# Phylogenetic Methods Come of Age: Testing Hypotheses in an Evolutionary Context

John P. Huelsenbeck* and Bruce Rannala

The use of molecular phylogenies to examine evolutionary questions has become commonplace with the automation of DNA sequencing and the availability of efficient computer programs to perform phylogenetic analyses. The application of computer simulation and likelihood ratio tests to evolutionary hypotheses represents a recent methodological development in this field. Likelihood ratio tests have enabled biologists to address many questions in evolutionary biology that have been difficult to resolve in the past, such as whether host-parasite systems are cospeciating and whether models of DNA substitution adequately explain observed sequences.

Evolutionary biology is founded on the concept that organisms share a common origin and have subsequently diverged through time. Phylogenies represent our attempts to reconstruct this evolutionary history, and there is probably more interest in phylogenetic reconstruction today than at any time in the past. For years phylogenetics played a relatively minor role in evolutionary biology, and it is only in the past decade that the importance of phylogeny in most branches of biology has been fully recognized (*1, 2*). Today it is not uncommon to see phylogenies applied in fields far removed from evolutionary biology. For example, they have found a practical use in tracing routes of infectious disease transmission and in identifying the relationship of pathogens, such as the New Mexico hantavirus (*3*).

With the realization that phylogeny can provide answers to many questions of interest in evolutionary biology, there has been an explosion in the number of statistical tests that take phylogeny into account. In part, this is because an essentially infinite number of possible tests can be applied to any biological question. A hypothesis test involves calculating a test statistic from the data and then determining the probability of the observed statistic if the hypothesis were true; the probability is obtained from the null distribution of the test statistic (that is, the distribution if the hypothesis is true). For hypothesis tests involving phylogeny, the null distribution is usually generated by either permuting data matrices or resampling from the original data. However, the statistical properties of many tests based on such procedures are known to be poor, and although permutation of data matrices is a common procedure, the null hypothesis for many such tests is often not well defined (*4*). Similarly, although nonparametric bootstrapping is widely used to evaluate the support of the data for a particular phylogeny, the statistical interpretation of bootstrap values remains problematic (*5*).

The past 5 years have seen remarkable advances in the use of parametric statistical tests of questions involving phylogeny. In particular, increased computing speed, more realistic models of DNA substitution, and improved computer programs have led to practical statistical tests using likelihood ratios and Monte Carlo simulation procedures. Although statistical tests can be constructed in many different ways (*1, 6*), we concentrate in this review on likelihood ratio tests (LRTs) for several reasons. First, LRTs have the same status in hypothesis testing as does maximum likelihood in parameter estimation. That is, just as maximum likelihood estimates (MLEs) are known to have desirable statistical properties such as consistency, LRTs are known to outperform other hypothesis tests under many conditions. For example, LRTs are known to be optimal (uniformly most powerful) when comparing simple hypotheses, and LRTs often perform well for cases in which no optimal test is known (*7*). Second, many applications of LRTs do not assume that the phylogeny is known. This is an advance over tests that assume that the phylogeny is known without error (*1*) because all existing methods of phylogeny reconstruction are subject to both systematic and random errors. In many cases, the error in phylogeny estimation can be large (*8*). Third, LRTs provide a unified framework for testing hypotheses.

## Maximum Likelihood and Hypothesis Testing

Maximum likelihood estimation of phylogenetic trees was first introduced by Edwards and Cavalli-Sforza in the early 1960s (*9*). Felsenstein (*10*) implemented the method for DNA sequence data, and most recent advances have focused on the analysis of DNA sequences. Stated simply, the MLE of phylogeny is the tree for which the observed data are most probable. For the present purposes, the data are aligned DNA sequences for $s$ species. The first step in a likelihood analysis is to calculate the probability of the observed sequences; this probability depends on an explicit mathematical model of evolution (*11*). The model consists of two parts: (i) a phylogenetic tree with branch lengths defined in terms of the expected number of substitutions per site, and (ii) a model of the process of DNA substitution (that is, specifying the probability of the occurrence of a nucleotide substitution at a particular site over the length of a branch). For many studies the phylogenetic tree is the only parameter of interest, but in the course of finding the maximum likelihood tree, other parameters are estimated that may also be of importance (such as the transition rate–transversion rate bias).

Much attention has focused on the accuracy of the phylogenetic trees reconstructed by maximum likelihood. Simulation studies suggest that maximum likelihood is typically more accurate (that is, more likely to predict the actual evolutionary tree) and robust (that is, less sensitive to incorrect models and assumptions) than other methods of phylogenetic inference (*12, 13*). Moreover, likelihood provides a natural means of hypothesis testing (*14*). The LRT statistic for comparing two hypotheses ($\Lambda$) is defined as

$$\Lambda = \frac{\max[L(\text{null hypothesis}\,|\,\text{data})]}{\max[L(\text{alternative hypothesis}\,|\,\text{data})]} \quad (1)$$

The likelihood $L$ is maximized under both the null and alternative hypotheses. The likelihood ratio provides a measure of the support of the data for one hypothesis versus another. If $\Lambda > 1$, the data are more probable under the null hypothesis, and this is favored; the alternative hypothesis is favored if $\Lambda < 1$. When nested hypotheses are examined (that is, the null hypothesis is a special case of the more general, alternative hypothesis), $\Lambda$ will always be $<1$ and $-2 \log$

The authors are in the Department of Integrative Biology, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed. E-mail: johnh@mws4.biol.berkeley.edu

$\Lambda$ is approximately $\chi^2$ distributed under the null hypothesis with $q$ degrees of freedom, where $q$ is the difference in the number of free parameters between the null and alternative hypotheses. Alternatively, the probability of observing a given $\Lambda$ if the null hypothesis were correct (the significance level) can be calculated by using Monte Carlo simulations, as explained below (*15*).

Although LRTs have a long history in statistics, they have had only a limited application in phylogenetics, with the first application of an LRT (a test of the molecular clock) proposed in 1981 (*10*). Why has it taken so long for LRTs to be applied in phylogenetic analysis? One problem concerns the use of topology as a model parameter. It is known that many of the standard results for LRTs do not apply to phylogenetic trees (*16*). For example, in considering nested phylogenetic hypotheses, the usual $\chi^2$ approximation to the distribution of the test statistic often cannot be used to determine the significance of the LRT statistic (*16*). This problem can be avoided, however, by generating null distributions using computer simulation (*16, 17*). In this procedure, known as parametric bootstrapping or Monte Carlo simulation, the null distribution of the test statistic is calculated by simulating many data sets (Fig. 1). Monte Carlo simulation has been widely used in statistics since the early 1960s (*15*). Model parameters for the simulations are estimated from the original data under the null hypothesis. The likelihood ratio is calculated for each simulated data set, and the proportion of the replicates in which the likelihood ratio calculated using the original data is exceeded for the simulated data represents the significance level of the test.

Table 1 lists several hypotheses involving phylogeny for which LRTs are available. LRTs have been applied to problems such as the relative fit of models of DNA substitution to sequence data and the evaluation of evidence for the monophyly of a taxonomic group. For many of the questions posed in

**Table 1.** Biological questions involving phylogeny that have been addressed using LRTs.

| Question | Assumptions | Results |
|---|---|---|
| Are DNA substitution rates constant among lineages [that is, does a molecular clock exist (*10*)]? | $H_0$: Assume that DNA substitution rates are equal among lineages.<br>$H_1$: Allow substitution rates to vary among lineages. | A molecular clock is most often rejected, suggesting that there is rate variation among lineages. |
| Is a DNA substitution model adequate to explain the data (*16*)? | $H_0$: Assume a particular model of DNA substitution.<br>$H_1$: Assume a multinomial distribution for the frequencies of site patterns. | Current models of DNA substitution fit the observed data poorly. Sequences from pseudogenes show the best fit. |
| Are DNA substitution rates biased for different nucleotides (*16*)? | $H_0$: Assume that substitution rates are equal among nucleotides (for example, the transition rate equals the transversion rate).<br>$H_1$: Allow transition rate–transversion rate bias. | The addition of unequal rate parameters to the substitution matrix usually provides an improved fit of the model. |
| Are DNA substitution rates constant among sites (*27*)? | $H_0$: Assume equal rates among sites.<br>$H_1$: Allow among-site rate heterogeneity. | The addition of parameters allowing among-site rate variation typically provides a significant improvement to the fit of the model. |
| Are DNA substitution rates constant among genomic regions [that is, in different genes or different codon positions (*21*)]? | $H_0$: Assume that substitution rates are the same in all data partitions (regions).<br>$H_1$: Assume an independent substitution rate for each partition (region). | Rates vary significantly among genomic regions (for example, at different codon positions). |
| Is the DNA substitution process identical among lineages (*22*)? | $H_0$: Assume a homogeneous substitution process among lineages.<br>$H_1$: Allow parameters of the substitution model to vary among lineages. | Base frequencies and the transition rate–transversion rate bias varied significantly among four of the major lineages that gave rise to present-day life forms (*22*). |
| Are the substitutions in stem regions of ribosomal DNA sequences correlated (*34*)? | $H_0$: Assume that substitution is independent among sites.<br>$H_1$: Allow correlated changes in nucleotide duplets in stem regions. | A model that allows for correlated substitutions at pair-bonded stem sites of ribosomal DNA sequences provides an improved fit of the model (*34*). |
| Is the DNA substitution process identical among genomic regions (*21*)? | $H_0$: Assume that the substitution parameters are the same among genomic regions.<br>$H_1$: Allow substitution parameters to vary among genomic regions. | Base frequencies and transition rate–transversion rate bias significantly varied in first, second, third, and transfer RNA partitions of mitochondrial data (*21*). |
| Is a prespecified taxonomic group monophyletic (*35*)? | $H_0$: Assume that a group is monophyletic.<br>$H_1$: Relax the constraint of monophyly. | Analysis of partial HIV sequences from the patients of a dentist supported the idea of multiple sources of infection for one of the patients (*39*). |
| Are phylogenies estimated from different data congruent (*31*)? | $H_0$: Assume that the same phylogeny underlies all data partitions.<br>$H_1$: Allow different phylogenies to underlie different data partitions. | This test has not been widely applied. |
| Are the phylogenies for hosts and parasites consistent with a common history (*25*)? | $H_0$: Assume an identical phylogeny for associated hosts and parasites.<br>$H_1$: Allow different phylogenies for hosts and parasites. | For 13 species of gophers and their associated lice, the phylogenies appear different; for a subset of these species, the phylogeny of hosts and parasites appears identical (*25*). |
| Are the speciation times for hosts and parasites the same (*25*)? | $H_0$: Assume that hosts and associated parasites speciated at the same time.<br>$H_1$: Allow speciation times to vary independently in hosts and parasites. | For five species of cospeciating gophers and lice, the speciation times appear to be identical (*25*). |

Table 1, alternative tests are available, some of which are claimed to be nonparametric. However, all statistical tests involving phylogeny require assumptions about the evolutionary process, even though an explicit model may not be used. Assumptions about the process of evolution are required, for example, when estimating a phylogenetic tree. One of the advantages of LRTs is that model assumptions can themselves be tested and potentially improved.

## Tests of Models of DNA Substitution

All phylogenetic methods make assumptions, whether explicit or implicit, about the process of DNA substitution. Systematists are in an awkward situation in that they know the assumptions of a phylogenetic method are imperfect. Yet they also know that the match between the process of nucleotide substitution generating the sequence variation and the substitution model assumed may be critical. The realism of substitution models is important because methods for inferring phylogeny may be less accurate, or may be inconsistent (that is, converge to an incorrect tree with increased amounts of data), in situations where the model is incorrect (*8, 13, 18*). Evolutionary biologists also have an intrinsic interest in accurately modeling the processes that produce variation in DNA sequences and thereby improving our understanding of molecular evolution. Molecular systematists interested in phylogenetic inference have long been troubled by the question of how to choose the optimal substitution model for a particular data set. Maximum likelihood provides a rational method for choosing substitution models for phylogenetic analysis through the use of LRTs.

Current models implemented in phylogenetic inference using maximum likelihood (and several other methods as well) assume that DNA substitutions follow a Poisson process. The most general model allows each type of nucleotide substitution to have an independent rate parameter (there are 12 rate parameters in total) (*19*). Also, rate heterogeneity among sites can be accommodated by assuming that rates are distributed among different sites according to some probability distribution (usually a gamma, Bernoulli, or log-normal distribution), or by assigning sites to different rate classes (for example, first, second, and third codon positions) and then estimating the substitution rate for each class (*20*). The models implemented in likelihood have also been modified to allow parameters to be estimated separately for different data partitions or for different branches of the phylogenetic tree (*21, 22*). In short, the substitution models used in a phylogenetic analysis can be made arbitrarily complex by the addition of parameters, each of which can be estimated using likelihood methods.

One approach to the choice of models in phylogenetic analysis is to use a very complicated (parameter-rich) model for which a large number of free parameters will result in a high likelihood. However, this approach has several disadvantages. First, because a large number of parameters must be estimated for complicated models, the analysis becomes computationally difficult. Second, the error associated with each parameter estimate is higher for more complicated models than for simple ones. This decrease in accuracy appears to apply to all parameters of the phylogenetic model, including the topology; in certain cases, the accuracy of the estimated phylogeny may be improved by using a simpler model (although this is not universal) (*12, 13*). Finally, an overly complicated model may not be needed to account for the observed data. Occam's razor provides a principle for choosing among hypotheses that explain a set of observations equally well; the simpler (most parsimonious) hypothesis is preferred. Although a complicated model may make the observed data more probable, it will not necessarily provide a significant improvement in the likelihood over a model with fewer parameters.

How can the model be chosen that best fits the data without introducing superfluous parameters? One approach is to compare the likelihoods of different models using an LRT (*10, 16, 23*). The significance

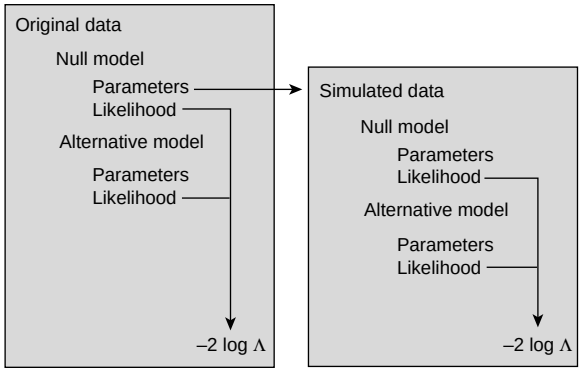


**Fig. 1.** A diagram illustrating the application of parametric bootstrapping to determine significance levels. Parameters from the original data are estimated using maximum likelihood. The MLEs of parameters under the null model are used to construct many simulated data sets of the same size as the original. For each simulated data set, the LRT statistic ($-2 \log \Lambda$) is calculated and compared with the value obtained for the original data.

**Table 2.** LRT results for the gopher-louse COI data set. LRTs were performed for three hypotheses of DNA substitution. The null hypothesis for the test of "equal transition and transversion rates" constrains the transition rate to be equal to the transversion rate. The null hypothesis for the test of "equal rates among sites" is that all sites have an equal rate of substitution, whereas the alternative hypothesis allows rates to be gamma-distributed random variables. The null hypothesis for the test of the molecular clock assumes that the rates among lineages are equal. LRTs reject the null hypotheses of an equal transition rate–transversion rate bias and equal rates among sites but do not reject the molecular clock null hypothesis. F81 indicates maximum likelihood estimation under the F84 model of DNA substitution, but with $\kappa = 0.0$ (*40*). Analyses were performed with the constraint of a molecular clock (c) or without the clock constraint (nc). Single and double asterisks indicate significance at $P < 0.05$ and $P < 0.005$, respectively.

| 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 |

of the LRT statistic ($\Lambda$) can be approximated using simulation or, if the models are nested, by comparing $-2 \log \Lambda$ to a $\chi^2$ distribution, with $q$ degrees of freedom, where $q$ is the difference in the number of free parameters between the null and alternative models of DNA substitution.

For illustrative purposes, we applied this procedure to mitochondrial cytochrome oxidase I (COI) DNA sequences gathered by Hafner *et al.* (*24*) for 13 species of gophers and their associated lice (Table 2). First, we examined the molecular clock hypothesis (*10*). This hypothesis is satisfied if DNA substitutions follow a Poisson process and the mean rate of substitution has remained constant in different lineages. The log likelihood calculated under the clock hypothesis is $\log L = -2243.26$ for the gophers and $\log L = -2782.23$ for the lice when a simple model of DNA substitution is used. A more general model assumes that each branch of the phylogenetic tree has a unique unconstrained rate of substitution. This introduces $s - 2$ additional parameters; the likelihood for this latter model is therefore higher than that under the molecular clock hypothesis ($\log L = -2227.98$ for the gophers and $\log L = -2776.18$ for the lice). Because the models are nested (that is, equal rates among lineages are a special case of the unrestricted model) and the phylogenetic tree is held constant, the statistic $-2 \log \Lambda$ can be compared with a $\chi^2$ distribution with $s - 2$ degrees of freedom to determine the significance of the test (*10*). In this case, the molecular clock hypothesis cannot be rejected for either the gophers or the lice. The same LRT procedure applied to the models of DNA substitution shows that the best-fitting model for the gophers and the lice allows for different rates for transitions and transversions, unequal base frequencies, and among-site rate heterogeneity (*25*).

The ability to choose among models in performing a phylogenetic analysis is one of the great strengths of a likelihood approach. For many widely used phylogenetic methods, there are no generally accepted criteria for choosing among possible evolutionary models [but see (*26*)]. For example, the maximum parsimony method allows many types of data to be analyzed under a large class of substitution models or "weighting schemes," but few criteria exist for choosing among weighting schemes. Methods for choosing models are important because different models may lead to different conclusions about phylogeny. Much of the arbitrary nature of model choice is eliminated by using a likelihood framework; when different substitution models provide different estimates of phylogeny, the tree associated with the best-fitting model is preferred.

The study of substitution models using LRTs has also provided molecular evolutionists with insights about how the process of DNA substitution operates. Application of LRTs indicates that some of the parameters of models of DNA substitution, which reflect the biology, are very important. For example, accounting for among-site rate heterogeneity almost always provides an improved fit of the model to the data [there is not as significant an improvement for pseudogenes, for which selection has been relaxed (*27*)]. The improvement in the likelihood obtained by adding among-site rate heterogeneity is usually so great that formal consideration of the significance level is unnecessary. However, LRTs also allow tests of much more subtle hypotheses, such as the way in which the process of substitution differs across the genome (*21*).

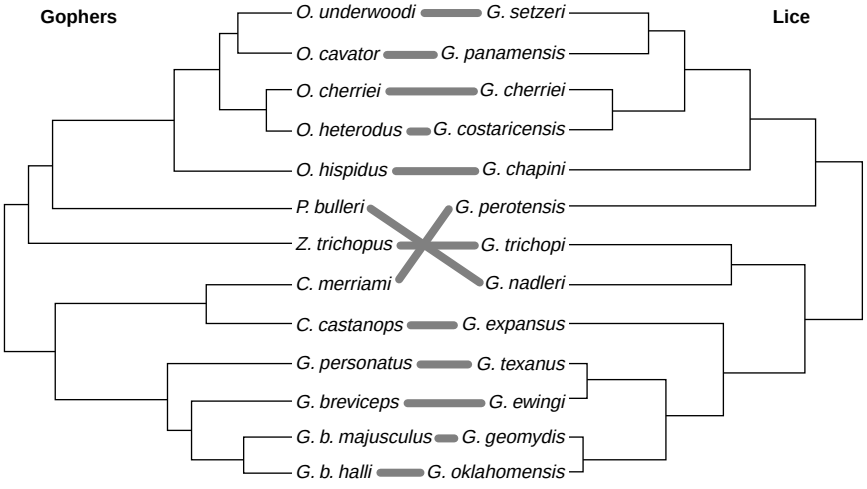


**Fig. 2.** The MLEs of phylogeny for 13 gopher and louse species for COI sequence data (13 pocket gopher species in the genera *Cratogeomys*, *Geomys*, *Orthogeomys*, *Pappogeomys*, and *Zygogeomys* and 13 louse species in the genera *Geomydoecus*; *Geomys bursarius* is abbreviated as *G. b.*). Maximum likelihood trees were estimated using the program PAUP*, version 4.0 (*37*). The substitution model assumed in the analysis allows unequal transition and transversion rates, unequal base frequencies, and among-site rate heterogeneity (*38*). The log likelihoods for the gopher and louse trees are $\log L = -1923.01$ and $\log L = -2352.55$, respectively.

## Tests for Phylogenetic Association

One of the most innovative and useful applications of phylogenies involves the comparison of topologies estimated for different partitions of a data set (for example, different genes) for different species. If the partitioned data share a common evolutionary history, then the topologies estimated from each should be congruent. A comparison of topologies from different data partitions has been used to identify horizontal gene transfer in bacteria and fungi (*28*); horizontal gene transfer may be suspected if the tree estimated using one gene is different from the tree estimated using another gene for the same set of species. Similarly, comparison of tree topologies has been used to examine the rate of reassortment of the RNA segments in the hantavirus (*29*). The hantavirus has three negative sense RNA segments; when more than one virus infects a cell, the opportunity exists for reassortment of the infecting viral segments among the progeny. If genetic reassortment plays an important evolutionary role in the hantavirus, then the trees estimated for the same set of viruses from different segments should be [and are (*29*)] different. Finally, a comparison of the phylogenies for hosts and parasites is a critical step in determining whether they have cospeciated. Cospeciation of hosts and associated parasites is invoked if the branching patterns and speciation times of the host and parasite trees agree (*30*).

Although many important questions can be addressed in the areas of evolutionary biology and epidemiology by comparing phylogenetic trees for different species or different genes, until recently there have been few statistical criteria for deciding whether the trees are in agreement. A likelihood approach uses a LRT of the hypothesis that trees estimated for different data partitions, or different species, are congruent [that is, the phylogenetic history is the same (*31*)]. The null hypothesis for the LRT of congruence is that the same topology underlies different data partitions; the likelihood is maximized under this constraint, but other parameters of the evolutionary model (such as the branch lengths or the transition rate–transversion rate ra-

tio) are estimated independently for each data partition. The likelihood under the alternative hypothesis relaxes the constraint that the same topology underlies all data partitions, although all other aspects of the model are the same.

The LRT of congruence has been successfully used to explore questions of host-parasite cospeciation (*25*). In closely associated host-parasite systems, an allopatric speciation event in a host lineage might be expected to isolate parasite populations associated with each incipient host species, thereby producing a simultaneous allopatric speciation event among parasites. A history of cospeciation in host and parasite lineages should then be reflected by congruent phylogenies for hosts and their associated parasites. What does application of the LRT of congruence indicate about cospeciation in the gopher-louse system? The LRT statistic for the null hypothesis (that the phylogenies for gophers and lice are congruent) is much smaller than would be expected if the null hypothesis were true (*32*). Hence, although the trees for the gophers and lice are similar (*24, 25, 33*), the gophers and lice did not strictly cospeciate; host-switching by the lice, persistence of multiple ancestral louse lineages, or both must be invoked to explain the differences between the phylogenetic trees.

Are there any portions of the gopher-louse tree that are congruent and suggest cospeciation? Analysis of a subset of the associated gopher and louse species (the top five gopher and louse species of Fig. 2) suggests that these gopher and louse species have cospeciated. A more refined LRT suggests that the speciation times of the associated gopher and louse species are also identical. The null hypothesis for a LRT of "temporal cospeciation" assumes that the tree and the relative branch lengths for host and parasite phylogenies are the same but that the overall rate of substitution for the two trees may differ (*25*). The alternative hypothesis relaxes the constraint that the branch lengths for the host and parasite trees are proportional. The null hypothesis that the branching times are identical cannot be rejected, which is consistent with a model of cospeciation for five of the associated gopher and louse species. Because these species appear to have cospeciated, we can also examine whether the substitution rate differs between gophers and lice (*24, 25, 33*). An LRT of the null hypothesis that the substitution rates are identical in hosts and parasites reveals that the substitution rate is much higher in lice than in gophers [3.01 ± 0.53 times the rate for gophers (*25*)]. This rate difference may have several biological explanations, including a higher mutation rate in lice or a shorter generation time (*24*).

## Prospects for Likelihood Ratio Tests in Phylogenetics

The field of phylogenetics has seen remarkable advances in the past 40 years; the principal aim has progressed from reconstructing phylogenies, with little concern for sources of error, to evaluating the reliability of trees and (more recently) addressing biological questions using phylogenies. Maximum likelihood and LRTs have played an important role in the development of phylogenetics and should continue to provide a source for advances. In many ways, testing evolutionary hypotheses that are dependent on phylogeny presents an unusual and difficult statistical problem to the evolutionary biologist. However, it appears that standard statistical approaches may be applied successfully. We have shown that LRTs can be used to study a wide range of biological questions, such as the fit of a substitution model to sequence data and the agreement of phylogenies estimated from different data sets. However, the application of LRTs in phylogenetics is a relatively recent phenomenon, and the range of questions that can be addressed by LRTs is currently limited (Table 1). For example, several questions of general interest in biology, such as whether two or more characters are correlated (*1*), can be addressed using LRTs only in restricted circumstances (*34*). Moreover, questions concerning morphological evolution are difficult to address using LRTs because realistic models of morphological evolution are generally lacking.

Although LRTs have proven useful for studying a variety of biological hypotheses, several unresolved questions remain concerning the general utility of the approach. Few studies have examined the power of LRTs for testing particular phylogenetic hypotheses, or whether such tests are biased (*16, 35*). Another problem involves the computational expense of the hypothesis testing procedure; the likelihood is repeatedly maximized for many simulated data sets, and this can quickly stress the computer resources of most research laboratories. A potential solution to this problem is to perform a small number of replicates and then fit a probability distribution, such as a $\chi^2$ or gamma, to the simulated likelihoods. Also, simple LRTs may not be appropriate in all situations. Methods of sequential analysis are needed when a hypothesis is originally tested using one data set and later reexamined using additional data (*36*).

Explicit model-based methods are a recent innovation in phylogenetics. One advantage of these approaches is that the exact hypothesis being tested is clear if the test is properly formulated. These methods also offer the possibility that evolutionary models may be gradually improved as new biological processes are discovered and incorporated into the models used for phylogenetic analysis. Statistical approaches to phylogenetic inference have led to many improvements in our understanding of the process of DNA substitution over the past decade, allowing a much broader range of biological questions to be examined in a rigorous way.

### REFERENCES AND NOTES


1. P. H. Harvey and M. D. Pagel, *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
2. J. Felsenstein, *Am. Nat.* **125**, 1 (1985).
3. C.-Y. Ou *et al.*, *Science* **256**, 1165 (1992); S. T. Nichol *et al.*, *ibid.* **262**, 914 (1993); B. Hjelle *et al.*, *J. Virol.* **68**, 592 (1994).
4. D. L. Swofford, J. L. Thorne, J. Felsenstein, B. M. Wiegmann, *Syst. Biol.* **45**, 575 (1996).
5. D. M. Hillis and J. J. Bull, *ibid.* **42**, 182 (1993).
6. J. W. Archie, *Syst. Zool.* **38**, 239 (1989); D. Maddison, *Evolution* **44**, 539 (1990).
7. J. A. Rice, *Mathematical Statistics and Data Analysis* (Duxbury, Belmont, CA, 1995).
8. J. Felsenstein, *Syst. Zool.* **27**, 401 (1978).
9. A. W. F. Edwards and L. L. Cavalli-Sforza, in *Phenetic and Phylogenetic Classification,* J. McNeill, Ed. (Systematics Association, London, 1964), pp. 67–76; L. L. Cavalli-Sforza and A. W. F. Edwards, *Evolution* **21**, 550 (1967).
10. J. Felsenstein, *J. Mol. Evol.* **17**, 368 (1981).
11. See the supplementary material on likelihood theory, available to *Science* Online subscribers at *Science*'s World Wide Web site, http://www.sciencemag.org.
12. M. K. Kuhner and J. Felsenstein, *Mol. Biol. Evol*. **11**, 459 (1994); Y. Tateno, N. Takezaki, M. Nei, *ibid.*, p. 261; B. S. Gaut and P. O. Lewis, *ibid*. **12**, 152 (1995); J. P. Huelsenbeck, *ibid.*, p. 843.
13. J. P. Huelsenbeck, *Syst. Biol.* **44**, 17 (1995).
14. A. W. F. Edwards, *Likelihood* (Johns Hopkins Univ. Press, Baltimore, MD, 1972).
15. D. R. Cox, *Math. Stat. Prob.* **1**, 105 (1961); *J. R. Stat. Soc. Ser. B* **24**, 406 (1962).
16. N. Goldman, *J. Mol. Evol.* **36**, 182 (1993).
17. J. Felsenstein, *Annu. Rev. Genet.* **22**, 521 (1988).
18. J. T. Chang, *Math. Biosci*. **134**, 189 (1996).
19. D. Barry and J. A. Hartigan, *Stat. Sci.* **2**, 191 (1987); Z. Yang, *J. Mol. Evol.* **39**, 105 (1994).
20. J. L. King and T. H. Jukes, *Science* **164**, 788 (1969); M. Hasegawa, H. Kishino, T. Yano, *J. Mol. Evol.* **22**, 160 (1985); S. R. Palumbi, *ibid.* **29**, 180 (1989); Z. Yang, *Mol. Biol. Evol.* **10**, 1396 (1993); *J. Mol. Evol.* **39**, 306 (1994).
21. Z. Yang, *J. Mol. Evol*. **42**, 587 (1996).
22. _____ and D. Roberts, *Mol. Biol. Evol.* **12**, 451 (1995).
23. J. Felsenstein, in *Statistical Analysis of DNA Sequence Data*, B. S. Weir, Ed. (Dekker, New York, 1983).
24. M. S. Hafner *et al., Science* **265**, 1087 (1994); M. S. Hafner and R. D. M. Page, *Philos. Trans. R. Soc. London Ser. B* **349**, 77 (1995).
25. J. P. Huelsenbeck, B. Rannala, Z. Yang, *Evolution* **51**, 410 (1997).
26. A. Rzhetsky and M. Nei, *Mol. Biol. Evol.* **12**, 131 (1995).
27. Z. Yang, N. Goldman, A. Friday, *ibid.* **11**, 316 (1994).
28. D. E. Dykhuizen and L. Green, *J. Bacteriol.* **173**, 7257 (1991); D. S. Hibbett, *Mol. Biol. Evol.* **13**, 903 (1996).
29. W. W. Henderson *et al.*, *Virology* **214**, 602 (1995).
30. D. R. Brooks, *Syst. Zool.* **30**, 229 (1981); D. Simberloff, *ibid*. **36**, 175 (1987); M. S. Hafner and S. A. Nadler, *Nature* **332**, 258 (1988).
31. J. P. Huelsenbeck and J. J. Bull, *Syst. Biol.* **45**, 92 (1996).
32. The LRT statistic for the test of congruence was

calculated assuming a F81 (*10*) model of DNA substitution. This model allows for different equilibrium nucleotide frequencies but assumes that the rate of transitions is equal to the rate of transversions. The significance of the LRT statistic $-2 \log \Lambda = 69.58$ was approximated using the parametric bootstrap procedure.
33. R. D. M. Page, *Syst. Biol.* **45**, 151 (1996).
34. M. Schöniger and A. von Haeseler, *Mol. Phylogenet. Evol.* **3**, 240 (1994).
35. J. P. Huelsenbeck, D. M. Hillis, R. Nielsen, *Syst. Biol.* **45**, 546 (1996).
36. A. Wald, *Sequential Analysis* (Chapman & Hall, London, 1947).
37. We used a tester version of the program PAUP* 4.0 [D. L. Swofford, *PAUP*: Phylogenetic Analysis Using Parsimony (*And Other Methods), Version 4.0* (Sinauer, Sunderland, MA, 1996)]. This program estimates trees under the parsimony, distance, and maximum likelihood criteria.
38. The F84+$\Gamma$ model of DNA substitution was used in the analysis of gophers and lice. The parameter estimates ($\kappa$ , transition rate–transversion rate ratio; $\alpha$, gamma shape parameter) for the gopher (G) and louse (L) trees are $\log L_G = -1923.01$, $\log L_L = -2352.55$; $\kappa_G = 4.63$, $\kappa_L = 7.17$; $\alpha_G = 0.15$, $\alpha_L = 0.18$.
39. D. M. Hillis, B. Mable, C. Moritz, in *Molecular Systematics*, D. M. Hillis, C. Moritz, B. Mable, Eds. (Sinauer, Sunderland, MA, 1996), pp. 515–543.
40. The Felsenstein 1984 (F84) model has been implemented in J. Felsenstein's DNAML program since 1984. This model allows different equilibrium nucleotide frequencies and a transition rate–transversion rate bias.
41. We thank J. Bull, J. Garza, D. Hillis, R. Nielsen, A. Meyer, M. Slatkin, D. Wake, T. Wiehe, and Z. Yang for critical review of the manuscript. Supported by NIH grant GM40282 to M. Slatkin, a Miller postdoctoral fellowship (J.H.), and a Natural Sciences and Engineering Research Council (NSERC) of Canada postdoctoral fellowship (B.R.).