detailed discussion of cophenetic matrices.

[TABLE 1]

3. A Stochastic Model for Leaf Data

Evolution has two components that may be modeled as stochastic processes: the branching created by speciation and extinction to form a phylogeny, and the propagation of characters along the branches of that phylogeny. We do not model the branching process stochastically, choosing instead to treat the phylogeny as a parameter in a model for the propagation of data along each lineage. We adopt standard Markov models for the second component (Goldman, 1990).

Typical character data on n taxa may be arranged as an $n \times N$ matrix, where N is the common number of sites, or positions, providing information from each taxa. We consider problems where elements of this matrix are discrete characters from a finite set $\mathcal{D}$ of size d. These data are viewed as a present-day snapshot of a realization of a stochastic process that has evolved along the branches of an unknown phylogeny Ψ. Modeling is reduced to a single site by assuming evolution among sites is independent.

A stochastic model describes the joint distribution of $y = \{y_v,\ v \in V = \mathcal{I} \cup \mathcal{L}\}$, the historical record at $\mathcal{I}$ and current status at $\mathcal{L}$ for a given site. The ancestral root state y_ρ is assigned an initial distribution π_0 on $\mathcal{D}$. Conditionally on y_ρ, two continuous-time, $\mathcal{D}$-valued Markov processes emanate independently from the root ρ along the corresponding branches of Ψ. As a given process reaches an internal node v, its value is recorded as y_v. Evolution continues by repeating this mechanism, with conditionally independent Markov processes emanating from every internal node $v \in \mathcal{I}$. Observable y_v for $v \in \mathcal{L}$ are simply the end products of this evolution. Such a model for y is specified by the phylogeny Ψ, the initial distribution π_0, and transition probabilities $p\left(y_v | y_{\sigma(v)}, t_v, \beta\right)$. Here $\sigma(v)$ is the parent node of v, t_v is the intervening branch length, and β is a parameter vector describing rates of change among states in the Markov process for a given branch. We consider two particular models in the examples in Sections 6 and 7. The probability of the particular realization y at a given site is

$$\pi_0(y_\rho) \prod_{v \in V \backslash \rho} p(y_v | y_{\sigma(v)}, t_v, \beta). \tag{1}$$

Both dependency between sites and site specific $\boldsymbol{\beta}$'s can be incorporated into the model (e.g. Yang, 1996). Specification of an initial distribution π_0 should reflect the underlying biology, or can be estimated from the observed frequencies in the data.

To calculate the likelihood function from leaf data at multiple sites, we must marginalize (1) over all values of the unobserved historical record $\{y_v,\ v \in \mathcal{I}\}$ for all sites. Straight summation is computationally prohibitive, requiring on the order of Nd^n calculations. The pruning method for likelihood evaluation requires on the order of Ndn computations because it takes advantage of the Markov property of the substitution model (e.g., Felsenstein 1983). Pruning produces a collection of fragmentary likelihoods, starting from the leaves and working recursively to the root, for each site. For each leaf v, $L_v(i) = 1[y_v = i]$ for state $i \in \mathcal{D}$ ($1[\cdot]$ is the indicator function). At an internal node v, the conditional probability of descendant leaf data given $y_v = i$ is

$$L_v(i) = \left(\sum_{j \in \mathcal{D}} L_u(j) p(j|i, t_u, \beta) \right) \left(\sum_{m \in \mathcal{D}} L_w(m) p(m|i, t_w, \beta) \right),$$

where $\sigma(u) = \sigma(w) = v$. The likelihood function becomes a product across sites

$$L(\Psi, \beta) = \prod_{k=1}^{N} \sum_{i \in \mathcal{D}} \pi_0(i) L_\rho^k(i) \tag{2}$$

where the superscript denotes root fragmentary likelihoods at the kth site.

4. The Bayesian Perspective and Metropolis-Hastings Algorithms

Bayesian analysis requires a prior distribution on the parameter space of the model. The prior should reflect the scientist's beliefs on how likely particular parameter values are before the data have been observed. The posterior distribution $\pi(\Psi)$ represents the uncertainty about the phylogeny in light of new evidence in the sequence data, and is proportional to the likelihood times the prior distribution. In each example we consider, a uniform prior is placed on the finite set of labeled histories and a flat prior density is assumed on a compact set of possible coalescent times, as well as any propagation parameters from the stochastic model.

For large problems, Monte Carlo techniques may be the only effective way to integrate $\pi(\Psi)$ so as to obtain posterior inferences. One such method is the Metropolis-Hastings algorithm, in which a transition mechanism proposes a new tree Ψ^* with density $Q(\Psi, \Psi^*)$, conditional on being at Ψ. We subject this draw to a randomized test, accepting it with probability

$$\min\left(1, \frac{\pi(\Psi^*)Q(\Psi^*, \Psi)}{\pi(\Psi)Q(\Psi, \Psi^*)}\right) \tag{3}$$

otherwise remaining at Ψ.

The Markov chain $\Psi_1, \Psi_2, \ldots$ so formed from an initial state Ψ_1 converges in distribution to $\pi(\Psi)$ when Q is irreducible (e.g., Tierney, 1994). The important theoretical point is that for almost every realization of the chain,

$$\lim_{K\to\infty} \frac{1}{K} \sum_{i=k_0+1}^{K+k_0} f(\Psi_i) = E_\pi[f(\Psi)] \tag{4}$$

where f is a function whose expectation is desired. For example, when f is the indicator function of a particular topology, equation (4) means that the empirical relative frequency of that topology in the Markov chain converges to its corresponding (marginal) posterior probability. Credible sets, the Bayesian counterparts of confidence regions, are collections of topologies having high relative frequency in the chain. The topology with the highest relative frequency is our reported reconstruction.

Some initial sample points are discarded as burn-in ($k_0 > 0$ in (4)) so as to reduce the bias in the Monte Carlo estimates (Besag and Green, 1993). We determine burn-in by inspecting time series plots of the log posterior. A difficult problem with the implementation of MCMC is to know whether or not the K used to approximate (4) is large enough. If K is too small then the Monte Carlo variance may overwhelm the signal (Geyer, 1992). Roughly speaking, a chain is said to mix well if it acts like an independent sample. Cowles and Carlin (1996) present a survey of convergence diagnostics commonly used to check for evidence of poor mixing. Though we emphasize the reproducibility of our results from random starts Ψ_1, we also analyze univariate statistics from single runs. Such diagnostic tools are helpful in ferreting out inefficient samplers, but provide little insight into finding efficient ones. By considering distances in tree space, we have developed an algorithm that extends Bayesian analysis to relatively large problems.

5. A Proposal Distribution for Trees

We consider a two-stage proposal distribution. The first stage randomly selects a canonical representation $\{\sigma, \mathbf{a}\}$ for the current tree Ψ, while the second stage perturbs the components of $\mathbf{a}$. In particular, the first stage Q_1 samples one of the 2^{n-1} canonical orderings of the current tree by independently flipping a fair coin at each internal node, thus selecting a particular superdiagonal $\{d_{i,i+1} : i = 1, \ldots, n-1\}$ of a canonical cophenetic matrix having times-to-coalescence $\{a_i = \frac{d_{i,i+1}}{2}\}$. The second stage Q_2 simultaneously and independently modifies the elements of $\mathbf{a}$. Specifically, $a_i^* = |U_i|$, where U_i is uniformly distributed on the interval $(a_i - \delta, a_i + \delta)$ for a tuning constant $\delta > 0$. The tuning constant 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 reflection of uniform probability mass onto the positive line is an efficient way to obtain symmetric proposals near the boundary. Symmetry of each component update can be seen by inspecting the transition density

$$Q_{2,i}(x, y) \quad = \quad \frac{1}{2\delta}\left(1[|x-y| < \delta] + 1[y < \delta - x]\right) = Q_{2,i}(y, x)$$

where $1[\cdot]$ is the indicator function, and $x, y > 0$. Regarding the composition $Q = \prod_i Q_{2,i} \circ Q_1$, the density at a possible proposed tree given the current tree equals

$$\frac{1}{2^{n-1}} \prod_i Q_{2,i}(a_i, a_i^*) = \frac{1}{2^{n-1}} \prod_i Q_{2,i}(a_i^*, a_i).$$

The symmetry of Q simplifies (3), making this a Metropolis algorithm.

We illustrate the action of the second stage $Q_2 = \prod_i Q_{2,i}$ in a second example, a six-leaved tree represented by $\{\sigma, \mathbf{a}\} = (\{(4, 6, 3, 2, 1, 5), (2.3, 4.6, 2.0, 4.0, 3.4)\}$. To recover the tree from the canonical cophenetic matrix, place the leaf nodes along a horizontal axis in their canonical order. From each interleaf midpoint append an internal node below the axis a distance equal to the corresponding time-to-coalescence. Working from top to bottom, draw branches from each internal node to the nearest parentless nodes to the left and right (Figure 2). Newton, Mau, and Larget (1998) discuss this same proposal mechanism from a different perspective.

[FIGURE 2]

A candidate phylogeny proposed by Q may differ from the current tree both in its labeled history and in its tree topology. In Figure 2 for example, when the perpendicular at $\{3, 2\}$ is maximally increased while that at $\{4, 6\}$ is simultaneously decreased, a new labeled history is proposed. Significantly, when a similar adjustment is made to the perpendiculars at $\{1, 5\}$ and $\{2, 1\}$, a different topology results, with $\{1\}$ coalescing with $\{2, 3\}$ instead of $\{5\}$. Had the canonical ordering (4,6,3,2,5,1) been chosen instead, the same realization by Q_2 would have coalesced $\{5\}$ with $\{2, 3\}$.

We now establish irreducibility of $Q = Q_2 \circ Q_1$. For a given permutation of labels, repeated applications of Q_2 clearly allow transit from any superdiagonal $\{2a_i\}$ to any other superdiagonal. Thus, it is sufficient to show that Q allows moves among any of the $n!$ permutations of labels. A single application of Q_1 can move any label to the first position of the permutation. Repeated applications of Q_2 ensure that $2a_1$ is the largest superdiagonal element, corresponding to a tree in which that first label is connected to the root by one long branch. Another application of Q_1 allows us to place any remaining label into the second position while fixing the first by choosing the current branch orientation at the root node. Successive applications of Q_2 yield a superdiagonal in which $a_1 > a_2 > a_i,\ i > 2$, so that the second label coalesces directly into the penultimate internal node. By continuing the process we are able to move to any permutation. Thus, Q is irreducible.

6. An Example with Binary Data

Sytsma and Gottlieb (1986) studied the evolutionary relationships among nine species of the genus Clarkia, plants indigenous to California. These botanists extracted chloroplast DNA (cpDNA) from the leaves of a young plant of each species, and exposed that cpDNA to twenty-nine restriction enzymes. A restriction enzyme acts on DNA by physically cutting the molecule wherever it recognizes a particular base pattern. At each position on the genome where a cut occurs, a restriction site is said to be present. Sytsma and Gottlieb determined 609 positions where restriction sites occurred, 490 of which were non-informative (sites at every species). Informative data were translated into a 9×119 matrix of zeroes and ones, representing the absence or presence of a restriction site at mapped positions in the chloroplast genome for each species.

We entertain a simple stochastic model for the evolution of restriction sites; a two-state continuous time Markov chain with infinitesimal rates λ and μ, representing the intensity of the instantaneous transition from 0 to 1 and 1 to 0, respectively. The generator matrix is

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

The matrix $P(t)$ of transition probabilities through time t satisfies the Chapman-Kolmogorov equation $P'(t) = P(t) \cdot A$, with solution $P(t) = \exp(At)$ having (i, j)th entry

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

where $r = \mu/\lambda$ and $\theta = \mu + \lambda$ is the mutation rate.

We derive an initial distribution for restriction sites at the root from biological principles. A uniform distribution of nucleotide bases at the root and a single restriction enzyme with recognition sequence of length six suggest an initial distribution of $\pi_0^*(1) = (\frac{1}{4})^6$ for the presence of a restriction site at each genomic site. Since recognition sequences differ from each other in at least two positions, the probability of a restriction site at each data site is approximately: $\pi_0(1) = 29\pi_0^*(1) = 0.00708$. Because a particular location in the genome is detected by the presence of a restriction site, only a minuscule fraction of the 170 kb (kilobases) enters the data. Following Felsenstein (1992), the likelihood (2) is conditioned on seeing at least one restriction site at each position

$$L(\Psi, \beta) / \left[1 - p_0(\Psi, \beta)\right]^N$$

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

A complete Bayesian specification requires a prior on the propagation parameters $\beta = (r, \theta)$. We place a uniform prior over $[1, 4^6]$ on r, as constrained (somewhat liberally) by the biology, assuming a six base cutter. The mutation rate θ is confounded with time, so the branch lengths are proportional to amount of evolution. Hence, we fix θ throughout our analysis. In the MCMC algorithm, we cycle between an update of the phylogeny Ψ and an update of r. The proposal distribution for this second update is a uniform window centered at the current value.

We implemented the MCMC algorithm of Section 5 in Fortran 77 for this model and data set. From random starting trees, chains of length 250,000 were subsampled at a rate of one in a hundred to reduce dependence in stored output, yielding 2500 phylogenies per run. Each run took approximately twenty minutes on a SPARC 10 work station. Burn-in was less than 200 iterations, affecting only the first two stored samples. Convergence of the chain was inferred from the high degree of reproducibility of the posterior. In dozens of repeated runs from random starts, the realized relative frequencies deviate by at most $\pm 3\%$, a measure of Monte Carlo error for samples of this size. Additional programming considerations for this data set are addressed in Mau and Newton (1997).

[FIGURE 3]

[TABLE 2]

Figure 3 and Table 2 summarize our analysis of the Clarkia data. Topologies IV through VII have unrooted trees distinct from those that form the 99% credible region. Several subtopologies within the clade: $\{5, 6, 7, 8, 9\}$ are weakly supported by the data, due to the placement of leaf $\{7\}$. Bootstrap analysis using parsimony attaches a weak 61% confidence coefficient to that branch in the optimal tree (Sytsma and Gottlieb, 1986, p.1257). By contrast, the posterior probability of an alternate attachment is a near negligible 0.4%. The weakness in the rooting is reinforced when we simulate from a posterior that includes the outgroup species used by Sytsma and Gottlieb as a tenth taxa. That outgroup is attached beneath the $\{5, 6, 7, 8, 9\}$ clade with high posterior probability, but not at the root.

Our assurances of convergence and adequate mixing of the chain are predicated on the verifiable reproducibility of the posterior. Visual tracking of a chain as it moves between the three predominant topologies provides further confirmation. That we are able to present fairly tight credible regions for the Clarkia data depends on both the nature and the amount of data, but most dramatically on the number of taxa. In the next section, we increase the complexity of all three factors, and provide additional diagnostics.

7. A Nucleotide Sequence Example

We have analyzed aligned protein coding mitochondrial DNA sequences obtained from 32 species of cichlid fishes (Kocher, *et al.*, 1995) using the HKY85 model of nucleotide substitution (Hasegawa, Kishino, and Yano, 1985). Table 3 shows some information about these species and their standard taxonomy. The HKY85 model is an example of the branching Markov substitution models discussed in Section 3. There are four base composition parameters π_A, π_C, π_G, and π_T which represent the long run relative frequency of each nucleotide base in a single population, and these are used as the probability distribution of base values at the root of the phylogeny. The overall rate of substitutions is θ. A parameter κ characterizes the difference in substitution rates between transitions (changes between A and T or between C and G) and transversions (any other change). Appendix I presents the generator matrix for the HKY85 Markov model, and the resulting transition probability matrix. Each DNA sequence contains 1044 sites that can be partitioned into three blocks of sites according to codon position, and our analysis allowed different parameter values across blocks. Across all species, 567 of these sites are constant. Rather than entertain a full Bayesian analysis of this problem, we fixed parameters at estimated values, and approximated the posterior distribution of the unknown phylogeny only (Appendix II discusses how we get parameter estimates).

[TABLE 3]

There are more than 10^{40} topologies in this problem, as compared to 2×10^6 in the Clarkia example, thus presenting a significant challenge to model-based analysis. After extensive program testing, we can report with some confidence a Monte Carlo approximation to the posterior distribution. Our summary of most probable topologies was calculated by combining results from ten independent realizations of the Metropolis chain. Each realization started at a phylogeny randomly sampled from the uniform prior distribution and proceeded for 1,100,000 steps. The first $k_0 = 10^5$ steps were discarded as burn in, and the remaining chain was subsampled every 200 steps to produce a stored sequence of 5000 phylogenies. The tuning parameter δ was selected adaptively during the initial stage of the run, but attained a fixed post-burn-in value of 0.00098 in all ten runs, equal to about 3% of the total height of a typical sampled tree. The acceptance rate of proposed trees was approximately 40% with this window. A C-language implementation using efficient likelihood evaluation and storage techniques took about 20 hours per realization on a Pentium 200 PC.

[Figure 4]

The choice of chain length, subsampling rate, and burn-in parameters was affected by the limitations of computer speed and memory, but the analysis of output from test runs was also helpful. Figure 4 shows a time-series plot of the log-likelihood of saved phylogenies from one of the ten final runs. It is typical to see a dramatic increase in log-likelihood during the burn-in period, with stabilization at what appears to be a stationary series after about 10^4 basic steps. The autocorrelation function in Figure 5 indicates that trees separated by about 20 storage steps present approximately independent log-likelihoods. We examined other one-dimensional summaries of individual realizations to diagnose mixing problems. A particularly useful measure is the frequency of topological changes within individual clades. We defer reporting this analysis until the subtopologies are defined.

[Figure 5]

Summarizing the empirical distribution of sampled phylogenies presented a challenge. In the combined runs, we found that 34 topologies accounted for half the posterior probability. A 90% credible region contained nearly 600 different tree topologies. Evidently, much of this variation was caused by uncertainty in subtopological branching structure in the presence of fairly well supported monophyletic groups or clades. We have identified six such clades of between four and seven taxa, four of which appeared in every saved tree. Figure 6 shows three probable subtopologies for a clade of five species. Other subtopologies are defined at the bottom of Table 4, and simplify the presentation of topology uncertainty in that table.

[Figure 6]

[Table 4]

Compelling evidence for the effectiveness of our MCMC algorithm is the reproducibility of probability estimates from independent realizations, especially since the starting positions arise from a uniform distribution. The last three columns in Table 4 quantify the variation from run to run in the posterior probability of individual topologies. Monte Carlo standard error is less than 0.3% for the top 13 topologies, for example.

[TABLE 5]

The rate at which a tree topology switches between its various subtopologies measures mixing. Table 5 tabulates these switching counts from one run on the cichlid data. Ideally, the number of switches would be close to the expected number assuming independent sampling. Indeed, Table 5 shows that the number of switches between the subtopologies within clade A for one typical run agrees remarkably well with the expected counts assuming independence shown here:

	A_1	A_2	A_3
A_1	3019	722	144
A_2	722	173	34
A_3	144	34	7

The other clades do not mix as rapidly, with the number of observed switches only a small fraction of that expected assuming independence. The rate of mixing within these clades is sufficient, however, to give us reproducible results over separate runs.

[TABLE 6]

Marginal probabilities are natural in Bayesian analysis, and suggest further effective summaries of the posterior distribution. Table 6 shows the posterior distribution of subtopologies within clades. Clearly, the designation of clade is appropriate for A, B, C, D, while F and especially E are not necessarily monophyletic. Note that B includes a member from the Limnochromini tribe, so it is somewhat surprising that this artificial clade is so unambiguously supported.

[TABLE 7]

A second posterior summary is the distribution over clade trees. That is, the uncertainty in how clades are connected to form the tree. Table 7 shows the most probable tree topologies ignoring variability within clades. Notice that clades A, B, C, and D, the four clades with unanimous support in our samples, also are connected to each other unambiguously.

Finally, we compare our estimate with two phylogenies reconstructed using traditional methodologies. Kocher *et al.* (1995) use a distance matrix method neighbor-joining on the third codon position to identify clades A, B, C, D, and F. The authors then apply maximum parsimony to four of those clades to obtain optimal subtopologies, rooting each with a taxon from an adjoining clade as an outgroup. We obtained a maximum likelihood estimate with the program Dnamlk from Phylip, version 3.572c, which, like our method, assumes a molecular clock that maintains a constant rate of mutation along each branch.

[TABLE 8]

Because of our codon position specific model and different methodology, it is not surprising that our answer does not agree in whole with either reconstruction. There is a fair degree of similarity in the different solutions arrayed in Table 8. Each estimate has clades A, B, C, D, and F in common, reinforcing the current taxonomic scheme based on geographic proximity. The greatest disparity between estimates involves the attachment of taxa from clade E. Except for the control pair $\{22, 23\}$, these species are dispersed throughout Lake Tanganyika. Interestingly, the three methods concur in placing the B clade closer to the Malawi flock, A, rather than to members from its own flock, and in preferring a B clade which admits a taxon from another tribe.

8. Simulation Study

Further support for the efficacy of our proposed algorithm came from a simulation study. We generated ten synthetic data sets analogous to the cichlid data set. Each one had 32 aligned sequences of length 1044, separated into blocks of sites corresponding to the three codon positions, and each one was obtained using the simulation software Seq-Gen version 1.04 from Rambaut and Grassly (1996). To mimic the complexity of the cichlid problem, we fixed the true phylogeny in this simulation equal to one sampled from the posterior in Section 7. In particular, the corresponding tree topology was the most probable one (Table 4). We also used the same substitution parameters as those obtained in the cichlid problem.

For each synthetic data set we ran our MCMC algorithm twice, using the same burn-in, subsampling rate, and total chain length as before. Having two runs enables us to compare variation both within and between data sets, and thus to gauge Monte Carlo error. The posterior distribution is fairly diffuse for all synthetic data sets, as it is in the cichlid problem. The posterior probability of the top ten tree topologies ranges from 0.22 to 0.79 across the twenty runs with a mean of 0.49, as compared to 0.32 in Table 4. The number of tree topologies in a 90% credible set ranges from 20 up to 409. This is smaller than the 600 odd topologies reported in the cichlid data analysis using all ten runs, but comparable to that from any single realization.

The variation between runs is small compared to the variation among synthetic data sets. Considering the probability assigned to the top ten tree topologies, the absolute difference between runs was 0.01 on average. For any tree topology, the difference between runs in estimated posterior probability was consistently smaller than 0.021.

The true tree topology was captured very well by the estimated posterior distributions: the modal tree topology was equal to the true one in one case, differed by a single branch placement in eight cases, and by two branch placements for one synthetic data set. In seven of the ten cases the true topology appeared in the top ten. This is about what we expect when the average posterior mass assigned to the top ten topologies is 0.49. Overall, the posterior probability of the true tree topology ranged from 0.0076 to 0.3103. Evidence is emerging that as we increase the number of sites of simulated data, more and more posterior mass is concentrated on the true tree topology.

9. Discussion and Extensions

The analysis of the Clarkia data set and restriction site model is straightforward, allowing us to introduce concepts and notation required in the more complex cichlid fish example. We therefore confine our remarks to the second, more challenging, example.

Table 8 compares the estimates from the three methods, but not the methods themselves. Both maximum parsimony and maximum likelihood have solutions that maximize objective functions determined by the data. By contrast, our method provides a distributional assessment where pockets of high posterior probability are located as the chain traverses tree space. Our estimate of the phylogeny is a simple by-product of that assessment. We simulate a chain of trees where the long-run relative frequency of hitting any particular tree topology is proportional to its marginal posterior probability. For moderate number of taxa and a discriminating set of data, almost all tree topologies have essentially zero probability mass, and as such are unseen in even extremely long chains. Nevertheless, reproducible samples from the posterior are generated efficiently. Other methods require bootstrapping to appraise the quality of their estimate. At best, a few hundred such phylogenies provide confidence coefficients on how well each node of the original estimate is supported by the data. Our method automatically provides posterior probabilities, as illustrated in Table 4 for the cichlids, from which confidence coefficients for a particular tree can be calculated if one so chooses.

For large numbers of taxa, other methods rely on incomplete heuristic searches and time consuming rearrangements that do not guarantee to produce an absolute maximum. We claim a clear advantage in this regard, since averaging over local maxima (tree islands in the phylogenetic context) goes to the very heart of Markov chain Monte Carlo technology. The likelihood surface for phylogeny may contain numerous local modes, depending on the number of taxa, the data, and the model. The proposal distribution we have developed appears to navigate reasonably well between modes in the two examples presented.

Our claim that we are accurately sampling from the posterior density of the cichlid fish example rests primarily on the close agreement in the estimates of posterior probabilities in different runs from widely dispersed random starting trees. We appreciate that researchers may not have the luxury of conducting multiple runs when run times approach a day in duration. To that end, time series diagnostics and intra-clade switching statistics from a single realization support our claim that the chain is mixing well given our level of subsampling. Notwithstanding, we echo the caveat of Cowles and Carlin (1996, p. 902) that no single convergence diagnostic is infallible, and that making comparisons between a few parallel runs from disparate starting trees is necessary.

Our experience with the fish data indicates that stickiness can occur when one employs models that exhibit a significant lack of fit. For example, application of the standard HKY85 model equally to all sites generated chains with strongly autocorrelated log posteriors that did not mix within the E pseudo-clade. Modification of our method may be necessary to increase the rate of mixing within clades to successfully solve a wider range of data sets and models.

The technique described herein supposes a molecular clock. This constraint is a particular concern for the cichlid fish data, where significantly higher mutation rates have been observed among members of the Ectodini tribe (Kocher *et al.*, 1995, p.425). An alternate viewpoint, considers additive trees in which leaves are not constrained to be contemporaneous, and branch lengths are measured in units of evolution instead of time. We have implemented such a non-clock model by decomposing additive trees into dendograms and star components (see Lapointe and Legendre, 1992). The proposal distribution described in this paper is applied to the dendogram. A variant of Q_2, indexed by leaf labels, is applied simultaneously to the star component. The subsequent hybrid driver has been applied successfully to the cichlid data, and we plan to report the results elsewhere.

Our reported calculations have assumed rate constancy within blocks of sites, although models allowing more general rate variation are available that may further improve the fit (Yang, 1996). Independence between sites is a more difficult assumption to relax. Schöniger and von Haeseler (1994) look at protein coding regions where the first two codons are assumed correlated. General dependent models are described in Tavaré and Feng (1995). Fortunately, our method is independent of the particular form of the likelihood (or the prior). Provided the likelihood is computable and the number of additional parameters in the model remains manageable, one should be able to substitute into the acceptance ratio and run the chain in a reasonable amount of time.

Since the original submission of this article, we have become aware of efforts of others to run Markov chains on the space of phylogenetic trees. Kuhner, Yamato, and Felsenstein (1997) use MCMC to sample genealogies to estimate the product of the effective population size and the mutation rate per site. Yang and Rannala (1997) and Li, Pearl, and Doss (1996) have proposed MCMC algorithms for Bayesian phylogenetic inference that differ greatly from the method we have developed. The global nature of the tree update, coupled with the existence of a tuning parameter that moderates the overall acceptance rate, distinguishes our approach.

Acknowledgements

The authors would like to thank Carter Denniston and David Baum for helpful discussions that provided the requisite biological background. Further thanks go to Tom Kurtz, Peter Donnelly, Simon Tavaré, Joe Felsenstein, Charlie Geyer, François Lapointe, Don Simon, and Brian Yandell for stimulating conversations during the evolution of this article. The suggestions of an anonymous reviewer regarding convergence diagnostics and computer simulated data led to improvements in the final version.

References

Besag, J., and Green, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society*, Series B **55**, 25-37.

Camin, J.H. and Sokal, R. R. (1965). A method for deducing branching sequences in phylogeny. *Evolution* **19**, 311-326.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Society* **91**, 883-904.

Evans, S. and Speed, T. (1993). Invariants of some probability models used in phylogenetic inference. *Annals of Statistics* **21**, 355-377.

Felsenstein, J. (1978). The number of evolutionary trees. *Systematic Zoology* **27**, 27-33.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368-376.

Felsenstein, J. (1983). Statistical inference of phylogenies. *Journal of the Royal Statistical Society*, Ser. A, **146**, 246-272.

Felsenstein, J. (1985a). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791.

Felsenstein, J. (1985b). Phylogenies and the comparative method. *The American Naturalist* **125**, 1-15.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22**, 1-65.

Felsenstein, J. (1992). Phylogenies from restriction sites: a maximum likelihood approach. *Evolution* **46**, 159-173.

Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155**, 279-284.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science* **7**, 437-511.

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**, 345-361.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**, 182-198.

Griffiths, R. C. and Tavaré, S. (1994a). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131-159.

Griffiths, R. C. and Tavaré, S. (1994b). Ancestral inference in population genetics. *Statistical Science*, **9**, 307-319.

Hasegawa, M. and Kishino, H. (1989) Confidence limits on the maximum likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution* **43**, 672-677.

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.

Huelsenbeck, J. P. and Rannala, B. (1997). Phylogenetic methods come of age: testing hypothesis in an evolutionary context. *Science* **276**, 227-231.

Kocher, T. D., Conroy, J. A., McKaye, K. R., Stauffer, J. R., and Lockwood, S. F. (1995). Evolution of NADH Dehydrogenase Subunit 2 in East African Cichlid Fish. *Molecular Phylogenetics and Evolution* **4**, 420-432.

Kuhner, M. K., Yamato, J., and Felsenstein, J. (1997). Applications of Metropolis-Hastings genealogy sampling, pp 183-192 in *Progress in Population Genetics and Human Evolution*, ed. P. Donnelly and S. Tavaré. IMA Volumes in Mathematics and its Applications, volume 87. Springer Verlag, Berlin.

Lapointe, F.- J. and Legendre, P. (1991). The generation of random ultrametric matrices representing dendograms. *Journal of Classification* **8**, 177-200.

Lapointe, F.- J. and Legendre, P. (1992). A statistical framework to test the consensus among additive trees (cladograms). *Systematic Biology* **41**, 158-171.

Li, S., Pearl, D. K., and Doss, H. (1996) Phylogenetic tree construction using Markov chain Monte Carlo. Department of Statistics, Ohio State University, Technical Report 583.

Mau, B. and Newton, M. A. (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* **6** 122-131.

Navidi, W. C, Churchill, G. A., and von Haeseler, A. (1993). Phylogenetic inference: linear invariants and maximum likelihood. *Biometrics* **49**, 543-555.

Newton, M. A. (1996). Bootstrapping phylogenies: Large deviations and dispersion effects. *Biometrika* **83**, 315-328.

Newton, M. A., Mau, B., and Larget, B. (1998). Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. *Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference on Statistics and Molecular Biology.* (to appear).

Rambaut, A. and Grassly, N. C. (1996). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *CABIOS* (in press).

Schöniger, M. and von Haeseler, A. (1994). A stochastic model for the evolution of autocorrelated DNA sequences. *Molecular Phylogenetics and Evolution* **3**, 240-247.

Sinsheimer, J. S., Lake, J. A., and Little, R. J. A. (1996). Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics* **52**, 193-210.

Smouse, P. E. and Li, W.-H. (1989). Likelihood analysis of mitochondrial restriction-cleavage patterns for the human-chimpanzee-gorilla trichotomy. *Evolution* **43**, 1162-1176.

Sokal, R. R. and Sneath, P. H. A. (1963). *Numerical Taxonomy.* San Francisco: Freeman.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. In *Molecular Systematics* (2nd edition), 407-511, ed. Hillis. D. M., Moritz, C., and Mable, B. K. Sinauer Associates, Inc., Sunderdland, Massachusetts.

Sytsma, K. J. and Gottlieb, L. D. (1986). Chloroplast DNA evolution and phylogenetic relationships in Clarkia sect *peripetasma* (onagraceae). *Evolution* **40**, 1248-1261.

Tavaré, S. and Feng, Y. (1995). Reconstructing phylogenetic trees when sites are dependent. *DIMACS Technical Report 95-48*, 55-57.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**, 1701-1762.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analysis. *TREE* **11**, 367-371.

Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14**, (in press).

Appendix I: Standard HKY85 Model

The generator matrix is:

$$A = \begin{pmatrix} -\theta(\kappa\pi_C + \pi_R) & \kappa\theta\pi_C & \theta\pi_A & \theta\pi_G \\ \kappa\theta\pi_T & -\theta(\kappa\pi_T + \pi_R) & \theta\pi_A & \theta\pi_G \\ \theta\pi_T & \theta\pi_C & -\theta(\kappa\pi_G + \pi_Y) & \kappa\theta\pi_G \\ \theta\pi_T & \theta\pi_C & \kappa\theta\pi_A & -\theta(\kappa\pi_A + \pi_Y) \end{pmatrix}$$

Transition probability matrix $P(t)$ via spectral decomposition of the above generator is:

$$\begin{pmatrix} \pi_T + \pi_T \frac{\pi_R}{\pi_Y} \mathrm{e}^{-\theta t} + \frac{\pi_C}{\pi_Y} \mathrm{e}^{-\gamma_1 t} & \pi_C(1 + \frac{\pi_R}{\pi_Y} \mathrm{e}^{-\theta t} - \frac{1}{\pi_Y} \mathrm{e}^{-\gamma_1 t}) & \pi_A(1 - \mathrm{e}^{-\theta t}) & \pi_G(1 - \mathrm{e}^{-\theta t}) \\ \pi_T(1 + \frac{\pi_R}{\pi_Y} \mathrm{e}^{-\theta t} - \frac{1}{\pi_Y} \mathrm{e}^{-\gamma_1 t}) & \pi_C + \pi_C \frac{\pi_R}{\pi_Y} \mathrm{e}^{-\theta t} + \frac{\pi_T}{\pi_Y} \mathrm{e}^{-\gamma_1 t} & \pi_A(1 - \mathrm{e}^{-\theta t}) & \pi_G(1 - \mathrm{e}^{-\theta t}) \\ \pi_T(1 - \mathrm{e}^{-\theta t}) & \pi_C(1 - \mathrm{e}^{-\theta t}) & \pi_A(1 + \frac{\pi_Y}{\pi_R} \mathrm{e}^{-\theta t}) + \frac{\pi_G}{\pi_R} \mathrm{e}^{-\gamma_2 t} & \pi_G(1 + \frac{\pi_Y}{\pi_R} \mathrm{e}^{-\theta t} - \frac{1}{\pi_R} \mathrm{e}^{-\gamma_2 t}) \\ \pi_T(1 - \mathrm{e}^{-\theta t}) & \pi_C(1 - \mathrm{e}^{-\theta t}) & \pi_A(1 + \frac{\pi_Y}{\pi_R} \mathrm{e}^{-\theta t} - \frac{1}{\pi_R} \mathrm{e}^{-\gamma_2 t}) & \pi_G(1 + \frac{\pi_Y}{\pi_R} \mathrm{e}^{-\theta t}) + \frac{\pi_A}{\pi_R} \mathrm{e}^{-\gamma_2 t} \end{pmatrix}$$

where $\pi_Y = \pi_T + \pi_C,\ \pi_R = \pi_A + \pi_G,\ \gamma_1 = \theta(\kappa\pi_Y + \pi_R),\ \gamma_2 = \theta(\kappa\pi_R + \pi_Y)$

Appendix II: Estimation of separate parameters for three HKY85 models, one for each codon position.

Exploratory data analysis of the cichlid DNA sequences by codon position shows striking differences in the percentage of variable sites (31.5%, 14.9%, and 90.5%, respectively) and in base composition. We first estimated the values of π separately for each codon position using observed base counts. To estimate different values of θ, we ran three separate simulations using data from each position alone, updating κ each cycle, and used the subsequent relative total tree heights to find values of $\theta_1 = 1.4$, $\theta_2 = 1.0$ and $\theta_3 = 8.3$. Finally, with these three values of θ, we ran a chain with all the data allowing κ to change separately for each position, estimating $\kappa_1 = 7.5$, $\kappa_2 = 2.5$, and $\kappa_3 = 10.75$. All subsequent runs were conducted with these parameter values fixed.

Table 1: A canonical cophenetic matrix for the tree in Figure 1, with canonical ordering (5,7,4,1,2,6,3). For specificity, coalescent times $\mathbf{T}$ in Figure 1 are set at (0.8,0.3,0.7,0.5,0.9,1.5), yielding a superdiagonal (9.4,1.6,4.6,6.4,3.6,2.2). Redundant lower triangular entries are suppressed. The shorthand notation $(\sigma, \mathbf{a})$ becomes $\{(5,7,4,1,2,6,3),(4.7,0.8,2.3,3.2,1.8,1.1)\}$.

Canonical Ordering of Leaf Labels	5	7	4	1	2	6	3
5	0	9.4	9.4	9.4	9.4	9.4	9.4
7		0	1.6	4.6	6.4	6.4	6.4
4			0	4.6	6.4	6.4	6.4
1				0	6.4	6.4	6.4
2					0	3.6	3.6
6						0	2.2
3							0

Table 2: Enumeration of the Clarkia topologies visited by the Markov chain, ranked by frequency of occurrence. Letters represent the clades defined in Figure 3. Topologies I through III constitute a 99% credible region.

Label	Parenthetic Representation of the Topology	Labeled Histories Types	Labeled Histories Counts	Frequencies Relative	Frequencies Cumulative
I	(A(B(C(7D))))	34	1621	0.649	0.649
II	((A(C(7D)))B)	28	447	0.179	0.828
III	((AB)(C(7D)))	28	419	0.168	0.996
IV	(A(B((C7)D)))	4	5	0.002	0.998
V	((AB)((C7)D))	3	3	0.001	0.999
VI	((A((C7)D))B)	2	2	0.0008	0.9996
VII	((AB)((CD)7))	1	1	0.0004	1.0000

Table 3: Tribal classification of 31 species of African cichlid fish. Taxa 1 through 5 form a flock from Lake Malawi. The remainder from Lake Tanganyika constitute a Tanganyikan flock. The Malawi, Ectodini, and Lamprologini tribes are represented by the letters A, C, and D, respectively. B consists of $\{6, 7, 8, 9\}$, a combination of most of Tropheini and one species of Limnochromini. E = $\{22, 23, 24, 26, 27\}$ and F =$\{28, 29, 30, 31\}$ are convenient conglomerations (pseudo-clades) of remaining tribes. Taxa $\{25\}$ is not grouped. Taxa $\{32\}$ is an outgroup from Central America.

Label	Species Name	Tribe	Clade
1	Pseudotropheus zebra	Malawi	A
2	Buccochromis lepturus	Malawi	A
3	Champsochromis spilorhynchus	Malawi	A
4	Lethrinops auritus	Malawi	A
5	Rhamphochromis sp.	Malawi	A
6	Lobochilotes labiatus	Tropheini	B
7	Petrochromis orthognathus	Tropheini	B
8	Gnathochromis pfefferi	Limnochromini	B
9	Tropheus moorii	Tropheini	B
10	Callochromis macrops	Ectodini	C
11	Cardiopharynx schoutedeni	Ectodini	C
12	Opthalmotilapia ventralis	Ectodini	C
13	Xenotilapia flavipinnus	Ectodini	C
14	Xenotilapia sima	Ectodini	C
15	Chalinochromis popeleni	Lamprologini	D
16	Julidochromis marlieri	Lamprologini	D
17	Telmatochromis temporalis	Lamprologini	D
18	Neolamprologus brichardi	Lamprologini	D
19	Neolamprologus tetracanthus	Lamprologini	D
20	Lamprologus callipterus	Lamprologini	D
21	Lepidiolamprologus elongatus	Lamprologini	D
22	Perissodus microlepis 1	Perissodini	E
23	Perissodus microlepis 2	Perissodini	E
24	Cyphotilapia frontosa	Tropheini	E
25	Tanganicodus irsacae	Eretmodini	unattached
26	Limnochromis auritus	Limnochromini	E
27	Paracyprichromis brieni	Cyprichromini	E
28	Oreochromis niloticus	Tilapiini	F
29	Tylochromis polylepis	Tylochromini	F
30	Boulengerochromis microlepis	Tilapiini	F
31	Bathybates sp.	Bathybatini	F
32	Cichlasoma citrinellum	Central America	outgroup

Table 4: Most probable tree topologies from the combined runs for the cichlid fish data set. The posterior probabilities are averaged over the ten separate runs. The standard deviation, minimum, and maximum of the computed posterior probabilities of the ten separate runs are in the last three columns. Letters indicate monophyletic groups and subscripts denote subtopologies within these clades. Subtopologies for the A clade are depicted in Figure 6

	Parenthetic Representation	Posterior		Variation among Runs		
Rank	of the Topology	Prob.	Cumul.	SD	min	max
1	$(((((((A_1B_1)E_1)C_1)D_1)25)F_1)32)$	0.0714	0.0714	0.0062	0.0630	0.0844
2	$(((((((A_1B_1)E_1)C_1)D_1)25)F_2)32)$	0.0567	0.1281	0.0080	0.0464	0.0708
3	$(((((((A_1B_1)E_1)C_1)D_2)25)F_1)32)$	0.0342	0.1623	0.0088	0.0216	0.0458
4	$(((((((A_1B_1)E_1)C_1)D_2)25)F_2)32)$	0.0293	0.1916	0.0062	0.0210	0.0422
5	$(((((((A_1B_2)E_1)C_1)D_1)25)F_1)32)$	0.0271	0.2187	0.0064	0.0166	0.0366
6	$(((((((A_1B_1)E_2)C_1)D_1)25)F_1)32)$	0.0238	0.2425	0.0024	0.0186	0.0264
7	$(((((((A_1B_2)E_1)C_1)D_1)25)F_2)32)$	0.0215	0.2640	0.0057	0.0104	0.0278
8	$(((((((A_1B_1)E_2)C_1)D_1)25)F_2)32)$	0.0200	0.2840	0.0041	0.0138	0.0274
9	$(((((((A_1B_1)E_1)C_1)D_1)25)F_3)32)$	0.0188	0.3028	0.0028	0.0156	0.0232
10	$(((((((A_2B_1)E_1)C_1)D_1)25)F_1)32)$	0.0171	0.3199	0.0022	0.0144	0.0210
11	$(((((((A_2B_1)E_1)C_1)D_1)25)F_2)32)$	0.0139	0.3337	0.0021	0.0110	0.0176
12	$(((((((A_1B_1)E_1)C_2)D_1)25)F_1)32)$	0.0124	0.3461	0.0041	0.0038	0.0170
13	$(((((((A_1B_1)E_2)C_1)D_2)25)F_1)32)$	0.0102	0.3563	0.0033	0.0044	0.0152
14	19 others	≥ 0.0050	0.4924			
33	117 others	≥ 0.0010	0.7367			
150	683 others	≥ 0.0001	0.9258			

$B_1 = (((0607)09)08)$ $\quad D_1 = (((15(1920))((1617)18))21)$ $\quad F_1 = ((2829)(3031))$

$B_2 = (((0607)08)09)$ $\quad D_2 = ((15(((1617)18)(1920)))21)$ $\quad F_2 = (((2829)30)31)$

$C_1 = ((10(1112))(1314))$ $\quad E_1 = (((2223)27)(2426))$ $\quad F_3 = (((2829)31)30)$

$C_2 = (10((1112)(1314)))$ $\quad E_2 = (((2223)(2426))27)$

Table 5: Swapping between subtopologies for a single run on the cichlid data. Each row shows the distribution of the subtopologies that immediately follow a given subtopology in the saved sequences of trees. For example, subtopology A_2 followed A_1 719 times in the 5000 saved tree topologies. Topologies which appear only infrequently have been collapsed. A tree in which a clade does not appear is tabulated under $-$.

	A_1	A_2	A_3
A_1	3024	719	142
A_2	715	180	34
A_3	146	30	9

	B_1	B_2	B_3	B_x
B_1	3219	53	122	55
B_2	54	1175	1	39
B_3	146	30	9	185
B_x	58	39	5	12

	C_1	C_2	C_3
C_1	4122	46	7
C_2	49	765	1
C_3	5	3	1

	D_1	D_2	D_3
D_1	3119	19	2
D_2	18	1836	1
D_3	2	1	1

	E_1	E_2	E_x	$-$
E_1	2579	106	37	228
E_2	114	734	21	111
E_x	35	28	148	28
$-$	222	112	32	464

	F_1	F_2	F_3	$-$
F_1	1602	366	234	5
F_2	364	1407	195	9
F_3	237	192	205	4
$-$	5	9	4	161

Table 6: Marginal posterior probabilities for the top three subtopologies within each clade of cichlids. Subtopologies within clades are numbered in order of their posterior probability. Their combined posterior probabilities are summed. The final column is the probability that the constituent species are not monophyletic.

Clade	X_1	X_2	X_3	Combined prob.	Prob. not a clade
A	0.776	0.183	0.041	1.000	0
B	0.709	0.239	0.034	0.982	0
C	0.856	0.141	0.003	1.000	0
D	0.661	0.338	0.001	1.000	0
E	0.608	0.191	0.023	0.822	0.161
F	0.446	0.393	0.125	0.964	0.036

Table 7: Variation in the inter-clade coalescence of cichlids. Although A and B always merge together first, we continue to differentiate them to emphasize their disparate geographic origins. Uncertainty in the location of taxa in clade E and taxon 25 cause almost all of the variability at this level of summary.

Clade Trees	Posterior Prob.	Cumulative Prob.
$((((((((AB)E)C)D)25)F)32)$	0.645	0.645
$((((((AB)E)C)(D25))F)32)$	0.102	0.747
$((((((((AB)(2426))((2223)27))C)D)25)F)32)$	0.040	0.789
$((((((((AB)((2223)27))(2426))C)D)25)F)32)$	0.033	0.821
$(((((((AB)E)C)D)F)25)32)$	0.027	0.848
$((((((((AB)((2223)(2426)))27)C)D)25)F)32)$	0.024	0.872
$(((((((AB)E)C)25)D)F)32)$	0.017	0.889
$((((((((AB)E)C)D)25)((2829)30))31)32)$	0.017	0.906
$((((((((AB)27)((2223)(2426)))C)D)25)F)32)$	0.015	0.921

Table 8: Comparison of the Markov chain Monte Carlo estimate of the phylogeny for the cichlids with estimates using other methods. In the neighbor-joining solution, the use of [,] indicates how species in Clade E have been redistributed across clades. An additional subtopology is $B_5 = ((06(0708))09)$.

Technique	Topology
Neighbor-joining plus parsimony	$((([[[(A_2 B_5)[C_2(2223)]](2426)]27)(D_2\ 25))F_3)$
Maximum likelihood	$(((((A_2 B_2)((((2223)24)27)26))C_2)(D_2\ 25))F_2)$
Markov chain Monte Carlo	$(((((A_1 B_1)(((2223)27)(2426)))C_1)(D_1\ 25))F_1)$

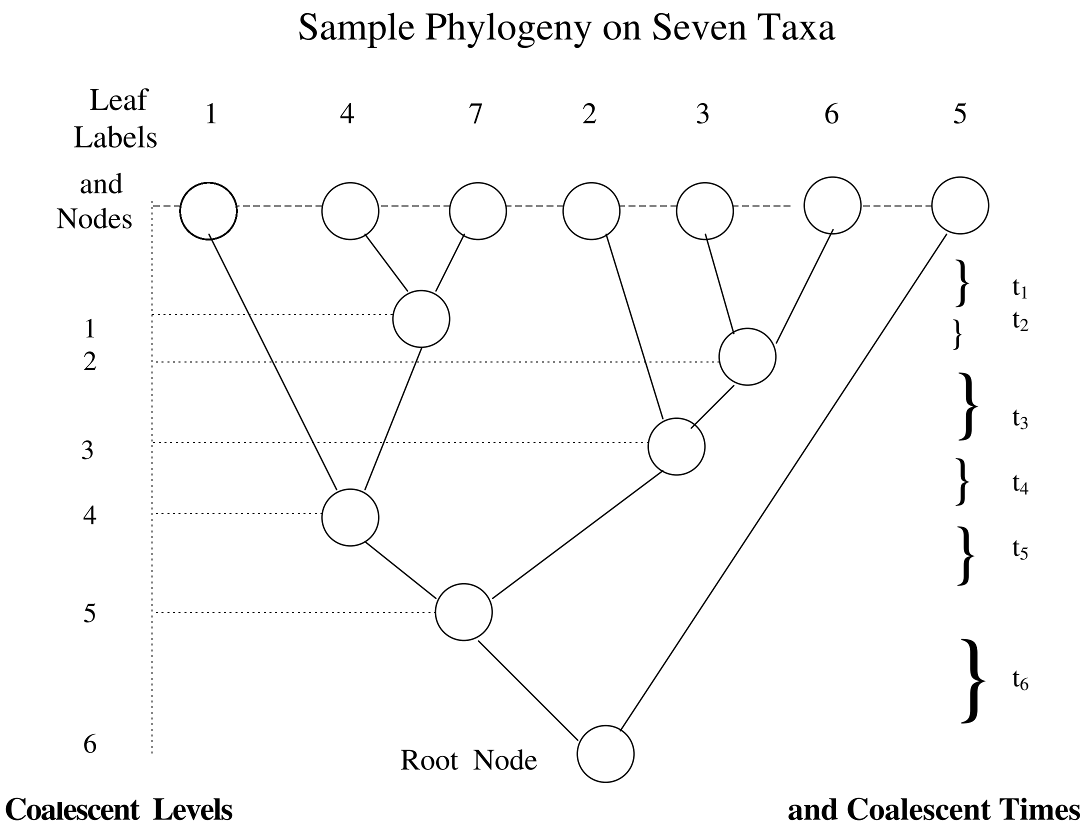

Figure 1: Labeled leaf nodes across the top identify present-day taxa. Internal nodes drawn below represent common ancestors, with connecting edges indicating lines of descent. $\{1, 4, 7\}$ and $\{2, 3, 6\}$ are examples of monophyletic groups (or clades), subsets of organisms whose most recent common ancestors have no other descendants among the considered organisms.

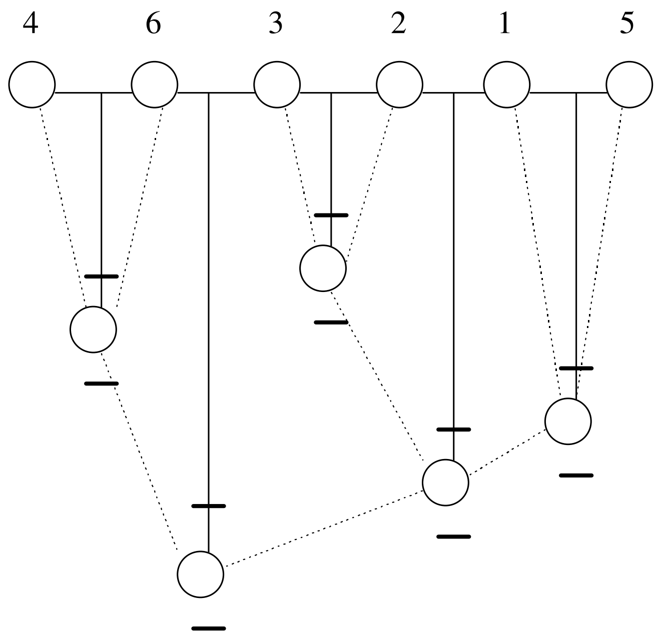

Figure 2: Candidate trees, attainable from the current tree, are characterized by intervals of size 2δ, centered at the current internal nodes, that constrain the repositioning of the internal nodes.

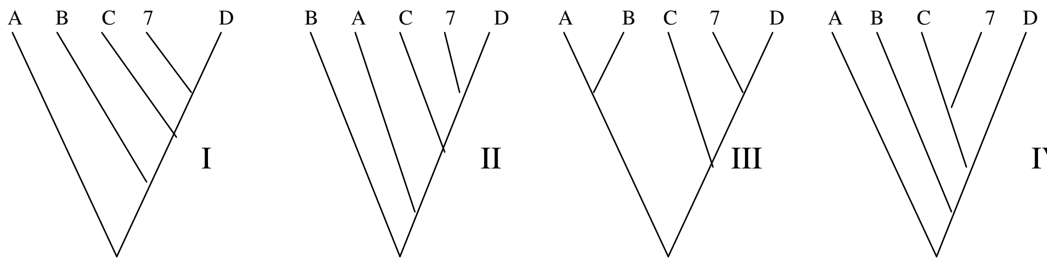

Figure 3: Shapes of the four most common topologies for the nine species of Clarkia, where $A = \{1, 2\}, B = \{3, 4\}, C = \{5, 6\}$, and $D = \{8, 9\}$ denote monophyletic pairs. The first three topologies differ only in the placement of the root.

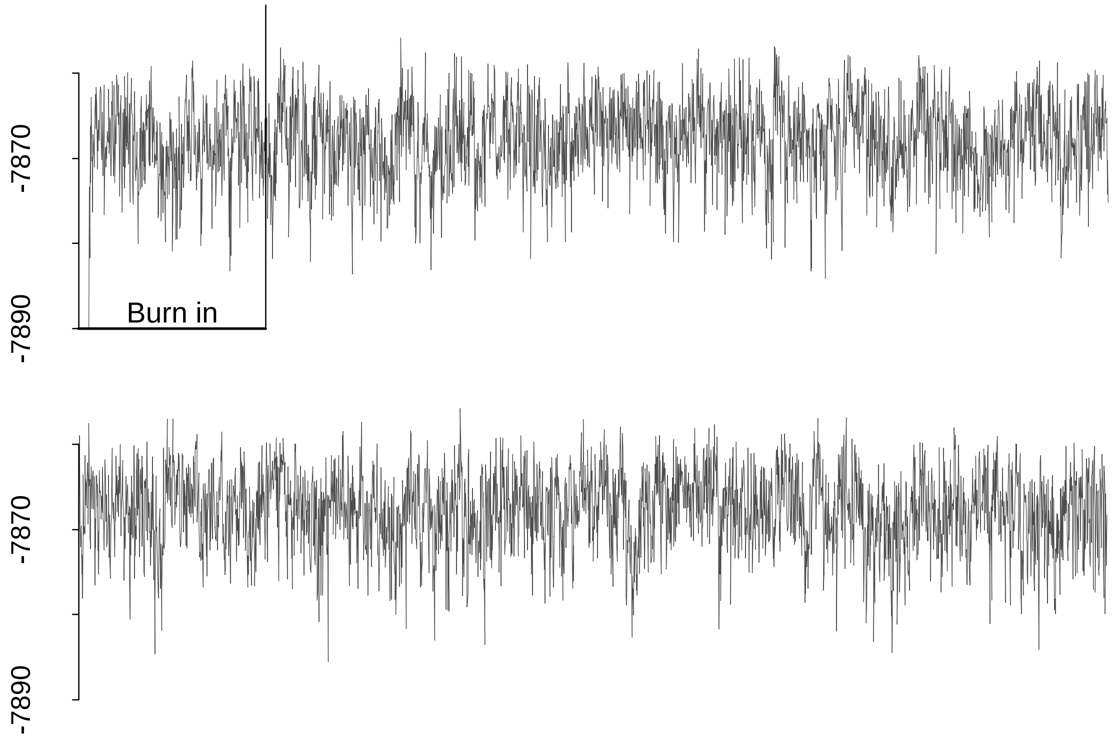

Figure 4: Time series plot of the log posterior for 5500 subsampled trees of cichlids. The log posterior plateaus at a mean value of -7868 after 75 sample points, indicating a burn-in of 500 is extremely conservative for this particular run. Lower panel shows second half of our run.

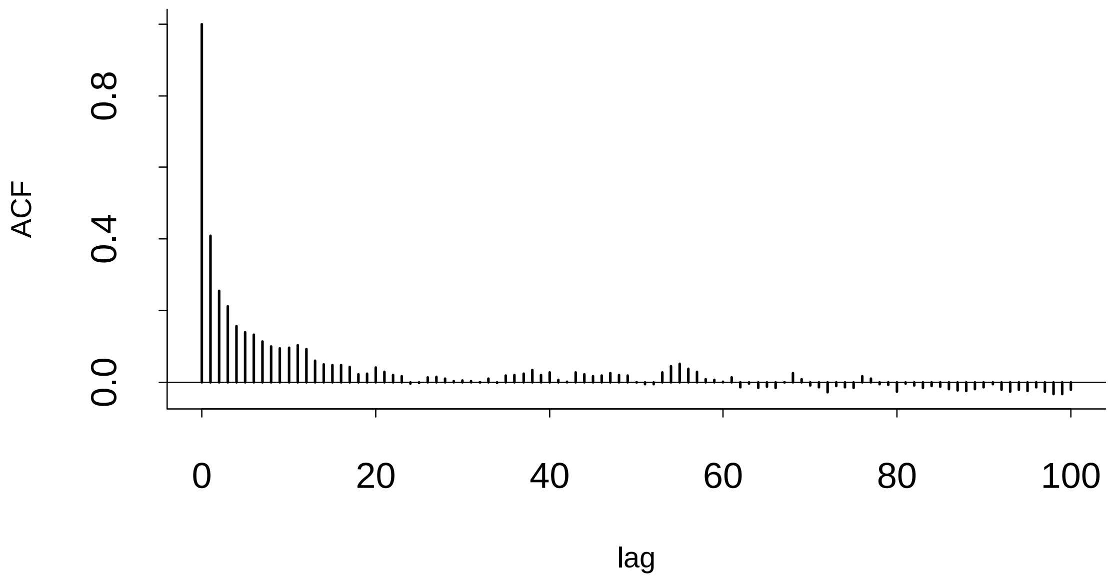

Figure 5: Autocorrelation function of the log posterior displayed in Figure 4. ACF drops below noise level at approximately 20 lags.

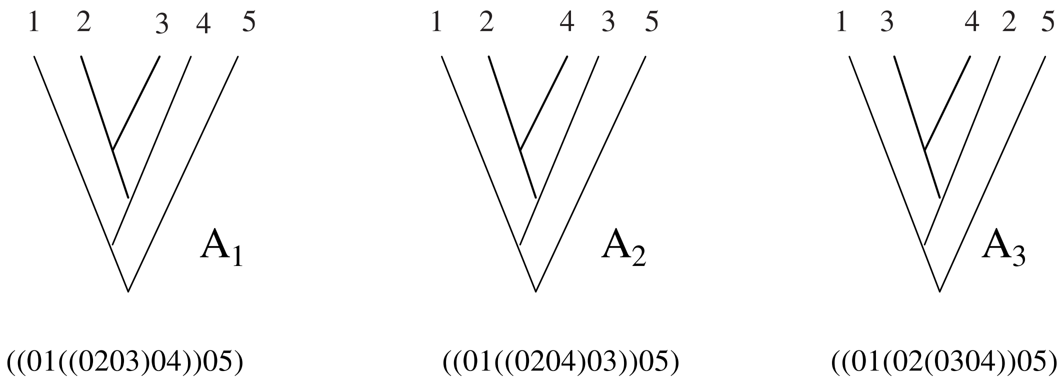

Figure 6: Supported subtopological variation within the Malawi flock $A = \{1, 2, 3, 4, 5\}$. Variability is confined to coalescences involving taxa $\{2, 3, 4\}$. Marginal posterior probabilities are 0.776, 0.163, and 0.041 respectively.