Evolution, 51(2), 1997, pp. 410–419

STATISTICAL TESTS OF HOST-PARASITE COSPECIATION

JOHN P. HUELSENBECK,[1] BRUCE RANNALA, AND ZIHENG YANG
Department of Integrative Biology, University of California, Berkeley, California 94720-3140
[1]*E-mail: johnh@mws4.biol.berkeley.edu*

Abstract.—A history of cospeciation (synchronous speciation) among ecologically associated, but otherwise distantly related, species is often revealed by a strong correspondence of their phylogenies. In this paper, we present several tests of cospeciation that use maximum-likelihood and Bayesian methods of phylogenetic estimation. The hypotheses tested include: (1) topological agreement of phylogenies for coevolving groups; (2) identical speciation times of associated species; and (3) identical evolutionary rates in genes of associated species. These tests are applied to examine a possible instance of host-parasite coevolution among pocket gophers and lice using mitochondrial COI DNA sequences. The observed differences between gopher and louse trees cannot be explained by sampling error and are consistent with a rate of host switching about one-third the host speciation rate. A subset of the gopher-louse data is consistent with a common history of evolution (i.e., the topologies and speciation times are identical). However, the relative rate of nucleotide substitution is two to four times higher in the lice than in the gophers.

Key words.—Birth-death process, conditional probability test, cospeciation, likelihood-ratio test, maximum likelihood, maximum-posterior probability.

Received May 16, 1996. Accepted October 16, 1996.

Systematists have often observed similar phylogenies for groups of species that have a close ecological association but are otherwise distantly related. In such cases, the associated species may occupy identical positions if the phylogeny of one group is superimposed on that of the other. This phenomenon is particularly frequent among host-parasite or host-symbiont systems, giving rise to generalizations such as Farenholz's rule (that parasite phylogeny mirrors host phylogeny; see Brooks 1979) and Szidat's rule (that the more primitive the host, the more primitive the parasite it harbors; Szidat 1956). Such phylogenetic agreement has also been observed between phytophagous insects and the plant species on which they feed (Ehrlich and Raven 1964).

Observations of phylogenetic similarity among ecologically associated taxonomic groups have frequently been cited as evidence of coevolution among the groups. Under this paradigm, a speciation event in one group initiates a subsequent speciation in the associated group. In the case of hosts and parasites, for example, a speciation event within a particular host lineage might be expected to isolate the parasite population associated with each incipient host species, and thus to produce an allopatric speciation event among parasites. The prediction, in this case, is that the parasite and host phylogenies should be similar, reflecting the association of speciation events in the two groups. Many recent studies have attempted to identify instances of coevolution between hosts and parasites using phylogenies based on allozyme data (Baverstock et al. 1985; Hafner and Nadler 1988; Rannala 1992) or DNA sequences (Hafner et al. 1994).

In this paper, we present several statistical tests designed to identify coevolution. The methods are intended for use with DNA sequence data, and exploit recent advances in maximum-likelihood and Bayesian phylogenetic estimation (Huelsenbeck and Bull 1996; Rannala and Yang 1996). Although we focus in this paper on the specific problem of identifying coevolution between hosts and parasites, the methods should prove useful in studying other instances of coevolution as well. Three hypotheses of cospeciation are examined: (1) agreement of topologies for host and parasite species; (2) identical speciation times in hosts and parasites; and (3) identical nucleotide substitution rates in sequences from associated species.

DATA, MODELS, AND ESTIMATION THEORY

The statistical tests presented in this paper depend on phylogenetic estimates obtained using either maximum-likelihood (ML; Felsenstein 1981) or maximum posterior probability (MAP; Rannala and Yang, in press) methods. In this section we review the statistical models underlying both methods of estimation.

Data

We assume that DNA sequences from homologous regions are available for both host and parasite taxa. For simplicity, we assume a one-to-one correspondence between host and parasite taxa. Let $\mathbf{X} = \{x_{kh}\}$ and $\mathbf{Y} = \{y_{ka}\}$ be the aligned nucleotide sequences for hosts and associated parasites, respectively, where $k = 1, 2, \ldots, s$, $h = 1, 2, \ldots, c_H$, and $a = 1, 2, \ldots, c_P$; s is the number of sequences sampled (equal for parasites and hosts), c_H is the number of nucleotide sites per sequence for hosts, and c_P is the number of nucleotide sites per sequence for parasites. Each column of the data matrix, $\mathbf{x}_h = \{x_{1h}, \ldots, x_{sh}\}'$ or $\mathbf{y}_a = \{y_{1a}, \ldots, y_{sa}\}'$, specifies the nucleotides for the s sequences at the hth host site or ath parasite site.

Models, Parameter Estimation, and Hypothesis Testing

To calculate the probability of observing a particular site pattern, it is necessary to specify a model of DNA substitution, a topology τ, and the branch lengths of the topology. Branch lengths are measured in units of expected number of substitutions per site and are denoted $\mathbf{v} = \{v_1, v_2, \ldots, v_b\}$, where b is the total number of branches ($2s - 3$ for unrooted topologies and $2s - 2$ for rooted topologies). Figure 1 provides an example of a rooted topology with branch lengths indicated.

We use the substitution model implemented in J. Felsen-

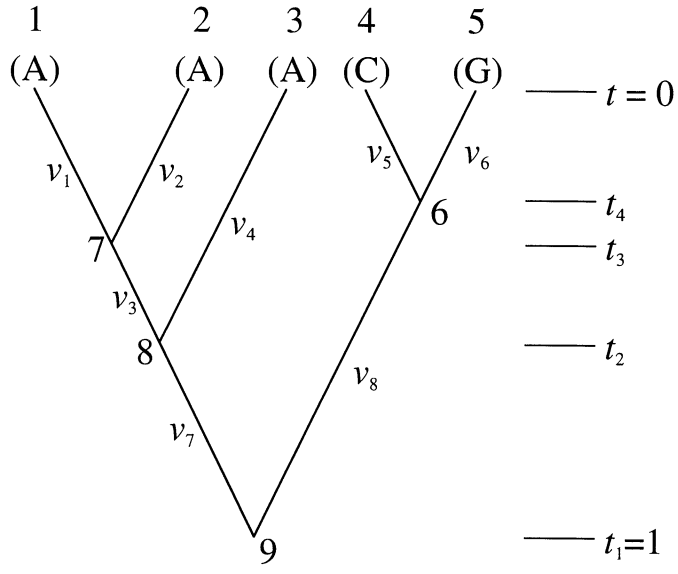

FIG. 1. A five-taxon tree with terminal nodes (host species) labeled 1 to 5 and interior nodes labeled 6 to 9. The branch lengths, in units of expected number of substitutions per site, are $\mathbf{v} = \{v_1, \ldots, v_9\}$. One possible site pattern for the host species is shown at the tips of the branches, $\mathbf{x}_h$ = (A, A, A,C, G)'. The speciation times are indicated by $t_1, \ldots, t_4$. The node times are standardized such that the root is at $t_1 = 1$ and the tips of the tree are at $t = 0$.

stein's DNAML program. This model has been implemented since 1984 and is denoted the F84 model in this paper. The instantaneous rate matrix of the substitution process is

$$\mathbf{Q} = \{q_{ij}\} = \begin{bmatrix} \cdot & \pi_C & (1 + \kappa/\pi_R)\pi_G & \pi_T \\ \pi_A & \cdot & \pi_G & (1 + \kappa/\pi_Y)\pi_T \\ (1 + \kappa/\pi_R)\pi_A & \pi_C & \cdot & \pi_T \\ \pi_A & (1 + \kappa/\pi_Y)\pi_C & \pi_G & \cdot \end{bmatrix}, \quad (1)$$

where q_{ij} $(i \neq j)$ is the substitution rate from nucleotide i to j, with the nucleotides ordered A, C, G, and T. The transition/transversion rate ratio is κ and π_j is the equilibrium frequency of nucleotide j, with $\pi_Y = \pi_C + \pi_T$ and $\pi_R = \pi_A + \pi_G$. When $\kappa > 0$, transitions are more frequent than transversions. The diagonals of the rate matrix q_{ii} are specified by the requirement that the row sums of $\mathbf{Q}$ are zero. The matrix is multiplied by a constant such that the average rate of substitution is one, and time (t) is then measured by the expected number of substitutions per site (v). Let $\Theta = \{\kappa, \pi_A, \pi_C, \pi_G, \pi_T\}$ and $\mathbf{P}(v,\Theta) = \{p_{ij}(v,\Theta)\}$ be the transition probability matrix, where $p_{ij}(v,\Theta)$ is the probability that nucleotide i changes into j over branch length v. $\mathbf{P}(v,\Theta)$ can be obtained from the rate matrix $\mathbf{Q}$ through the operation $\mathbf{P}(v,\Theta) = e^{\mathbf{Q}v}$.

Among site rate heterogeneity can be accommodated by allowing rates at different sites to be random variables drawn from a gamma distribution. The shape parameter α of the distribution is inversely related to the rate variation and can be estimated by maximum likelihood. Yang (1994) provides formulae for likelihood calculation under the F84 model with either equal or gamma distributed rates among sites.

The probability of observing a particular site pattern given a topology, branch lengths, and a model of DNA substitution is a sum over all possible nucleotide assignments to the internal nodes of the tree. For the topology of Figure 1, the probability of observing data at site h, say $\mathbf{x}_h$ = {A, A, A, C, G}, is

$$f(\mathbf{x}_h \mid \boldsymbol{\tau},\mathbf{v},\Theta) = \sum_{n_6}\sum_{n_7}\sum_{n_8}\sum_{n_9} \pi_{n_9}\cdot p_{n_7A}(v_1,\Theta)\cdot p_{n_7A}(v_2,\Theta) \cdot p_{n_8n_7}(v_3,\Theta)\cdot p_{n_8A}(v_4,\Theta)\cdot p_{n_6C}(v_5,\Theta) \cdot p_{n_6G}(v_6,\Theta)\cdot p_{n_9n_8}(v_7,\Theta)\cdot p_{n_9n_6}(v_8,\Theta), \quad (2)$$

where the summations are over all possible nucleotide states n_6, n_7, n_8, and n_9 for the ancestral nodes.

The likelihood is the probability of observing the data given the specified topology and the model of nucleotide substitution. We assume independent substitution at sites, and the likelihood for the host sequences is

$$L(\boldsymbol{\tau},\mathbf{v},\Theta \mid \boldsymbol{\chi}) = \prod_{h=1}^{c_H} f(\mathbf{x}_h \mid \boldsymbol{\tau},\mathbf{v},\Theta). \quad (3)$$

The likelihood for the parasite sequences can be calculated similarly. Parameters in the model ($\boldsymbol{\tau}$, $\mathbf{v}$, and Θ) are estimated by maximizing the likelihood function.

The likelihood-ratio statistic for comparing two models (Λ) is defined as

$$\Lambda = \frac{L_0(\text{Null Hypothesis} \mid \text{Data})}{L_1(\text{Alternative Hypothesis} \mid \text{Data})}. \quad (4)$$

The ratio of the likelihoods calculated under the null and alternative hypotheses is a measure of the relative merit of the hypotheses. If Λ is less than one, then the alternative hypothesis is favored. If Λ is greater than one, the null hypothesis is favored. For the special case in which nested hypotheses are considered (i.e., the null hypothesis is a special case of the alternative hypothesis), $\Lambda < 1$ and $-2\log\Lambda$ is asymptotically χ^2 distributed under the null hypothesis with q degrees of freedom, where q is the difference in the number of parameters between the general and restricted hypotheses (Cox and Hinkley 1974). Topology is not a standard statistical parameter, however, and many of the usual results from statistics may not apply (Goldman 1993). We therefore use Monte Carlo simulation (also known as parametric bootstrapping) to generate the null distribution of $-2\log\Lambda$ when the likelihood is maximized over topologies. Simulated data are generated under the null hypothesis using maximum-likelihood estimates of parameters.

The MAP method (Rannala and Yang 1996) uses a birth-death process as the prior distribution of topologies and branch lengths. The speciation and extinction rates of the process are λ and μ, respectively. Tree topologies and speciation times are regarded as random variables, and the likelihood function is calculated by summing over topologies and integrating over node times. Estimates of parameters are then used to evaluate the conditional probabilities of different topologies given the data. These probabilities provide significance measures for tests involving the topology.

The current implementation of the MAP method assumes a molecular clock. Thus, rooted trees are used and the node

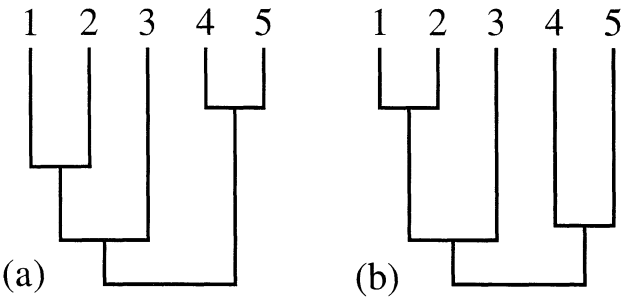

FIG. 2. An example of two labeled histories with the same topology. Labeled history *a* differs from labeled history *b* in that the speciation event producing taxa 4 and 5 occurred after the the speciation event producing taxa 1 and 2 for tree *a*.

times are ordered, with speciation events occurring at times $t_1 > t_2 > \ldots > t_{s-1}$ (Fig. 1). A phylogenetic tree with relative node times ordered is referred to as a labeled history (Edwards 1970). For example, the two trees of Figure 2 are different labeled histories even though their topology is the same. There are a total of $s!(s-1)!/2^{s-1}$ labeled histories for s sequences (as opposed to the $(2s-3)!/2^{s-2}(s-2)!$ distinct rooted topologies). The time of the first bifurcation is set to one ($t_1 = 1$) and parameter estimates are then relative to this time scale. The lengths of the branches are completely determined by the node times and the overall substitution rate m. For example, the length of branch 3 (v_3) from Figure 1 is calculated as $v_3 = m(t_2 - t_3)$.

HYPOTHESIS TESTS

We formulate statistical tests of three distinct hypotheses: (1) that the host and parasite phylogenies are consistent with one common history (i.e., the topologies are in perfect agreement); (2) that hosts and parasites speciated at similar times (i.e., the labeled histories and speciation times agree); and (3) that the rates of nucleotide substitution in hosts and parasites are the same. These hypotheses are illustrated in Figure 3. Figure 3a shows an example in which host and parasite phylogenies are in complete agreement with one another. Note that the node times and branch lengths for the two trees do not agree. Figure 3b shows an example in which the topology and branch points are identical, although the two topologies differ in their overall rate of nucleotide substitution. Figure 3c shows an example in which the topologies and node times agree and the overall rate of nucleotide substitution is the same. Table 1 summarizes the tests.

Tests of Identical Topology

Likelihood-Ratio Test for Identical Topology.—Huelsenbeck and Bull's (1996) test for heterogeneity of trees from different data partitions can be used to test whether host and parasite phylogenies are congruent. The likelihood of each topology under the null hypothesis (that host and parasite topologies are identical) is calculated assuming the constraint that the same topology underlies both host and parasite sequences, but with the other parameters (e.g., $\mathbf{v}$, κ, and α) optimized independently for hosts and parasites. The likelihood of the host and parasite sequences under the alternative

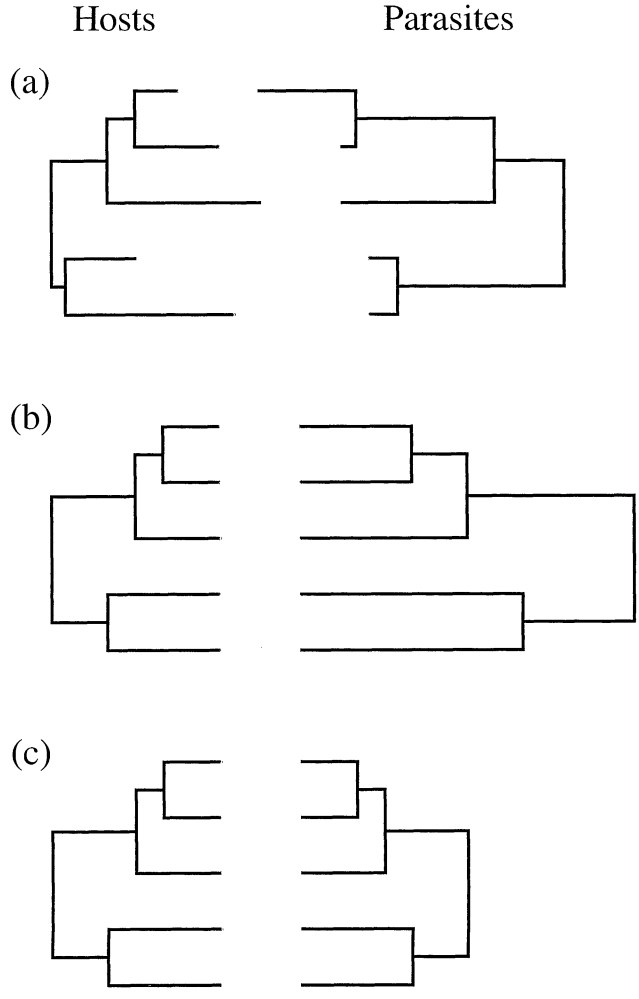

FIG. 3. Diagramatic representations of the three null hypotheses considered in this paper. Under the null hypothesis of identical topology (*a*), the topologies are the same for host and parasite trees although the node times and branch lengths may differ. Under the null hypothesis of temporal cospeciation (*b*), the topologies and speciation times of host and parasite trees are identical. The overall rate of nucleotide substitution, however, may differ for the host and parasite trees so that one tree may be proportionally larger than the other. Under the null hypothesis of identical substitution rate (*c*), the topologies, node times, and overall rate of nucleotide substitution are identical for host and parasite trees.

hypothesis (that host and parasite topologies may not be identical) allows the possibility that different trees underlie the host and parasite sequences.

The likelihood ratio for a test of congruence between host and parasite topologies will always be less than or equal to one because the alternative hypothesis is more general than the null hypothesis. Because the likelihood is maximized over topology, however, the usual statistical properties of nested hypothesis tests may not hold and we use Monte Carlo simulation to determine the null distribution of $-2\log\Lambda$, instead of the χ^2 distribution typically used for such cases.

Conditional-Probability Test for Identical Topology.—The MAP method estimates the (conditional) probability that each tree is correct for the host and parasite nucleotide sequences. The probability that host and parasite topologies are in com-

plete agreement is a sum of the products of the posterior probabilities for all possible identical pairs of host and parasite topologies. This is the probability that strict cospeciation has occurred between hosts and parasites.

Likelihood-Ratio Test of Identical Speciation Times

In the preceding test, only the topologies of the host and parasite trees were considered and the relative branch lengths for the two groups were unconstrained. In this section we consider a test of the null hypothesis that the coevolving host and parasite taxa speciated at the same time given that their topologies are the same.

We describe this test assuming that sequences evolve according to a molecular clock, although the test can be performed without the clock assumption. Under the clock constraint, the branch lengths (**v**) are completely determined by the standardized node times (**t**) and the overall rate of substitution (m). Under the null hypothesis of identical speciation times, the same labeled history (τ) and node times (**t**) underlie the host and parasite sequences. The host and parasite sequences may differ in their pattern of nucleotide substitution, $\boldsymbol{\Theta}$, however, as well as in their overall rate of nucleotide substitution, m. The alternative hypothesis allows different node times for host and parasite sequences. The only difference between the two hypotheses is whether the trees have proportional branch lengths; under the null hypothesis, the host and parasite trees are constrained to have proportional branch lengths (**t**) whereas under the alternative hypothesis $\mathbf{t}_H$ and $\mathbf{t}_P$ may be different (where the subscripts denote host [H] and parasite [P] node times, respectively).

The likelihood ratio is calculated as in the previous test. Since the topology is fixed, $-2\log\Lambda$ is expected to have an asymptotic χ^2 distribution with $s - 2$ degrees of freedom under the null model. The null distribution may also be generated via simulation under the null hypothesis.

Likelihood-Ratio Tests of Identical Substitution Rates

The most restrictive null hypothesis considered in this paper assumes identical topologies, node times, and overall rates of substitution. Not only are host and parasite species strictly coevolving, but their rate of nucleotide substitution is the same. The likelihood under the null hypothesis is calculated with the constraint that the overall rate of nucleotide substitution in host and parasite sequences is the same. Under the alternative hypothesis, overall rates of substitution in host and parasite sequences may differ. The likelihood ratio provides a test of the null hypothesis of equal substitution rates for hosts and parasites. As before, the significance of $-2\log\Lambda$ can be determined using a χ^2 distribution with one degree of freedom.

A likelihood-ratio test of identical substitution rates can also be performed in which a prior distribution of node times and topologies is assumed (Rannala and Yang 1996). For such a test, the likelihood would be calculated by summing over all topologies and integrating over node times. The advantage of this test would be that it takes into account uncertainty in topology.

TABLE 1. Tests of host-parasite cospeciation. Symbols are defined in the text.

Test of Identical Topology for Hosts and Parasites.—Are the topologies for the hosts and parasites consistent with one evolutionary history?

Likelihood-ratio test:

H_0: The topologies are identical for hosts and parasites

$$L_0(\tau, \mathbf{v}_H, \mathbf{v}_P, \Theta_H, \Theta_P|\mathbf{X}, \mathbf{Y}) = \max[L(\tau, \mathbf{v}_H, \Theta_H|\mathbf{X})\cdot L(\tau, \mathbf{v}_P, \Theta_P|\mathbf{Y})]$$

H_1: Different topologies are allowed to underlie the host and parasite sequences

$$L_1(\tau_H, \tau_P, \mathbf{v}_H, \mathbf{v}_P, \Theta_H, \Theta_P|\mathbf{X}, \mathbf{Y}) = \max[L(\tau_H, \mathbf{v}_H, \Theta_H|\mathbf{X})]\cdot\max[L(\tau_P, \mathbf{v}_P, \Theta_P|\mathbf{Y})]$$

Significance: The significance of the likelihood-ratio test statistic $-2\log\Lambda = -2(\log L_0 - \log L_1)$ is determined using Monte Carlo simulation under the null hypothesis.

Conditional-probability test:

The probability that the trees for hosts and parasites are the same (i.e., that H_0 is true) is

$$\Sigma_\tau f(\tau \mid \mathbf{X}, \lambda_H, \mu_H, m_H, \Theta_H)\cdot f(\tau \mid \mathbf{Y}, \lambda_P, \mu_P, m_P, \Theta_P)$$

Test of Temporal Cospeciation among Hosts and Parasites.—Given that the topologies of the host and parasite taxa are the same, did the speciation events among associated taxa occur at similar times?

Likelihood-ratio test:

H_0: The speciation times for host and parasite trees are identical, although the overall rate of nucleotide substituition may be different for the two trees. In other words, the trees have proportional branch lengths.

$$L_0(\tau, \mathbf{t}, m_H, m_P, \Theta_H, \Theta_P|\mathbf{X}, \mathbf{Y}) = \max[L(\tau, \mathbf{t}, m_H, \Theta_H|\mathbf{X})\cdot L(\tau, \mathbf{t}, m_P, \Theta_P|\mathbf{Y})]$$

H_1: The hosts and parasites speciated at potentially different times.

$$L_1(\tau, \mathbf{t}_H, \mathbf{t}_P, m_H, m_P, \Theta_H, \Theta_P|\mathbf{X}, \mathbf{Y}) = \max[L(\tau, \mathbf{t}_H, m_H, \Theta_H|\mathbf{X})\cdot L(\tau, \mathbf{t}_P, m_P, \Theta_P|\mathbf{Y})]$$

Significance: When the molecular clock is assumed in both models, the likelihood-ratio test statistic $-2\log\Lambda = -2(\log L_0 - \log L_1)$ is approximately χ^2 distributed with $s - 2$ degrees of freedom.

Test of Identical Substitution Rates for Hosts and Parasites.—Given that the topologies and node times for host and parasite species agree, is the rate of substitution identical in hosts and parasites?

Likelihood-ratio test:

H_0: The rate of nucleotide substitution in hosts and parasites is identical.

$$L_0(\tau, \mathbf{t}, m, \Theta_H, \Theta_P|\mathbf{X}, \mathbf{Y}) = \max[L(\tau, \mathbf{t}, m, \Theta_H|\mathbf{X})\cdot L(\tau, \mathbf{t}, m, \Theta_P|\mathbf{Y})]$$

H_1: The rates of nucleotide substitution in hosts and parasites may be different.

$$L_1(\tau, \mathbf{t}, m_H, m_P, \Theta_H, \Theta_P|\mathbf{X}, \mathbf{Y}) = \max[L(\tau, \mathbf{t}, m_H, \Theta_H|\mathbf{X})\cdot L(\tau, \mathbf{t}, m_P, \Theta_P|\mathbf{Y})]$$

Significance: The likelihood-ratio test statistic $-2\log\Lambda = -2(\log L_0 - \log L_1)$ is approximately χ^2 distributed with one degree of freedom.

EXAMPLE USING GOPHERS AND LICE

To illustrate the methods presented in this paper, we applied the likelihood-ratio and conditional-probability tests to the gopher and louse data of Hafner et al. (1994). These authors collected cytochrome oxidase I (COI) sequence data for pocket gophers (15 species in the genera *Cratogeomys*, *Geomys*,

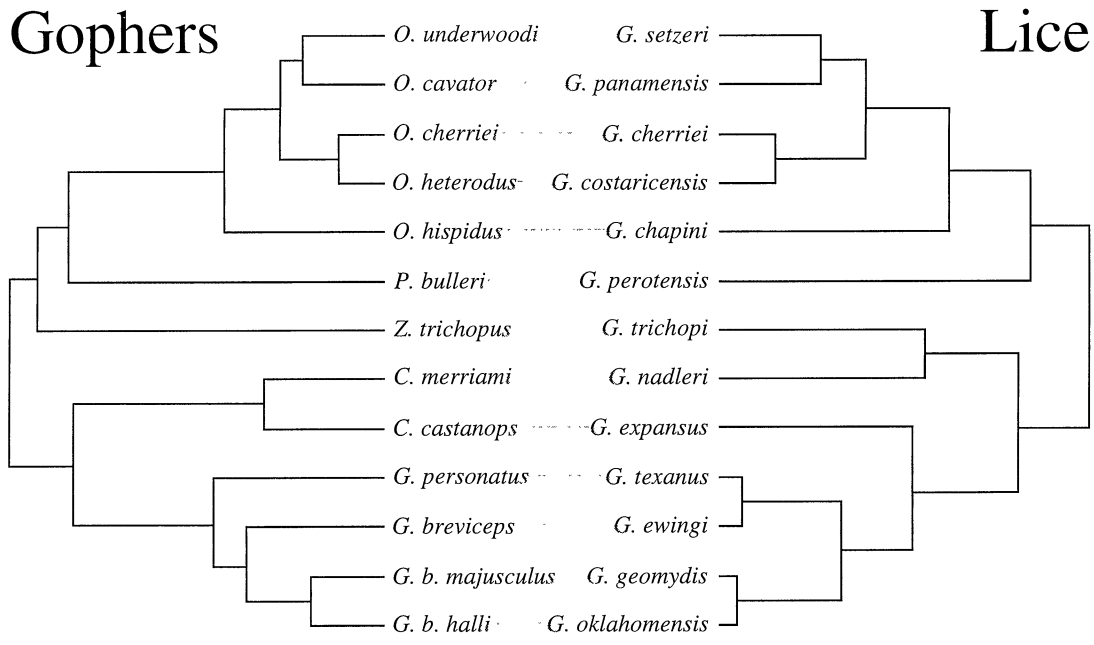

FIG. 4. The phylogenies estimated for the large dataset for the gopher and louse sequences. Hosts are joined to the parasites they harbor by the grey lines. *Geomys bursarius* is abbreviated as "*G. b.*" Maximum likelihood was used, with the F84 + Γ model of DNA substitution and the molecular clock assumed. This model provides the best statistical fit to the data without superflous parameters. The log likelihoods and parameter estimates (κ, transition/transversion rate ratio; α, gamma shape parameter) for the gopher (G) and louse (L) trees are: $\log L_G = -1923.01$, $\log L_L = -2352.55$, $\kappa_G = 4.63$, $\kappa_L = 7.17$, $\alpha_G = 0.15$, $\alpha_L = 0.18$.

Orthogeomys, *Pappogeomys*, *Thomomys*, and *Zygogeomys*) and for their ectoparasitic lice (17 species in the genera *Geomydoecus* and *Thomomydoecus*). There was a one-to-one correspondence between gopher and louse species for 13 species pairs. We consider two subsets of the taxa. The "large" dataset contains the 13 species pairs for which there was a one-to-one correspondence between hosts and parasites. The "small" dataset contains five pairs of associated gopher and louse species for which phylogenies appear to agree (*O. underwoodi/G. setzeri*, *O. cavator/G. panamenis*, *O. cherriei/G. cherriei*, *O. heterodus/G. costaricensis*, and *O. hispidus/G. chapini*).

Analysis of the Large Dataset

The trees estimated for the 13 associated species of host and parasite taxa considered in this section are not identical. Figure 4 shows the estimated phylogenies for the gopher and louse species obtained using maximum likelihood and assuming the F84+Γ model of DNA substitution and the existence of a molecular clock. The alpha test version of the program PAUP* 4.0 was used to estimate phylogenies (Swofford 1996). In addition to parsimony analysis, the new version of PAUP estimates trees using maximum-likelihood and distance methods. In this analysis, we adopted a statistical model-fitting approach to phylogeny estimation (Goldman 1993); the model of DNA substitution was chosen that provided the best fit to the data without introducing superfluous parameters. Models that allow for a transition/transversion rate bias and among-site rate variation (as modeled by a gamma distribution) provided significant improvements in the likelihood for both gopher and louse sequences (Table 2). A molecular clock could not be rejected for gophers or lice when the optimal model of DNA substitution (F84+Γ) was considered (Table 2; also see Page 1990). We therefore used maximum likelihood with a F84+Γ model of DNA substitution and the constraint of a molecular clock to estimate the phylogeny of the gophers and lice. The estimated phylogenies are similar to those published by Hafner et al. (1994). The topological distance between the gopher and louse trees is $d_T = 8$ (Robinson and Foulds 1981) meaning that the trees agree in six of the 10 taxon bipartitions (where a taxon bipartition removes one internal branch from the unrooted tree and divides the taxa into those on each side of the deleted branch). The host and parasite trees are more similar than would be expected by chance under a Markov branching model ($P < 0.01$, using Page's [1988] component test; also see Hafner et al. 1994), suggesting there is some degree of coevolution among gophers and lice.

There are several possible explanations for the observed lack of agreement between host and parasite trees: one possibility is that the same topology underlies both host and parasite sequences and different estimates of phylogeny were obtained due to the limited number of nucleotide sites sampled (379 base pairs for both the gopher and louse species). An alternative hypothesis is that different trees underlie the gopher and louse sequences because of host switching by the lice or that multiple parasite lineages existed in ancestral hosts (Page 1993). We tested the first hypothesis, that host and parasite topologies are in complete agreement and sampling error has produced different trees, by using the likelihood-ratio test for identical topology. We used a program written in the C computer language to maximize the likelihood under the null and alternative models. The program does not estimate the parameters κ and α of the substitution model

TABLE 2. Likelihood-ratio test results for the large dataset. F81 indicates maximum-likelihood estimation under the F84 model of DNA substitution but with $\kappa = 0.0$. Analyses were performed with the constraint of a molecular clock (c) or without the clock constraint (nc).

Data	Model of DNA Substitution	log L_0	log L_1	$-2 \log \Lambda$
Test of equal transition/transversion rate				
Gophers (all positions)	F81 vs. F84 (nc)	−2227.98	−2102.14	251.68**
Lice (all positions)	F81 vs. F84 (nc)	−2776.18	−2637.11	278.14**
Gophers (all positions)	F81 vs. F84 (c)	−2243.26	−2114.91	256.70**
Lice (all positions)	F81 vs. F84 (c)	−2782.23	−2643.62	277.22**
Test of equal rates among sites				
Gophers (all positions)	F84 vs. F84 + Γ (nc)	−2102.14	−1913.33	377.62**
Lice (all positions)	F84 vs. F84 + Γ (nc)	−2637.11	−2345.76	582.70**
Gophers (all positions)	F84 vs. F84 + Γ (c)	−2114.91	−1923.01	383.80**
Lice (all positions)	F84 vs. F84 + Γ (c)	−2643.62	−2352.55	582.14**
Test of molecular clock				
Gophers (all positions)	F81 (c vs. nc)	−2243.26	−2227.98	30.56**
Lice (all positions)	F81 (c vs. nc)	−2782.23	−2776.18	12.10
Gophers (all positions)	F84 (c vs. nc)	−2114.91	−2102.14	25.54*
Lice (all positions	F84 (c vs. nc)	−2643.62	−2637.11	13.02
Gophers (all positions)	F84 + Γ (c vs. nc)	−1923.01	−1913.33	19.36
Lice (all positions)	F84 + Γ (c vs. nc)	−2352.55	−2345.76	13.58

* $P < 0.05$; ** $P < 0.005$.

and these parameters were set to zero and ∞, respectively. The likelihood-ratio test statistic for topological congruence was $-2\log\Lambda = 69.58$. The significance of this value was determined using simulation under the null hypothesis with maximum-likelihood estimates of model parameters. Figure 5 shows the simulated null distribution of $-2\log\Lambda$. The null hypothesis is clearly rejected; the differences in the topologies are greater than would be expected through sampling error. Because of the large number of sequences, the conditional probability test of topological congruence was not performed for these data. Also, because topological congruence was rejected, the additional tests outlined above were not performed as these tests all require that the topologies perfectly agree.

To summarize, gopher and louse topologies agree more than would be expected under the null hypothesis of random agreement for a Markov branching process (Hafner et al. 1994) and this is evidence for cospeciation. The agreement between host and parasite topologies is not perfect, however, and we can reject the hypothesis that the observed topological differences are due to sampling error. The data, therefore, suggest some host switching by the parasitic lice or multiple parasite lineages in ancestral hosts.

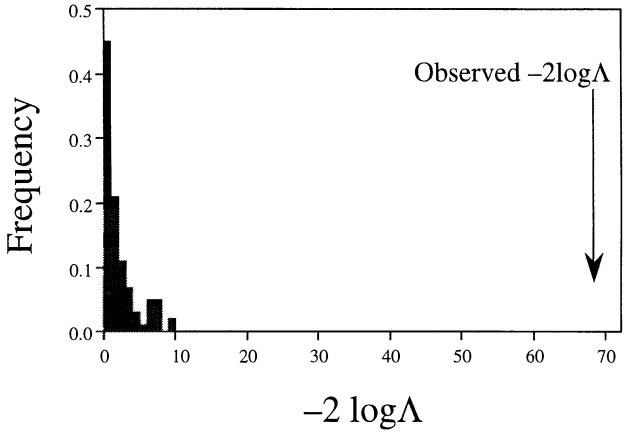

FIG. 5. The null hypothesis of identical topologies for the 13 host and parasite sequences of the large dataset is rejected using the likelihood-ratio test. The null frequency distribution of $-2\log\Lambda$ generated using Monte Carlo simulation is shown. Maximum-likelihood estimates of the parameters were used for the simulation.

Model of Host Switching by Parasites.—How much host switching would need to occur to account for the observed level of difference between phylogenies of gophers and lice? We consider a model-based method for estimating the rate of host switching. A birth-death process is used to model speciation and extinction of host lineages with random host switching by the parasites. During a small interval of time Δt when there are n species, host speciation occurs with probability $n\lambda\Delta t$, host extinction occurs with probability $n\mu\Delta t$, and a host-switching event occurs with probability $n\lambda\Delta t$. The probability that two or more events occur is of order $o(\Delta t)$. A speciation event in a host lineage produces a corresponding speciation in its associated parasite, and the extinction of a host results in extinction of its parasite. When a host-switching event occurs, the affected parasite lineage speciates; one of the two newly formed lineages remains associated with its current host and the other attacks another extant host lineage (with each host having an equal probability of being attacked). The parasite previously associated with the attacked host becomes extinct. This model is simple but provides a reasonable starting point for studying the effects of host switching on patterns of host-parasite coevolution. In many cases these assumptions will be violated and other models may be more appropriate. For example, rates of host switching might depend on geographic distance between hosts, and host switching by one parasite may not cause extinction of a second. We emphasize that such models are very dependent on the biology of the species considered in particular cases (i.e., whether ectoparasites or endoparasites, and whether the parasitic life cycles are direct, or involve addi-

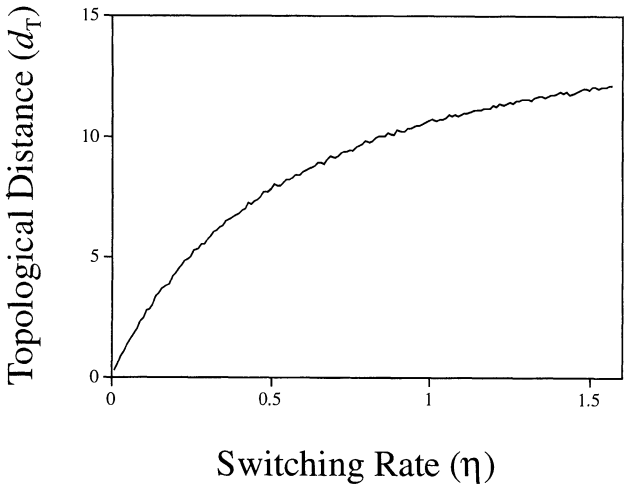

FIG. 6. The expected topological distance (d_T) between the host and parasite phylogenies as a function of the host-switching rate. The graph was generated under a birth-death process of host cladogenesis with parasite cospeciation and host switching. As the host-switching rate increases, the distance between the host and parasite trees increases monotonically. The observed topological distance between the host and parasite trees for the large dataset analysis was $d_T = 8$, implying a host-switching rate of about $\eta = 0.5$ (one-third the gopher speciation rate).

tional intermediate hosts). The basic motivation of using a stochastic model for the process of host switching, however, should be generally useful.

We estimated the speciation and extinction rates of the pocket gophers using the method of Nee et al. (1994). This method uses maximum-likelihood estimates of the labeled history and node times (**t**) under the assumption of a molecular clock. The errors associated with the estimated node times are ignored and these are treated as observations to estimate λ and μ under a birth-death process using maximum likelihood. A tree for the gophers was estimated using maximum likelihood assuming a molecular clock and the F84 + Γ model of DNA substitution (Fig. 4). The node times (**t**) of the tree were scaled between zero and one and estimates of λ and μ obtained using maximum likelihood. The estimates are $\hat{\lambda} = 1.56$ and $\hat{\mu} = 0.00$ ($\log L(\lambda,\mu|t) = -6.09$). These estimates are consistent with a Yule process model of cladogenesis in which only speciation occurs (Yule 1924).

These estimates were used in computer simulations to generate random trees under the birth-death process model using different values of the host switching rate parameter (η). Ten thousand replications were performed for each switching rate and the average distance (using the Robinson and Foulds metric) between simulated host and parasite phylogenies was calculated. The results are shown in Figure 6. When the rate of host switching is zero, the trees are topologically identical; as η increases the topological distance between the host and parasite phylogenies increases monotonically. The observed topological distance between gopher and louse trees for the large dataset analysis ($d_T = 8$) suggests a host-switching rate of about $\eta = 0.5$ under the model. This implies that the rate of host switching is approximately one-third the rate of host speciation.

Analysis of the Small Dataset

The estimated phylogenies of gophers and lice for the small dataset (those five species associations at the top of Fig. 4) are consistent with a common history. The likelihood-ratio test of identical topology failed to reject the null hypothesis ($-2\log\Lambda = 0.0$, $P = 1.0$). The probability that gopher and louse phylogenies are identical was calculated as $P = 0.96$ by the conditional probability test. For both the likelihood-ratio and conditional-probability calculations, the F84 model of DNA substitution was assumed with κ estimated and $\alpha = \infty$. Note that the two probabilities have different meanings, and indicate an interesting difference between tests using the likelihood ratio and the conditional probability. The likelihood-ratio test failed to reject the null hypothesis of topological agreement. The conditional probability test is more powerful in confirming that the null hypothesis is correct.

The posterior probabilities of the host and parasite labeled histories are particularly revealing; most of the probability is associated with the same labeled history for both parasites and hosts (0.983 and 0.976 for the gopher and louse histories, respectively). The estimate of λ (the speciation rate) for the gophers obtained using the MAP method was 1.30 ± 1.63, which is similar to the value obtained for the large dataset using the method of Nee et al. (1994).

Not only are the topologies congruent for the small dataset, but the node times of the gopher and louse phylogenies are consistent with cospeciation. The null hypothesis of identical node times could not be rejected using the likelihood-ratio test of temporal cospeciation (Table 3). The likelihood-ratio test of temporal cospeciation assumes the molecular clock; for the small dataset analysis, the molecular clock assumption was found to hold for both the gopher and louse data (Table 3). The results of likelihood ratio tests of the molecular clock presented here are consistent with the results of Hafner et al. (1994) who also used a likelihood-ratio to test the molecular clock hypothesis.

The Pattern of Nucleotide Substitution in Hosts and Parasites

The tests of identical topology and temporal cospeciation using the small dataset support the idea that, at least for this subset of taxa, gophers and lice are strictly coevolving. This part of the tree should provide opportunities for examining the processes of nucleotide substitution in two distantly related groups (mammals and insects) in a situation where the underlying tree is identical. Several authors (Hafner et al. 1994; Hafner and Page 1995; Page 1996) have previously examined the process of nucleotide substitution in gopher and louse sequences and have concluded that rates of nucleotide substitution were significantly higher (three times higher) in lice than in gophers. They suggested this may be evidence for a generation time effect (Wu and Li 1985) on nucleotide substitution rates in the two groups. Hafner et al. (1994) suggested that the rate of synonymous substitution is about 10 times higher in louse sequences than in gopher sequences. The test employed by Hafner et al. (1994) examined all possible pairs of branch lengths for cospeciating hosts and parasites and determined the significance of differences using Wilcoxon sign-rank tests. Page (1996) applied

TABLE 3. Likelihood-ratio test results for the small dataset.

Data	Model of DNA Substitution	log L_0	log L_1	$-2 \log \Lambda$
Test of molecular clock				
Gophers (all positions)	F84	−934.82	−934.15	1.34
Gophers (all positions)	F84 + Γ[a]	−919.92	−919.57	0.70
Gophers (third codon position)	F84	−438.36	−437.95	0.81
Lice (all positions)	F84	−1233.51	−1232.44	2.14
Lice (all positions)	F84 + Γ[b]	−1186.82	−1187.78	1.92
Lice (third codon position)	F84	−640.57	−638.66	3.82
Test of temporal cospeciation				
Gophers/Lice (all positions)	F84	−2169.24	−2168.33	1.82
Gophers/Lice (all positions)	F84 + Γ	−2108.42	−2107.70	1.44
Gophers/Lice (third codon position)	F84	−1080.23	−1078.93	2.60
Test of identical rates				
Gophers/Lice (all positions)	F84[c]	−2195.45	−2178.27	34.36*
Gophers/Lice (all positions)	F84 + Γ[d]	−2128.46	−2116.13	24.66*
Gophers/Lice (third codon position)	F84[e]	−1120.77	−1098.55	44.44*

* $P < 0.001$.
[a] $\alpha = 0.19$.
[b] $\alpha = 0.16$.
[c] Parasite/host rate = 2.15 ± 0.29.
[d] Parasite/host rate = 3.48 ± 0.93.
[e] Parasite/host rate = 2.82 ± 0.46.

a likelihood-ratio test of the hypothesis that the louse tree has the same branch lengths as the gopher tree; his approach assumes that the branch lengths of the host phylogeny are estimated without error. We apply the methods described in this paper to test for a difference in respective rates of DNA substitution in cospeciating gophers and lice.

The overall rate of nucleotide substitution differs between the gopher and louse sequences. For example, the null hypothesis of identical substitution rates for hosts and parasites can be rejected using a likelihood-ratio test for data of the third position only, or for the combined first, second, and third positions (Table 3). The rate of substitution in the louse sequences was estimated to be 3.02 ± 0.53 times that of the gopher sequences when all positions are considered. This result is similar to the finding of Hafner et al. (1994) and Page (1996) that the overall rate of nucleotide substitution in lice was three times and 2.6 times that of gophers, respectively. Contrary to the findings of Hafner et al. (1994), we found that the rate ratios at the three codon positions are also similar in gophers and lice. This difference might be due to the fact that Hafner et al. (1994) use fixed transition/transversion rate ratios (that is, $\kappa = 4$ for gophers and $\kappa = 10$ for lice, whereas our maximum-likelihood estimates of κ from the data are about 4.4 and 4.1 for the gopher and louse sequences, respectively).

DISCUSSION

Cospeciation has captured the interest of systematists because the evidence of cospeciation may often be detected by reconstructing the phylogenies of the coevolved species. Previous workers have devised tests to detect nonrandom topological similarity and/or identical speciation times (see Brooks 1981, 1986; Page 1988, 1990, 1991; Hafner and Nadler 1990; Lapointe and Legendre 1990). The tests presented in this paper are based on a different null hypothesis from previous tests. The null hypothesis for the test of identical topology, for example, assumes that the hosts and parasites share a common history and therefore allows the rejection of strict cospeciation. This is quite different from tests of random similarity between host and parasite phylogenies based on a Markov branching process (Page 1988). A test of complete topological agreement between host and parasite phylogenies determines whether complete cospeciation has occurred, whereas other tests investigate less strict hypotheses about the degree of cospeciation. One advantage of the likelihood ratio or conditional probability is that they can be easily modified to accommodate different null hypotheses. For example, one might want to test the null hypothesis of one or fewer host-switching events. Such a modified test would require (1) a model relating host switching to conflicts in topology; and (2) a method that maximizes the likelihood of the sequence data under the model.

Alternative tests of the null hypothesis of identical topology for host and parasite data that do not use likelihood or Bayesian methods of phylogeny estimation are also possible. For example, Farris et al. (1994) proposed a parsimony-based test that examines the incongruence of trees estimated from partitioned data. This test was designed for the case in which data from the same species are partitioned into subsets (e.g., the data are partitioned by gene, codon position, functional domain, etc.). However, the Farris et al. (1994) test could just as easily be applied to the question of host-parasite cospeciation. The null hypothesis under the Farris et al. (1994) test is not clear, however, and because the assumptions are not explicitly stated, it is easy to confound the effects of different topologies with those of different substitution rates or different character transition probabilities in the data partitions.

A problematic aspect of the tests proposed here (and most others as well) is the necessity of accommodating situations

in which there is not a one-to-one correspondence between associated species. For the gopher/louse data, for example, two of the 15 gopher species were infested with two species of lice. The solution used in this paper was to eliminate taxa from the analysis that had more than one parasite. Another possible approach is to replicate host sequences, assigning a replicated host species to each of the multiple parasite species. The expectation is that the coevolved parasites would group with replicated hosts. Yet another possibility is to initially include all associated species in the likelihood analysis and to remove species from multiple associations in a stepwise fashion, maximizing the likelihood at each step with the assumption that a pattern of strict cospeciation underlies the collection of taxa as a whole with excess species representing subsequent invaders.

In standard statistical problems, the likelihood or the posterior probability provide the information necessary to construct statistical tests, and these tests often have well-known properties. For example, likelihood-ratio tests are known to be optimal for testing a pair of simple hypotheses, and likelihood-ratio tests of compound hypotheses often perform well in cases where no optimal test exists (Rice 1995). The phylogeny estimation problem is atypical, however, because topology is not a regular statistical parameter (Goldman 1993; Yang et al. 1995). Some standard results in statistics are invalid for the phylogeny estimation problem. It is known, for example, that for likelihood-ratio tests that involve maximization over topologies, $-2\log\Lambda$ is not χ^2 distributed even if the hypotheses are nested (Goldman 1993; Yang et al. 1995). Future studies of the properties of tests of coevolution should clarify the power of the competing tests; simulation studies should be particularly useful in this context.

Program Availability

Computer programs written in C are available to perform the likelihood-ratio tests of identical node times and substitution rates and to calculate posterior probabilities of trees (PAML). PAML is available from the web page at http://mw511.biol.berkeley.edu/homepage.html.

Acknowledgments

We thank M. Hafner for kindly providing us with the aligned COI sequences for both gophers and lice. This paper benefited from discussions with D. Parks and J. Bull, and from constructive comments from M. Hafner and R. Page. This research was supported by a Miller Postdoctoral Fellowship awarded to JH and a Natural Sciences and Engineering Research Council (NSERC) of Canada postdoctoral fellowship awarded to BR. Support for ZY was provided by a grant from the National Institutes of Health (GM40282) to M. Slatkin.

Literature Cited

Baverstock, P. R., M. Adams, and I. Beveridge. 1985. Biochemical differentiation in bile duct cestodes and their marsupial hosts. Mol. Biol. Evol. 2:321–337.

Brooks, D. R. 1979. Testing the context and extent of host-parasite coevolution. Syst. Zool. 28:299–307.

———. 1981. Hennig's parasitological method: A proposed solution. Syst. Zool. 30:229–249.

———. 1986. Analysis of host parasite coevolution. Int. J. Parasitol. 17:291–300.

Cox, D. R., and D. V. Hinkley. 1974. Theoretical statistics. Cambridge Univ. Press, Cambridge.

Edwards, A. W. F. 1970. Estimation of branching points of a branching diffusion process. J. R. Stat. Soc. B Biol. Sci. 32: 155–174.

Ehrlich, P., and P. H. Raven. 1964. Butterflies and plants: A study in coevolution. Evolution 18:586–608.

Farris, J. S., M. Källersjö, A. G. Kluge, and C. Bult. 1994. Testing for significance of incongruence. Cladistics 10:315–319.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Hafner, M. S., and S. A. Nadler. 1988. Phylogenetic trees support the coevolution of parasites and their hosts. Nature 332: 258–259.

———. 1990. Cospeciation in host-parasite assemblages: Comparative analysis of rates of evolution and timing of cospeciation events. Syst. Zool. 39:192–204.

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.

Huelsenbeck, J. P., and J. J. Bull. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. Syst. Biol. 45:92–98.

Lapointe, F.-J., and Legendre. 1990. A statistical framework to test the consensus of two nested classifications. Syst. Zool. 39: 1–13.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994. Extinction rates can be estimated from molecular phylogenies. Phil. Trans. R. Soc. Lond. B 344:77–82.

Page, R. D. M. 1988. Quantitative cladistic biogeography: Constructing and comparing area cladograms. Syst. Zool. 37:254–270.

———. 1990. Temporal congruence and cladistic analysis of biogeography and cospeciation. Syst. Zool. 39:205–226.

———. 1991. Clocks, clades, and cospeciation: Comparing rates of evolution and timing of cospeciation events in host-parasite assemblages. Syst. Zool. 40:188–198.

———. 1993. Genes, organisms, and areas: The problem of multiple lineages. Syst. Zool. 42:77–84.

———. 1996. Temporal congruence revisited: Comparison of mitochondrial DNA sequence divergence in cospeciating pocket gophers and thier chewing lice. Syst. Biol. 46:151–167.

Rannala, B. 1992. Comparative evolutionary genetics of trematode parasites (Plagiochiidae) and their anuran hosts. Can. J. Zool. 70:993–1000.

Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Rice, J. 1995. Mathematical statistics and data analysis. Duxbury Press, Belmont, CA.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Swofford, D. L. 1996. PAUP*: phylogenetic analysis using parsimony (*and other methods). Vers. 4.0. Sinauer, Sunderland, MA.

Szidat, L. 1956. Der marine Charakter der Parasitenfauna der Susswasserfisches des Stromssystems des Rio de la Plata und ihre Deutung als Reliktfauna des Tertiaren Tethys-Meere. Proc. 14 Int. Con. Zool. 1953:128–138.

Wu, C.-I., and W.-H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Nat. Acad. Sci. 82:1741–1745.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.

———. In press. Maximum likelihood models for combined analysis of multiple sequence data. J. Mol. Evol.

YANG, Z., N. GOLDMAN, AND A. E. FRIDAY. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Syst. Biol. 44:384–399.

YULE, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. Phil. Trans. R. Soc. Lond. A 213:21–87.

Corresponding Editor: D. Fairbairn