Statistics in Molecular Biology
IMS Lecture Notes - Monograph Series (1999) Volume 33

MARKOV CHAIN MONTE CARLO FOR THE BAYESIAN ANALYSIS OF EVOLUTIONARY TREES FROM ALIGNED MOLECULAR SEQUENCES

BY MICHAEL A. NEWTON, BOB MAU AND BRET LARGET[1]

University of Wisconsin-Madison and Duquesne University

We show how to quantify the uncertainty in a phylogenetic tree inferred from molecular sequence information. Given a stochastic model of evolution, the Bayesian solution is simply to form a posterior probability distribution over the space of phylogenies. All inferences are derived from this posterior, including tree reconstructions, credible sets of good trees, and conclusions about monophyletic groups, for example. The challenging part is to approximate the posterior, and we do this by constructing a Markov chain having the posterior as its invariant distribution, following the approach of Mau, Newton, and Larget (1998). Our Markov chain Monte Carlo algorithm is based on small but global changes in the phylogeny, and exhibits good mixing properties empirically. We illustrate the methodology on DNA encoding mitochondrial cytochrome oxidase 1 gathered by Hafner *et al.* (1994) for a set of parasites and their hosts.

1. Introduction. Stochastic models have long been considered useful for describing variation in the molecular sequences of extant populations (e.g., Jukes and Cantor, 1969; Felsenstein, 1973; Kimura, 1980). Parameters in such models include the phylogeny, which encodes the pattern of evolutionary relationships among populations, and substitution rates, which describe how molecules change over time within populations. It seems quite natural to infer these parameters using the induced likelihood function in some way, but such inference has been difficult in practice because computations can be prohibitively expensive. Owing to the Markovian nature of the standard models, evaluation of the likelihood function follows straightforward recursive equations, and so evaluation is not the difficult part. The difficulty arises with optimization, since the likelihood resides over a complicated parameter space, and seems to admit no simple representation (Felsenstein, 1981, 1983; Goldman, 1990; Yang, Goldman, Friday, 1995). Nevertheless, computer code is available for approximate maximum likelihood calculation (Olson, *et al.*, 1994; Felsenstein, 1995; Swofford, 1996).

Beyond estimation, practitioners have demanded some way to assess uncertainty in aspects of the estimated phylogeny, just as error bars accompany simpler kinds of point estimates. A standard and appealingly simple calculation

[1]B. Larget acknowledges the support of the National Science Foundation.
AMS 1991 *subject classifications.* Primary 62F99; secondary 92D20.
Key words and phrases. Cospeciation, Metropolis-Hastings algorithm, phylogeny.

is to apply Efron's bootstrap (Felsenstein, 1985), and although this method may accurately approximate sampling distributions, its role for statistical inference about the phylogeny has been a matter of some debate (e.g., Felsenstein and Kishino, 1993; Newton, 1996; Efron, Halloran, and Holmes, 1996; Chernoff, 1997). Of course, bootstrapping a complicated estimator serves to compound the computational problem. Thus, bootstrapping a full-blown maximum likelihood estimator is practically impossible with today's implementations. A common practice is to bootstrap a much simpler estimator.

An alternative, model-based, assessment of uncertainty was postulated some time ago by J.F.C. Kingman, in the discussion of Joe Felsenstein's 1983 paper on statistical issues in evolutionary biology:

> In view of the difficulties of the maximum likelihood approach, it seems worth asking what a Bayesian analysis would look like. The author has shown us how to write down the likelihood function, and this has only to be multiplied by a suitable prior. ... The result is a set of posterior probabilities for collections of possible phylogenies, not just a single estimate, and it may well be that there are tractable approximations of the probabilities of some compound events. Has this approach been explored?

Until recently, this Bayesian approach had not been explored. Sinsheimer et al. (1996) developed exact Bayesian calculations for the four-species problem. Several groups have been pursuing Markov chain Monte Carlo (MCMC) approximations. Mau and Newton (1997) described an MCMC method for models satisfying a molecular clock, and presented calculations for binary, restriction-sites data. Mau, Newton, and Larget (1998) have extended these calculations to problems with more taxa and nucleotide sequence data. Yang and Rannala (1997), and Li, Pearl and Doss (1996) have developed different Markov chain Monte Carlo strategies for the same general problem. In fact, the MCMC method of Kuhner, Yamato, and Felsenstein (1995) can be modified to produce approximate Bayesian phylogenetic inferences, even though their model considers within population sampling of sequences. The purpose of the present article is to review the Mau, Newton, Larget approach and to illustrate the calculations in an example.

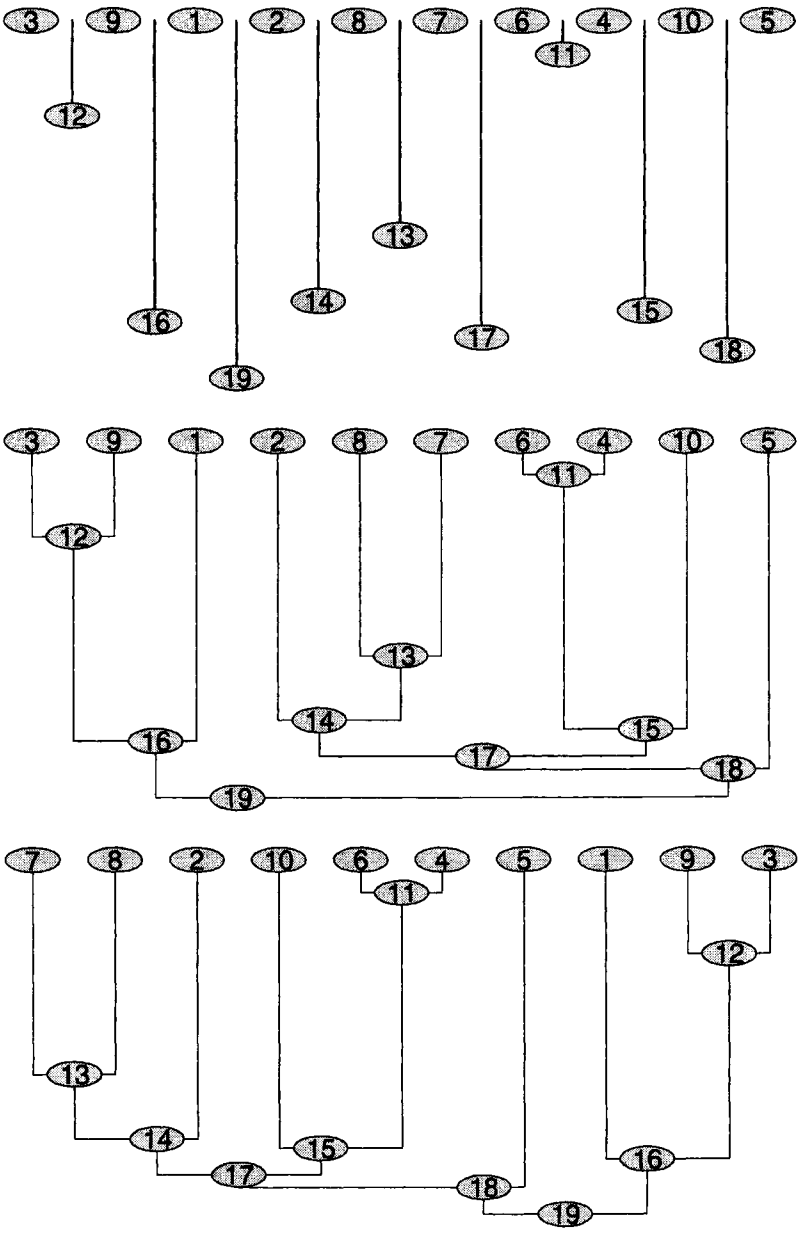
Of course Bayesian analysis provides more than assessments of uncertainty about phylogenies, the focus of this work. The array of inference problems presented in Huelsenbeck and Rannala (1997), for example, all may be approached from a Bayesian perspective. We anticipate that future research will clarify the role of Bayesian analysis for evolutionary biology, but first some essential computational problems must be addressed.

2. Phylogeny. A phylogeny or evolutionary tree, τ, admits various representations. For the present discussion, it will be convenient to treat τ as a pair (t, σ) where $t = (t_1, \ldots, t_{s-1}) \in \mathbb{R}^{s-1}$ is a vector of positive speciation times, s is the number of species under consideration, and σ is a permutation of $\{1, 2, \ldots, s\}$. The path of evolution corresponding to τ can be envisioned by processing t and σ in a manner as illustrated in Figure 1. Species labels $\{1, 2, \ldots, s\}$ become leaf nodes of a tree upon being arranged horizontally in the order determined by σ. Moving from left to right, we drop a vertical line of length t_i from a point in between the ith and $(i+1)th$ leaf node, and we call the lower endpoint an internal node. In Figure 1, internal nodes are labeled by increasing speciation time. We draw a tree by moving downwards from the species labels, establishing in turn each internal node as the parent of two descendant nodes. Acting on a given internal node j, say, an edge is formed between node j and the parentless node $k < j$ having horizontal position closest to node j from the left, and a second edge is formed to the nearest such parentless node from the right. Eventually, all nodes are connected, and a tree results. The node corresponding to the largest t_i is called the root.

Several remarks are in order regarding this construction. The horizontal axis serves only to organize the information, and has no intrinsic scale. On the other hand, the vertical axis records time into the past. As drawn, our trees are rooted and have contemporaneous tips, and will be considered parameters of models which satisfy the molecular clock hypothesis. In work to extend our methods to models where evolutionary rates vary among branches of the tree, a somewhat different tree representation is more appropriate than the one just described. It is noteworthy, however, that the essential elements of the Monte Carlo algorithm to be described carry over readily to this more general case. Note that the drawing algorithm has not been defined when two times are equal, and so we omit this case (i.e., assume $t_i \neq t_j$ for all i, j.) and thus consider only binary trees. This is not a serious restriction because the likelihood function (Section 3) is continuous in t, and hence the likelihood of a polytomy may be arbitrarily close to the likelihood of a binary tree produced by resolving the polytomy with tiny branch lengths.

The phylogeny $\tau = (t, \sigma)$ records the path of evolution from a single ancestral population to the present array of s populations under study. Any point on the tree drawn according to the rules above thus represents a population at some time in the past. Evidently, different (t, σ) pairs determine the same path of evolution, noting again that the horizontal axis in Figure 1 has no scale. For example, rearranging species 3 and 9 does not change the path of evolution. A given unordered set of times $\{t_1, t_2, \ldots, t_{s-1}\}$ induces the same path of evolution in 2^{s-1} different ways. To see this, take a phylogeny and rotate the graph by

FIG. 1. *Phylogeny: The top panel shows the raw ingredients in the (t, σ) representation of a phylogeny. Internal nodes appear at the lower end of the vertical line dropped from the horizontal. The middle panel shows the tree formed by processing the first panel, that is by moving down from the leaves towards the root, establishing connections each time an internal node is reached. The bottom panel shows a second version of the same phylogeny.*



180 degrees above any of the $s-1$ internal nodes. Strictly speaking, therefore, the phylogeny τ is an equivalence class containing 2^{s-1} different versions (t, σ). The third panel in Figure 1 shows a second version of the preceding tree.

The representation of a phylogeny as 2^{s-1} points (t, σ) is particularly conducive to Markov chain Monte Carlo (MCMC), as we discuss in Section 4. A key feature is that the tree is part of a continuum, and the branching pattern of the tree is induced by the permutation σ and the relative ordering of the times $t_1, \ldots, t_{s-1}$. Indeed, the branching pattern inherent in τ may be of interest, but we do not represent that pattern directly, choosing instead to work with more elementary objects which combine to produce the pattern. This somewhat indirect approach leads to very simple MCMC steps (Section 4) and may be associated with the efficiency of the algorithm.

Different summaries of τ may be of interest to the biologist. The *labeled history* describes the branching pattern of τ obtained by ignoring the magnitude of the times $t_1, t_2, \ldots, t_{s-1}$, but respecting their ordering. Incidentally, counting labeled histories is quite simple given our construction, as there are $s!$ ways to arrange the s species labels, $(s-1)!$ orderings of the times, but we have overcounted by 2^{s-1}, leaving $s!(s-1)!/2^{s-1}$ distinct labeled histories. The *tree topology* corresponding to τ is another property of its branching pattern, where we record only the sequence of connections, but ignore details of their time ordering. The tree topology may be characterized by nested parentheses, such as

$$\mathsf{top}(\tau) = ((1,(3,9)),(((2,(7,8)),((4,6),10)),5)) \tag{2.1}$$

for the phylogeny shown in Figure 1. Here we have taken the convention that when two groups of organisms are merged, we place on the left that group containing the smallest species label. Assessing uncertainty in the tree topology is often of interest and will be the focus of our application in Section 5.

3. Modeling substitutions. The probability of data given a tree τ is derived from a model of DNA evolution along the branches, and many such models have been studied. We follow convention here and take the same general assumptions as those characterizing many standard models. That is, we consider the extant DNA sequences to be aligned into n sites, and we suppose that the evolution of different sites is independent. At any given site, a stochastic process associates a DNA base to each point on the branches of τ, and observed data are the s DNA bases at the leaf nodes. The standard models assume that base substitutions occur at points of Poisson processes, with independent evolution among branches. These restrictions still leave us some flexibility in the modeling of substitutions (Yang, Goldman, and Friday, 1994). It is noteworthy that the

MCMC algorithm discussed in Section 4 is not linked to the particular model of evolution. As long as likelihood evaluation is a feasible calculation, we can readily implement a posterior simulation. This is in contrast to Gibbs sampler algorithms, for example, whose very structure is determined by the likelihood function under consideration.

In Section 5 we report calculations for the model of Hasegawa, Kishino, and Yano (1985), which, being richer in parameters, subsumes the earlier models of Jukes and Cantor (1969), Kimura (1980), and Felsenstein (1981). The generator matrix for a process governed by HKY85 contains the following infinitesimal rates of change (the diagonal is determined because rows sum to 0):

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} \cdot & \kappa\pi_g & \pi_c & \pi_t \\ \kappa\pi_a & \cdot & \pi_c & \pi_t \\ \pi_a & \pi_g & \cdot & \kappa\pi_t \\ \pi_a & \pi_g & \kappa\pi_c & \cdot \end{pmatrix} \end{array}$$

The π's indicate long-run probabilities of each base along one very long branch and κ allows different substitution rates for transitions (changes between A and G or between C and T) and transversions (any other changes). The infinitesimal rates determine transition probabilities from one base to another over any extended time period, and these further involve a mutation rate parameter θ. We omit details here.

By the independent-sites assumption, the likelihood is a product of n factors, one from each site in the aligned sequences. This naturally collapses to a product over $m \leq n$ unique observed patterns of s bases, and so fixing s, the likelihood evaluation takes $O(m)$ operations. Furthermore, the Poisson process assumption implies that evolution is Markovian, and thus that the probability of a given pattern can be calculated recursively in $O(s)$ steps. This *pruning* algorithm is a critical component of our procedure, and so we review it briefly (see also, Felsenstein, 1983). Let u_i denote the unknown base at the site of interest in the ancestral sequence associated with internal node i of the tree τ. Note that $s+1 \leq i \leq 2s-1$ and $u_i \in \{A, G, T, C\}$. Each internal node partitions the descendant species into two distinct groups, whose observed DNA data we label $\mathsf{A}(i)$ and $\mathsf{B}(i)$. By the assumed independence of substitutions among branches, the conditional probability of all data descending from i, given u_i, is

$$p\{\mathsf{A}(i), \mathsf{B}(i) \mid u_i, \tau\} = p\{\mathsf{A}(i) | u_i, \tau\} \times p\{\mathsf{B}(i) | u_i, \tau\}. \tag{3.2}$$

These probabilities are important because the likelihood contribution from a

site with the given pattern is

$$\sum_{u_{\text{root}}} p\{\mathsf{A}(\text{root}),\ \mathsf{B}(\text{root})\ |u_{\text{root}}, \tau\}\ p(u_{\text{root}}). \tag{3.3}$$

That is, it is a mixture of transition probabilities against the distribution of the unknown base at the root node. By taking these initial base probabilities equal to the stationary base probabilities π_a, π_c, π_t, or π_g, the Markov process becomes reversible. To implement the pruning algorithm, one observes that by the Markov property, probabilities in (3.2) may be obtained recursively, moving from the leaves of the tree to the root. In our labeling system, the recursion moves successively through internal nodes $i = s + 1$ to $i = 2s - 1$.

On the phylogeny in Figure 1, for example, $p\{\mathsf{A}(12),\ \mathsf{B}(12)\ |u_{12}, \tau\}$ is the product of the conditional probability of data for species 3 and species 9 given u_{12}, the base at internal node 12. These four conditional probabilities are used subsequently to evaluate the conditional probability of data descending from node 16, given u_{16}, which finally enters the likelihood calculation (3.3).

We note that the (t, σ) representation of τ is not the one most conducive to the pruning calculation which relies directly on relationship information in τ. Our software uses a second representation in which every internal node is associated with its descendant nodes.

In summary, likelihood evaluation is a straightforward calculation when we fix the data, the tree, and parameters governing the substitution model.

4. The posterior and MCMC. In contrast to other forms of statistical inference, Bayesian inference centers on the extent to which opinion about an unknown is affected by data. Furthermore, probability is the sole medium for transmitting uncertainty and opinion (e.g., Bernardo and Smith, 1994). To implement an analysis, a Bayesian evolutionary biologist must therefore begin with a probability distribution over the set of possible phylogenies. This might be derived from a model of speciation, or from the analysis of existing data. To our knowledge, little work has been done on the assessment of *prior* probabilities for trees, but this certainly represents an important problem if Bayesian analysis is to be helpful in evolutionary biology. In the present work, we illustrate calculations with a particularly simple *flat* prior, and note that the algorithm proceeds easily with any user-supplied prior distribution. The flat prior we assume is relative to the (t, σ) representation of the phylogeny, in suggestive notation:

$$p(\tau) = p(t)\,p(\sigma) = \left(\prod_{i=1}^{s-1} p(t_i)\right) \frac{1}{s!} \propto \left(\prod_{i=1}^{s-1} 1[0 \le t_i \le t_{\max}]\right) \tag{4.4}$$

$$\propto 1\,[0 \le t_i \le t_{\max}, \text{ for all } i]\,.$$

Here $t_{\max}$ bounds the time to the root node. One consequence of this prior is that, like the Kingman coalescent (Kingman, 1982), we induce a uniform probability distribution over labeled histories, and thus a non-uniform distribution over topologies. This prior favors balanced tree topologies because they have many more labeled histories than unbalanced ones (e.g., Lapointe and Legendre, 1991; Brown, 1994). Checking the sensitivity of our calculations to the choice of prior will be critical in applications, but we do not pursue such sensitivity analysis here.

Parameters of the substitution model also are unknown and a full Bayesian analysis requires a prior distribution for them. In this section we focus on the phylogeny, and thus we consider all other parameters to be known. For example, the stationary base probabilities $\pi_a, \pi_c, \pi_t, \pi_g$ can be estimated by the relative frequency of the different bases in the observed sequences.

In light of the data, and taking as reasonable the stochastic model of evolution, inference about the phylogeny τ must be based on the posterior distribution, having density

$$p(\tau|\mathsf{data}) \propto p(\mathsf{data}|\tau) \times p(\tau). \tag{4.5}$$

Monte Carlo appears to be the only effective method for summarizing this distribution, even though the pruning algorithm enables evaluation of the posterior up to a constant. Inference about monophyletic groups, most probable topologies, and the uncertainty in certain branch points, for example, all are based on expectations with respect to this posterior. Within the class Monte Carlo algorithms, Markov chain methods present the most promising integration methods, and we review here the proposal of Mau, Newton, and Larget (1998).

An MCMC algorithm realizes a Markov chain $\tau^1, \tau^2, \ldots, \tau^B$ that has (4.5) as its stationary distribution (e.g., Tierney, 1994). Empirical averages in the chain converge as B grows to posterior expectations by the law of large numbers for Markov chains. We construct our Markov chain using the Metropolis-Hastings approach. That is, we move from $\tau^i = \tau$ to the next state τ^{i+1} by first proposing a candidate phylogeny τ^* generated according to a proposal distribution that has transition density $q(\tau, \tau^*)$. Next we compute the Metropolis-Hastings ratio

$$r = \frac{p(\tau^*|\mathsf{data})\, q(\tau^*, \tau)}{p(\tau|\mathsf{data})\, q(\tau, \tau^*)}. \tag{4.6}$$

If $r \geq 1$, then $\tau^{i+1} = \tau^*$. Otherwise, we move to τ^* with probability r and stay put with probability $1 - r$. The power of this approach resides both in its simplicity and in its great flexibility, because the choice of q, which affects the Monte Carlo efficiency of the algorithm, is almost arbitrary.

Monte Carlo approximations of posterior probabilities can be biased if the distribution of τ^1 is far from the target posterior distribution, and so it is common practice to let the chain run for a burn-in period before using any of the sampled states. Determining the length of the burn-in and the total chain length B to ensure accurate approximations is difficult in advance, and is typically based on realizations of the chain that are monitored using a range of diagnostic checks (e.g., Cowles and Carlin, 1996).

In most implementations of the Metropolis-Hastings algorithm, a collection of proposal distributions determine the complete algorithm (e.g., Besag *et al.*, 1995). We have found that a single proposal distribution works for the phylogeny problems considered so far. This proposal distribution is global in that τ^* can differ from τ in *all* respects, and so, in a sense, we have attempted to design efficiency into the algorithm. Inefficient algorithms are ones which traverse the parameter space slowly and thus exhibit significant positive correlations on one-dimensional summaries. Local, single-site updating proposals change parts of the parameter at a time, and are at risk for low efficiency. One risk of a global proposal distribution, on the other hand, is that we may reject candidates too frequently, and thus produce an inefficient algorithm. We avoid this in two ways: by making our global changes small in magnitude, and by basing changes on distance within the tree, so that proposed trees are close in posterior density to the current tree.

More specifically, our proposal distribution works like this. We obtain at random from the equivalence class defining the current tree τ one of its 2^{s-1} versions, thus identifying a pair (t, σ). Fixing the leaf label permutation σ, we generate a new vector t^* of times by

$$t_i^* = t_i \oplus \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, s-1$$

where ϵ_i are independent and identically distributed Uniform$(-\delta, \delta)$ random variables for some tuning parameter $\delta > 0$, and $\oplus$ indicates addition reflected into the interval $(0, t_{\max})$. For example, $\oplus$ returns $|t_i + \epsilon|$ if $t_i + \epsilon < 0$. Thus the proposal is to perturb the speciation times of a version of the current tree.

When the tuning parameter δ is small, the candidate tree is close to the current tree in terms of pairwise distance between species, and so we expect the likelihood of the candidate tree to be close to that of the current tree. Similarity in likelihood is derived by the similar distance structure, and not by a direct appeal to the model, making the proposal method independent of the model form. Interestingly, the candidate tree can be quite different from the current tree in terms of branching structure.

In Figure 2, a version of τ from Figure 1 has had its species times perturbed,

FIG. 2. *Proposal: This graph shows how a candidate phylogeny τ^* is obtained from one version of the current tree by perturbing speciation times in the (t, σ) representation. The shaded boxes indicate the range of the uniform perturbations. The dark circles indicate times within the current tree, and crosses indicate times in a particular candidate τ^*.*

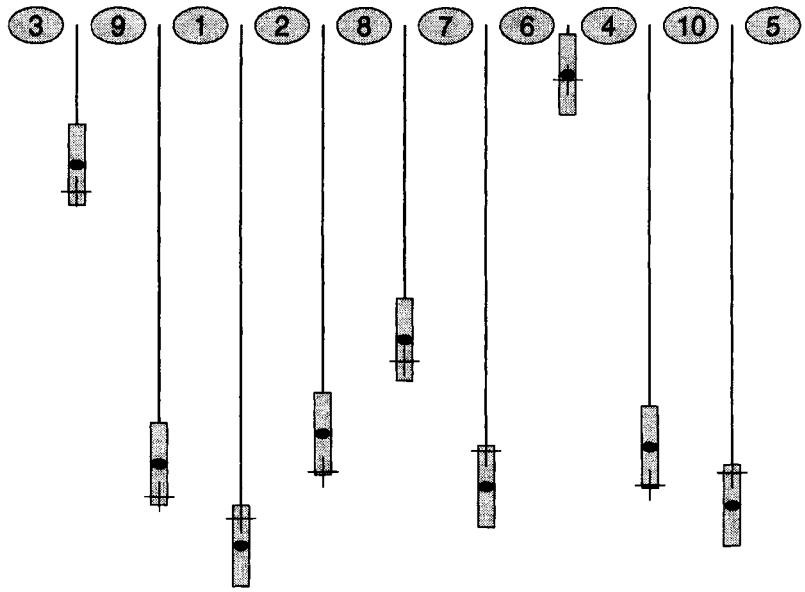


TABLE 1
Gopher/Lice Species Labels

Label	Louse Species	Label	Gopher Species
1	G. texanus	1	G. personatus
2	G. ewingi	2	G. breviceps
3	G. oklahomensis	3	G. bursarius (a)
4	G. geomydis	4	G. bursarius (b)
5	G. nadleri	5	P. bulleri
6	G. chapini	6	O. hispidus
7	G. panamensis	7	O. cavator
8	G. setzeri	8	O. underwoodi
9	G. cherriei	9	O. cherriei
10	G. costaricensis	10	O. heterodus
11	G. expansus	11	C. castanops
12	G. perotensis	12	C. merriami
13	G. trichopi	13	Z. trichopus

leading to a tree τ^* with a different topology:

$$\mathsf{top}(\tau^*) = ((1,(3,9)),((2,((4,6),(7,8))),(5,10)))$$

Compare with (2.1). Certainly movements induced by this proposal mechanism are restricted, unless δ is very large. Mau, Newton, and Larget (1998) established irreducibility; i.e., that starting at any phylogeny τ, and given any other phylogeny τ_0, there exists $K < \infty$ such that there is positive probability density of moving to τ_0 after K applications of the proposal. Mau, Newton, and Larget (1998) also established symmetry which means that the Metropolis-Hastings ratio (4.6) reduces to a ratio of posterior densities, and thus, under a flat prior, to a ratio of likelihoods.

5. An example: Host parasite evolution.

5.1. *Data and model structure.* We illustrate the MCMC calculations with data reported by Hafner *et al.* (1994) regarding a study of molecular evolution in hosts and their parasites. To facilitate a comparison, we focus on a subset that was analyzed by Huelsenbeck, Rannala, and Yang (1997) (hereafter HRY97). The data consist of 26 aligned DNA sequences, $n = 379$ bases long, encoding mitochondrial cytochrome oxidase I (COI) for 13 lice species and the corresponding 13 species of their gopher hosts (Table 1). There are $m = 156$ and $m = 130$ unique site patterns in the lice and gopher data, respectively. Table 2 shows summary frequencies of the different nucleotide bases, as well as the numbers of sites at which all bases are the same and sites exhibiting base variation.

Because the life cycle of the parasitic lice occurs exclusively in the fur of the host gophers, a natural hypothesis is that the organisms have coevolved, and thus have a common branching structure in their phylogenies. On the other hand, factors such as interaction among gopher species could produce differences between louse and gopher trees. Using the HKY85 model of DNA substitution discussed in Section 3, we compare the topological structure of host and parasite phylogenies. Our approach is to integrate results from separate analyses of the louse and gopher data.

The overall amount of DNA variation differs between gophers and lice, but within each group a molecular clock assumption is reasonable (HRY97, p. 414). By splitting the 26 taxa into two groups, we overcome the need to specifically model violations of a molecular clock.

We allow the mutation rate θ to vary over codon position because of the significant rate variation among sites (Table 2). It turns out that a model with codon-specific mutation rates fits somewhat better than the one used by HRY97

TABLE 2

Summary Statistics: Middle four columns show the observed base frequencies broken down by codon position for both louse and gopher data sets. The total number of sites is decomposed into n_c sites that have constant base value among all species, and n_v sites that exhibit variation.

data	codon	$\hat{\pi}_a$	$\hat{\pi}_g$	$\hat{\pi}_c$	$\hat{\pi}_t$	n_c	n_v
lice	1	0.281	0.362	0.101	0.254	101	25
	2	0.136	0.196	0.202	0.464	120	6
	3	0.321	0.185	0.112	0.388	2	125
	all	0.246	0.248	0.138	0.369	223	156
gopher	1	0.314	0.300	0.121	0.262	113	13
	2	0.160	0.172	0.230	0.435	122	4
	3	0.386	0.051	0.232	0.337	14	113
	all	0.287	0.174	0.195	0.345	249	130

in which mutation rates follow a discretized Gamma distribution. Observed base frequencies also vary over codon position, and so we similarly allow codon-specific relative frequency parameters. In the calculations reported below, we consider the base frequency parameters (Table 2) and the mutation rates to be fixed at their estimated values. Actually, we use some preliminary MCMC runs to estimate the mutation rate parameters, $(0.136, 0.026, 2.838)$ for the lice and $(0.154, 0.021, 2.825)$ for the gophers. These mean 1 vectors are empirical averages taken from long preliminary MCMC runs in which both the phylogeny and the mutation rate parameters were updated. Small posterior variance lead us to fix these rates at their estimated values.

HRY97 found evidence of transition/transversion bias, and so we follow suit, allowing a free parameter κ for each data set. Rather than fixing an estimated value, we augment the MCMC algorithm, including κ as an additional unknown, having a flat prior on the positive line. We considered a model with codon-specific κ, but this did not significantly improve fits.

5.2. *MCMC implementation.* For each data set, our Monte Carlo estimate of the posterior distribution over phylogenies is based on realizing four independent Markov chains of length 1,020,000. In the first 20,000 cycles of each run, only the phylogeny is updated, starting from a random tree drawn from the uniform prior distribution, and κ is fixed at a rough estimate obtained from preliminary runs (9.87 for lice, 11.45 for gophers). Subsequently, each cycle alternates between an update of τ given κ and an update of κ given τ, the latter based on a simple uniform window proposal distribution. During the initial cycles, we adaptively change the window size δ for the τ update, ultimately fixing values of

$\delta = 0.00625$ and $\delta = 0.003125$ for lice and gophers, respectively. Our automated method increases δ if the recent acceptance rate is high, and decreases it otherwise, but only adapts during the early burn-in phase. We routinely monitor the loglikelihood of sampled trees throughout this burn-in period, noting a typical pattern of dramatic increase followed by stability at some plateau. The pattern is consistent across independent runs.

After burn in, each production run is subsampled every 20 cycles to reduce the size of output used in estimating posterior probabilities. We calculated the tree topology and loglikelihood of all subsampled phylogenies. Figures 3 and 4 show some diagnostic plots from a further one out of 20 subsampling of these 50,000 phylogenies from one of the four runs. At this level of subsampling, there is very little within chain dependence both for the loglikelihood series and for the binary series indicating whether or not $\mathsf{top}(\tau)$ equals the most frequently observed topology. Rather than show a simple time-series plot of this binary variable, we use the cusum diagnostic suggested in Yu (1995). We simply plot the cumulative sum of the binary time series, centered by a cumulative sum of ones times the overall mean. Slow mixing can be diagnosed when we compare the cusum plot to a similar plot calculated on a random permutation of the binary series. Slow mixing is characterized by long excursions and a fairly smooth plot. Rapid mixing is indicated by these plots.

It is more relevant to consider dependence properties within each production run of 50,000 phylogenies than within the subsamples in Figures 3 and 4 because our Monte Carlo approximations arise from the former. Nevertheless, the plots provide some indication of sampler behavior and demonstrate adequate mixing on several summary quantities.

That four independent runs produce comparable results suggests that any posterior multimodality is not adversely affecting the sampler. Furthermore, the four independent runs provide simple Monte Carlo standard error estimates in place of the somewhat more complicated within-chain methods (Geyer, 1992).

5.3. *Posterior summaries.* Tables 3 and 4 summarize our Monte Carlo estimates of the posterior distribution over tree topologies separately for the lice and gophers. In each run, the posterior probability of a topology is calculated simply as its empirical relative frequency. Over the four runs, the average of these proportions is our reported Monte Carlo estimate, and the standard deviation divided by two is our reported Monte Carlo standard error. Tables 3 and 4 take advantage of clear subtopological structure and also report only topologies in an 80% credible set.

The most probable tree topologies that we find agree with those determined by approximate maximum likelihood in HRY97 (their Figure 4). The best louse

TABLE 3

Posterior Distribution over Topologies, Lice: Subtopologies are $A_1 = (((1,2),(3,4)),11)$, $A_2 = ((1,2),((3,4),11))$, $B_1 = (5,13)$, $C_1 = (7,8)$, *and* $D_1 = (9,10)$.

Rank	Topology top(τ)	p[top(τ) | data] $\pm$ se	cumulative
1	$((A_1,B_1),((6,(C_1,D_1)),12))$	0.514 ± 0.005	0.514
2	$(((A_1,B_1),(6,(C_1,D_1))),12)$	0.101 ± 0.002	0.615
3	$(((A_1,B_1),12),(6,(C_1,D_1)))$	0.073 ± 0.001	0.688
4	$((A_1,B_1),(((6,D_1),C_1),12))$	0.044 ± 0.001	0.732
5	$((A_2,B_1),((6,(C_1,D_1)),12))$	0.043 ± 0.002	0.775
6	$((A_1,(6,(C_1,D_1))),(B_1,12))$	0.027 ± 0.003	0.803

TABLE 4

Posterior Distribution over Topologies, Gophers: Subtopologies are $E_1 = (1,(2,(3,4)))$, $F_1 = (6,((7,8),(9,10)))$, *and* $G_1 = (11,12)$.

Rank	Topology top(τ)	p[top(τ) | data] $\pm$ se	cumulative
1	$((E_1,F_1),((5,G_1),13))$	0.118 ± 0.003	0.118
2	$(((E_1,F_1),5),(G_1,13))$	0.084 ± 0.003	0.202
3	$(((E_1,F_1),(5,G_1)),13)$	0.082 ± 0.001	0.284
4	$((E_1,F_1),(5,(G_1,13)))$	0.062 ± 0.003	0.346
5	$((((E_1,F_1),5),G_1),13)$	0.055 ± 0.001	0.400
6	$((E_1,F_1),((5,13),G_1))$	0.050 ± 0.002	0.451
7	$(((E_1,F_1),13),(5,G_1))$	0.043 ± 0.001	0.494
8	$((((E_1,F_1),5),13),G_1)$	0.033 ± 0.002	0.526
9	$((E_1,13),((5,G_1),F_1))$	0.032 ± 0.003	0.558
10	$((E_1,(5,F_1)),(G_1,13))$	0.027 ± 0.003	0.585
11	$((E_1,13),((5,F_1),G_1))$	0.024 ± 0.001	0.609
12	$(E_1,((5,G_1),(F_1,13)))$	0.023 ± 0.001	0.632
13	$((E_1,((5,G_1),F_1)),13)$	0.021 ± 0.002	0.653
14	$(E_1,(((5,F_1),G_1),13))$	0.020 ± 0.001	0.673
15	$(E_1,(((5,G_1),F_1),13))$	0.019 ± 0.001	0.692
16	$(E_1,(((5,G_1),13),F_1))$	0.018 ± 0.002	0.710
17	$((E_1,((5,F_1),G_1)),13)$	0.015 ± 0.001	0.725
18	$(((E_1,F_1),(5,13)),G_1)$	0.015 ± 0.001	0.740
19	$(E_1,((5,(G_1,13)),F_1))$	0.011 ± 0.001	0.752
20	$(((E_1,(5,F_1)),G_1),13)$	0.011 ± 0.001	0.763
21	$((E_1,(5,G_1)),(F_1,13))$	0.010 ± 0.001	0.773
22	$((E_1,(F_1,13)),(5,G_1))$	0.010 ± 0.001	0.783
23	$(((E_1,F_1),(G_1,13)),5)$	0.009 ± 0.001	0.792
24	$(((E_1,(5,F_1)),13),G_1)$	0.009 ± 0.001	0.801

topology is well supported, with a posterior probability of 51.4%. Here, most of the uncertainty involves the placement of taxon 12, and to a lesser extent, taxon 6. Marginally, the most probable subtopology for the remaining 11 taxa has probability 74.3%, and only ten subtopologies are in a 99% credible set. Monophyletic groups, or clades, $A = \{1,2,3,4,11\}$, $B = \{5,13\}$, $C = \{7,8\}$, and $D = \{9,10\}$, occur with probability exceeding 99.4%. Collapsing subtopologies within clades is another type of marginalization that leads to succinct summaries of the posterior. Ignoring taxa 6 and 12, these clades are grouped as either $((A,B),(C,D))$, $((A,(C,D)),B)$, or $(A,(B,(C,D)))$ with probabilities 88.1%, 9.5%, and 2.3% respectively (and a 0.1% probability that not all these clades appear).

Somewhat greater uncertainty is present in the gopher phylogeny (Table 4), with 24 topologies in the 80% credible set, and the most probable one of probability only 11.8%. Much of the uncertainty lies in the positioning of taxa 5 and 13. Marginally for the remaining 11 taxa, the best gopher subtopology has probability 65.6%, and only six subtopologies are necessary to form a 99% credible set. Three clades are identified with posterior probability 1: $E = \{1,2,3,4\}$, $F = \{6,7,8,9,10\}$, and $G = \{11,12\}$. Ignoring variation in subtree topology within clades and the placement of taxa 5 and 13, the clades are joined as either $((E,G),F)$, $(E,(F,G))$ or $((E,F),G)$ with probabilities 67.8%, 27.5%, and 4.7% respectively.

While there is substantial structural similarity for most of the sampled trees from both posterior distributions, there is no single tree topology which appears in both samples. The absence of posterior overlap provides evidence against strict coevolution of the hosts and parasites, similar to the conclusion in HRY97. From our posterior sample, we can infer more. In particular, we are able to isolate and quantify those species pairs where coevolution appears to fail. The unanimous placement of taxa 11 and 12 as nearest relatives with probability 1 in the gopher posterior contrasts markedly with the highly variable placement of taxa 12 in the louse posterior. Similarly, taxa 5 and 13, bound together in the louse posterior, vary wildly between clades in the gopher posterior. These gopher/louse pairs are the most likely candidates for an alternate evolutionary pathway. In contrast, we can identify the largest set of gopher/louse pairs supporting strict cospeciation. The common subtopology of seven stably attached taxa is $((1,(3,4)),((7,8),(9,10)))$ and has marginal posterior probability 99.5% for lice and 99.6% for gophers. Having a large Monte Carlo sample makes such a determination fairly straightforward.

Without pursuing it further, we note that parameter estimation and model building are effectively carried out with the help of an MCMC sampler. We settled upon codon-specific mutation rates and a single transition/transversion

FIG. 3. *Output Analysis, Lice: Panels on the right are autocorrelation functions for two summaries of the phylogeny sequence sampled by MCMC: loglikelihood, and indicator of best topology. For the loglikelihood series, the left panel shows simple time series plots of the output. A cusum plot is given in the left panel for the binary indicator of best topology. The dotted curve is the cusum plot of a random permutation of the series.*

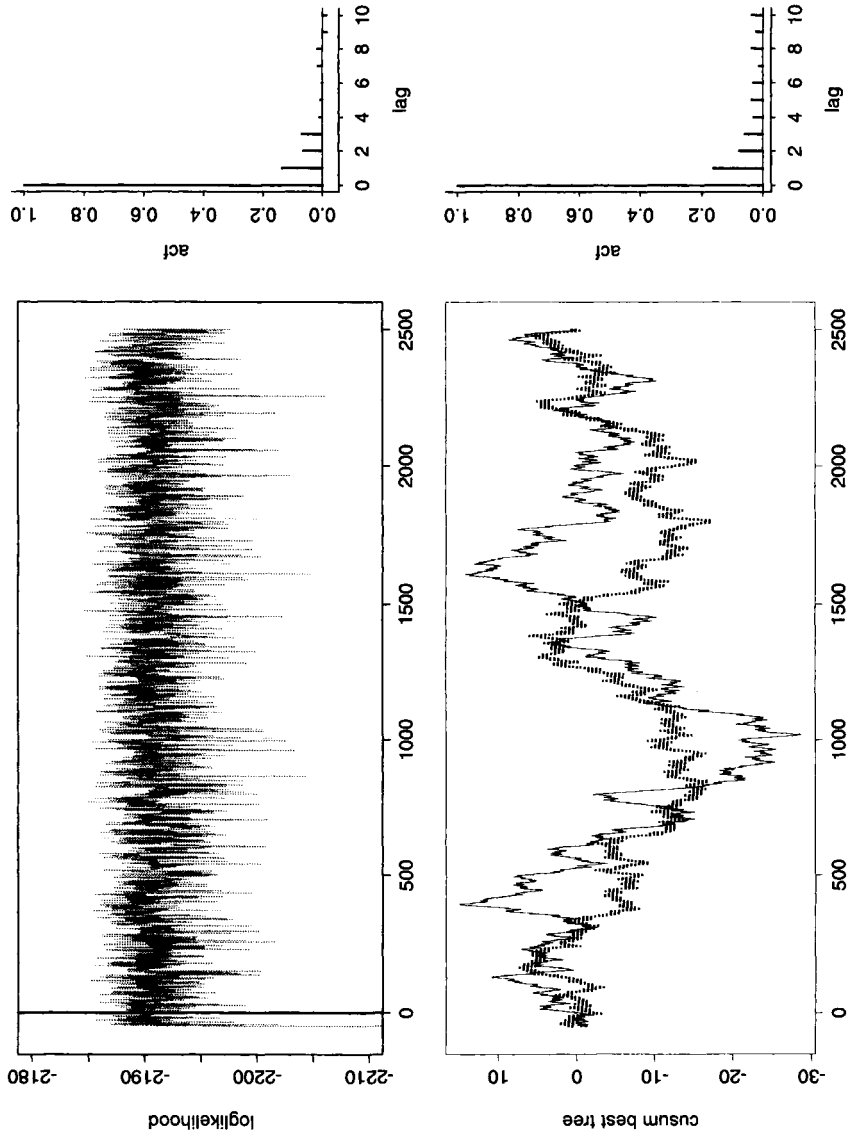
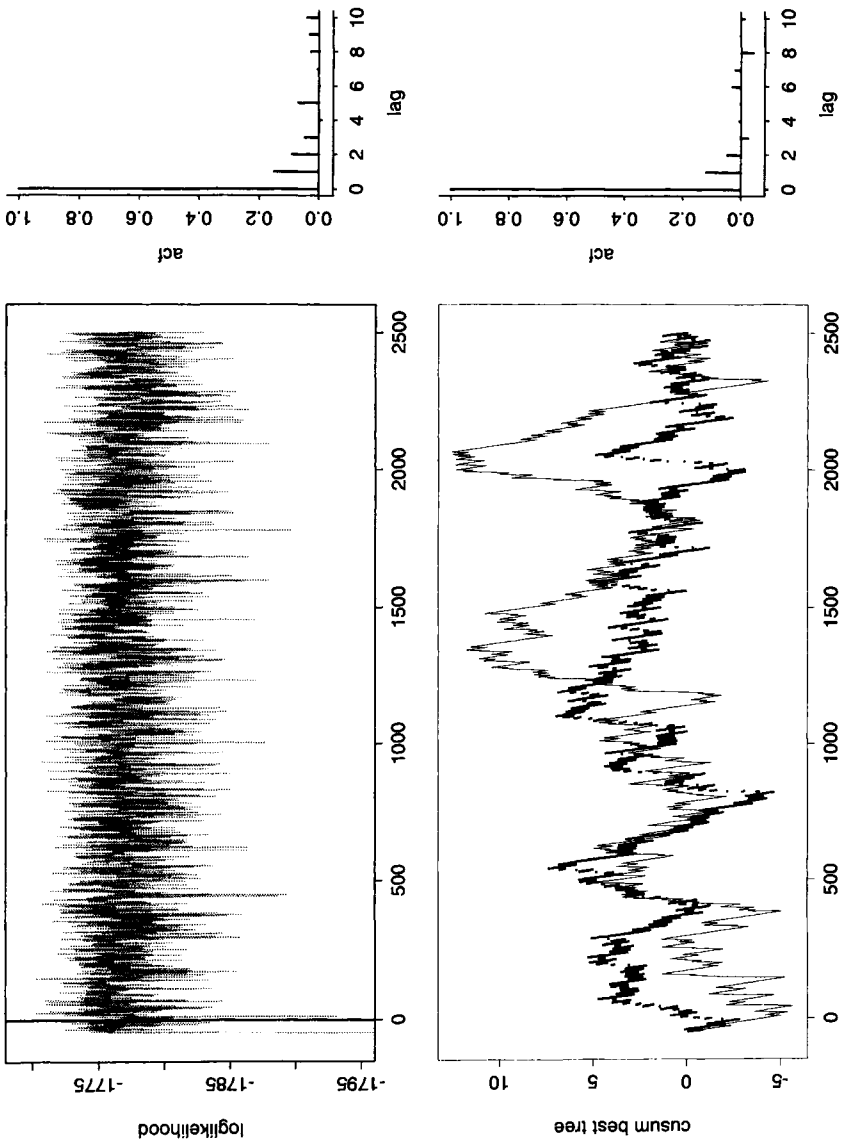


FIG. 4. *Output Analysis, Gophers: Same diagnostics as Figure 3.*



bias parameter by simply rerunning our chains using different likelihood evaluation routines, and allowing simultaneous parameter updating. Importantly, final parameter estimates do not condition on an estimated phylogeny, and no optimization methods are used.

6. Concluding remarks. Much remains to be done before we understand the utility of Bayesian methods for evolutionary biology. They may be helpful for studying cospeciation because the scientific questions of interest relate to topological structure of the phylogeny, and more classical statistical methods do not provide completely satisfactory results. Inference based on likelihood ratio tests may be effective, but frequency calibration typically requires fixing parameter estimates and phylogenies, and hence exact significance levels are unknown. Other tests ask if the similarity between host and parasite estimated phylogenies is more than can be attributed to chance, under a model of phylogenesis (e.g., Page 1988), but the reference measure here appears to have little to do with the cospeciation hypothesis. A Bayesian approach, on the other hand, allows us to make direct probabilistic statements concerning relevant aspects of phylogeny structure.

Even the most simple questions require sophisticated computation, and so we have started our investigation by trying to approximate the posterior distribution over phylogeny space within the context of a standard parametric model of evolution. Initial experimentation with these computations is cause for some optimism. The problem with Markov chain Monte Carlo is not so much in developing an algorithm as it is in developing a reasonably efficient algorithm, and we think that our simple technique of perturbing speciation times shows promise. Further research is needed to uncover the relative merits of competing algorithms. In addition, it may be helpful to clarify the relationship of Bayesian and bootstrap methodology, so that biologists will with more confidence be able to assess uncertainty in evolutionary hypotheses.

REFERENCES

BERNARDO, J.M. and SMITH, A.F.M (1994). *Bayesian Theory.* Wiley, New York.

BESAG, J., GREEN, P.J., HIGDON, D. and MENGERSEN, K. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10** 3–66.

BROWN, J.M.K. (1994). Probabilities of evolutionary trees. *Systematic Biology* **43** 78–91.

COWLES, M.K. and CARLIN, B.P. (1996). MCMC convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, **91** 883–904.

CHERNOFF, H. (1997). Invited Lecture. IMA Summer Research Program on Statistics in the Health Sciences. Minneapolis, July, 1997.

EFRON, B., HALLORAN, B. and HOLMES, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the USA* 13429–13434.

FELSENSTEIN, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22** 240–249.

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 (with discussion). *Journal of the Royal Statistical Society, Series A*, **146** 246–272.

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

FELSENSTEIN, J. (1995). PHYLIP (phylogeny inference package) version 3.5c. Computer program distributed by the University of Washington.

FELSENSTEIN, J. and KISHINO, H. (1993). Is there something wrong with the bootstrap? A reply to Hillis and Bull. *Systematic Biology* **42** 193–200.

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.

HAFNER, M.S., SUDMAN, P.D., VILLABLANCA, F.X., SPRADLING, T.A., DEMASTES, J.W. and NADLER, S.A. (1994). Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* **265** 1087–1090.

HASEGAWA, M., KISHINO, H. and YANO, T. (1985). Dating the Human-Ape Splitting by a Molecular Colck of Mitochondrial DNA. *Journal of Molecular Evolution* **22** 160–174.

HUELSENBECK, J.P. and RANNALA, B. (1997). Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276** 227–232.

HUELSENBECK, J.P., RANNALA, B., and YANG, Z. (1997). Statistical tests of host-parasite cospeciation. *Evolution* **51** 410–419.

JUKES, G.H. and CANTOR, C.R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, Munroe, H.N. (ed.), pp. 21–132. Academic Press, New York.

KIMURA, M. (1980). A simple method for estimating rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16** 111–120.

KINGMAN, J.F.C. (1982). The Coalescent. *Stochastic Processes and their Applications* **13** 235–248.

KUHNER, M.K., YAMATO, J. and FELSENSTEIN, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140** 1421–1430.

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

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

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.

MAU, B., NEWTON, M.A. and LARGET, B. (1998). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, to appear.

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

OLSEN, G.J., MATSUDA, H., HAGSTROM, R. and OVERBECK R.(1994). FastDNAml: a tool for the construction of phylogenetic trees of DNA sequences using maximun likelihood. *CABIOS* **10** 41–48.

PAGE, R.D.M. (1988). Quantitative cladistic biogeography: Constructing and comparing area cladograms. *Systematic Zoology* **37** 254–270.

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.

SWOFFORD, D.L. (1996). PAUP: Phylogenetic analysis using parsimony and other methods, Sinauer, Sunderland, MA.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of*

Statistics **22** 1701–1762.

YANG, Z., GOLDMAN, N. and FRIDAY, A. (1995). Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Systematic Biology* **44** 384–399.

YANG, Z. and RANNALA, B.(1997). Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method. *Molecular Biology and Evolution* **14** 717–724.

YU, B.(1995). Comment: extracting more diagnostic information from a single run using the cusum path plot. *Statistical Science* **10** 54–58.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN–MADISON
1210 WEST DAYTON ST.
MADISON WI, 53706-1685
NEWTON@STAT.WISC.EDU

DEPARTMENT OF GENETICS
UNIVERSITY OF WISCONSIN–MADISON
445 HENRY MALL
MADISON WI, 53792
ROBERTM@GENETICS.WISC.EDU

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
DUQUESNE UNIVERSITY
440 COLLEGE HALL
PITTSBURGH PA 15282
LARGET@MATHCS.DUQ.EDU