Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology

John P. Huelsenbeck,[1]* Fredrik Ronquist,[2] Rasmus Nielsen,[3] Jonathan P. Bollback[1]

As a discipline, phylogenetics is becoming transformed by a flood of molecular data. These data allow broad questions to be asked about the history of life, but also present difficult statistical and computational problems. Bayesian inference of phylogeny brings a new perspective to a number of outstanding issues in evolutionary biology, including the analysis of large phylogenetic trees and complex evolutionary models and the detection of the footprint of natural selection in DNA sequences.

The idea that species are related through a history of common descent is an old one, predating Darwin. Yet the idea provides an organizing principle in biology that has profound importance for a number of fundamental questions. These questions range from the basic, What is the phylogeny of life?, or the more esoteric, How can the association in traits caused by a common history be accommodated?, to the practical, How did a virus spread through a population? Today, in fact, any study of DNA sequences sampled from different species or from different individuals in a population is likely to start with a phylogenetic analysis. The widespread use of phylogenies today is largely driven by the fundamental importance of phylogenies to questions in biology, the immense quantity of sequence data produced by biologists, and the fact that sophisticated analyses can now be performed by using fast desktop computers.

Perhaps the most frustrating aspect of phylogenetic analysis to the uninitiated is the bewildering variety of inference methods that could be performed and that are actively propounded by different experts. This article might be misconstrued as describing yet another such method—Bayesian inference of phylogeny, a method that has only recently found its way to the field despite its long tenure in statistics (*1*–*3*). Although Bayesian inference of phylogeny uses the same models of evolution as many other methods of analysis, it represents a powerful tool for addressing a number of long-standing, complex questions in evolutionary biology (Table 1). Here we describe Bayesian inference of phylogeny and illustrate applications for inferring large trees, detecting natural selection, and choosing among models of DNA substitution.

Table 1. The Bayesian approach to problems in phylogeny.

Problem	Bayesian approach	Ref.
Inferring phylogeny	Find tree with maximum posterior probability; evaluate features in common among the sampled trees	(*1*–*3*)
Evaluating uncertainty in phylogenies	Evaluate clade probabilities; form credible set containing trees whose cumulative probability sums to 0.95	(*3*, *40*)
Detecting selection	Model substitution process on the codon and calculate probability of being in purifying or positively selected class; sample substitutions and count number of synonymous and nonsynonymous changes	(*29*, *32*)
Comparative analyses	Perform analysis on many trees, and weight results by the probability that each tree is correct	(*41*–*43*)
Divergence times	Use fossils as a calibration. Infer divergence times by using a strict or relaxed molecular clock	(*44*)
Testing molecular clock	Calculate Bayes factor for the clock versus no branch length restrictions	(*24*)

Bayesian Inference of Phylogeny

Bayesian inference of phylogeny is based on a quantity called the posterior probability of a tree (Fig. 1). Bayes's theorem

$$\Pr[\text{Tree} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]}$$

(where the vertical bar should be read as "given") is used to combine the prior probability of a phylogeny (Pr[Tree]) with the likelihood (Pr[Data | Tree]) to produce a posterior probability distribution on trees (Pr[Tree | Data]). The posterior probability of a tree can be interpreted as the probability that the tree is correct. Inferences about the history of the group are then based on the posterior probability of trees. For example, the tree with the highest posterior probability might be chosen as the best estimate of phylogeny (*1*). Usually all trees are considered a priori equally probable, and the likelihood is calculated under one of a number of standard Markov models of character evolution.

The posterior probability, although easy to formulate, involves a summation over all trees and, for each tree, integration over all possible combinations of branch length and substitution model parameter values. It is all but impossible to do this analytically. Fortunately, a number of numerical methods are available that allow the posterior probability of a tree to be approximated, the most useful of which is Markov chain Monte Carlo [MCMC (*4*)]. MCMC has revolutionized Bayesian inference, with recent applications to Bayesian phylogenetic inference (*1*–*3*) as well as many other problems in evolutionary biology (*5*–*7*). The basic idea is to construct a Markov chain that has as its state space the parameters of the statistical model and a stationary distribution that is the posterior probability distribution of the parameters. For the phylogeny problem, the MCMC algorithm involves two steps: (i) A new tree is proposed by stochastically perturbing the current tree. (ii) This tree is then either accepted or rejected with a probability described by Metropolis *et al.* (*8*) and Hastings (*9*). If the new tree is accepted, then it is subject to further perturbation. It turns out that for a properly constructed and adequately run Markov chain, the proportion of the time that any tree is visited is a valid approximation of the posterior probability of that tree (*10*). Although MCMC has made analysis of many complex models possible, it is not a panacea, as chains can fail to converge to the stationary distribution for a number of reasons (e.g., a poor

[1]Department of Biology, University of Rochester, Rochester, NY 14627, USA. [2]Department of Systematic Zoology, Evolutionary Biology Centre, Uppsala University, Norbyv. 18D, SE-752 36 Uppsala, Sweden. [3]Department of Biometrics, Cornell University, Ithaca, NY 14853–1643, USA.

*To whom correspondence should be addressed. E-mail: johnh@brahms.biology.rochester.edu

mechanism for proposing new states or failure to run the chain long enough).

Inferring Large Trees

Phylogenetic inference is difficult primarily because of the large number of trees that may describe the relationships of a group of species and the vagaries of the substitution process. When rates of DNA substitution are high, for example, multiple substitutions at a site can obscure the history of a character. In fact, under some branch-length conditions, phylogenetic methods may converge to the wrong tree, a situation in which the method is said to be inconsistent (*11*). Phylogenetic methods that explicitly model the substitution process, thereby correcting for multiple substitutions, can often overcome problems of statistical inconsistency. Unfortunately, the most powerful methods (e.g., maximum likelihood) can only be used on relatively small data sets and many of the faster methods (e.g., many distance methods) do not take full advantage of the information contained in the DNA sequences.

Bayesian inference takes a view of the phylogeny problem that makes analysis of large data sets more tractable: Instead of searching for the optimal tree, one samples trees according to their posterior probabilities. Once such a sample is available, features that are common among the trees can be discerned. For example, the sample can be used to construct a consensus tree, with the posterior probability of the individual clades indicated on the tree. This is roughly equivalent to performing a maximum likelihood analysis with bootstrap resampling (*3*), but much faster.

To illustrate this, we wrote a computer program implementing the MCMC algorithm (*8*, *9*). In particular, we implemented a variant of MCMC called Metropolis-coupled MCMC that is less prone to entrapment in local optima (*12*). We applied the method to four large phylogenetic data sets that span the size range of many problems faced by systematists today (*13*–*16*). The smallest data set included 106 *wingless* sequences sampled from insects, whereas the largest included 357 *atpB* sequences sampled from plants. We assumed a general model of DNA substitution in the analyses (*17*, *18*). This model allowed each nucleotide change to have its own rate and the nucleotide bases to have different frequencies. We allowed rates to vary across sites either by assuming that the rate at a site is a random variable drawn from a gamma distribution or by dividing the sites into first, second, and third codon positions and estimating their rates of substitution separately. At least two chains were run for each data set. All chains were started from random trees (*19*).

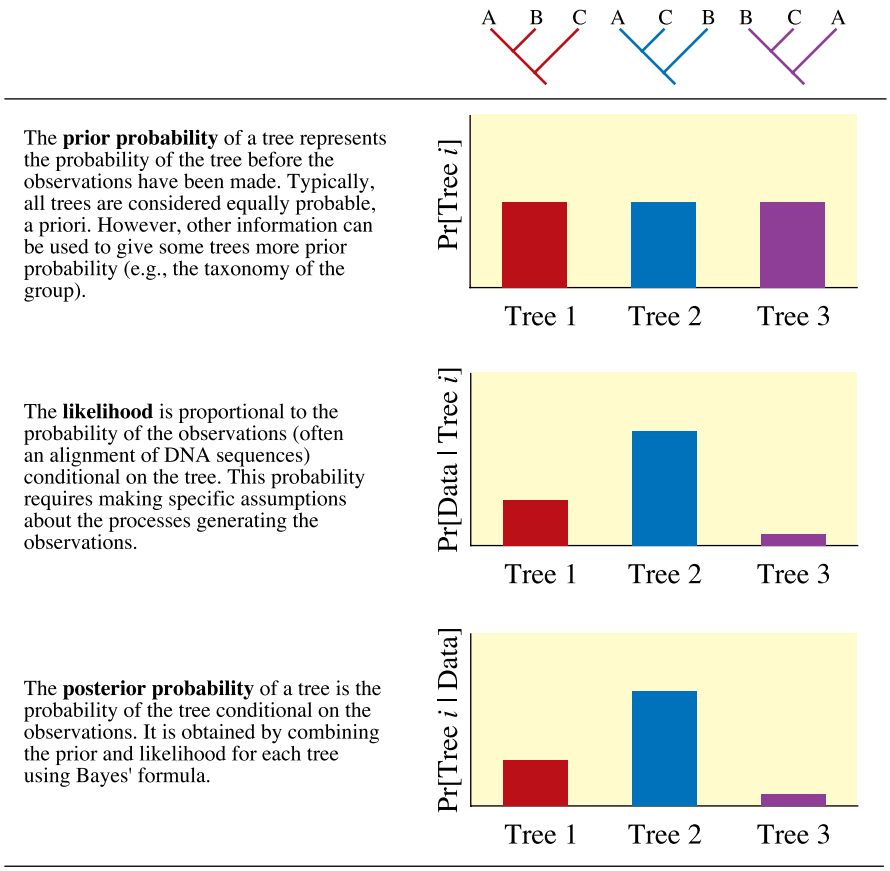

Fig. 1. The main components of a Bayesian analysis.

The greatest practical problems associated with the use of MCMC are determining how long to run a chain to obtain a good approximation of the posterior probabilities of trees and identifying pathological cases where the MCMC algorithm fails to converge. We examined a number of diagnostics to check convergence of multiple chains [see supplemental material (*20*)]. Most important, we checked that inferences made from independent chains were indistinguishable. The posterior probabilities of individual clades are highly correlated for independent chains, as shown in Fig. 2. Together the results of the various diagnostics suggest that the chains converged and that the inferences from the chains are valid.

The Bayesian analyses, run in the course of a few weeks on a fast desktop computer, were largely concordant with the parsimony analyses (*20*). However, the support for the deeper divergences was generally higher in the Bayesian analysis. Notably, for the plant *atpB* data the Bayesian tree differed in the placement of *Ceratophyllum*. Parsimony placed this enus sister to monocots with low support, whereas the Bayesian analysis placed *Ceratophyllum* more basally and in a position that is more congruent with the results from another gene (*rbcL*) (*14*) and a study that included 560 species and three genes (*21*). The Bayesian analysis of *Astragalus* was also similar to the parsimony analysis, with the exception that the support for the *Neo-Astragalus* clade was more similar to the corrected parsimony bootstrap proportions rather than the uncorrected values. The Bayesian analysis also provided information on the substitution model parameters (*20*). The estimates of the substitution rates were typically higher than the corresponding parsimony estimate; this is necessarily true as the parsimony method minimizes the number of changes at a site and must underestimate the total number of changes.

Choosing Appropriate Models

The results of any phylogenetic analysis, including those discussed above, are conditional on the assumptions made in the analysis. Modeling assumptions that poorly fit the observations can lead to erroneous inferences. For example, a phylogenetic model that assumes equal rates across sites can result in inconsistent inferences when rates differ, even if all other parameters of the model are correct (*22*). It is important, then, that a careful consideration of alternative models of

evolution be made so that the most appropriate is used in the phylogenetic analysis.

But how does one go about selecting the most appropriate model? In the past decade, practitioners in phylogenetics have become more sophisticated when choosing among evolutionary models. A common approach uses the likelihood ratio test, with the null distribution relying on asymptotic theory or computer simulation (*23*). Likelihood ratio tests, and other similar methods, are very useful, but can depend on the tree used to perform the test. A number of Bayesian approaches can also be used to choose among evolutionary models. For example, Bayes factors—comparing the marginal likelihoods of two models—have proven to be useful in choosing among evolutionary models (*24*). One advantage of these methods is that the results are not conditional on an assumed topology being correct. The Markov chain simulation effectively treats the topology as a nuisance parameter by summing over trees. An exhaustive description of Bayesian methods of model choice is not feasible here, but we will illustrate one method that uses predictive densities–posterior predictive simulation (*25*).

If an evolutionary model does a good job of explaining the observed DNA sequences, then data simulated under that model should be similar to the observations. Posterior predictive simulation tests the adequacy of a model by comparing a test statistic with the posterior predictive distribution of that statistic generated under the assumption that the model is correct. The test statistic should measure how well a model performs in predicting the observations. The posterior predictive distribution is approximated by simulating new observations by using parameter values sampled from the posterior distribution of the model being scrutinized. Uncertainty in the tree and substitution model parameters is accommodated by sampling from the posterior distribution. The test statistic for the simulated data is then compared with that for the actual data. If a model provides a good fit to the data, then the original test statistic should fall within the central region of the simulated distribution. For a poorly fitting model, the test statistic will fall outside the tails of the predictive distribution.

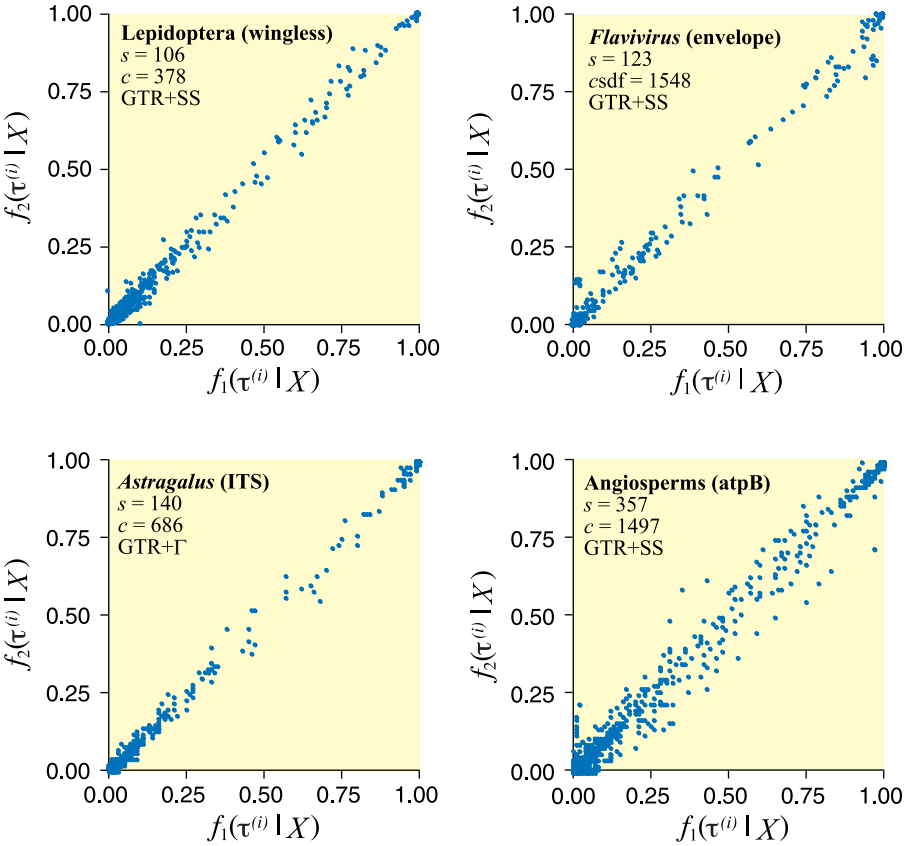

Fig. 2. Convergence of independent Markov chains. The figures show the posterior probability of a clade [or subtree; $\tau^{(i)}$] conditional on the observed DNA sequences (X) for two chains, each of which started from different random trees. Note that the posterior probabilities of individual clades found in different chains [$f_1(\tau^{(i)} \mid X)$ versus $f_2(\tau^{(i)} \mid X)$] is highly correlated, and that there are no instances in which a particular clade found with high probability in one chain is not found in the other. All analyses assumed the general time reversible (GTR) model of DNA substitution. Rate variation across sites was accommodated by using the gamma (+Γ) model for the ITS data and the site-specific (+SS) model for the protein-coding genes. The analyses included from $s = 106$ to $s = 357$ sequences that were from $c = 378$ to $c = 1497$ sites in length.

We illustrate the use of posterior predictive simulation for measuring the overall adequacy of a phylogenetic model in simulation and for testing the homogeneity of nucleotide frequencies at the *Drosophila alcohol dehydrogenase* (*Adh*) locus (*26*). For the first case, we compared a simple model of DNA substitution (*27*) with a more general model (*17*) using data simulated under the latter. As expected, the inadequacy of the simple model is revealed (Fig. 3B; $p_T = 0.008$) while the more parameter-rich model provides a good description of the underlying process (Fig. 3A; $p_T = 0.556$). For the empirical example involving 58 *Adh* sequences (*26*), we use a test statistic that measures the general deviation in nucleotide frequencies among the sequences (*28*). The predictive distribution of this test statistic was evaluated by using MCMC (*29*) and compared with the observed value (Fig. 3C). Because the observed value is well outside the predictive distribution, the hypothesis of constant base frequencies among species is easily rejected. Although this method of inference is in the classical tradition of hypothesis testing, the Bayesian approach adds the ability to deal appropriately with uncertainty in the phylogeny.

Bayesian Inference of Functional Importance in Molecular Evolution

In studies of the evolution of biological molecules and their functions, researchers are often interested in substitution patterns. Typical questions include the following: (i) At what rate do various types of substitutions occur? (ii) Has the mode of evolution changed along the phylogeny? (iii) Which parts of the protein are functionally constrained or under positive selection? And (iv) is the evolution of amino acid residues correlated? The two most common approaches to these questions are to infer substitutions on a fixed phylogeny by using parsimony and to analyze the inferred substitutions as if they were real data, or, to develop likelihood models that can be compared by using likelihood ratio tests.

The parsimony approach has been extensively used in studies of molecular evolution. However, it suffers from the drawback that the number of substitutions will be underestimated and that a large part of the statistical uncertainty is ignored when concentrating on only one possible history of substitutions on the phylogeny (*29*). The approach using likelihood ratio tests has a solid statistical foundation, but every time a new hypothesis is to be tested, a new likelihood model must be

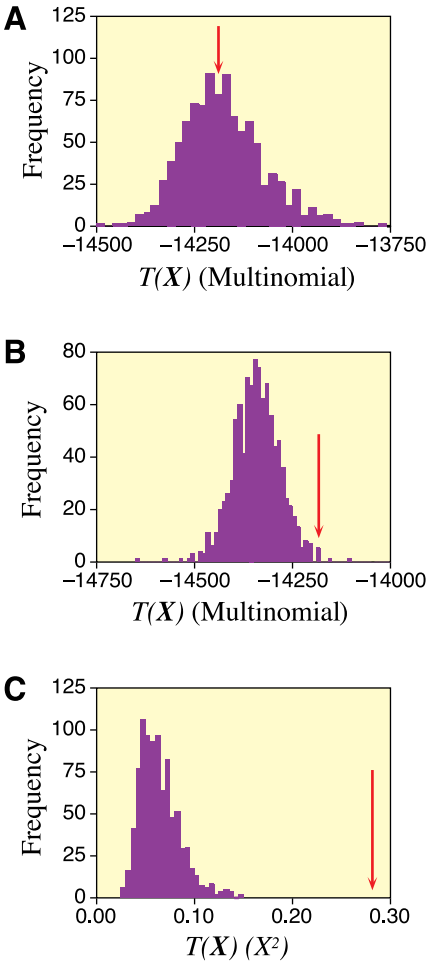

Fig. 3. The posterior predictive distributions for tests of (**A**) the adequacy of the GTR model, (**B**) of the adequacy of the Jukes-Cantor model, and (**C**) the hypothesis of constant nucleotide frequencies over time. The arrows above the distributions show the observed value of the test statistics.

implemented. In addition, both methods typically assume that the phylogeny is known without error.

Ideally, we would like to study molecular evolution by making inferences regarding the type and distribution of substitutions on the phylogeny, while at the same time accommodating the inherent uncertainty in the tree and associated history of substitutions. Consider an alignment of DNA sequences involving four species in which "A" was observed at a site for three of the species and "C" for the fourth. One history of substitution that could explain these observations involves a single change along the branch leading to the C. However, there are infinitely many other such histories that could explain the observations, all involving more changes. The parsimony approach considers only a single history—the history of substitution that involves the fewest number of changes. In a Bayesian framework one considers many possible histories of substitution, weighted by their probability of occurring under a specific model of evolution. Character histories can be sampled in proportion to their probability by using simulation (*29*). Moreover, the uncertainty in the tree and model parameters can be accounted for by sampling trees with MCMC.

The approach of mapping substitutions on a tree can be used to detect positively selected residues. Positive selection occurs when natural selection increases the frequency of new amino acid mutations. Positive selection at the molecular level is shown by an increase in the rate of nonsynonymous substitution over the rate of synonymous substitution. Positive selection occurs in many systems (*30*), but particular attention has focused on viral DNA sequences. A statistical approach for identifying sites undergoing positive selection is based on first testing for the presence of positively selected sites with a likelihood ratio test, and then, if the test is significant, identifying positively selected sites by using an empirical Bayes approach (*31*, *32*). Empirical Bayes approaches differ from other Bayesian methods in that the prior distribution is determined, in part, by the data. The empirical Bayes approach has been useful in identifying positively selected residues in a number of systems (*29*, *32*–*34*).

An alternative approach is to use the posterior distribution of substitutions to examine the pattern of nonsynonymous substitutions. Of particular interest are amino acid residues in which more nonsynonymous substitutions occurred than expected under neutrality (i.e., equal nonsynonymous and synonymous rates). We illustrate this method on a data set containing 28 influenza sequences of the hemagglutinin gene (*35*, *31*). Hemagglutinin is an envelope gene of the virus and is a potential target for the host immune system. Previous studies based on maximum likelihood, and other methods, have demonstrated that positive selection is acting on this set of sequences (*31*, *35*, *36*). To identify positively selected sites, we used the posterior expectation of the number of nonsynonymous substitutions in a site, E_{NS}. Using MCMC, we estimated E_{NS} for each site under the hypothesis that the rate of nonsynonymous substitutions equals the rate of synonymous substitutions (*37*, *38*). The predictive distribution of E_{NS} was also evaluated by using MCMC. Seven residues were identified with a value of E_{NS} larger than 4 (Fig. 4). All of these residues were located in proximity to each other on the globular head of the molecule. The posterior predictive probability of $E_{NS} > 4$ in a residue is approximately 0.002 if the rate of nonsynonymous substitution equals the rate of synonymous substitution. Presumably these seven sites have increased levels of nonsynonymous variation because of positive selection. This observation is confirmed by the fact that all seven residues are located within known antigenic sites, and four of them are located within the very same antigenic site (*39*). Strong positive selection appears to have been occurring in the history of these sequences so as to avoid immune recognition. Moreover, a majority of the positively selected substitutions in the hemagglutinin gene tends to be conservative amino acid changes (*37*), implying that even though there is strong selection pressure for changing binding affinities in these sites, some selection must also occur in the same sites to maintain the structure and function of the protein.

The Future of Bayesian Phylogenetic Inference

There are many reasons to believe that the success of Bayesian inference will continue as it is applied to a wider range of problems in evolutionary biology. These reasons include the ease with which complex evolutionary models can be examined, the accommodation of phylogenetic uncertainty (an advantage conferred by using MCMC), and the fact that the method concentrates attention on the evolutionary models. Bayesian analysis should also prove useful in addressing some of the outstanding problems in phylogenetics, such as detecting and accommodating horizontal gene transfer (a process that complicates phylogenetic analysis of bacteria), per-

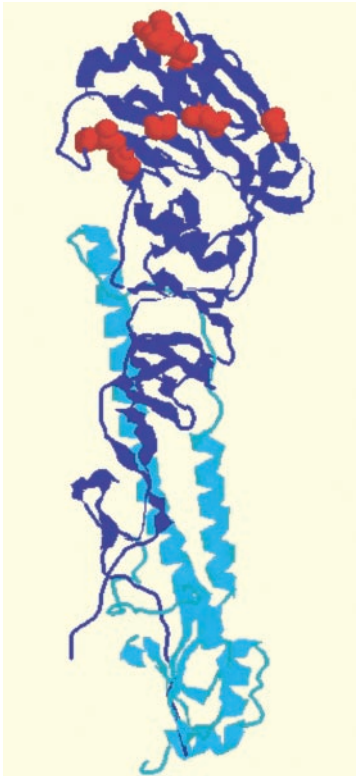

Fig. 4. The protein structure of the influenza hemagglutinin protein, chains A and B. The seven positively selected residues are marked in red.

forming phylogenetic analyses by using whole-genome data and understanding the evolution of the genome in the context of phylogeny, and constructing large trees by combining the results of smaller and overlapping analyses.

References and Notes

1. B. Rannala, Z. Yang, *J. Mol. Evol.* **43**, 304 (1996).
2. B. Mau *et al.*, *Biometrics* **55**, 1 (1999).
3. B. Larget, D. Simon, *Mol. Biol. Evol.* **16**, 750 (1999).
4. W. R. Gilks *et al.*, Eds., *Markov Chain Monte Carlo in Practice* (Chapman & Hall, London, 1996).
5. M. K. Kuhner, J. Yamato, J. Felsenstein, *Genetics* **140**, 1421 (1995).
6. P. Beerli, J. Felsenstein, *Genetics* **152**, 763 (1999).
7. R. Nielsen, *Genetics* **154**, 931 (2000).
8. N. Metropolis *et al.*, *J. Chem. Phys.* **21**, 1087 (1953).
9. W. Hastings, *Biometrika* **57**, 97 (1970).
10. L. Tierney, *Ann. Stat.* **22**, 1701 (1994).
11. J. Felsenstein, *Syst. Zool.* **27**, 401 (1978).
12. C. J. Geyer, in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, E. M. Keramidas, Ed. (Interface Foundation, Fairfax Station, VA, 1991), pp. 156–163.
13. A. Brower, *Proc. R. Soc. Lond. B* **267**, 1201 (2000).
14. V. Savolainen *et al.*, *Syst. Biol.* **49**, 306 (2000).
15. P. M. Zanotto *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 548 (1996).
16. M. J. Sanderson, M. F. Wojciechowski, *Syst. Biol.* **49**, 671 (2000).
17. S. Tavaré, *Lect. Math. Life Sci.* **17**, 57 (1986).
18. Z. Yang, *Mol. Biol. Evol.* **10**, 1396 (1993).
19. J. P. Huelsenbeck, F. Ronquist, *Bioinformatics* **17**, 754 (2001).
20. Supplementary material is available on *Science* Online at www.sciencemag.org/cgi/content/full/294/5550/2310/DC1
21. D. Soltis *et al. Bot. J. Linn. Soc.* **133**, 381 (2000).
22. B. Gaut, P. Lewis, *Mol. Biol. Evol.* **12**, 152 (1995).
23. N. Goldman, *J. Mol. Evol.* **36**, 182 (1993).
24. M. A. Suchard, R. E. Weiss, J. S. Sinsheimer, *Mol. Biol. Evol.* **18**, 1001 (2001).
25. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, *Bayesian Data Analysis* (Chapman & Hall, London, 1995).
26. S. Atrian, L. Sánchez-Pulido, R. González-Duarte, A. Valencia, *J. Mol. Evol.* **47**, 211 (1998).
27. T. H. Jukes, C. R. Cantor, in *Mammailian Protein Metabolism*, H. N. Munro, Ed. (Academic Press, New York, 1969).
28. DNA sequences were simulated under the GTR model (*17*) on a tree of ten taxa and 2000 nucleotides. The MCMC analysis, conducted by using MrBayes, was under the GTR and Jukes-Cantor (*27*) models. 1000 samples were randomly drawn from the joint posterior distribution of model parameters and topology, and new data sets of 2000 sites were simulated under both models. The test statistic was calculated as the multinomial probability of observing the data. We used

$$X^2 = \sum_{i=1}^{58} \sum_{j=\{A,C,G,T\}} \frac{(f_{ij} - \bar{f}_j)^2}{\bar{f}_j}$$

as a test statistic for examining the constancy of nucleotide frequencies over time. The summations are over species and nucleotides, and $(\bar{f})_j$ is the average frequency of nucleotide j in the pooled data.
29. R. Nielsen, in preparation.
30. Z. Yang, J. P. Bielawski, *Trends Ecol. Evol.* **15**, 496 (2000).
31. Z. Yang *et al.*, *Genetics* **155**, 431 (2000).
32. R. Nielsen, Z. Yang, *Genetics* **148**, 929 (1998).
33. W. J. Swanson *et al.*, *Mol. Biol. Evol.* **18**, 376 (2001).
34. W. J. Swanson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 2509 (2001).
35. W. M. Fitch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 7712 (1997).
36. R. M. Bush, W. M. Fitch, C. A. Bender, N. J. Cox, *Mol. Biol. Evol.* **16**, 1457 (1999).
37. R. Nielsen, J. P. Huelsenbeck, *Proc. Pac. Symp. Biocomp.*, in press.
38. A Markov chain was simulated for 1,000,000 cycles under a GTR model (*17*) with MrBayes. After the first 100,000 cycles were discarded as burn-in, the chain was sampled every 1000 updates. The posterior predictive distribution is approximated by simulating a new data set for each of the sampled parameter values and then evaluating the chi-square statistic for each of the simulated data sets.
39. Locations of antigenic sites were obtained from the Influenza Sequence Database, Los Alamos National Laboratory, Los Alamos, NM.
40. J. Felsenstein, thesis, University of Chicago (1968).
41. J. P. Huelsenbeck, B. Rannala, J. P. Masly, *Science* **288**, 2349 (2000).
42. J. B. Losos, D. B. Miles, in *Ecological Morphology: Integrative Organismal Biology*, P. C. Wainwright, S. Reilly, Eds. (Univ. of Chicago Press, Chicago, 1994), pp. 60–98.
43. M. Pagel, *Proc. R. Soc. London Ser. B Biol. Sci.* **255**, 37 (1994).
44. J. Thorne, H. Kishino, I. S. Painter, *Mol. Biol. Evol.* **15**, 1647 (1998).

45. This research was supported by a Swedish National Research Council grant (F.R.) and NSF grants DEB-0075406 (J.P.H.) and DEB-0089487 (R.N.).

So instant, you don't need water...

NEW! *Science* Online's Content Alert Service

There's only one source for instant updates on breaking science news and research findings: *Science*'s Content Alert Service. This free enhancement to your *Science* Online subscription delivers e-mail summaries of the latest research articles published each Friday in *Science* – ***instantly***. To sign up for the Content Alert service, go to *Science* Online – and save the water for your coffee.

Science
www.sciencemag.org

For more information about Content Alerts go to www.sciencemag.org. Click on Subscription button, then click on Content Alert button.