A BAYESIAN FRAMEWORK FOR THE ANALYSIS OF COSPECIATION

Jᴏʜɴ P. Hᴜᴇʟsᴇɴʙᴇᴄᴋ,[1,2] Bʀᴜᴄᴇ Rᴀɴɴᴀʟᴀ,[3] ᴀɴᴅ Bʀᴇᴛ Lᴀʀɢᴇᴛ[4]
[1]*Department of Biology, University of Rochester, Rochester, New York 14627*
[2]*E-mail: johnh@brahms.biology.rochester.edu*
[3]*Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11794*
[4]*Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, Pennsylvania 15282*

Abstract.—Information on the history of cospeciation and host switching for a group of host and parasite species is contained in the DNA sequences sampled from each. Here, we develop a Bayesian framework for the analysis of cospeciation. We suggest a simple model of host switching by a parasite on a host phylogeny in which host switching events are assumed to occur at a constant rate over the entire evolutionary history of associated hosts and parasites. The posterior probability density of the parameters of the model of host switching are evaluated numerically using Markov chain Monte Carlo. In particular, the method generates the probability density of the number of host switches and of the host switching rate. Moreover, the method provides information on the probability that an event of host switching is associated with a particular pair of branches. A Bayesian approach has several advantages over other methods for the analysis of cospeciation. In particular, it does not assume that the host or parasite phylogenies are known without error; many alternative phylogenies are sampled in proportion to their probability of being correct.

Key words.—Cospeciation, host switching, Markov chain Monte Carlo, molecular biology, molecular evolution, phylogeny.

Host species and their associated parasites often exhibit a pattern of concordant phylogeny. That is, the host and parasite phylogenies are largely congruent if the parasite tree is superimposed on the host tree. A pattern of concordance is expected if host speciation isolates the parasite associated with each incipient host species causing it to also speciate via allopatric speciation. Analysis of several datasets has suggested potential instances of cospeciation between hosts and parasites. For example, in analyses of DNA sequences sampled from gophers and their louse parasites, cospeciation has been suggested as an important process for producing the louse phylogeny (Hafner and Page 1995; Huelsenbeck et al. 1997). Similarly, phylogenetic analysis of rodents and the Adenaviridae (single-stranded RNA viruses that cause hemorrhagic fever) has shown cospeciation to be a possible mechanism for producing the viral genealogy (Bowen et al. 1997).

Although it is true that some degree of cospeciation can be invoked to explain the similarity between host and parasite phylogenies, it is rarely the case that the host and parasite trees are identical. Two null hypotheses about the host and parasite phylogenies can currently be tested: (1) that host and parasite trees are identical (Huelsenbeck et al. 1997); and (2) that host and parasite trees are more similar than expected for two independent phylogenies generated under a random branching model of cladogenesis (Page 1988). Rejection of the first hypothesis allows the biologist to dismiss the idea that stochastic error in estimating host and/or parasite phylogenies is responsible for the differences in the trees. Rejection of the second hypothesis allows the biologist to infer that the hosts and parasites are not independent of one another (i.e., that the phylogenies of hosts and parasites are not generated by independent evolutionary processes). Rejection of both hypotheses implies that either some degree of host switching by the parasites has occurred or that other factors, such as within-population coalescence processes (Pamilo and Nei 1988), have led to different gene trees for hosts and parasites despite a history of strict cospeciation. This is the case with the gophers and lice; the trees are more similar than would be expected under a random branching model of cladogenesis (Hafner et al. 1994; Huelsenbeck et al. 1997), but are not identical (Huelsenbeck et al. 1997).

How can one infer the frequency of host switching and the lineages involved in host switches? Previous work on the problem of inferring where host-switching events occurred has relied on maximizing the number of cospeciation events relative to host switching and sorting events that are needed to explain the differences between the host and parasite trees (Page 1994). Implicitly, this type of analysis assumes that the host and parasite phylogenies are estimated without error. It also assumes that the host-switching rate is low enough that all switching events are evident from an analysis of particular regions of disagreement between host and parasite phylogenies. That is, all of the weight in the analysis is put on the reconstruction that maximizes the number of cospeciation events; other, perhaps less likely, possibilities are not considered in the analysis. In this paper, we assume a simple stochastic model for host switching and perform Bayesian estimation of the host switching rate as well as the possible events of host switching that could explain the differences between the host and parasite tree.

Mᴇᴛʜᴏᴅs

A Simple Model of Host Switching

We assume that the phylogeny of the host species is represented by a rooted binary tree (Fig. 1). The tips of the host phylogeny, τ_H, are labeled n_1 to n_s, where s is the number of host species. The internal nodes of the host tree, which represent speciation events in the host that eventually give rise to the s species, are labeled n_{s+1} to n_{2s-1}. The times of the nodes on the tree are denoted $\mathbf{t}_H = (t_1, t_2, \ldots, t_{2s-1})$. The node times are scaled such that the tips of the tree are at time 0 and the root of the tree is at time 1. Thus, $t_1 = t_2 = \ldots = t_s = 0$ and $t_{2s-1} = 1$, with all other nodes having $0 \leq t_i$
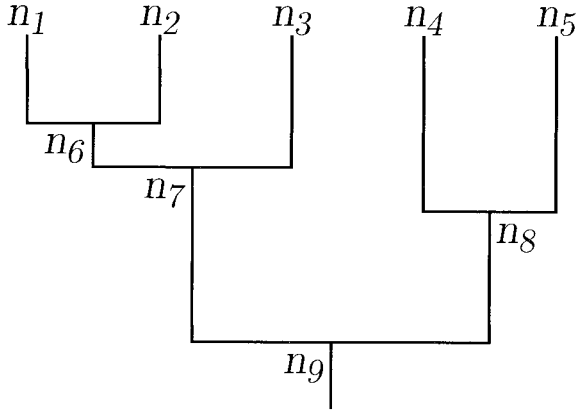
FIG. 1. The phylogeny for the host species is represented by a rooted binary tree, τ_H. Speciation times on the host tree are contained in the vector $\mathbf{t}_H$. The tips of the tree are at time $t = 0$ and the root of the tree is at time $t = 1$.

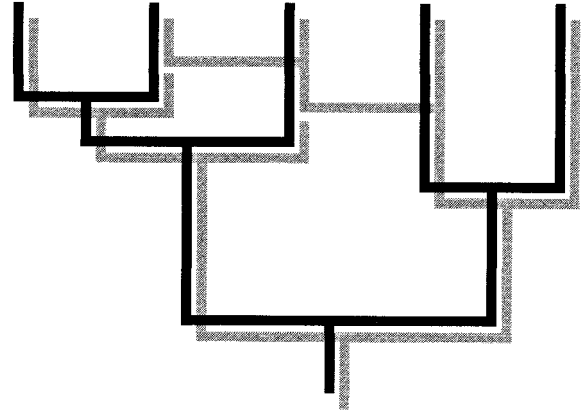FIG. 2. An example of a host tree (dark lines) with two events of host switching. The parasite tree (lighly shaded phylogeny) is completely determined by the host tree and the events of host switching.

≤ 1. All parameter estimates in this paper are relative to this time scale. The sum of the scaled branch lengths is referred to as the total tree length (T). The number of host lineages that leave present-day descendents and were in existence at time t is denoted $b(t)$ ($b[t] > 1$). The root time of the parasite tree is constrained by the host tree and the number of host-switching events. In the absence of any host switching, the parasite tree will have a root time of 1.0. However, depending on the number and position of the host switches, the root time of the parasite tree may be less than one.

We assume that in the absence of host switching, the parasite tree (τ_P) is completely determined by the host tree. That is, if there were no host-switching events, the host and parasite trees would be identical ($\tau_H = \tau_P$; $\mathbf{t}_H = \mathbf{t}_P$). Host switching by the parasites occurs on the host phylogeny according to a Poisson process with rate λ. During a small interval of time, dt, during which there are $b(t)$ lineages, the probability that a host-switching event occurs is $b(t)\lambda dt$ and the probability that two or more host-switching events occur is of order $o(dt)$. If a host-switching event occurs at time t, a parasite species colonizes a new host (one of the $b[t] - 1$ host species, excluding its current host, that existed at time t and that left descendents in the sample). Under the simplest model we consider, each of the $b(t) - 1$ host lineages that exist at the time of the switching event (and leave descendents) is colonized with equal probability. This is the same model used by Huelsenbeck et al. (1997) for estimating host-switching rates using a simulation method and ignores the effect of geography and the potential physiological similarities of more closely related parasites. We also considered a more complicated model that allows the probability of host switching to depend on the phylogenetic distance of the target host from the current host, which potentially takes account of this second effect.

Each host-switch event has a source (γ_i, the place on the host tree where the ith switching event originated), a target (δ_i, the host lineage to which the parasite switches), and a time (z_i, the time at which the ith event occurred). The collection of events is denoted $\mathbf{e} = (\xi, \mathbf{z}, \boldsymbol{\gamma}, \boldsymbol{\delta})$, where ξ is the number of host-switching events on the tree, $\mathbf{z}$ is a vector of

event times, $\boldsymbol{\gamma}$ is a vector of sources, and $\boldsymbol{\delta}$ is a vector of targets. When $\xi = 0$, $\mathbf{z}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$ are empty. Figure 2 provides an example of a host tree with $\xi = 2$ host-switching events.

The prior distribution for $\mathbf{e}$ given the host tree (τ_H, $\mathbf{t}_H$) with a total tree length T, and switching rate, λ, is described as follows. There is a probability $e^{-\lambda T}$ of no host-switching events. The probability density of realizing ξ host-switching events at times $\mathbf{z}$ for any collection of ξ source-target pairs is

$$\frac{e^{-\lambda T}(\lambda T)^{\xi}}{\xi!} \times \left(\frac{1}{T}\right)^{\xi} \times \prod_{i=1}^{\xi} \frac{1}{b(z_i) - 1}. \quad (1)$$

Host-switching events are placed on the host phylogeny according to a Poisson process with rate λ. The probability density that a host-switching event is placed at any point on the host tree is $1/T$. When a host-switching event occurs, the probability that it attacks any lineage that existed at time t (excluding the current host) and left descendents is $1/(b[t] - 1)$.

Model of DNA Substitution for Hosts and Parasites

In this paper, we analyze cytochrome oxidase I (COI) DNA sequences from 13 gophers and their parasitic lice (Hafner et al. 1994). We will describe here the model of DNA substitution assumed for analysis of these sequences. However, the methodology developed in this paper works with any of the stochastic models of character change commonly used in phylogenetic analyses.

We assume that aligned DNA sequences are available for the host and parasite taxa. In this paper, we assume a one-to-one correspondence between hosts and parasites (i.e., $s_H = s_P = s$). The lengths of the sequences are c_H and c_P for the host and parasite sequences, respectively. The aligned DNA sequences are contained in the matrices $\mathbf{X} = \{x_{kh}\}$ and $\mathbf{Y} = \{y_{ka}\}$ for the hosts and parasites, respectively, where $k = 1, 2, \ldots, s$, $h = 1, 2, \ldots, c_H$, and $a = 1, 2, \ldots, c_P$.

The rates of substitution on the host and parasite phylogenies are specified by the tree height (m_H and m_P, for host and parasite trees, respectively). Each branch of the phylogenetic tree is multiplied by the tree height giving the number of substitutions per site that are expected to occur along the

branch. For example, if $m_H = 0.1$ and the two ends of the branch are at times 0.53 and 0.27, the total number of substitutions expected to occur along the branch is $v = 0.1 (0.53 - 0.27) = 0.026$.

We assume that substitutions occur according to a time-homogeneous Poisson process with instantaneous rate matrix, **Q**. In particular, we assume the HKY85 model of DNA substitution, which was first proposed by Hasegawa, Kishino, and Yano (Hasegawa et al. 1984, 1985). The instantaneous rate matrix for the HKY85 model is

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} \cdot & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & \cdot \end{pmatrix}, \qquad (2)$$

where κ is the transition/transversion rate bias and $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ are the equilibrium base frequencies. When $\kappa > 1$, transitions occur at a higher rate than transversions. We allow the transition/transversion rate ratio to be different for hosts (κ_H) and parasites (κ_P). The rows of the instantaneous rate matrix sum to zero. Moreover, the constraint that $-\Sigma q_{ii}\pi_i = 1$ is added; this ensures that the branch lengths of the phylogenetic tree are in terms of expected number of substitutions per site, v. The probability that nucleotide i changes into j over a branch of length v is contained in the matrix $\mathbf{P} = \{p_{ij}\}$. **P** can be obtained from the rate matrix **Q** through the operation $\mathbf{P} = e^{\mathbf{Q}v}$. We accommodate rate variation across sites by assuming that the rate at a site is a random variable drawn from a mean-one gamma distribution, with shape parameter α (Yang 1993, 1994). We allow the shape parameter to be different for the host (α_H) and parasite (α_P) data. Substitution models that assume gamma-distributed rate variation with K discrete categories are denoted "$+d\Gamma_K$". Parameters of the model of DNA substitution are contained in the vectors $\theta_H = (m_H, \kappa_H, \alpha_H, \pi_H)$ and $\theta_P = (m_P, \kappa_P, \alpha_P, \pi_P)$ for the hosts and parasites, respectively.

The probability of observing the data for the ith site for the host sequences ($\mathbf{x}_i$) is a sum over all possible assignments of nucleotides to the internal nodes of the tree. The branch whose child is node k and whose parent is $\sigma(k)$ has v_k expected substitutions. Also, let **w** be a data vector of states at the internal nodes of the host tree. The probability of observing the host data at the ith site given the tree is then

$$f(\mathbf{x}_i \,|\, \tau_H, \mathbf{t}_H, \theta_H)$$
$$= \sum_{j=1}^{K} \left[\sum_{\mathbf{w}} \pi_{w_{2s-1}} \left(\prod_{k=1}^{s} p_{w_{\sigma(k)}x_k}(v_k r_j) \right) \left(\prod_{k=s+1}^{2s-2} p_{w_{\sigma(k)}w_k}(v_k r_j) \right) \right] \times \frac{1}{K},$$
$$(3)$$

where r_j is the rate multiplier for the jth gamma category. The summation is over all possible combinations of nucleotides that can be assigned to the internal nodes of the tree. The probability of observing the ith parasite site ($\mathbf{y}_i$) is calculated in a similar manner. Felsenstein (1981) describes a pruning algorithm for efficiently calculating the probability of observing a site. Assuming independence of the substitutions across sites, the probability of observing the full host sequence dataset is

$$f(\mathbf{X} \,|\, \tau_H, \mathbf{t}_H, \theta_H) = \prod_{i=1}^{c_H} f(\mathbf{x}_i \,|\, \tau_H, \mathbf{t}_H, \theta_H) \qquad (4)$$

and the probability of observing the full parasite sequence data is

$$f(\mathbf{Y} \,|\, \tau_P, \mathbf{t}_P, \theta_P) = \prod_{i=1}^{c_P} f(\mathbf{y}_i \,|\, \tau_P, \mathbf{t}_P, \theta_P). \qquad (5)$$

Because the parasite tree is completely determined by the host phylogeny and the events of host switching on the host phylogeny, the probability of observing the full parasite sequence data can also be written as

$$f(\mathbf{Y} \,|\, \tau_H, \mathbf{t}_H, \mathbf{e}, \theta_P) = \prod_{i=1}^{c_P} f(\mathbf{y}_i \,|\, \tau_H, \mathbf{t}_H, \mathbf{e}, \theta_P). \qquad (6)$$

Bayesian Estimation Using Markov Chain Monte Carlo

In a Bayesian analysis, inferences are based on the posterior density of the parameter of interest. In this study, we are interested in estimating the rate of host switching and the parameters of the substitution model (λ, θ_H, θ_P). The likelihood function for λ, θ_H, θ_P is

$$\ell(\lambda, \theta_H, \theta_P)$$
$$= \int f(\mathbf{X} \,|\, \tau_H, \mathbf{t}_H, \theta_H) f(\mathbf{Y} \,|\, \tau_H, \mathbf{t}_H, \mathbf{e}, \theta_P) \, dF(\mathbf{e} \,|\, \tau_H, \mathbf{t}_H, \lambda)$$
$$\times dF(\tau_H, \mathbf{t}_H), \qquad (7)$$

where the single integral denotes a summation over all possible host trees and all possible events of host switching on the host tree and integration over all possible host speciation times and event positions. Integration with respect to the probability measure is used to denote summation for discrete random variables and integration for continuous random variables. The posterior probability density of the parameters λ, θ_H, and θ_P is

$$f(\lambda, \theta_H, \theta_P \,|\, \mathbf{X}, \mathbf{Y}) = \frac{\ell(\lambda, \theta_H, \theta_P) f(\lambda, \theta_H, \theta_P)}{f(\mathbf{X}, \mathbf{Y})}, \qquad (8)$$

where

$$f(\mathbf{X}, \mathbf{Y}) = \int \ell(\lambda, \theta_H, \theta_P) \, dF(\lambda, \theta_H, \theta_P) \qquad (9)$$

and integration is over the space for λ, θ_H, and θ_P. We assume independent uniform priors for most parameters of the model: λ, m_H, m_P, κ_H, κ_P, α_H, α_P. Moreover, the speciation times on the host tree, $\mathbf{t}_H$, are distributed as the order statistics drawn from a uniform $(0,1)$ distribution, conditional on agreement with the tree topology, τ_H. Table 1 provides more information on the priors used in this analysis.

We use Markov chain Monte Carlo (MCMC) to perform the high dimensional summation and integration that is involved in evaluating equations (8) and (9). Specifically, we use the Metropolis-Hastings-Green (MHG) algorithm (Green 1995). The MHG algorithm is an extension of the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970) that allows the process to move between parameter subspaces of differing dimension. Appendix 1 provides a simple ex-

TABLE 1. A table showing the priors that were used for many of the parameters of the model of DNA substitution and host switching.

Parameter	Prior
ξ	Poisson (λT)
m_H	uniform (0, 100)
m_P	uniform (0, 100)
κ_H	uniform (0, 1000)
κ_P	uniform (0, 1000)
α_H	uniform (0, 10)
α_P	uniform (0, 10)
π_H	Dirichlet (π_H)
π_P	Dirichlet (π_P)
λ	uniform (0, 1000)

ample of MCMC applied to the problem of estimating the probability of heads for a coin.

In this study, we use 16 different proposal mechanisms to update the state of the chain. These proposal mechanisms insure that every state can be reached from every other state (i.e., that the Markov chain is irreducible). Appendix 2 gives details on the proposal mechanisms with the acceptance probability for each.

<center>RESULTS</center>

Examining Cospeciation for Gophers and Lice

We applied the Bayesian method developed in this paper to COI DNA sequence data sampled from gophers and their ectoparasitic lice (Hafner et al. 1994). The data include 15 species of gophers from the genera *Cratogeomys, Geomys,*

Orthogeomys, Pappogeomys, Thomomys, and *Zygogeomys* and 17 species of lice from the genera *Geomydoecus* and *Thomomydoecus*. For 13 of the pairs, there was a one-to-one correspondence between gopher and louse species; we excluded species that had more than one louse species associated with a gopher species. This left us with a total of 13 gopher and louse species. This dataset has been analyzed many times by others to illustrate methods for analysis of cospeciation (Page 1994; Huelsenbeck et al. 1997). We use the gopher and louse dataset here to facilitate comparison with earlier work.

Figure 3 shows the maximum-likelihood estimates of phylogeny for the gophers and lice. The trees were estimated assuming the HKY85 + dΓ_4 model of DNA substitution under a molecular clock using the program PAUP* (Swofford 1998). The log likelihood for the gopher sequences was $\max[\log_e f(\mathbf{X} \mid \tau_H, \mathbf{t}_H, \theta_H)] = -1908.94$ and the log likelihood of the louse sequences was $\max[\log_e f(\mathbf{Y} \mid \tau_P, \mathbf{t}_P, \theta_P)] = -2344.93$. Maximum-likelihood estimates of the parameters of the substitution model were $m_H = 0.20$, $\kappa_H = 11.64$, $\alpha_H = 0.13$, and $\pi_H = (0.33, 0.23, 0.12, 0.33)$ for the gophers and $m_P = 0.68$, $\kappa_P = 18.45$, $\alpha_P = 0.17$, and $\pi_P = (0.27, 0.13, 0.21, 0.40)$ for the lice. Note that the topologies of the host and parasite trees agree in some parts (e.g., for the species pairs *O. underwoodi–G. setzeri, O. cavator–G. panamensis, O. cherriei–G. cherriei, O. heterodus–G. costaricensis,* and *O. hispidus–G. chapini*), but disagree in other parts of the tree. Huelsenbeck et al. (1997) tested the null hypothesis that the gopher and louse trees are identical and that the incongruence is due to stochastic error in estimating the phylogenies. This null hypothesis can be rejected for the entire
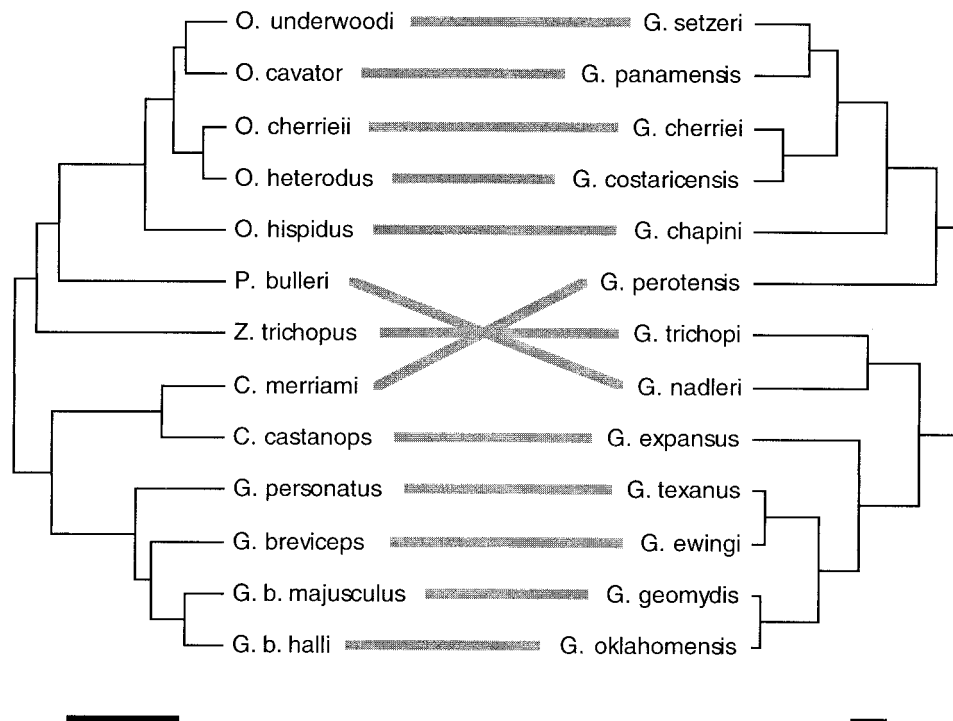

FIG. 3. The maximum-likelihood estimates of phylogeny for the gophers and lice. The HKY85 + dΓ_4 model of DNA substitution was assumed in the analysis. The molecular clock assumption was enforced. The horizontal bars at the base of each tree represent 0.1 substitutions/site.
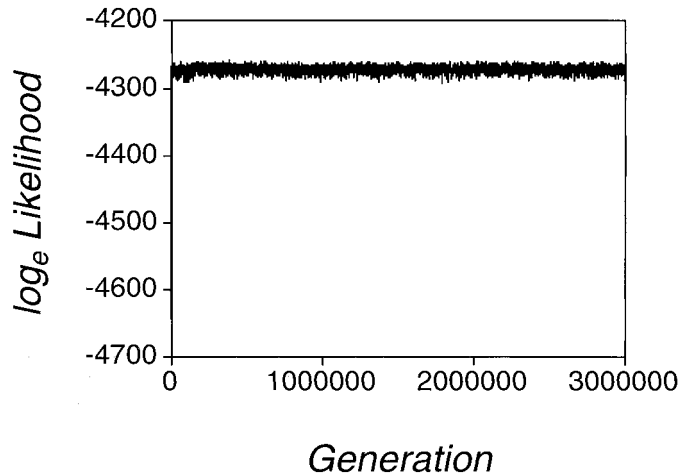
FIG. 4. The Markov chain appeared to reach stationarity quickly (after about 50,000 generations). The log of the likelihood function is tracked through time in this figure.

FIG. 5. The posterior probability of the number of host switches (A) and of the host-switching rate (B).

dataset (all 13 sequences), but not for the five species pairs that have a concordant pattern of phylogeny.

As outlined in this paper, our method assumes the molecular clock. We used a likelihood-ratio test to examine the null hypothesis that the rates across lineages are constant for the gopher and louse sequences (Felsenstein 1981). The molecular clock hypothesis could not be rejected for either the gopher or the louse datasets ($-2 \log_e \Lambda = 18.88$ for the gopher sequences, $P = 0.091$; $-2 \log_e \Lambda = 12.28$ for the louse sequences, $P = 0.423$).

Figure 4 shows the log-probability of the observed host and parasite sequences for successive iterations of our MCMC analysis. The chain was run for 3×10^6 generations and sampled every 100 generations. The chain appeared to quickly reach stationarity after only a few tens of thousands of generations. We discarded the first 100,000 generations as the "burn-in" of the chain. Inferences were based on the remaining 29,000 sampled points. We ran several chains that started with different numbers of events (0, 20, and 50 events of host switching). The chains converged to the same posterior distribution regardless of the initial number of host-switching events; inferences based on any of these chains would be essentially the same.

Figure 5 shows the posterior probability distribution for the number of host-switching events (Fig. 5A) and the rate of host-switching, λ (Fig. 5B). There were an average of $\xi = 9.20$ (4, 20) host-switching events on the tree and the rate of host switching was $\lambda = 1.50$ (0.42, 3.50) (where the interval represents the 95% credibility region for the parameters).

The inferences on host switching are not dependent on any single tree being correct for hosts and parasites because the method visits many trees in proportion to their probability. The posterior probabilities of the host and parasite trees can be obtained by noting the proportion of the time the Markov chain visited different rooted trees. Figure 6 summarizes the results for the gopher and louse data. The figure shows the most probable trees of the 29,000 trees saved duing the MCMC analysis. The numbers at the internal nodes of the
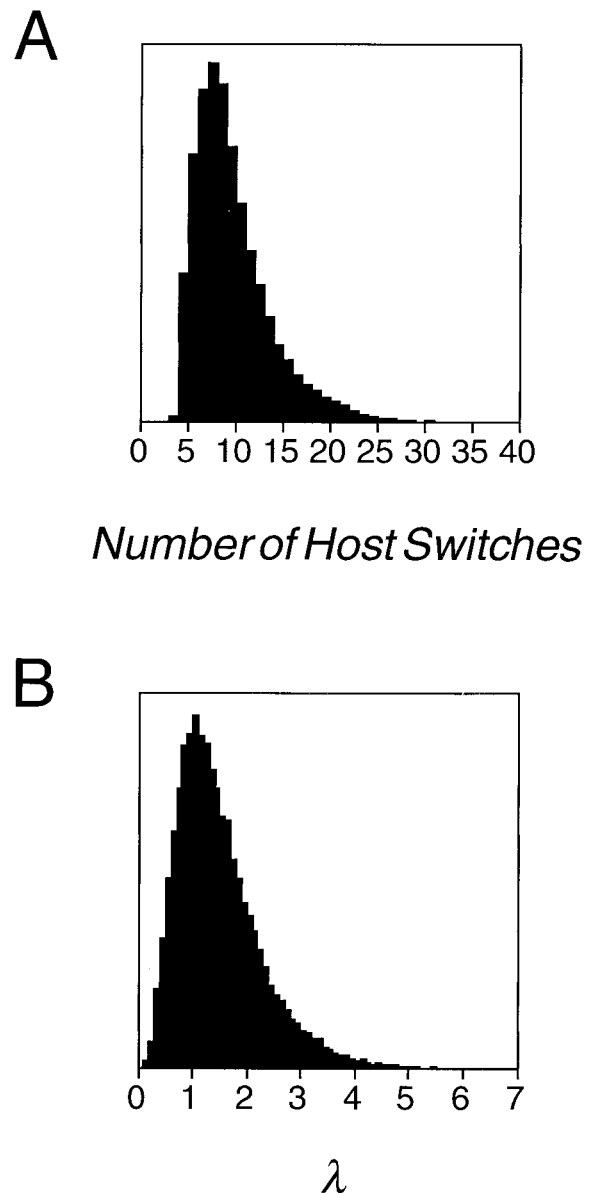
trees do not represent nonparametric bootstrap proportions, as is typical in phylogenetic studies, but rather represent the posterior probability that the clade is correct.

The upper numbers on each branch of the trees of Figure 6 represent the posterior probability of the clade being correct under the model of host switching developed in this paper. Importantly, the model of host switching links the host and parasite data; even though different trees are allowed for the host and parasite sequences, the host-switching model makes some parasite trees a priori more probable than others (i.e., given the host tree, host speciation times, and host-switching rate). Consider, for example, the case where λ is fixed to be zero. In this case, the host and parasite trees are identical and the analysis would be the same as combining host and parasite sequences and performing the Bayesian analysis as-
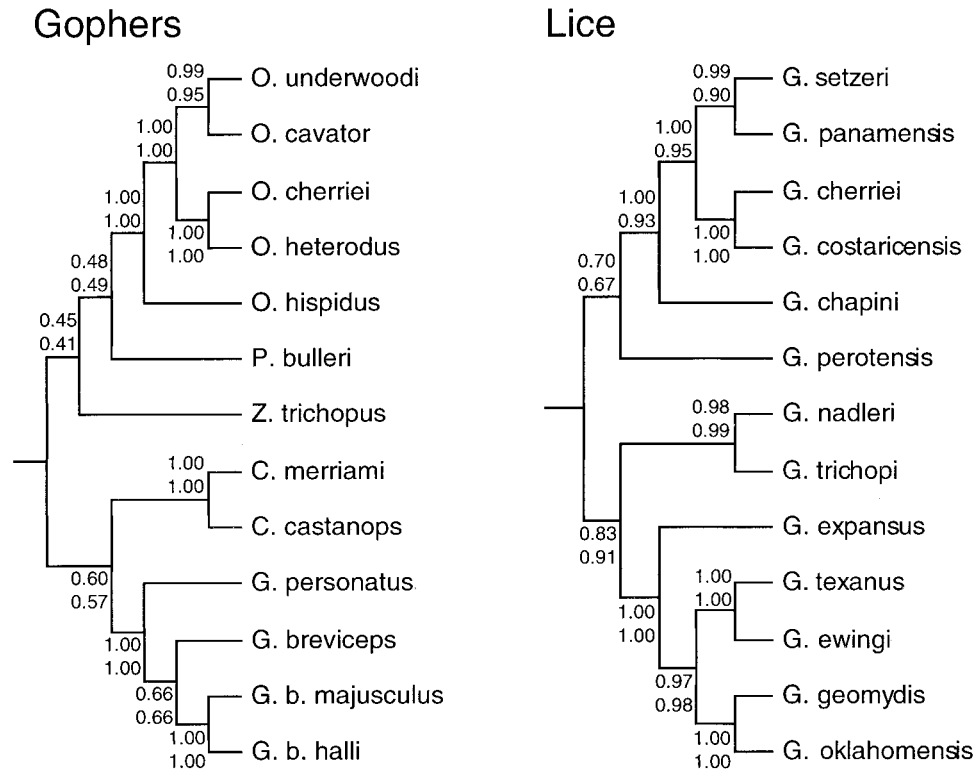
FIG. 6. The most probable trees for the gophers and the lice. The numbers on the internal branches of the trees represent the posterior probability that the clade is correct. The upper number on each branch was obtained by allowing the two trees to be related by a model of host switching. The lower number on each branch was obtained by analyzing each dataset independently.

suming the same tree for hosts and parasites. The lower numbers on each branch of the trees of Figure 6 represent the posterior probability of the clade being correct when the gopher and louse sequences are analyzed independently. The program BAMBE (Simon and Larget 1998) was used to perform the Bayesian analysis under the HKY85 + dΓ_4 model of DNA substitution. The posterior probabilities were very similar for the gopher data regardless of whether a model of host switching was assumed. However, for the louse data, relating the host and parasite sequences through a model of host switching seemed to change the support for several clades. For example, for that part of the tree where the hosts and parasite sequences are congruent in the maximum-likelihood analysis (*O. underwoodi–G. setzeri, O. cavator–G. panamensis, O. cherriei–G. cherriei, O. heterodus–G. costaricensis,* and *O. hispidus–G. chapini*), the support for the louse clades was generally higher under the model of host switching than when the sequences are analyzed separately.

We also performed an analysis that examined where on the tree host-switching events occurred. To simplify the presentation, we present here the results obtained when the host tree was fixed (the maximum-likelihood tree was assumed to be correct for the gophers). The analysis was then performed by allowing all other parameters to be random variables. The Markov chain was run for 2×10^6 generations and the first 50,000 generations were discarded as the burn-in. Figure 7 shows the posterior probability density of the parameters of the substitution model for the gophers and the lice. The rate of substitution was $m_H = 0.19$ (0.15, 0.25) for gophers and

$m_P = 0.95$ (0.51, 1.79) for lice; the rate of substitution was several times higher in the lice, which is consistent with earlier work (Hafner et al. 1994; Hafner and Page 1995; Huelsenbeck et al. 1997). The transition/transversion rate ratio was also higher in the lice ($\kappa_H = 11.72$ [8.47, 15.93], $\kappa_P = 20.65$ [12.54, 34.96]). The parameter of the mean-one gamma distribution for among-site rate variation was approximately the same in the gophers and lice ($\alpha_H = 0.14$ [0.10, 0.19], $\alpha_P = 0.17$ [0.14, 0.20]).

The posterior probability of the number of host switches and the rate of host switching did not change substantially when the gopher phylogeny was fixed (Fig. 8). The estimates of the number of host switches and the host-switching rate were $\xi = 9.94$ (5, 19) and $\lambda = 1.58$ (0.51, 3.39); these estimates are very similar to the values obtained when the host tree was considered a random variable.

The Bayesian analysis of host switching also provides information on the placement of the host-switching events on the host tree and on the direction of the changes. Figure 9 shows the posterior probability density of the speciation times for the gopher data. (Table 2 provides a key to the taxon names for Figs. 9–13.) Figure 10 shows the posterior probability of having a source associated with a particular branch and time (the width of the branches are proportional to the posterior probability density). Figure 11 shows the posterior probability of having a target associated with a particular branch and time. Note that very few events (sources or targets) were associated with the part of the phylogenetic tree that was congruent for the hosts and parasites (branches 1–
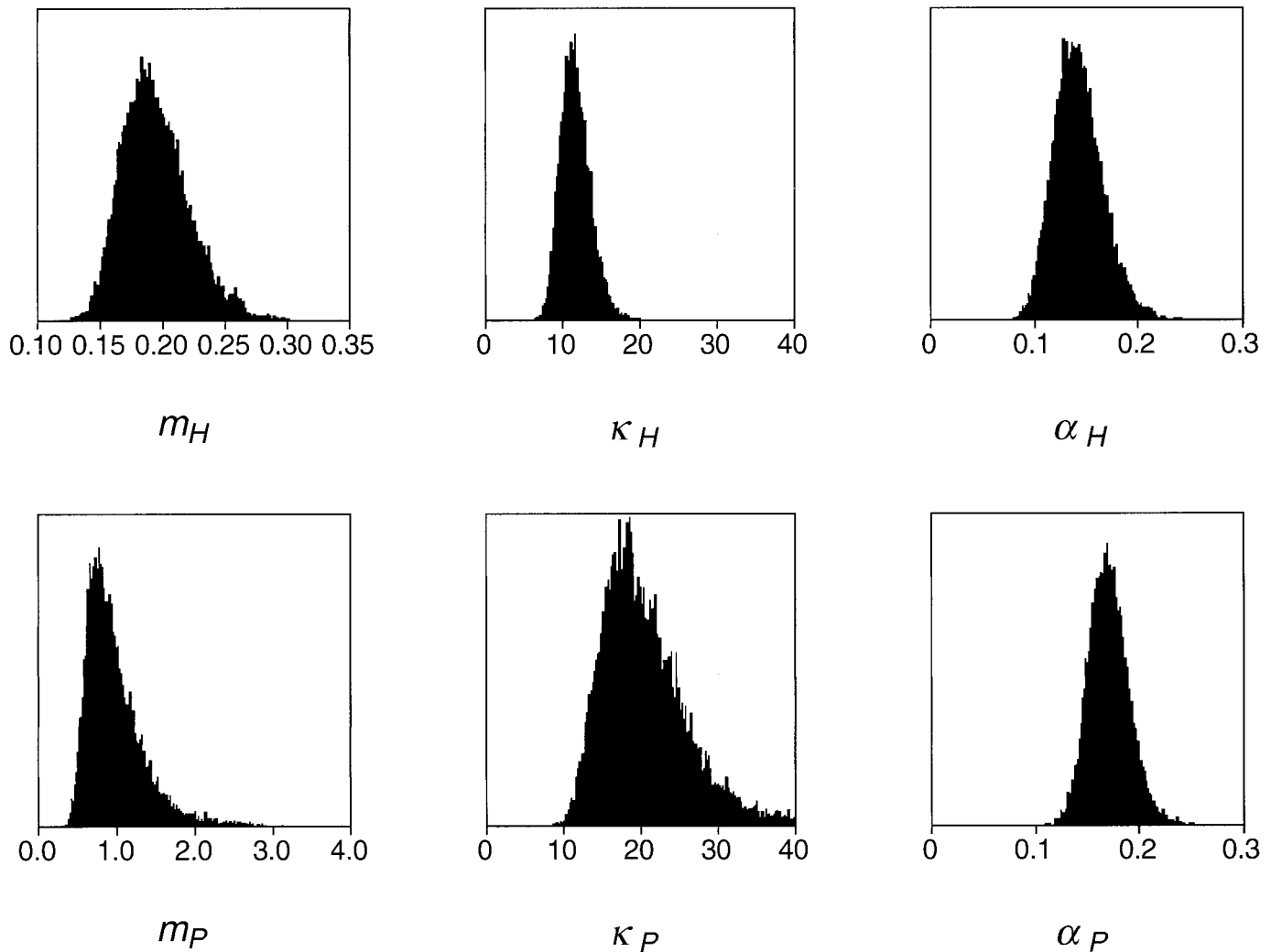
FIG. 7.   The posterior probability density of the parameters of the model of DNA substitution for the gophers and the lice. The posterior probability distributions for substitution rate (m), transition/transversion rate ratio (κ), and gamma shape parameter (α) for hosts (H) and parasites (P) are plotted.

5, and 14–16). For many of the branches, it appears that the probabilities that the branch contained a source or a target were roughly equal. However, for several branches, it was more likely that the branch contained a source or target. For example, branches 11 and 13 were more likely to have sources whereas branches 9, 10, and 12 were more likely to be targets (Fig. 12).

Figure 13 shows the specific branches involved in events of host switching. The figure shows the probability that a host-switching event had a source on branch i and a target on branch j. Events of host switching were especially prevalent between $n_6 \leftrightarrow n_7$, $n_{12} \leftrightarrow n_{13}$, $n_{11} \rightarrow n_{10}$, and $n_{10} \rightarrow n_9$.

We also compared the model of host switching used in this analysis (M_1) to an alternative model of host switching. The alternative model (M_2) was the same in all respects to M_1 except that, when a host switch occurs, the probability that a parasite switches to a closely related host is higher than the probability of a switch to a distantly related host. The probability that a switch was made to a specific target branch was proportional to the inverse of the distance to the

target branch. The model was motivated by the observation that when parasites infect more than one host, there is a tendency to infect closely related hosts (Norton and Carpenter 1998). Figure 14 provides an example of how host-switch probabilities were calculated for M_2. In a Bayesian framework, model choice is often guided by the Bayes factor. The Bayes factor is the ratio of the posterior odds to the prior odds in favor of M_2.

$$BF = \frac{\dfrac{f(M_2 \mid \text{data})}{f(M_1 \mid \text{data})}}{\dfrac{f(M_2)}{f(M_1)}} = \frac{f(M_2 \mid \text{data})f(M_1)}{f(M_1 \mid \text{data})f(M_2)}. \qquad (10)$$

The Bayes factor can also be interpreted as the ratio of the marginal likelihoods of the competing models, where the marginal likelihood is the probability of the data with all model parameters integrated out. Twice the logarithm of the Bayes factor is on roughly the same scale as the likelihood-ratio test statistic. Lavine and Schervisch (1999, p. 120) argue
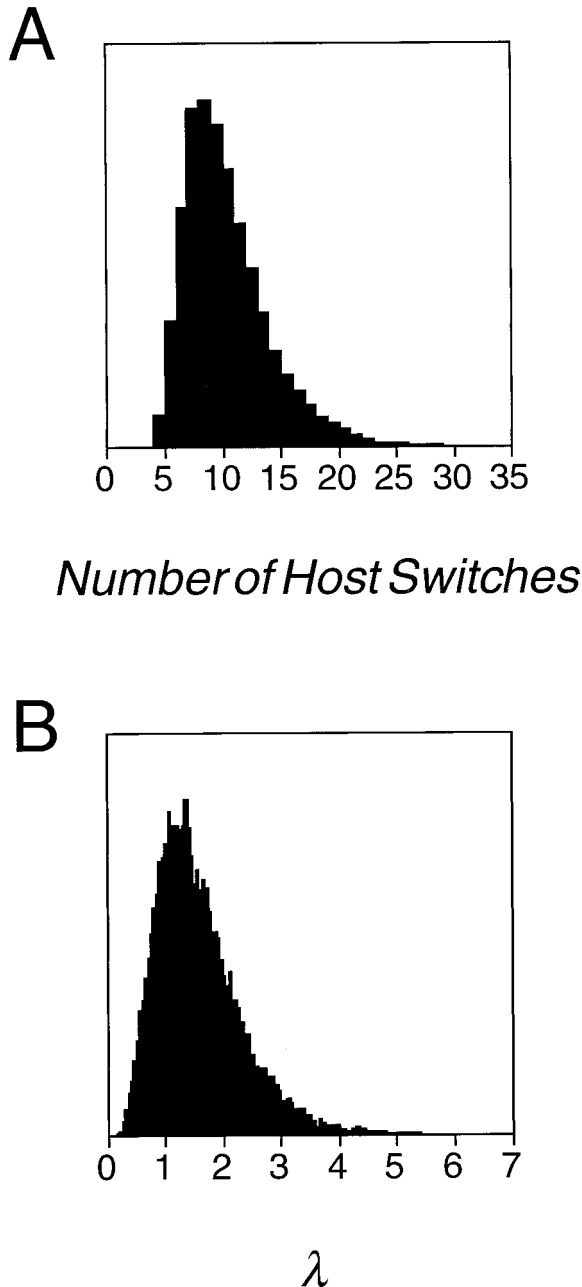
FIG. 8. The posterior probability density of the number of host-switching events (A) and of the rate of host switching (B) did not change substantially when the gopher tree was fixed.
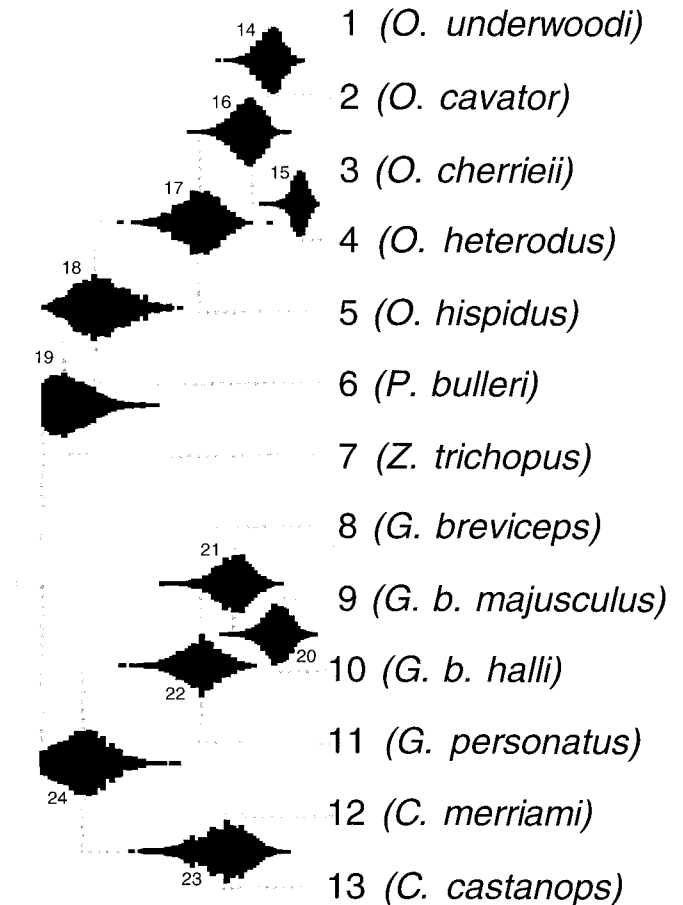

FIG. 9. The posterior probability density for host speciation times.

not fit as well as a model that assumes equally likely branch switches. Other models that modify the probabilities of host switching, however, may fit better than either model considered in this paper.

DISCUSSION

The gopher and louse sequences collected by Hafner et al. (1994) are one of the paradigm examples of cospeciation and have been examined many times. Page (1996) applied a meth-

that the Bayes factor should be interpreted as measuring "the *change* in the odds in favor of the hypothesis when going from the prior to the posterior."

We estimated the Bayes factor for a comparison of M_1 to M_2 using MCMC. The Markov chain was allowed to jump between competing models. The proportion of the time that the chain stayed in M_1 or M_2 is an approximation of the Bayes factor. The Bayes factor was equal to 1.03×10^{-3}, which represents very strong support in favor of M_1. This means that the specific model of host switching that assigns a higher probability of switches between closely related hosts does

TABLE 2. Key to the taxon numbers used in Figures 9–13.

| Host species	Associated parasite	Number
Orthogeomys underwoodi	Geomydoecus setzeri	1
Orthogeomys cavator	Geomydoecus panamensis	2
Orthogeomys cherrieii	Geomydoecus cherrieii	3
Orthogeomys heterodus	Geomydoecus costaricensis	4
Orthogeomys hispidus	Geomydoecus chapini	5
Pappogeomys bulleri	Geomydoecus nadleri	6
Zygogeomys trichopus	Geomydoecus trichopi	7
Geomys breviceps	Geomydoecus ewingi	8
Geomys b. majusculus	Geomydoecus geomydis	9
Geomys b. halli	Geomydoecus oklahomensis	10
Geomys personatus	Geomydoecus texanus	11
Cratogeomys merriami	Geomydoecus perotensis	12
Cratogeomys castanops	Geomydoecus expansus	13
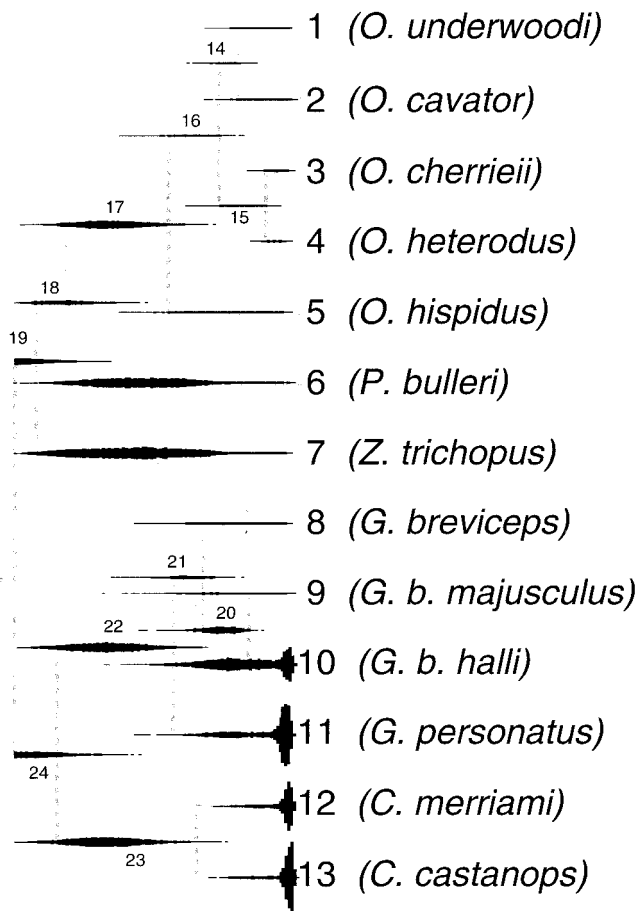
FIG. 10. The posterior probability density for having a source on a particular branch.
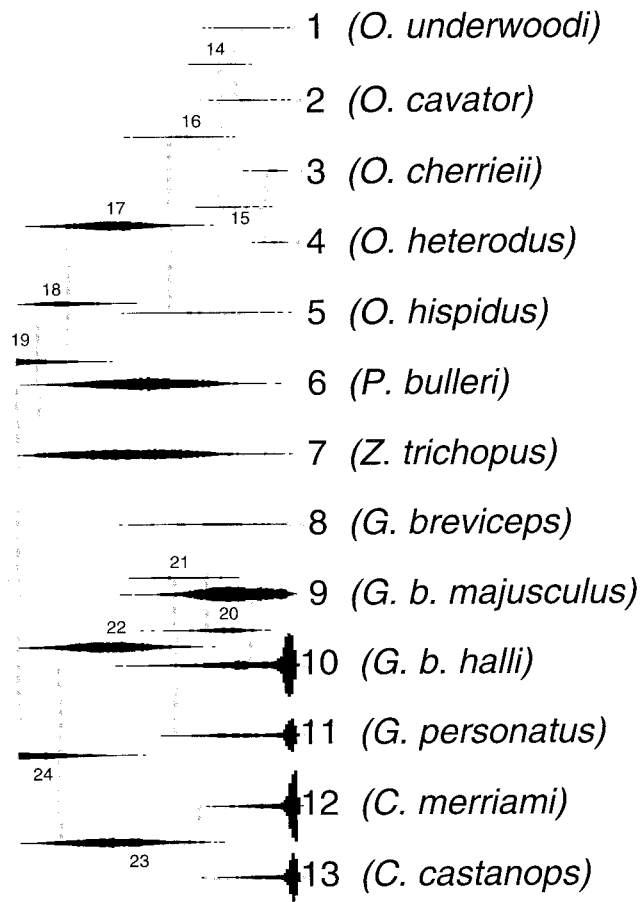

FIG. 11. The posterior probability density for having a target on a particular branch.

od that attempts to maximize the number of cospeciation events between hosts and parasites using as the observations the estimated trees of gophers and lice. When Page's TreeMap method was applied to the maximum-likelihood trees for the gophers and lice, the maximum number of cospeciation events between the gophers and lice was 10, with seven sorting events and one event of host switching required to explain the differences between the hosts and parasites; eight events are required to map the parasite tree onto the host tree. This estimate is in the credible region of the number of host-switching events found in this paper. However, the Bayesian method put almost no weight on the smallest numbers of host switches (zero to three events). The probabilities of having four, five, and six events of host switching were 0.054, 0.098, and 0.122, respectively.

Given that current methods provide a quick and reasonable estimate of the number of host switches required to explain a host and parasite tree, why use a statistical method that assumes a stochastic model? The method presented in this paper has several advantages over earlier methods. First, the method does not assume that the phylogenetic trees of hosts and parasites are known without error; the method treats the host and parasite phylogenies as random variables and integrates over uncertainty in phylogeny. Second, by assuming a stochastic model of host switching and performing standard

statistical estimation of the parameters of the model, one can easily obtain credibility intervals on the parameters and examine other, perhaps less likely, reconstructions of host switching. Finally, one can change the model of host switching and perform hypothesis tests of competing models of host switching. Bayesian methods of model choice can provide powerful tools for determining what aspects of the model of host switching are important. Statistical methods of model choice have been used with success to scrutinize models of DNA substitution. The result has been a gradual improvement in models of DNA substitution in terms of their ability to explain observed DNA sequences. It is conceivable that a similar research program can be implemented for the study of host switching with the overall goal being to find biological processes that provide significant improvements to the fit of stochastic models of host switching.

As presented here, our method does not include several important biological processes. Those that should eventually be accommodated include: (1) parasite speciation and/or extinction in the absence of speciation/extinction by the host (this will allow multiple parasites to be associated with each host); and (2) host extinction and sampling (hidden or missing hosts can serve as a source for host switching). Extensions of the numerical approaches outlined in this paper should eventually allow speciation by the hosts and parasites to be
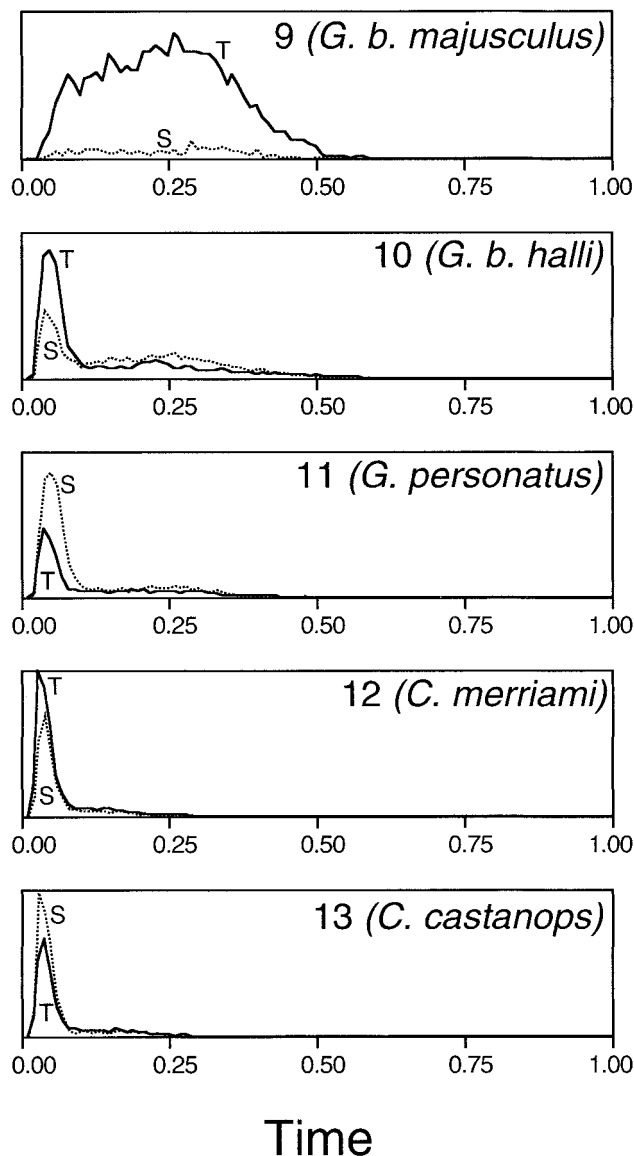
FIG. 12. Some branches are more likely to contain sources or targets. The posterior probability of having a source or target associated with branches 9–13. The tips of the tree are at 0.00 and the root at 1.00.
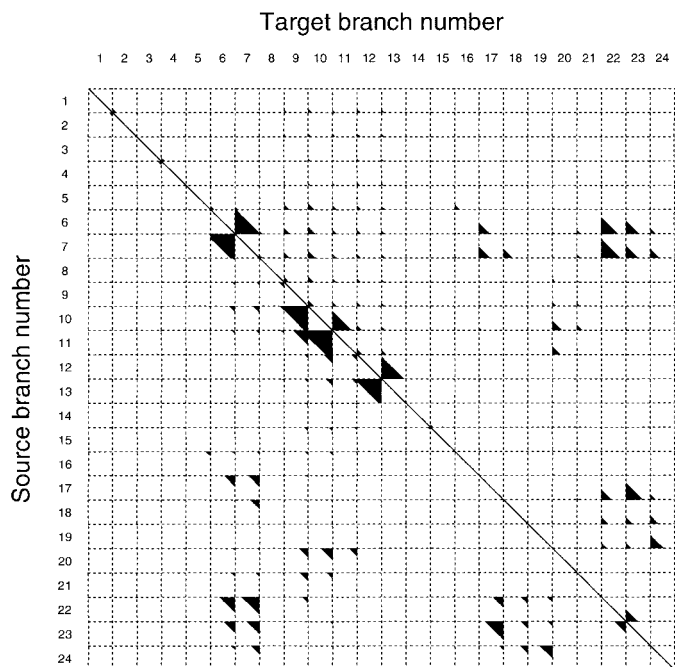

FIG. 13. The probability of having an event of host switching that starts on branch i (source) and goes to branch j (target). The amount of the box that is shaded is proportional to the probability of a switch from the source to the target branch.

models. Host switching, then, would be analagous to migration to new areas by species or horizontal gene transfer. Thus, the method presented here has other potential applications. For example, the method could be used to estimate rates of genetic transfer among bacteria. In this case, the host phylogeny would be the phylogeny of the bacteria and the parasites would be the multiple homologous genes sampled from the bacteria. The object would be to explain the incongruence among the gene trees via horizontal transfer. Similarly, the model of host switching proposed here, or a similar model,
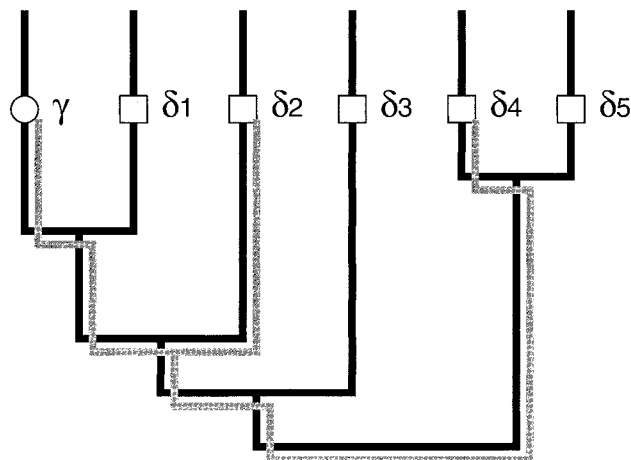
accommodated and also allow the constraint that only one parasite species is associated with a given host to be relaxed. For example, a birth-death prior on the times of host-speciation events could provide a way of allowing host-switching events to originate on extinct or unsampled host lineages. Information on the speciation and extinction rates is contained in the times of the speciation events leading to extant species (Nee et al. 1994). Moreover, the molecular clock assumption can be relaxed by assuming that rates of substitution change at speciation events (Thorne et al. 1998) or on the tree as a step function (Huelsenbeck et al. 2000).

Page (1994) has pointed out that there are instances of associations that are similar to cospeciation of hosts and parasites. For example, species within biogeographic areas and gene trees within species trees can be described by similar

FIG. 14. In the modified model of host switching (M_2), the probability of a switch between two host lineages was inversely proportional to the distance between the source branch (γ) and the target branch (δ_i). This figure shows the paths that must be taken to calculate the distance between the source and targets 2 and 4.

might be useful to estimate phylogeny in the presence of occasional genetic transfer among species. Finally, biogeography might usefully be studied in a framework similar to the one adopted here; the area cladogram would be considered fixed and the multiple species trees on the area cladogram would be considered variable.

We see this paper as providing a framework for the statistical analysis of host switching using Bayesian inference. We hope that this paper stimulates research that has as its goal the improvement and testing of models of host switching.

PROGRAM AVAILABILITY

A computer program, written in C, is available to perform the analyses described in this paper. The program is available from http://brahms.biology.rochester.edu.

LITERATURE CITED

Bowen, M. D., C. J. Peters, and S. T. Nichol. 1997. Phylogenetic analysis of the Arenaviridae: patterns of virus evolution and evidence for cospeciation between Arenaviruses and their rodent hosts. Mol. Phylogenet. Evol. 8:301–316.
Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.
Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82: 711–732.
Hafner, M. S., and R. D. M. Page. 1995. Molecular phylogenies and host-parasite cospeciation: gophers and lice as a model system. Phil. Trans. R. Soc. Lond. B 349:77–83.
Hafner, M. S., P. D. Sudman, F. X. Villablanca, T. A. Spradling, J. W. Demastes, and S. A. Nadler. 1994. Disparate rates of molecular evolution in cospeciating hosts and parasites. Science 265:1087–1090.
Hasegawa, M., T. Yano, and H. Kishino. 1984. A new molecular clock of mitochondrial DNA and the evolution of Hominoids. Proc. Japan Acad. Ser. B 60:95–98.
Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating the human-ape split by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.
Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.
Huelsenbeck, J. P., B. Rannala, and Z. Yang. 1997. Statistical tests of host-parasite cospeciation. Evolution 51:410–419.
Huelsenbeck, J. P., B. Larget, and D. L. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. Genetics. *In press.*
Lavine, M., and M. J. Schervish. 1999. Bayes factors: what they are and what they are not. Am. Statistician 53:119–122.
Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1091.
Nee, S., R. M. May, and P. H. Harvey. 1994. The reconstructed evolutionary process. Phil. Trans. R. Soc. Lond. B 344:305–311.
Norton, D. A., and M. A. Carpenter. 1998. Mistletoes as parasites: host specificity and speciation. Trends Ecol. Evol. 13:101–104.
Page, R. D. M. 1988. Quantitative cladistic biogeography: constructing and comparing area cladograms. Syst. Zool. 37: 254–270.
———. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. 43:58–77.
———. 1996. Temporal congruence revisited: Comparison of mitochondrial DNA sequence divergence in cospeciating pocket gophers and thier chewing lice. Syst. Biol. 46:151–167.
Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Simon, D., and B. Larget. 1998. Bayesian analysis in molecular biology and evolution (BAMBE), version 1. 01 beta. Department of Mathematics and Computer Science, Duquesne University, Duquesne, PA.
Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony and other methods. Sinauer Associates, Sunderland, MA.
Thorne, J., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.
Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.
———. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Corresponding Editor: W. T. Starmer

APPENDIX 1

A Simple Example of Bayesian Analysis Using Markov Chain Monte Carlo

A simple example illustrates how the Metropolis-Hastings-Green algorithm (Metropolis et al. 1953; Hastings 1970; Green 1995) can be used to evaluate the posterior distribution of a parameter. Consider the case of tossing a coin with the object of estimating the parameter θ, the probability of observing a head. The likelihood function, $l(\theta)$, in the case when there are five heads observed in 10 coin tosses is displayed in Figure A1. In this simple example, the likelihood is a function of a single parameter, whereas the model of host switching developed in this paper has many parameters, but the principle of MCMC is the same. In maximum-likelihood inference, the value of θ that maximizes $l(\theta)$ is chosen as the best estimate. Bayesian inference of the parameter θ, however, is based on the posterior distribution of the parameter (or, in other words, the probability of the parameter given the data).

The posterior distribution of θ is obtained using Bayes formula and requires that a prior be specified for the parameter θ. A uniform
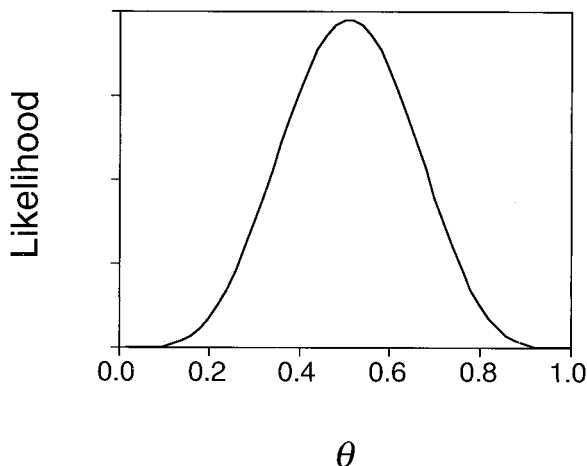

FIG. A1. The likelihood function when $x = 5$ heads are observed in $n = 10$ tosses of a coin. The parameter θ is the probability of heads for the coin. The maximum-likelihood estimate of θ is obtained by picking that value of θ that maximizes the likelihood function; $\theta = 0.5$ is the maximum-likelihood estimate for this case.

distribution can be assumed as a prior for θ, which is analogous to the uniform priors used in this study. A uniform prior on a parameter implies that, before the analysis, the investigator has no strong opinion on what value the parameter takes. The advantage of a uniform prior is that it does not bias the results of the analysis; the posterior distribution will be largely determined by the likelihood function. With a uniform prior, the posterior distribution of θ after observing the data is proportional to the likelihood function in Figure A1, but rescaled so that the total area under the curve is one. MCMC would work as follows to produce a sample of points drawn from the posterior distribution of θ. (1) The current state of the chain is θ. If this is the first generation of the chain, we set θ to some arbitrarily chosen value in the interval (0, 1). (2) Propose a new state θ' uniformly at random from the interval ($\theta - \epsilon$, $\theta + \epsilon$). (3) Calculate the likelihood ratio, $l(\theta')/l(\theta)$ and a random number U distributed uniformly between 0 and 1. (4) If U is less than the likelihood ratio (which will always be the case when the proposed value θ' is more likely), then the new state is accepted and we set $\theta = \theta'$. Otherwise, the chain does not change state. In either case, the value of θ (which may not have changed) is added to the sample. Steps 1–4 are repeated a large number of times.

Notice that the sequence of sampled points is not independent— sampled points are likely to be close to their predecessor. However, the relative frequency with which the chain stays in different sub-intervals of [0,1] is an approximation of the posterior probability distribution of θ in that subinterval. Figure A2 displays the MCMC approximation of the posterior density of θ and the true posterior density as determined by calculus. Note that the estimated and true posteriors distributions closely match when the chain is run for many generations.

APPENDIX 2

We use the Metropolis-Hastings-Green (MHG) algorithm to evaluate the posterior distributions of the parameters λ, θ_H, and θ_P. The acceptance probability for the MHG algorithm takes the general form:

$$R = \min[1, (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio})]. \quad (A1)$$

We use 16 move types to update the state of the Markov chain. These move types: (1) add an event of host switching to the tree; (2) delete an event of host switching from the tree; (3) change the position of the ith event (z_i) by a small amount; (4) change the branch of the ith source (γ_i); (5) change the branch of the ith target (δ_i); (6) change the host tree topology (τ_H); (7) change the host branch lengths ($\mathbf{t}_H$); (8) change m_H; (9) change m_P; (10) change κ_H; (11) change κ_P; (12) change α_H; (13) change α_P; (14) change λ, (15) change change π_H; and (16) change π_P. The ith move type is made with probability y_i.

Adding an event of host switching.—With probability ψ_1, the addition of an event of host switching is attempted. (The reverse move, deleting an event, is attempted with probability ψ_2.) The prior ratio for the addition of a single point of host switching ($z*$, $\gamma*$, $\delta*$) to the current state $\mathbf{e} = (\xi, \mathbf{z}, \gamma, \delta)$ is

$$\frac{\dfrac{e^{-\lambda T}(\lambda T)^{\xi+1}}{(\xi+1)!} \times \left(\dfrac{1}{T}\right)^{\xi+1} \times \prod_{i=1}^{\xi}\left[\dfrac{1}{b(z_i)-1}\right] \times \left[\dfrac{1}{b(z*)-1}\right]}{\dfrac{e^{-\lambda T}(\lambda T)^{\xi}}{(\xi)!} \times \left(\dfrac{1}{T}\right)^{\xi} \times \prod_{i=1}^{\xi}\left[\dfrac{1}{b(z_i)-1}\right]}$$

$$= \frac{\lambda}{(\xi+1)[b(z*)-1]} \quad (A2)$$

and the proposal ratio is

$$\frac{\psi_2 \times \dfrac{1}{(\xi+1)}}{\psi_1 \times \dfrac{1}{(\xi+1)} \times \dfrac{1}{T} \times \dfrac{1}{b(z*)-1}} = \frac{T\psi_2[b(z*)-1]}{\psi_1}. \quad (A3)$$
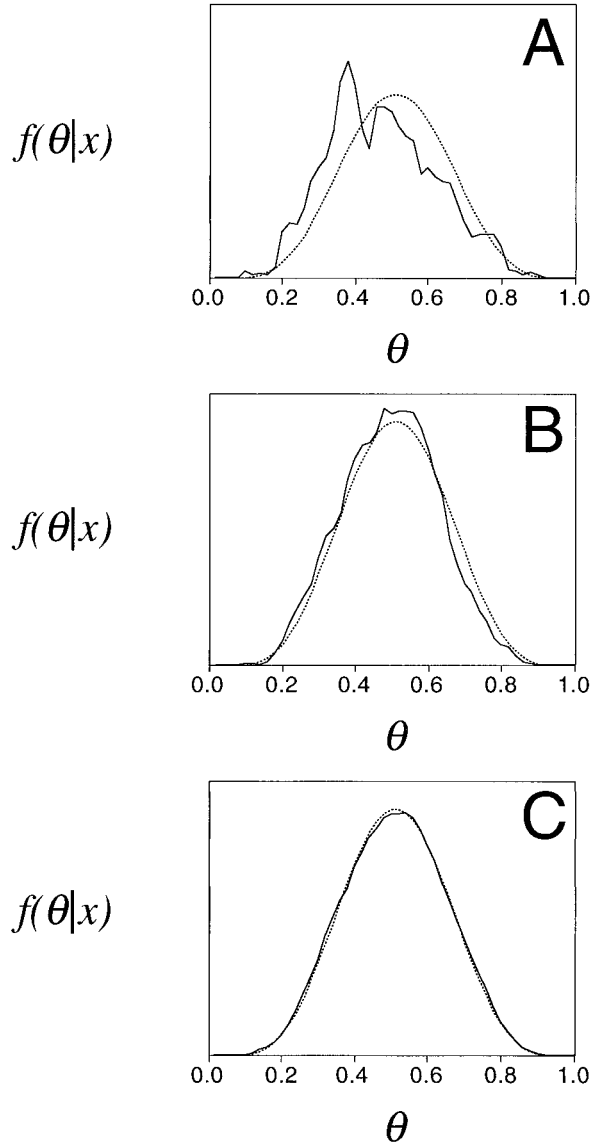
The acceptance probability is then

FIG. A2. The posterior probability of θ for the coin-tossing problem. The dotted line is the posterior probability calculated analytically. The solid line is the approximation of the posterior probability obtained using Markov chain Monte Carlo. The chain was run for 5000 generations (A), 50,000 generations (B), and 500,000 generations (C).

$$R = \min\left(1, (\text{likelihood ratio}) \times \frac{\lambda T\psi_2}{(\xi+1)\psi_1}\right). \quad (A4)$$

Deleting an event of host switching.—With probability ψ_2, an attempt is made to delete one of the ξ events from the tree. The acceptance ratio for this step is simply the inverse of equation (A4).

Changing the time of an event.—A move that changes the time of a source and target on the host tree is made with probability ψ_3. Each of the ξ events on the tree had an equal probability of being chosen. The time of an event is increased or decreased by adding to the current time (z_i) a uniformly distributed random variable on the interval $[-\epsilon_3, +\epsilon_3]$. The new time is denoted $z*$. The acceptance probability is then

$$R = \min\left(1, (\text{likelihood ratio}) \times \frac{b(z_i)-1}{b(z*)-1}\right). \quad (A5)$$

If the proposed time for an event crosses an internal node on the host tree, then each of the two possible branches is chosen with equal probability. This means that if the time of the event is being increased (i.e., moved toward the root of the tree) and the event must pass a node, that it either continues to move toward the root or that it changes direction and moves up the other branch.

Changing the source branch.—With probability ψ_4, the source branch of an event is changed. Each of the ξ sources has equal probability of being chosen. Of the $b(z) - 1$ branches available, each has equal probability of being chosen as the new source branch. The acceptance ratio for this move is

$$R = \min(1, \text{likelihood ratio}). \tag{A6}$$

Changing the target branch.—A move that changes the target is attempted with probability ψ_5. The details of this move are identical to move type 4.

Changing the host tree topology.—A move that changes the phylogeny of the hosts was made with probability ψ_6. One of the $s - 2$ internal nodes of the tree (excluding the root) was chosen; each internal node had an equal probability of being chosen. The new time of the speciation event of the node was chosen by drawing a uniform random variable on the interval (t_A, t_D), where t_A is the time of the ancestor of the chosen node and t_D is the time of the younger descendent node of the chosen node. A potentially new topology was constucted by randomly choosing two of the three descendent nodes to be descendents of the chosen node.

Sources that were associated with the branches in the area of rearrangement remain associated with the same branch. However, the times of the sources were changed such that their placement on the branch was proportional to their original placement. Targets associated with sources in the area of rearrangement had their times changed to reflect the new source times. The acceptance probability for this move is

$$R = \min\left(1, (\text{likelihood ratio}) \times \frac{e^{-\lambda T^*}(\lambda T^*)^\xi}{e^{-\lambda T}(\lambda T)^\xi} \times \frac{T}{T^*}\right.$$
$$\left. \times \prod_{i=1}^{\xi} \frac{b(z_i) - 1}{b(z_i^*) - 1}\right), \tag{A7}$$

where T^* is the total tree length after the node time has been adjusted.

Changing the host branch lengths.—With probability ψ_7, the time

of one of the $s - 2$ internal nodes (excluding the root) is changed. One of the internal nodes is chosen at random. The time of this internal node is increased or decreased by adding a uniformly distributed random variable on the interval $[-\epsilon_7, +\epsilon_7]$. The acceptance probability for this move is

$$R = \min\left(1, (\text{likelihood ratio}) \times \frac{e^{-\lambda T^*}(\lambda T^*)^\xi}{e^{-\lambda T}(\lambda T)^\xi} \times \frac{T}{T^*}\right.$$
$$\left. \times \prod_{i=1}^{\xi} \frac{b(z_i) - 1}{b(z_i^*) - 1}\right), \tag{A8}$$

where T^* is the total tree length after the node time has been adjusted.

Changing the value of a parameter by a small amount.—The parameters m_H, m_P, κ_H, κ_P, α_H, α_P, and λ had uniform priors (see Table 1). A move was made with probability ψ_8, ψ_9, ψ_{10}, ψ_{11}, ψ_{12}, ψ_{13}, or ψ_{14} that changed m_H, m_P, κ_H, κ_P, α_H, α_P, or λ, respectively. The current value of the parameter was increased or decreased by adding a uniformly distributed random variable on the interval $[-\epsilon_i, +\epsilon_i]$, where $i = 8, 9, 10, 11, 12, 13,$ or 14. The acceptance probability for moves 8–13 is

$$R = \min(1, \text{likelihood ratio}) \tag{A9}$$

and the acceptance probability for move 14 is

$$R = \min\left(1, (\text{likelihood ratio}) \times \frac{e^{-\lambda T^*}(\lambda T^*)^\xi}{e^{-\lambda T}(\lambda T)^\xi}\right). \tag{A10}$$

Changing the host or parasite base frequencies.—With probability ψ_{15} a move is attempted that changes the equilibrium base frequencies, π. The sum of the base frequencies is constrained to equal one and new values are proposed from a Dirichlet distribution with expected values at the current values. The Dirichlet distribution has probability density

$$f(\pi \mid \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i \in S} \Gamma(\alpha_i)} \prod_{i \in S} \pi_i^{(\alpha_i - 1)}, \tag{A11}$$

where S is the state space (A, C, G, or T), α_i is the Dirichlet parameter for the ith nucleotide, $\alpha_0 = \Sigma_{i \in S} \alpha_i$, and π_i is the frequency of the ith nucleotide. New base frequencies are drawn from the Dirichlet distribution with $\alpha_i = \pi_i \alpha_0$. We set $\alpha_0 = 100.0$ in all of the analyses of this study.