Syst. Biol. 53(1):47–67, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490264699

Bayesian Phylogenetic Analysis of Combined Data

JOHAN A. A. NYLANDER,[1] FREDRIK RONQUIST,[1] JOHN P. HUELSENBECK,[2] AND JOSÉ LUIS NIEVES-ALDREY[3]

[1]*Department of Systematic Zoology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18 D, SE-752 36 Uppsala, Sweden; E-mail: johan.nylander@ebc.uu.se (J.A.A.N)*
[2]*Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California–San Diego, La Jolla, California 92093-0116, USA*
[3]*Departamento de Bioversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales, José Gutiérrez Abascal 2, 28006 Madrid, Spain*

Abstract.— The recent development of Bayesian phylogenetic inference using Markov chain Monte Carlo (MCMC) techniques has facilitated the exploration of parameter-rich evolutionary models. At the same time, stochastic models have become more realistic (and complex) and have been extended to new types of data, such as morphology. Based on this foundation, we developed a Bayesian MCMC approach to the analysis of combined data sets and explored its utility in inferring relationships among gall wasps based on data from morphology and four genes (nuclear and mitochondrial, ribosomal and protein coding). Examined models range in complexity from those recognizing only a morphological and a molecular partition to those having complex substitution models with independent parameters for each gene. Bayesian MCMC analysis deals efficiently with complex models: convergence occurs faster and more predictably for complex models, mixing is adequate for all parameters even under very complex models, and the parameter update cycle is virtually unaffected by model partitioning across sites. Morphology contributed only 5% of the characters in the data set but nevertheless influenced the combined-data tree, supporting the utility of morphological data in multigene analyses. We used Bayesian criteria (Bayes factors) to show that process heterogeneity across data partitions is a significant model component, although not as important as among-site rate variation. More complex evolutionary models are associated with more topological uncertainty and less conflict between morphology and molecules. Bayes factors sometimes favor simpler models over considerably more parameter-rich models, but the best model overall is also the most complex and Bayes factors do not support exclusion of apparently weak parameters from this model. Thus, Bayes factors appear to be useful for selecting among complex models, but it is still unclear whether their use strikes a reasonable balance between model complexity and error in parameter estimates. [Bayes factors; Bayesian analysis; combined data; Cynipidae; gall wasps; MCMC; model heterogeneity; model selection.]

Increasingly, phylogenetic problems are being addressed using data from several different sources: morphology and molecules, DNA and protein, mitochondrial and nuclear genes, coding and noncoding sequences. Previously, it has been common to address such mixed data sets using the parsimony method. Where parametric methods have been applied, they have typically excluded some data (such as morphology) because of a lack of appropriate stochastic models, and they have often ignored obvious heterogeneity across data partitions because of the computational complexity of the maximum likelihood (ML) approach (for exceptions, see Yang, 1996b; DeBry, 1999; Pupko et al., 2002; Thorne and Kishino, 2002).

The recent development of Bayesian inference of phylogeny using Markov chain Monte Carlo (MCMC) estimation of posterior probability distributions has made it easier to address complex, parameter-rich stochastic models within a statistical framework, opening up the possibility for combined data analysis recognizing among-partition heterogeneity in data source and in properties of the evolutionary process. Recent stochastic models developed for new types of data, such as morphology (Lewis, 2001a; Ronquist and Huelsenbeck, in prep.), now make it possible to include virtually any kind of character used today to infer phylogeny in such analyses, and the computational efficiency of the Bayesian MCMC approach allows each data partition to be treated using more realistic evolutionary models. However, combined statistical analysis using Bayesian MCMC techniques introduces a whole range of questions that have not been addressed previously, while providing a new perspective on others. Here, we describe a Bayesian MCMC approach to combined data analysis, using empirical results from one combined data set to address some of these questions.

Bayesian MCMC Approach to Combined Data

Bayesian phylogenetic inference based on heterogeneous data is a straightforward extension of the methods already described for homogeneous data (see recent reviews by Huelsenbeck et al., 2001; Lewis, 2001b; Holder and Lewis, 2003). Assume that the data set X consists of two distinct partitions X_a and X_b and allow the substitution model parameters, θ_a and θ_b, respectively, to be completely different for the two partitions. In the models we explored, we further assumed that the two data subsets evolve on the same topology, τ, with the same set of branch lengths, ν, but that the overall rate differs across partitions according to a rate multiplier, denoted m_a and m_b for the two partitions. In other words, effective branch lengths are potentially different but proportional across partitions, as in the ML model proposed by Yang (1996; note that Yang used c instead of m for the multiplier). Using Bayes's rule (see for instance Huelsenbeck et al., 2001), the joint posterior probability distribution for this model becomes

$$f(\tau, \nu, \theta_a, \theta_b, m_a, m_b \mid X) = \frac{f(\tau, \nu, \theta_a, \theta_b, m_a, m_b)\, f(X \mid \tau, \nu, \theta_a, \theta_b, m_a, m_b)}{f(X)},$$

where $f(\tau, v, \theta_a, \theta_b, m_a, m_b)$ is the prior probability of the model parameters, $f(X \mid \tau, v, \theta_a, \theta_b, m_a, m_b)$ is the probability of the data given model parameters (the likelihood function), and $f(X)$ is the model likelihood (also called the integrated or predictive likelihood), which is a multidimensional sum and integral of the probability of the data over all parameter values.

The posterior probability distribution, which is the central quantity in Bayesian inference, is typically estimated using MCMC techniques instead of being derived analytically. The procedure is started with an arbitrary set of parameter values. In each cycle (generation) of the Markov chain, one parameter or a block of parameters is updated using a stochastic proposal mechanism. The most common mechanism used in Bayesian phylogenetic inference, Metropolis sampling, involves the proposal of a new state based on an arbitrary proposal distribution, q, and then acceptance of this state with a probability determined by the product of three ratios: the prior ratio, the likelihood ratio, and the proposal ratio. Assume, for instance, that we wish to update the substitution model parameters for partition a from θ_a to θ_a^*. The acceptance probability r would then become

$$r = \min\left[1, \frac{f(\theta_a^*)}{f(\theta_a)} \times \frac{f(X_a \mid \tau, v, m_a, \theta_a^*)}{f(X_a \mid \tau, v, m_a, \theta_a)} \times \frac{q(\theta_a \mid \theta_a^*)}{q(\theta_a^* \mid \theta_a)}\right].$$

When updating a homogeneous model or a parameter shared across all partitions, the calculation of the likelihood ratio (the second ratio in the product) always involves the entire data set. However, updating a partitioned parameter only requires consideration of the affected data partition, X_a in this case. The calculation of the likelihood ratio is by far the most computationally complex operation in MCMC analysis, and the speed of the calculation is roughly proportional to the size of the data set. Thus, the increase in the number of parameters in a partitioned model over that in a similar homogeneous model is largely offset by the speed gained in each cycle of the chain. The net result is that the time required for updating all model parameters a given number of times will remain roughly constant regardless of model partitioning. However, more complex models will of course have more dimensions in their parameter spaces, which might cause difficulties for the MCMC sampling procedure.

Convergence and Mixing

Theory predicts that a properly constructed Markov chain, if run long enough, will produce a valid sample from the posterior probability distribution (Tierney, 1994). However, the greatest practical problem in MCMC analysis is to determine when the chain is sufficiently close to its target distribution (the posterior distribution of interest) for the samples to provide a good approximation of this distribution. One of the most powerful approaches used to address this question is comparison of the results from independent runs started from different points in parameter space. In the phylogenetic context, we expect integration over topology to be particularly difficult; therefore, starting the independent runs from different, randomly chosen topologies should provide a good test of whether the chains are providing valid samples from the posterior probability distribution (Huelsenbeck et al., 2002).

It is useful to distinguish two potential sources of problems with MCMC estimation of a target distribution: convergence and mixing. The difference between them is best explained if we consider a posterior distribution with two separate regions, each containing roughly half of the total probability. Typically, a MCMC run starts sampling from a region with extremely low posterior probability because starting values are set arbitrarily or chosen randomly. When the chain has settled into the high-density regions of the distribution, it can be said to have converged, and the overall likelihood will tend to vary less than during the initial burn-in period. However, we still do not know how long it will take the chain to adequately sample both regions of high density in the posterior distribution; this is determined by the mixing behavior of the chain. The slower the mixing, the longer it will take the chain to move from one to the other of the high-density regions. Whereas the generation plot of the overall likelihood gives a preliminary idea of whether convergence might have occurred, assessment of the mixing behavior requires examination of the plots of all model parameters. This is particularly true when Metropolis coupling is used, because this technique allows the chain to jump between different regions in parameter space with little effect on overall likelihood (Huelsenbeck et al., 2001).

Bayesian Model Selection

Analyzing combined data using Bayesian MCMC methods allows us to specify partition-specific substitution models. As more partitions are being considered, the complexity of the joint model increases as does the complexity of the issue of model selection. One strategy for model selection for Bayesian MCMC analysis is to fit a substitution model to each partition prior to the analysis using, for example, a hierarchical likelihood-ratio test (hLRT; Huelsenbeck and Crandall, 1997; Posada and Crandall, 2001), the Akaike information criterion (AIC; Akaike, 1973), or the Bayesian information criterion (BIC; Schwartz, 1978), all of which are based on ML estimates. The Markov chain is then run using a composite 'supermodel' that consists of several submodels.

It is not self-evident, however, that such an approach will necessarily lead to an optimal composite model. Most importantly, the selection of an optimal model for one partition should not ignore information from other partitions. For example, the methods mentioned above depend on point estimates of the topology and other parameters, and it is well known that different topologies might rank models differently (Sanderson and Kim, 2000; Posada and Crandall, 2001). Thus, selecting an optimal model for each partition separately, on the best tree implied by the data from that partition, might result in a

combination of models that could not be optimal on the same topology. Furthermore, considering each partition separately may result in overparameterization, because such an approach makes it difficult to discover when it is appropriate for two partitions to share parameters.

Unfortunately, computational problems make it difficult to apply directly the methods described above to parameter-rich partitioned models. Furthermore, Bayesian statisticians often object in general to model testing based on point estimates, because such methods are not taking the uncertainty of the topology and other parameters into account. The argument is that a model with substantial posterior probability for a large range of parameter values could have a higher marginal (total) likelihood than a model with a narrow peak in its likelihood, even though the latter model may have the highest ML value. In such situations, Bayesian statisticians have argued, it would be unwise to compare models only based on the merits of a single point; instead, we should consider the entire parameter space and prefer the model with the largest total likelihood (Bollback, 2002; Holder and Lewis, 2003). An additional problem with the ML approach is that it favors the more parameter-rich model in comparisons of nested models unless the parameter-rich models are penalized as in AIC or BIC. That is, the favored model might contain parameters that have little or no explanatory value (Burnham and Anderson, 2002). The Bayesian approach does not always favor the more parameter-rich of two nested models; on the contrary, there is some concern that Bayesian methods may, under some circumstances, put too much emphasis on the simpler model. This phenomenon is known as Lindley's paradox, and it can occur with large data sets when the estimate from the complex model is close to the simple model (Bartlett, 1957; Lindley, 1957).

Because of the problems with the likelihood approach, we explored Bayesian model comparison based on Bayes factors. Assume that we wish to compare how well two models, M_0 and M_1, describe the processes generating a data set X. The Bayes factor in favor of model 1 over model 0, B_{10}, is calculated as the ratio of the model likelihoods $f(X \mid M_i)$:

$$B_{10} = \frac{f(X \mid M_1)}{f(X \mid M_0)}.$$

The model likelihoods, $f(X \mid M_i)$, are the same as the $f(X)$ denominator of Bayes's rule; the conditioning on a model is implicit in the latter.

The Bayes factor can be interpreted as the posterior odds of model 1 to model 0 in a Bayesian inference problem where we start with equal probability of the two models being true (Kass and Raftery, 1995; Wasserman, 2000). Alternatively, the Bayes factor can be viewed simply as a comparison of the predictive likelihoods of the models (Gelfand and Dey, 1994; Kass and Raftery, 1995; Wasserman, 2000) or a comparison of the ability of the models to update the priors (Lavine and Schervish, 1999; Wasserman, 2000). Both the latter comparisons would be valid even, although strictly speaking none of the models is likely to be an exact (true) description of the process under study. The Bayes factor comparison can be applied to any set of models, regardless of whether they are nested or not (as can AIC and BIC but not hLRT), and it is based on integration over the uncertainty in all parameter values rather than on ML point estimates (as opposed to AIC, BIC, and hLRT).

TABLE 1. Interpretation of the Bayes factor (B_{10}) (taken from Kass and Raftery, 1995).

$2 \log_e(B_{10})$	B_{10}	Evidence against M_0
0 to 2	1 to 3	not worth more than a bare mention
2 to 6	3 to 20	positive
6 to 10	20 to 150	strong
>10	>150	very strong

The Bayes factor is not used in a normal statistical test of whether a hypothesis should be rejected or accepted given some subjective cutoff value. Instead, the Bayes factor evaluates the relative merits of competing models, and the interpretation is left to the scientist. Jeffreys (1961) originally provided some guidelines for this interpretation, which have been modified by other workers. We use a version originally presented by Kass and Raftery (1995) (Table 1).

Questions Regarding Combined Phylogenetic Analysis

We applied combined Bayesian MCMC analysis to an empirical data set consisting of morphological and nucleotide data for 32 exemplar species of gall wasps (Hymenoptera: Cynipidae) and outgroups. The exemplars span the entire diversity of the family and include phytophagous guests in galls (inquilines) and gall inducers on a variety of both herbaceous and woody host plants (Table 2; Ronquist, 1999).

The morphological data consisted of 166 characters, which have previously been shown to partly resolve the phylogeny with strong support values using parsimony methods (Liljeblad and Ronquist, 1998). The nucleotide data are almost entirely original to this study and consisted of a total of 3,080 aligned base pairs (bp) from four genes: two nuclear protein-coding genes (elongation factor 1α F1 copy [EF1α] and long-wavelength opsin [LWRh]), one mitochondrial protein-coding gene (cytochrome oxidase *c* subunit I [COI]), and nuclear 28S ribosomal DNA (rDNA). We analyzed the data using a range of models of varying complexity (dimensionality) and explored the following questions.

What is the relationship between model complexity and computational complexity?—It is difficult to predict how MCMC estimation of the posterior probability distribution is affected by an increase in model complexity. The chain can be updated faster in those generations where model parameters affecting only some of the partitions are changed; however, more parameters also means that each parameter will be visited more rarely. More parameters will also affect the complexity and the shape of the posterior distribution, which might slow convergence

TABLE 2. Taxa of gall wasps (Cynipidae) and outgroups (Figitidae, Liopteridae, Ibaliidae) used in the analysis. Brief biological data are given for each exemplar genus. GenBank accession numbers are given for all sequences; a dash indicates missing data.

Taxon	Morphology[a]	Host plant[b]	Biology[c]	GenBank nos.			
				COI	28S	EF1α	LWRh
Cynipidae							
Synergini							
Synergus crassicornis		*Quercus* (Fg)	inquiline	AY368909	AY368936	AY368962	AY371051
Ceroptres cerri	*C. clavicornis*	*Quercus* (Fg)	inquiline	AY368910	AY368935	—	AY371052
Periclistus brandtii		*Rosa* (Ro)	inquiline	AF395181	AF395152	AF395173	AF395189
Synophromorpha sylvestris	*S. rubi*	*Rubus* (Ro)	inquiline	AY368911	AY368937	AY368961	—
"Aylacini"							
Xestophanes potentillae		*Potentilla* (Ro)	galler	AY368912	AY368938	AY368963	—
Diastrophus turgidus		Rosaceae	galler	AY368913	AY368939	AY368964	—
Gonaspis potentillae		*Potentilla* (Ro)	galler	AY368914	AY368940	AY368965	—
Liposthenes glechomae		*Glechoma* (La)	galler	AY368915	AY368941	AY368966	AY371053
Liposthenes kerneri		*Nepeta* (La)	galler	AY368916	AY368942	AY368967	AY371054
Antistrophus silphii	*A. pisum*	Asteraceae	galler	AY368917	AY368943	AY368968	AY371055
Rhodus oriundus		*Salvia* (La)	galler	AY368918	AY368944	AY368969	AY371056
Hedickiana levantina		*Salvia* (La)	galler	AY368919	AY368945	AY368970	AY371057
Neaylax verbenaca	*N. salviae*	*Salvia* (La)	galler	AY368920	AY368946	AY368971	AY371058
Isocolus rogenhoferi		Asteraceae	galler	AY368921	AY368947	AY368972	AY371059
Aulacidea tragopogonis		Asteraceae	galler	AY368922	AY368948	AY368973	AY371060
Panteliella bicolor	*P. fedtschenkoi*	*Phlomis* (La)	galler	AF395180	AF395153	AF395172	AF395188
Barbotinia oraniensis		*Papaver* (Pa)	galler	AF395179	AF395150	AF395171	AF395187
Aylax papaveris		*Papaver* (Pa)	galler	AY368923	AY368949	AY368974	AY371061
Iraella luteipes		*Papaver* (Pa)	galler	AY368924	AY368950	AY368975	—
Timaspis phoenixopodos		Asteraceae	galler	AY368925	AY368951	AY368976	AY371062
Phanacis hypochoeridis		Asteraceae	galler	AY368926	AY368952	AY368977	—
Phanacis centaureae		Asteraceae	galler	AY368927	AY368953	AY368978	—
Eschatocerini							
Eschatocerus acaciae		*Acacia* (Fb)	galler	AY368928	AY368954	AY368979	AY371063
Diplolepidini							
Diplolepis rosae		*Rosa* (Ro)	galler	AF395174	AF395157	AF395166	AF395182
Pediaspidini							
Pediaspis aceris		*Acer* (Sa)	galler	AY368929	AY368955	AY368980	AY371064
Cynipini							
Plagiotrochus quercusilicis[d]		*Quercus* (Fg)	galler	AF395178	AF395154	AF395162	AF395186
Andricus kollari	*A. quercusradicis*	*Quercus* (Fg)	galler	AF395176	AF395156	AF395168	AF395184
Neuroterus numismalis		*Quercus* (Fg)	galler	AY368930	AY368956	AY368981	—
Biorhiza pallida		*Quercus* (Fg)	galler	AY368931	AY368957	AY368982	AY371065
Figitidae							
Parnips nigripes		—	parasitoid	AY368932	AY368958	AY368983	AY371066
Liopteridae							
Paramblynotus virginianus	*P. zonatus*	—	parasitoid	AY368933	AY368959	AY368984	—
Ibaliidae							
Ibalia rufipes		—	parasitoid	AY368934	AY368960	AY368985	—

[a]Species coded for morphology if different from the species sequenced.

[b]Genus or family of host plant attacked by the exemplar genus if phytophagous. A few rarely used host plants have been omitted; see Ronquist and Liljeblad (2001) for more information. If all members of the genus attack the same host-plant genus, then the family to which that genus belongs is indicated in brackets: Fb = Fabaceae; Fg = Fagaceae; La = Lamiaceae; Pa = Papaveraceae; Ro = Rosaceae; Sa = Sapindaceae.

[c]Cynipidae are either inquilines (phytophagous guests) in galls or gall inducers. The outgroups are endoparasitoids attacking various insect larvae.

[d]Species name recently designated a senior synonym of *P. fusifex*.

and mixing. However, more realistic models may lead to posterior distributions that are easier to traverse using MCMC, despite the increase in the number of parameters. We examined the computational speed, time to convergence, and mixing over the entire range of models to examine these questions empirically.

Do morphological data influence multigene analyses?—Morphological data are potentially important in phylogenetic inference for many reasons. For instance, morphological characters are crucial in placing fossils in phylogenies and thus in dating branching events. However, the ability to combine morphological and molecular data in a single analysis is particularly important if it can be shown that morphology has significant influence on the phylogenetic estimate even when combined with multigene data sets. This question has remained largely unexplored with parametric methods, because only recently were stochastic models seriously considered for morphological data (Lewis, 2001a). We used an extended version of Lewis's models (Ronquist and Huelsenbeck, in prep.) in evaluating whether the 166 morphological characters in our data set significantly affected the phylogenetic estimate when combined with the 3,080 nucleotide characters from the four different genes.

Are composite models better?—When it becomes possible to analyze partitioned models easily, an obvious

question is how important it is to recognize across-partition heterogeneity in evolutionary processes. To examine this question, we used Bayes factor comparisons to look at the increase in model likelihood associated with the introduction of different model components accounting for within-partition or across-partition heterogeneity in the molecular portion of the data set.

Are complex models associated with increased variance of topology estimates?—Complex models are generally associated with more error variance in parameter estimates. If the error variance is excessive, it becomes a problem known as overparameterization or overfitting (Burnham and Anderson, 2002). However, overly simple models can also be problematic. In particular, oversimplified evolutionary models might lead to dramatically lowered topological variance and exaggerated clade probability values in Bayesian phylogenetic inference (Suzuki et al., 2002). To examine the relationship between model complexity and the precision of parameter estimates, we compared topology and tree-length estimates across models. We also looked at the effect of model complexity on the conflict between the morphological and molecular partitions.

Is the Bayesian MCMC approach sensitive to the inclusion of superfluous parameters in a complex model?—It may be difficult to design complex models that adequately explain a process under study without including one or a few parameters that are superfluous in the sense that (1) the data are not powerful enough to significantly alter their prior probability distribution or (2) the posterior probability distribution coincides with a less parameter-rich submodel. Such "superfluous" parameters might cause problems with MCMC estimation of the posterior distribution. We searched the posterior distributions of more complex models for such parameters to see whether they were present and, if so, whether there was any apparent effect on convergence or on the posterior distributions of other model parameters. If the Bayesian MCMC approach were sensitive to superfluous parameters, it might be difficult to design appropriate composite models that would result in successful combined analysis.

Do Bayes factors strike a reasonable balance between model complexity and error variance?—The ability to allow heterogeneity across data partitions in model parameters opens up a Pandora's Box of model choice problems, which are difficult to address without good model selection criteria and procedures. Standard likelihood ratio tests have a tendency to prefer complex models (Gelfand and Day, 1994; Burnham and Anderson, 2002) and various procedures have been developed to punish parameter-rich models (Akaike, 1973; Schwartz, 1978). In theory, the Bayes factor comparison does not suffer from this problem; a simple model can be favored over a more parameter-rich model even if the models are nested. We looked for instances of simple models winning over more complex ones and cases where the Bayes factor would favor model reduction by supporting the exclusion of weak parameters.

Materials and Methods

Data

We assembled DNA and morphological data for 29 gall wasp exemplars and three outgroup exemplars, the latter representing the families Figitidae, Liopteridae, and Ibaliidae (Table 2). Previous phylogenetic analyses indicate that Figitidae is the sister group to Cynipidae and that the Liopteridae and Ibaliidae are successively more distant outgroups (Ronquist, 1999). The gall wasp sample included representatives of all described tribes of the only extant subfamily. All major wasp genera of phytophagous guests in cynipid galls, also known as inquilines, were represented except for the genus *Saphonecrus*, which is considered close to if not embedded within *Synergus* (Ronquist, 1994, 1999; Nieves-Aldrey, 2001; Ronquist and Liljeblad, 2001). A broad selection of gall inducers attacking herbaceous and woody host plants was also included. At least half the described genera were included for all tribes except the Cynipini, or the oak gallers. This tribe, comprising more than 40 described genera, was represented by only four genera but is widely thought to be monophyletic (Kinsey, 1920; Askew, 1984; Ronquist, 1994, 1999; Liljeblad and Ronquist, 1998; Nieves-Aldrey, 2001; Ronquist and Liljeblad, 2001; Stone et al., 2002).

The morphological data were taken from Liljeblad and Ronquist (1998) and consist of 166 parsimony-informative discrete characters: 164 external morphological characters and two ecological characters (alteration of sexual/asexual generations, and host-plant choice) (Liljeblad and Ronquist, 1998: appendix 1). Some multistate characters were treated as ordered and others as unordered, as specified by Liljeblad and Ronquist (1998).

As far as possible, DNA data were collected from the same species for which we had morphological data. In a few cases, an exact match could not be obtained, but DNA sequences were obtained, from a close relative and these taxa were combined into a single terminal in the final analyses (Table 2). We sequenced parts of four genes: COI (1,078 bp), the nuclear protein-coding genes LWRh (481 bp) and EF1α, (367 bp), and the nuclear 28S rDNA (1,154 bp) (GenBank accession numbers in Table 2). Details of the DNA amplification protocols and primers were given by Rokas et al. (2002). The protein-coding genes (COI, LWRh, and EF1α) were easily aligned by eye. The ribosomal sequences (28S) differed in length, and some of the more variable regions were difficult to align manually. We used ClustalW 1.81 (Thompson et al., 1994) for this alignment. We applied a range of costs for the gap opening and gap extension penalties, and the individual alignments were subjected to parsimony bootstrap (Felsenstein, 1985) analyses using PAUP* (Swofford, 1998). Supported groups were largely congruent among the resulting trees. The alignment resulting from the use of the default settings in ClustalW is available from TreeBase (http://www.treebase.org, accession S970).

Parsimony Analyses of Morphological Data

For comparison with the Bayesian analysis, the morphological data set was subjected to parsimony analysis with equal character weights (standard parsimony) and with characters weighted according to degree of homoplasy using Goloboff's (1993) concave weighting function with the constant of concavity (k) set to 2 (implied weights parsimony). These analyses were completed using PAUP* (Swofford, 1998). A bootstrap majority rule consensus tree was calculated using 1,000 pseudoreplicates, each with five random addition sequences followed by tree bisection–reconnection (TBR) branch swapping and saving only one tree per pseudoreplicate (multrees = no).

Models of Character Evolution

For single DNA data partitions, we used standard substitution models: Jukes–Cantor (JC; Jukes and Cantor, 1969), Hasegawa–Kishino–Yano (HKY; Hasegawa et al., 1985), and general time reversible (GTR; Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990). Rates were either assumed to be equal or to vary across sites according to a gamma distribution (Γ; Yang, 1994) with or without a proportion of invariable sites (I; Gu et al., 1995). For morphological data, we used the Mk (Markov k) model of Lewis (2001a) extended to deal with ordered multistate characters and a new type of coding bias (only parsimony-informative characters scored) (Ronquist and Huelsenbeck, in prep.). The Mk model assumes equal state frequencies; it is possible to extend it to deal with unequal state frequencies, but we did not do so. The Mk model could be combined with equal or gamma-distributed rates across "sites" (i.e., characters) but not with a proportion of invariable "sites" because constant morphological characters were absent in the data matrix, making it impossible to estimate the proportion of invariable sites/characters.

We combined these elementary models in different ways to explore a range of composite models of different complexity (Table 3). First, we analyzed morphological and nucleotide data separately using 1-MkΓ and 1-GTRIΓ models (where 1 denotes a single data partition). Then we combined morphology and nucleotide data in models with either a single morphological and a single nucleotide partition (2-Mk-JC, 2-Mk-GTR, 2-MkΓ-HKYIΓ, and 2-MkΓ-GTRIΓ models) or a single morphological and four nucleotide partitions, each of the latter corresponding to a different gene (5-Mk-JC, 5-Mk-GTR, 5-MkΓ-HKYIΓ, and 5-MkΓ-GTRIΓ models). We allowed different data partition to evolve at a different rate, but branch lengths were assumed to be proportional across partitions (Yang, 1996b; Ronquist and Huelsenbeck, 2003). In the five-partition models, we also allowed the nucleotide models to be unique for each partition, i.e., we allowed stationary state frequencies and all other substitution model parameters to be independent across partitions. The number of free parameters in the examined evolutionary models ranged from 1 to 45 (Table 3).

TABLE 3. Summary of the models under which data were analyzed. The models contained one (either morphology or DNA), two (one morphology and one DNA), or five (one morphology and four DNA) data partitions. All parameters were allowed to be partition specific. The character-substitution models are given for the morphological and nucleotide (DNA) data. The number of free parameters is given excluding branch-length and topology parameters. Mk = Markov k model; JC = Jukes–Cantor model; HKY = Hasegawa–Kishino–Yano model; GTR = general time reversible model; I = proportion of invariant sites; Γ = gamma rate variation; m = rate multiplier.

Model	Partitions	Morphology model	DNA model	No. free parameters
1-MkΓ	1	Mk+Γ		1
1-GTRIΓ	1		GTR+I+Γ	10
2-Mk-JC	2	Mk, m	JC, m	1
2-Mk-GTR	2	Mk, m	GTR, m	9
2-MkΓ-HKYIΓ	2	Mk+Γ, m	HKY+I+Γ, m	8
2-MkΓ-GTRIΓ	2	Mk+Γ, m	GTR+I+Γ, m	12
5-Mk-JC	5	Mk, m	JC, m	4
5-Mk-GTR	5	Mk, m	GTR, m	36
5-MkΓ-HKYIΓ	5	Mk+Γ, m	HKY+I+Γ, m	29
5-MkΓ-GTRIΓ	5	Mk+Γ, m	GTR+I+Γ, m	45

Priors on Model Parameters

For the prior on topology, we assumed that all labeled trees are equally likely. We used an exponential prior with inverse scale parameter 1.0, Exponential(1.0), for branch lengths. For stationary state frequencies, we used a flat Dirichlet prior, Dirichlet(1,1,1,1). For the five nucleotide substitution rate ratios of the GTR model (scaled to the G-T rate) and the transition/transversion rate ratio of the HKY model, we used Exponential(ln 2) priors. These put 50% prior probability on rate ratios <1.0. We used a Uniform(0,50) prior on the shape parameter of the gamma distribution of rate variation and a Uniform(0,1) prior on the proportion of invariable sites.

Estimation of the Posterior Probability Distribution

We used Metropolis-coupled MCMC (Metropolis et al., 1953; Hastings, 1970; Geyer, 1991), as implemented in MrBayes 3.0 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), to estimate the posterior probability distribution. The gamma distribution of rate variation across sites was approximated by a discrete distribution with four categories, each category being represented by its mean rate. All chains, including coupled chains in the same run, were started from different, randomly chosen trees. Starting values for other parameters were set arbitrarily: branch lengths to 0.1, stationary base frequencies to 0.25, rates in rate sets to 1.0, invariant proportion to 0.0, and gamma shape to 0.5.

We randomly picked a combination of a parameter (or block of parameters) and an updating mechanism in each generation of the chain. The relative probability, or proposal rate, of picking each parameter–mechanism combination was determined by a probability factor associated with the updating mechanism. Assume that a model contained a tree updated by a mechanism with a proposal rate of 15.0 and two shape parameters updated by a mechanism with a proposal rate of 1.0. Then we

would update the tree with probability 15/17 and each shape parameter with probability 1/17 in every generation of the chain.

We used the following Metropolis proposals. Topology and branch lengths were updated using the LOCAL mechanism (proposal rate 15.0, tuning parameter 2.0*$\log_e$[1.1]; Larget and Simon, 1999), a stochastic TBR mechanism (proposal rate 3.0, tuning parameter 2.0*$\log_e$[1.61], extension probability 0.5), and a branch multiplier (proposal rate 3.0, tuning parameter 2.0*$\log_e$[2.59]). All other mechanisms were assigned a proposal rate of 1.0. Base frequencies were updated using a Dirichlet proposal (Dirichlet tuning parameter 300), rate ratios (transition/transversion rate ratio, GTR rate ratios scaled to the G-T rate, and partition rate multipliers scaled to the rate of the first partition) were updated using a uniform proposal within a sliding window (size 1.0), the gamma shape parameter was updated using a uniform proposal within a sliding window (size 0.5), and the proportion of invariable sites was updated using a uniform proposal within a sliding window (size 0.1).

We used four Metropolis-coupled chains with incremental heating, and the heating parameter was set to 0.2. In every generation, we randomly picked two chains and used a Metropolis mechanism to swap their states (Huelsenbeck and Ronquist, 2001). For each model, we ran four different runs of 2,000,000 generations each, sampling values every 100th generation.

Convergence Monitoring

For the initial determination of burn-in (the number of generations before apparent stationarity), we examined the plot of overall model likelihood against generation of the chain to find the point where the likelihood plot leveled off and started to fluctuate around a stable value. To provide additional confirmation of convergence and appropriate mixing, we compared results from the four independent runs. We checked that all runs had similar mean and variance of model likelihood after burn-in. We also compared posterior distributions and generation plots for all substitution model parameters, including total tree length, to check that the runs were producing similar marginal posterior distributions and that they were mixing appropriately over these distributions. We also compared majority rule consensus trees from the independent runs to check that topology and clade credibility values were similar. Convergence after the initially determined burn-in phase was confirmed in all cases by these additional tests. Final results were based on the pooled samples from the stationary phases of the four independent runs.

Estimation of Model Likelihood

The critical element that must be estimated to calculate Bayes factors is the model likelihood $f(X \mid M_i)$. This likelihood is usually impossible to evaluate analytically when the parameter space is large, but it can be estimated in a number of ways. In phylogenetics, different estimators have been used to calculate Bayes factors for comparing models. Suchard et al. (2001) used a method called the Savage–Dickey ratio (Verdinelli and Wasserman, 1995), which is restricted to comparing nested models. Aris-Brosou and Yang (2002) used a controversial method described by Aitkin (1991) that uses the arithmetic mean of the likelihood values sampled from the posterior distribution. We used the estimator proposed by Newton and Raftery (1994), which is the harmonic mean of the likelihood values sampled from the stationary phase of the MCMC run. This estimator is given by

$$\hat{f}(X \mid M_i) = \left[\frac{1}{n}\sum_{j=1}^{n} f(X \mid \tau_j, v_j, \theta_j)^{-1}\right]^{-1},$$

where $f(X \mid \tau_j, v_j, \theta_j)$ is the likelihood for a sample j out of a total of n from the joint posterior distribution. The harmonic mean estimator is less sensitive to the occasional occurrence of high likelihood values and more sensitive to low values than the arithmetic mean estimator of Aitkin (1991). Because high extremes are more likely to be a problem than low extremes, the harmonic mean should perform better than the arithmetic mean. Although some workers have questioned the general stability of the harmonic mean estimator, it should be sufficiently accurate for comparison of models with distinctly different model likelihoods given that the sample from the posterior distribution is large (Newton and Raftery, 1994; Gamerman, 1997). The value of the harmonic mean estimator was calculated using MrBayes, scaling all likelihood values to the smallest value sampled and taking full advantage of the numerical range of double-precision floating point values.

RESULTS

Computational Efficiency

Because many of the proposals changed parameters that affected only one or a few of the data partitions in a complex model, the computational time per MCMC cycle was shorter in partitioned analyses (Table 4). For all but the simplest model (Mk-JC), analyzing the data in five instead of two partitions provided a drastic increase in computational speed per generation, as expected. For example, the average computing speed was 20.0 generations/sec for the 45-parameter model 5-MkΓ-GTRIΓ but only 11.2 generations/sec for the 12-parameter model 2-MkΓ-GTRIΓ. Thus, analysis under the more parameter-rich model was nearly twice as fast as that under the less parameter-rich model, despite an almost fourfold increase in the number of parameters. Another striking effect is that allowing gamma-distributed rate variation across sites slows down the analysis considerably because the chain is integrating out the gamma distribution using the discrete four-category approximation, essentially quadrupling the number of computational operations needed to evaluate the likelihood ratio.

TABLE 4. Effect of model complexity on computational speed and time to apparent stationarity. All values are the average of four independent analyses on a 1.0 GHz AMD Athlon processor running Linux. Time to convergence (burn-in) was estimated as the number of generations required before the overall model likelihood reached apparent stationarity.

Model	No. free parameters	Speed (generations/sec)	Update cycle (sec)	Time to convergence (sec)[a]
Two-partition models				
2-Mk-JC	1	51.0	0.43	206 (157, 254)[b]
2-MkΓ-HKYIΓ	8	12.2	2.21	984 (820, 1,148)
2-Mk-GTR	9	47.0	0.51	553 (277, 851)[c]
2-MkΓ-GTRIΓ	12	11.2	2.41	2,410 (1,964, 3,125)
Five-partition models				
5-Mk-JC	4	51.3	0.43	171 (136, 214)
5-MkΓ-HKYIΓ	29	20.3	1.92	443 (394, 493)
5-Mk-GTR	36	60.1	0.50	237 (150, 333)
5-MkΓ-GTRIΓ	45	20.0	1.95	1,038 (650, 1,700)

[a]Values are mean (min, max) for the four independent runs.
[b]Based on only two runs; the remaining runs took 8,600 and 10,400 sec to converge.
[c]Based on three runs; the last run took 8,300 sec to converge.

A more relevant comparison than computational time per generation is the update cycle, i.e., the computational time required for a single update of all model parameters or parameter blocks. Because update mechanisms were selected randomly in the MCMC implementation, the actual update cycle for any parameter could vary considerably over short time spans. The times given here are the expectations over long runs (Table 4). Going from two to five data partitions leads to a slight but noticeable decrease in the update cycle in the five-partition models. This decrease could be due to improved caching efficiency because the small number of data required for many updates in the five-partition models might increase the probability of cache hits, leading to the processor accessing the data much faster. Without such effects, the update cycle would presumably have been similar regardless of the number of partitions. Again, the most striking effect is associated with allowing gamma-distributed rates across sites (Table 4).

Even though the update cycle remains constant, the increase in dimensionality of the posterior probability distribution might cause prolonged burn-in periods for the more complex models. However, this effect was not observed. On the contrary, the five-partition models reached apparent stationarity more quickly than did the corresponding two-partition models (Table 4). For instance, apparent stationarity occurred on average more than twice as fast in the 45-parameter 5-MkΓ-GTRIΓ model as in the 12-parameter 2-MkΓ-GTRIΓ model. Furthermore, convergence was unpredictable for the simplest models. For the 2-Mk-JC model, only two of four runs converged in a time period comparable to that of other models; the remaining two runs took 40–50 times longer to converge (Fig. 1). The slightly more complicated 5-Mk-JC model had one of four runs with a 20-time longer convergence than the others.

Morphology Versus DNA

For the comparison between the morphological and nucleotide data, we considered analyses under three models: 1-MkΓ (morphology only), 1-GTRIΓ (DNA only), and 2-MkΓ-GTRIΓ (all data combined).

Results from analysis of the morphological data under the 1-MkΓ model (Fig. 2a) are similar to those obtained from standard parsimony analysis of the same data (not shown), but they conform even more closely to the results from bootstrapped implied-weights parsimony analysis (constant of concavity, $k = 2$) (Fig. 2b). From this implied-weight analysis, the Bayesian results differ only in being slightly more resolved and having support values that are generally higher, especially for some of the weaker groupings in the implied-weights analysis. This observation supports the intuitive analogy between homoplasy-weighted parsimony analysis and parametric analysis allowing rate variation across sites.

Analysis of the molecular data under the 1-GTRIΓ model confirms some morphological results but suggests different groupings in many cases (Fig. 2c). There are two areas of particularly strong conflict. First, the molecular data suggest that the woody non-oak gallers (*Diplolepis*, *Pediaspis*, and *Eschatocerus*) form basal lineages in the Cynipidae instead of being closely related

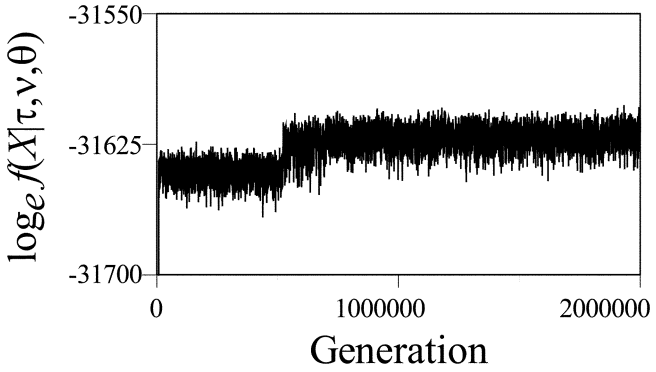

FIGURE 1. Generation plot of the marginal log likelihood, $f(X \mid \tau, v, \theta)$, for the combined data analyzed under the 2-Mk-JC model.

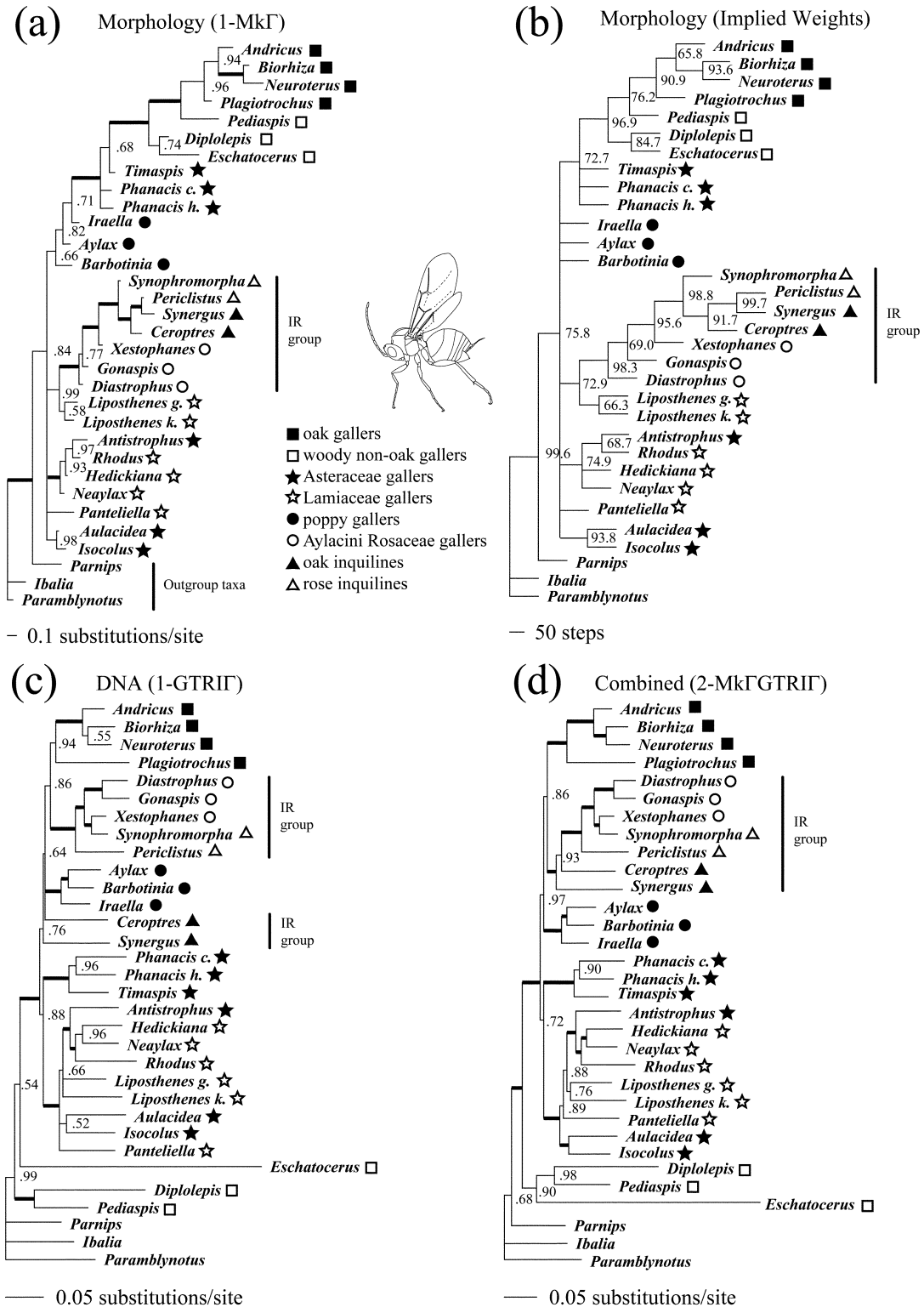

FIGURE 2. Comparison of the phylogenetic information in the morphological and molecular partitions and their influence on the combined analysis. Groups discussed in the text are indicated on the trees. A bold branch indicates a posterior probability of 1.0. IR group = inquilines + Aylacini Rosaceae gallers. (a) Majority-rule consensus tree based on the morphological data analyzed under the 1-MkΓ model. (b) Bootstrap consensus tree from an implied-weights parsimony analysis (constant of concavity, $k = 2$) of the morphological data. (c) Majority-rule consensus tree based on the molecular data analyzed under the 1-GTRIΓ model. (d) Majority-rule consensus tree based on the combined data analyzed under the 2-MkΓ-GTRIΓ model.

to the oak gallers (*Plagiotrochus*, *Andricus*, *Neuroterus*, and *Biorhiza*), as indicated by the morphological data (Figs. 2a, 2b). Second, the molecular data suggest that the oak inquilines (*Synergus* and *Ceroptres*) are not related to the rose inquilines (*Periclistus* and *Synophromorpha*) and the Aylacini Rosaceae gallers (*Diastrophus*, *Gonaspis*, and *Xestophanes*) (Fig. 2c), whereas the morphological data strongly support the monophyly of this entire assemblage, referred to here as the IR group (Inquilines + Aylacini Rosaceae gallers). In these cases, one can possibly argue that the morphological result is more likely than the molecular result (see Discussion). In most other cases where the molecules and morphology conflict, however, the molecular result is much more reasonable. For instance, the grouping of *Phanacis* and *Timaspis* is more in line with previous expectations and the results of standard parsimony analysis of morphological data (Liljeblad and Ronquist, 1998) than is the 1-MkΓ result; the placement of *Liposthenes* with other Lamiaceae gallers, such as *Panteliella* and *Rhodus*, appears more likely than the placement of this genus basal to the IR group as suggested by morphology; and the grouping of all poppy gallers (*Aylax*, *Iraella*, and *Barbotinia*) in a single clade is not in strong conflict with morphology and is in line with the observation that gall wasps in general are extremely conservative in their host-plant preferences (Ronquist and Liljeblad, 2001).

In the results of the combined analysis (Fig. 2d), there is ample evidence of influence from both the morphological (166 characters) and the molecular (3,080 bp) data. In many cases, the two data sets support each other, such as the increased support for the *Liposthenes* clade (0.58 with morphology, <0.50 with molecules, 0.76 in combined analysis), the *Aulacidea–Isocolus* clade (0.98 with morphology, 0.52 with molecules, 1.00 in combined analysis), and the oak gallers (*Plagiotrochus*, *Andricus*, *Neuroterus*, and *Biorhiza*; 0.96 with morphology, 0.94 with molecules, 1.00 in combined analysis). In many cases of conflict, the molecular result prevails, as expected. However, the morphological data are strong enough to change the position of the oak inquilines (*Synergus* and *Ceroptres*), such that the IR group becomes monophyletic in the combined analysis. This is quite a dramatic change compared with the molecular result (Fig. 2c).

Are Partitioned Models Better?

The estimated model likelihoods indicate a dramatic increase in model fit when going from two-partition models to their five-partition equivalents. The increase ranged from almost 600 to >900 log likelihood units (Tables 5, 6). However, other model components were even more important than allowing across-partition heterogeneity. Allowing within-partition rate variation was by far the most important model component (accounting for an increase of roughly 3,000 log likelihood units in the two available comparisons; Table 6). Allowing rate variation across molecular partitions but not within them (5-Mk-GTR) was far less successful than allowing rate variation within a single molecular partition (2-MkΓ-GTRIΓ); the difference was more than 2,000 log likelihood units in favor of the simpler model (Table 5). Allowing more realistic substitution models was also important, but the increase was dramatic only when going from JC to GTR (1,345 or 1,585 log likelihood units; Table 6), not when going from HKY to GTR (139 or 278 log likelihood units; Table 6), indicating that accounting for unequal base frequencies and unequal transition and transversion rates was more important than allowing all six substitution types to have their unique rate.

TABLE 5. Estimated model likelihood (predictive likelihood), $\hat{f}(X \mid M_i)$, for the different models.

Model	No. parameters	$\log_e \hat{f}(X \mid M_i)$
Two-partition models		
2-Mk-JC	1	−31,634
2-MkΓ-HKYIΓ	8	−27,396
2-Mk-GTR	9	−30,289
2-MkΓ-GTRIΓ	12	−27,121
Five-partition models		
5-Mk-JC	4	−30,962
5-MkΓ-HKYIΓ	29	−26,682
5-Mk-GTR	36	−29,377
5-MkΓ-GTRIΓ	45	−26,543

Do Complex Models Have Increased Topological Variance?

Complex models are associated with more topological uncertainty than are simple models. This is clearly seen when looking at the 95% credible sets of trees, i.e., the set obtained by starting with the sampled tree having the highest posterior probability and then adding trees in order of decreasing probability until the cumulative probability is 0.95. For instance, the simplest model (2-Mk-JC) had only three trees in its 95% credible set, whereas the most complex model (5-MkΓ-GTRIΓ) had 472 (Table 7). The most important factor by far is whether or not within-partition rate variation is accounted for in the model. Models allowing rate variation within partitions (8–45 parameters) had 236–472 trees in their 95% credible sets, whereas models that did not allow rate variation (1–36 parameters) had only 1–31 trees. Even if this effect is controlled for, however, the introduction of more parameters usually seems to increase topological uncertainty. For instance, the 5-Mk-JC model had 11 trees in its 95% credible set whereas the 5-Mk-GTR model had 31, and the 5-MkΓ-HKYIΓ model had 389 trees whereas the 5-MkΓ-GTRIΓ model had 472.

For models with rate variation, there also seems to be a positive correlation between topological uncertainty and tree length uncertainty (Table 7). The normalized SD of tree length (the SD divided by the mean) was 0.083 for the most complex of the rate-variation models (5-MkΓ-GTRIΓ) but only 0.041 for the simplest (2-MkΓ-HKYIΓ). There was also a tendency for more complex rate-variation models to have trees with longer branches. Models without rate variation showed no differences either in tree length or in tree length variation.

The majority rule consensus trees obtained under the eight different models were relatively similar (Figs. 3, 4). Of 29 possible clades, 18 were supported in all consensus

TABLE 6. Effect of various model components on model likelihood.

Model component	Model comparison (M_1/M_0)	Model likelihood		Bayes factor	
		$\log_e \hat{f}(X \mid M_1)$	$\log_e \hat{f}(X \mid M_0)$	$\log_e B_{10}$	$2\log_e B_{10}$
Data partition	5-Mk-JC/2-Mk-JC	−30,962	−31,634	672	1,344
	5-MkΓ-HKYIΓ/2-MkΓ-HKYIΓ	−26,682	−27,396	714	1,428
	5-Mk-GTR/2-Mk-GTR	−29,377	−30,289	912	1,824
	5-MkΓ-GTRIΓ/2-MkΓ-GTRIΓ	−26,543	−27,121	578	1,156
Rate variation	2-MkΓ-GTRIΓ/2-Mk-GTR	−27,121	−30,289	3,168	6,336
	5-MkΓ-GTRIΓ/2-Mk-GTR	−26,543	−29,377	2,834	5,668
Substitution model	2-Mk-GTR/2-Mk-JC	−30,289	−31,634	1,345	2,690
	5-Mk-GTR/5-Mk-JC	−29,377	−30,962	1,585	3,170
Substitution rates	2-MkΓ-GTRIΓ/2-MkΓ-HKYIΓ	−27,121	−27,396	275	550
	5-MkΓ-GTRIΓ/2-MkΓ-HKYIΓ	−26,543	−26,682	139	278

trees. An additional five were supported in nearly all trees, and there was more variation in the remaining six. The trees obtained under models allowing within-partition rate variation were identical with two exceptions: the 5-MkΓ-GTRIΓ tree (Fig. 4d) was unusual in grouping *Parnips* with *Paramblynotus* instead of with the Cynipidae, and the 2-MkΓ-GTRIΓ tree was unusual in placing the *Phanacis–Timaspis* clade more basally (Fig. 3d) than the other rate-variation trees did. Trees obtained under models with equal rates within partitions were more heterogeneous, differing among themselves in the placement of the poppy galler clade (*Iraella*, *Barbotinia*, and *Aylax*), the *Phanacis–Timaspis* clade, and the inquilines *Synergus* and *Ceroptres* (Figs. 3a, 3c, 4a, 4c). A characteristic feature of the equal-rates models was that they resolved the woody non-oak galling clade as (*Pediaspis*(*Diplolepis*, *Eschatocerus*)), partly supporting the morphology tree, whereas the rate-variation models resolved the clade as (*Eschatocerus*(*Diplolepis*, *Pediaspis*)). Except for this difference, the complex model trees were more congruent with the morphology tree than were the trees from simpler models (Figs. 2–4), even though many differences still persisted.

Differences between model extremes were quite striking for some taxa (Figs. 3, 4). For instance, *Synergus* groups with *Ceroptres* in a basal clade supported by a posterior probability of 1.0 under the simplest model (2-Mk-JC; Fig. 3a). In contrast, the posterior probability for *Synergus* + *Ceroptres* was only 0.02 under the most complex model (5-MkΓ-GTRIΓ; Fig. 4d), which placed both genera within a terminal IR clade supported by a posterior probability of 1.0. Under the simplest model, the posterior probability of the IR group being monophyletic was <0.00005. However, such extreme topological differences were uncommon among the more complex models.

Sensitivity to Superfluous Parameters

Among the models we examined, we had difficulties finding superfluous parameters, i.e., parameters for which the data were weak and the posterior distribution mainly reflected the prior. The parameters that came closest were the gamma shape and proportion of invariant sites for the two protein-coding nuclear gene fragments (EF1α and LWRh) in the five-partition models (5-MkΓ-HKYIΓ and 5-MkΓ-GTRIΓ). These fragments were short, many sites were constant, and there was little evidence of rate variation in the remaining sites. These factors led to a posterior probability distribution with a density throughout most of the parameter space for both

TABLE 7. Effect of model structure on tree uncertainty. Topological uncertainty increases when model complexity grows, as indicated by the increasing number of trees contained in the 95% and 99% credible sets and the decreasing average clade probability in the consensus tree. However, allowing rate variation has a much stronger effect than does the total number of parameters per se. Tree length uncertainty is positively correlated with increased uncertainty concerning tree topology for models allowing within-partition rate variation.

Model	No. Parameters	Credible tree sets		Mean support[a]	Tree length	
		95%	99%		Mean	NSD[b]
Two-partition models						
2-Mk-JC	1	3	6	0.971	1.935	0.016
2-MkΓ-HKYIΓ	8	236	797	0.953	3.727	0.041
2-Mk-GTR	9	1	9	0.994	1.923	0.016
2-MkΓ-GTRIΓ	12	378	1,032	0.946	2.835	0.033
Five-partition models						
5-Mk-JC	4	11	35	0.920	1.920	0.016
5-MkΓ-HKYIΓ	29	389	1,032	0.941	5.060	0.060
5-Mk-GTR	36	31	65	0.949	1.854	0.017
5-MkΓ-GTRIΓ	45	472	1,223	0.936	7.324	0.083

[a] Arithmetic mean of the posterior clade probabilities on the majority-rule consensus tree or, when this tree was not fully resolved (the 5-Mk-JC model), on a fully resolved tree based on the majority-rule consensus but with compatible groups included.

[b] Normalized SD.

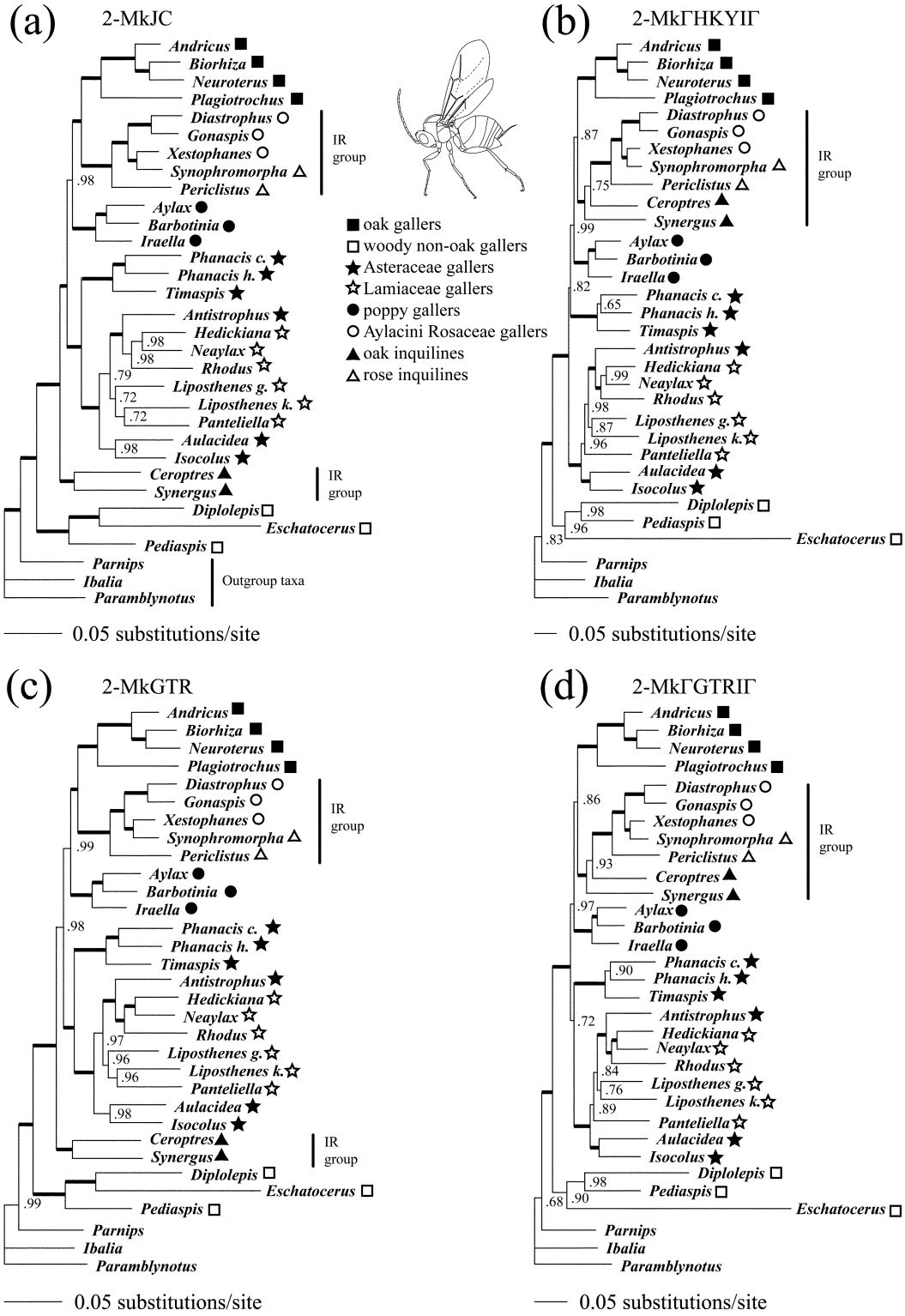

FIGURE 3. Majority-rule consensus trees based on the combined data analyzed under two-partition models. Groups discussed in the text are indicated on the trees. A bold branch indicates a posterior probability of 1.0. (a) 2-Mk-JC model (one free substitution model parameter). (b) 2-MkΓ-HKYIΓ model (eight parameters). (c) 2-Mk-GTR model (nine parameters). (d) 2-MkΓ-GTRIΓ model (12 parameters).

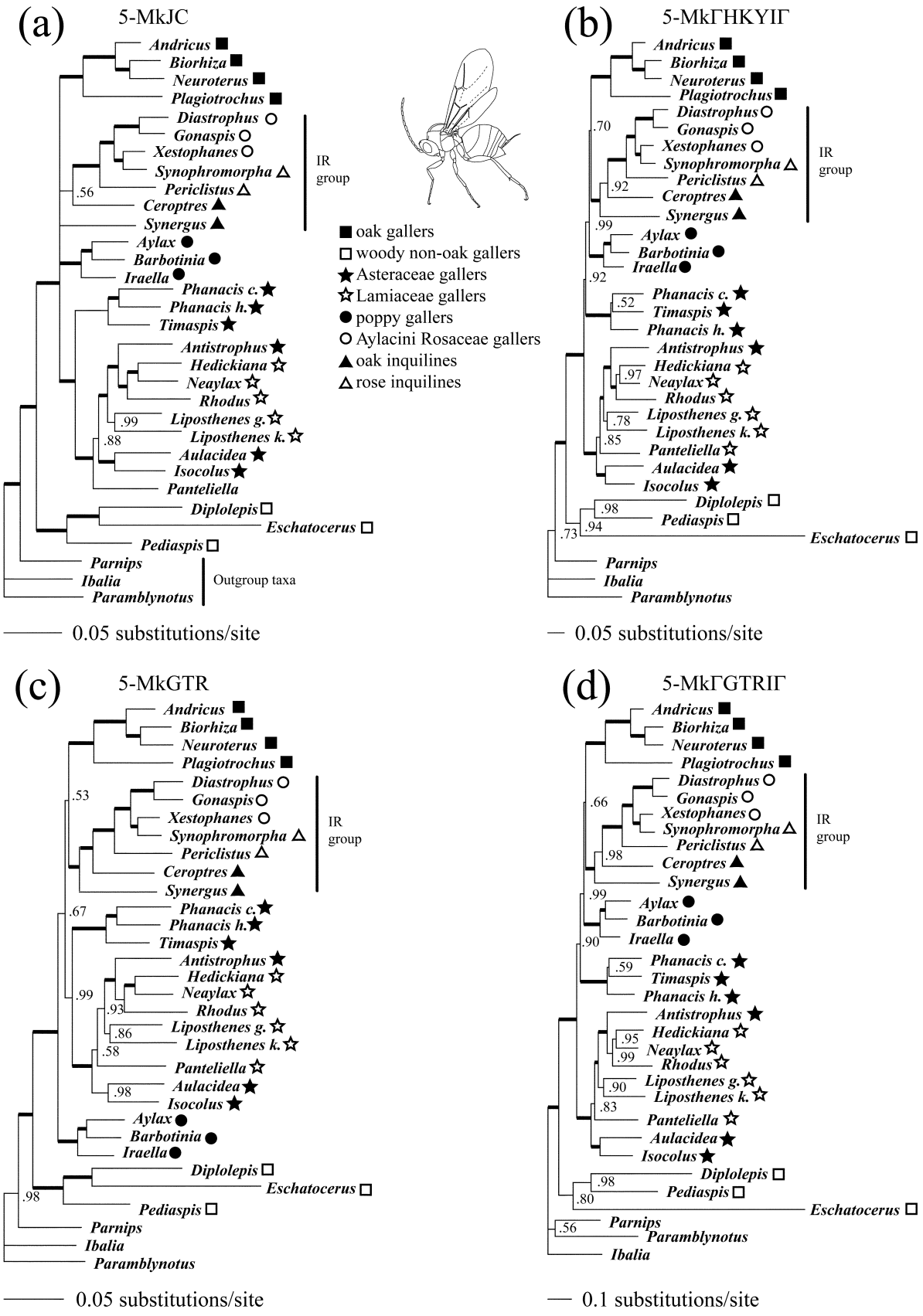

FIGURE 4. Majority-rule consensus trees based on the combined data analyzed under five-partition models. Groups discussed in the text are indicated on the trees. A bold branch indicates a posterior probability of 1.0. (a) 5-Mk-JC model (four free substitution model parameters). (b) 5-MkΓ-HKYIΓ model (29 parameters). (c) 5-Mk-GTR model (36 parameters). (d) 5-MkΓ-GTRIΓ model (45 parameters).

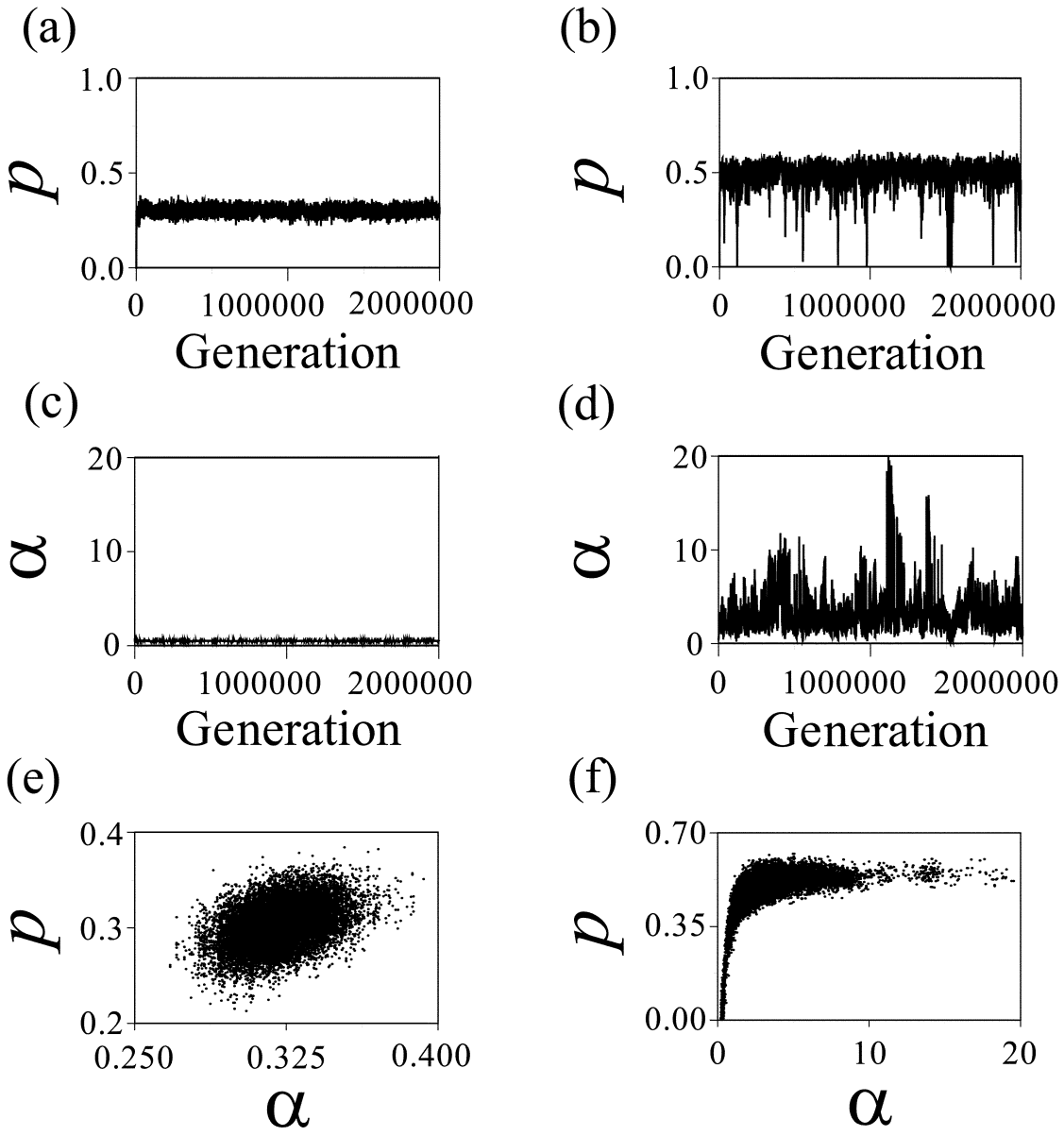

FIGURE 5. Generation plots and correlations of some of the parameters in the 5-MkΓ-GTRIΓ model. (a) Proportion of invariant sites (p) for the COI partition. (b) Proportion of invariant sites (p) for the LWRh partition. (c) Gamma shape (α) for the COI partition. (d) Gamma shape (α) for the LWRh partition. (e) Correlation plot between (α) and p for the COI partition. (f) Correlation plot between α and p for the LWRh partition.

the proportion of invariants and gamma shape (Figs. 5b, 5d). This result was in stark contrast with the focused distributions seen for the other two genes in the five-partition models or for all genes combined in the two-partition models. When examined closely, the diffuse distributions appear to be due to a concentration of the posterior density to two opposite combinations of parameter values: either the proportion of invariant sites is high and the rate variation moderate (high α), or the proportion of invariant sites is low and the rate variation considerable (low α). The correlation between the parameters is obvious when comparing the correlation plots between proportion of invariant sites and gamma shape for the COI and LWRh partitions (Figs. 5e, 5f). The marginal distributions of gamma shape and proportion of invariant sites show that the posterior distribution is highly peaked in the region with a high proportion of invariant sites and moderate rate variation (high α) even though there is a significant tail expanding into the region with the reverse parameter combination (Fig. 6). The long tails in the marginal posterior distributions led to slower mixing than typically observed, although the posterior distribution appeared to be adequately covered over the entire length of the run. The slower mixing did

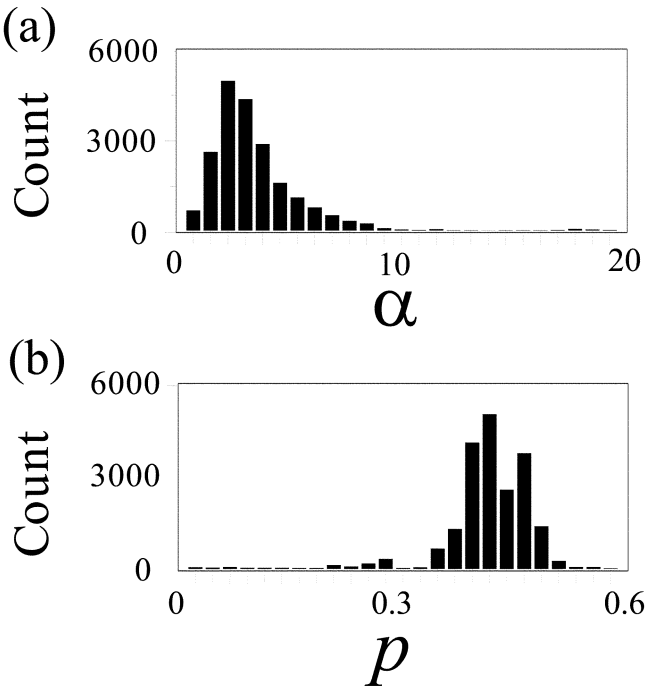

FIGURE 6. Marginal posterior distributions of two parameters of the 5-MkΓ-GTRIΓ model, for which the mixing of the chain is relatively slow. (a) Gamma shape (α) for the LWRh partition. (b) Proportion of invariant sites (p) for the LWRh partition.

not appear to affect convergence and mixing for other parameters. For instance, the marginal posterior distributions of gamma shape and proportion of invariants for the other gene fragments (COI and 28S) remained focused, and the chain mixed rapidly over them (Figs. 5a, 5c). Apparent convergence of overall model likelihood also remained unaffected: the number of update cycles to convergence was roughly the same as those for the simpler five-partition models (cf. Table 4).

We also found some examples of parameter sets for which the marginal posterior distributions broadly coincided with a submodel. The best examples were the six substitution rates of the GTR model for the nuclear protein-coding gene fragments in the most complex five-partition model. For both partitions, the posterior distribution suggested that the rates fell into two distinct classes, transitions and transversions, with only moderate variation within classes (Fig. 7). Thus, these partitions might have been more appropriately analyzed using an HKY model. However, the use of the complex GTR model did not seem to cause problems with overall convergence or mixing of other model parameters. Furthermore, going from a five-partition HKY to a five-partition GTR model had relatively little influence on tree uncertainty (cf. Table 7).

Avoiding Overparameterization

In general, Bayes factor comparisons favored the model with the highest number of parameters (Table 8), but there were two exceptions. The second most parameter-rich model (5-Mk-GTR), having 36 parameters, lost to both of the two-partition models allowing within-partition rate variation, one with eight parameters (2-MkΓ-HKYIΓ) and the other with 12 parameters (2-MkΓ-GTRIΓ). Again, this illustrates the importance of modeling rate variation (cf. Table 6) but also shows that Bayesian model selection can help the investigator avoid an inappropriate sequence of parameter addition. However, in all comparisons of nested models, the more complex model was favored; twice the log of the Bayes factor ranged from 279 to 10,183, far above the critical threshold of 10 required for the Bayes factor to be considered very strong evidence in favor of the better model (Table 1). The best model overall (5-MkΓ-GTRIΓ) was also the most parameter-rich of the models we tried.

To further test whether the Bayes factor might favor model reduction in some cases, we examined the weakest parameters in the most complex model. We estimated the model likelihood for six additional models obtained by reduction from the 5-MkΓ-GTRIΓ model: (1) proportion of invariant sites for the EF1α partition was removed; (2) gamma shape for the EF1α partition was removed; (3) proportion of invariant sites for the LWRh partition was removed; (4) gamma shape for the LWRh partition was removed; (5) an HKY model was used instead of a GTR model for the EF1α partition; and (6) an HKY model was used instead of a GTR model for the LWRh partition. In all these cases, the model likelihood decreased; the decrease ranged from 6 to 17 log likelihood units. Thus, Bayes factor comparisons still provided strong support (Table 1) for the inclusion of all of these parameters.

TABLE 8. Bayes factor comparisons between all models. Entries are twice the log of the Bayes factor in the comparison between models M_0 and M_1 ($2\log_e B_{10}$). The column models are arbitrarily labeled M_0; thus, positive values indicate support for the row model over the column model. Underlined entries indicate comparisons where the less parameter-rich model is favored by the Bayes factor.

	2-Mk-JC	2-MkΓ-HKYIΓ	2-Mk-GTR	2-MkΓ-GTRIΓ	5-Mk-JC	5-MkΓ-HKYIΓ	5-Mk-GTR	5-MkΓ-GTRIΓ
2-Mk-JC	0							
2-MkΓ-HKYIΓ	8,477	0						
2-Mk-GTR	2,691	−5,786	0					
2-MkΓ-GTRIΓ	9,026	549	6,335	0				
5-Mk-JC	1,344	−7,133	−1,347	−7,682	0			
5-MkΓ-HKYIΓ	9,905	1,428	7,214	878	8,560	0		
5-Mk-GTR	4,515	−3,962	1,824	−4,512	3,170	−5,390	0	
5-MkΓ-GTRIΓ	10,183	1,706	7,493	1,157	8,839	279	5,669	0

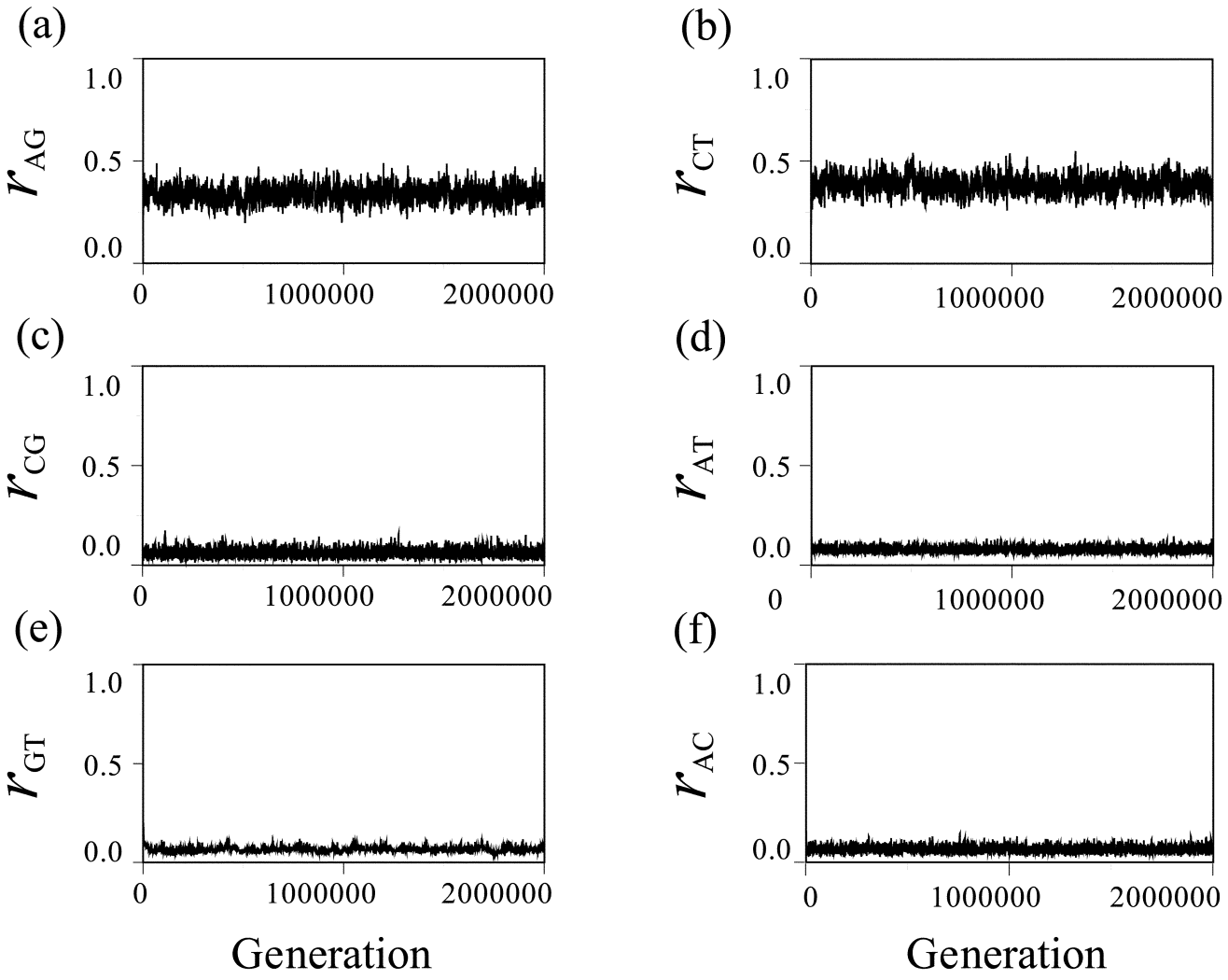

FIGURE 7. Generation plots of the substitution rates for the EF1α partition under the 5-MkΓ-GTRIΓ model. The rates are given as proportions of the rate sum. (a, b) The two transition rates, r_{AG} and r_{CT}. (c–f) The four transversion rates, r_{CG}, r_{AT}, r_{GT}, and r_{AC}.

DISCUSSION

Computational Feasibility of Combined Analysis

Our results demonstrate the computational efficiency of the Bayesian MCMC approach in dealing with composite models for heterogeneous data sets. The update cycle is virtually unaffected by model partitioning, and partitioned models reach apparent stationarity faster and more reliably than do unpartitioned models (Table 4). The occasional failure of two of the simplest models to converge rapidly suggests that the shape of the posterior distribution might be more important than its dimensionality in determining convergence. In other words, the simplest models seem to be associated with posterior distributions that have a very complex shape, which occasionally traps the Markov chains on their way to the region of high posterior density, even though Metropolis coupling is used to accelerate convergence and escape local maxima (Fig. 1). This phenomenon could be coupled with the low topological variance in the posterior probability distribution of the simplest models. Adjustments of the tuning parameters of the MCMC runs may have sped up convergence for the simplest models, but it is unlikely that simple fine tuning would have alleviated the problem completely.

The most complex models included both parameters with relatively diffuse posterior distributions and parameter sets whose posteriors nearly coincided with a submodel. Mixing was slower for parameters with diffuse posterior distributions (Figs. 5b, 5d), but the posterior remained focused and mixing was rapid for other parameters (Figs. 5a, 5c), suggesting that the Bayesian MCMC approach is robust to the inclusion of a modest number of weak parameters in models.

Despite these encouraging results, there must be a limit to the number of parameters that can be successfully included in a model even under Bayesian MCMC analysis. Our preliminary observations from analysis under a

radically overparameterized model with 12 partitions and 121 substitution model parameters indicate that this is indeed the case. Many of the parameters in this model had diffuse posterior distributions, and the chain did not reach stationarity or mix rapidly enough over all these distributions to provide a reasonable sample during the run length used for the other models (2,000,000 generations) (Figs. 8a, 8b). The overall likelihood reached apparent stationarity rapidly (Fig. 8c) and the chain seemed to sample successfully over some marginal posterior distributions that remained focused (Fig. 8d). However, the estimates of overall likelihood and the samples of the focused model parameters should be regarded with caution in the lack of appropriate indications of convergence and adequate mixing for the remaining model parameters. The results obtained under this model illustrate well how important it is to monitor convergence and mixing for all model parameters; it is not sufficient to look just at the overall likelihood (Fig. 8c).

Because theoretical considerations predict that the update cycle will be unaffected by model partitioning, we expect our results concerning computational speed to apply generally. However, the convergence and mixing behavior is likely to be influenced by the peculiarities of each data set and the details of the MCMC run; therefore, more data are needed before it can be safely concluded that convergence is generally faster for moderately complex models than for simple models.

Combining Morphology and Molecules

Previously, morphology has been ignored in parametric inference of phylogeny for several reasons. Most importantly, there has been skepticism directed toward the appropriateness of probabilistic models for morphological data, and only recently have such models been considered seriously (Lewis, 2001a; Ronquist and Huelsenbeck, in prep.). However, the slow progress in the development of parametric methods for morphology may also be attributed to a widely held belief that morphology would contribute little to parametric inference of phylogeny. Our results clearly show that this is not necessarily the case, not even when multigene data sets are available. The morphological data contributed <5% of the characters in our data set but still had significant influence on the tree from the combined analysis. Of course, the influence of a morphological data set will depend on its size and the strength of the phylogenetic signal in it. Although our data set was fairly large and did provide strong support for several groupings,

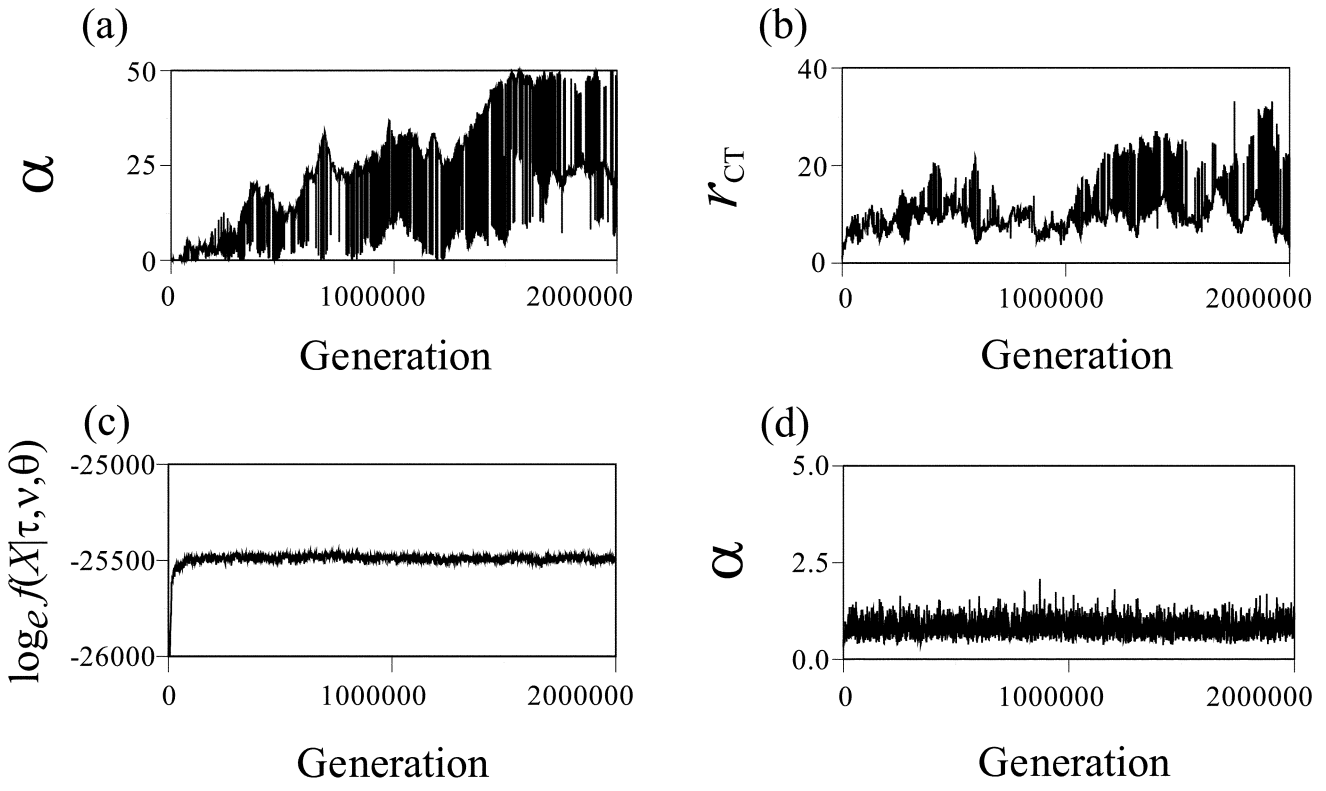

FIGURE 8. MCMC analysis under a radically overparameterized model with 121 free substitution model parameters (12-M$\kappa\Gamma$-GTRIΓ). (a) Generation plot of the gamma shape parameter (α) for the EF1α partition (first codon positions only). (b) Generation plot of the C-T rate parameter (r_{CT}) for the COI partition (first codon positions only). (c) Generation plot of the marginal log likelihood, $f(X \mid \tau, v, \theta)$. (d) Generation plot of α for the COI partition (first codon positions only).

several parts of the phylogeny were not well resolved, so this data set is not extremely "clean." Even some of the weaker signals in our morphological data, such as the support for the monophyly of the genus *Liposthenes*, did show through in the combined analysis (Fig. 2d), which suggests that there is generally a good potential for morphological phylogenetic signal to contribute to the result of a combined statistical analysis.

Although statistical analysis of morphological phylogenies is controversial, the difference compared with parsimony analysis may not be so dramatic in practice. Our comparisons between parametric and standard/implied-weights parsimony analysis of the same morphological data actually indicate that the two methods tend to give similar phylogenetic results (Figs. 2a, 2b).

As molecular data sets increase in size, morphological data will have less and less influence on the results of a combined analysis. However, conflicts between morphological and molecular signal could still contribute important information, either about shortcomings in molecular models or about interesting aspects of morphological evolution. Furthermore, morphological data will remain important for placing fossils and thus for dating splitting events in phylogenies. We hope that these reasons are sufficient to make combined analysis of morphology and molecules, where applicable, a common practice in statistical phylogenetics.

Recognizing Across-Partition Heterogeneity in Molecular Data

Our results indicate that evolutionary models for multigene data sets can be improved considerably by recognizing across-partition heterogeneity in model parameters such as overall rate, individual substitution rates, base frequencies, and gamma shape. This improvement in fit associated with partitioned models has been noted before in the ML framework (Yang, 1996b; DeBry, 1999; Pupko et al., 2002). However, even though across-partition heterogeneity is significant, other model components seem to be even more important, particularly those that deal with within-partition rate variation (for similar conclusions, see Wakely, 1994; Sullivan et al., 1995; Yang, 1996a) and some of the substitution rate and base frequency parameters. Until these model components have been accounted for, it might not be worthwhile considering partitioned models. We suspect that these conclusions are valid for most of the combined data sets used in phylogenetic analysis, but only explicit model comparison for each data set can guarantee that it conforms to the general pattern.

Priors in Complex Models

So far, there has been little discussion about priors in Bayesian phylogenetic inference. Rather than noninformative, many of the commonly used priors have been counterinformative in the sense that they put a lot of probability on unlikely parameter values. For instance, most workers would consider a tree with any branch length above 1.0 as suspect. Across a branch of that length, we would expect one change in every character and such massive amounts of change would make it difficult to even recognize homology between molecular sequences. However, it has been common to associate branch lengths with uniform prior distributions from 0 to a large, arbitrary cutoff value. Such priors place considerable emphasis on trees with long branches. Consider for instance a uniform(0,10) prior on branch lengths for a 30-taxon problem with $2 * 30 - 3 = 57$ branches: such a prior places less than $(1/10)^{57} = 10^{-57}$ probability on trees with all branches shorter than 1.0. When the data are strong enough, even such an extreme prior will have negligible impact on the posterior distribution. One could even argue that if the data overrode an extreme prior, it would provide additional evidence for the inference based on the posterior distribution.

However, selecting appropriate priors becomes more important when dealing with complex models. Parameter-rich models inevitably contain fewer data per parameter and it may be difficult to exclude all parameters for which the posterior distribution will reflect the prior to some extent. If appropriate priors are chosen, one can hope that weak-data parameters do not cause problems with the rest of the analysis, but an extreme prior showing through would destroy the credibility of the results. As expected, pilot runs conducted when this study was initiated showed several cases where counterinformative or extreme priors affected posterior distributions. This was particularly the case for uniform priors on branch lengths and uniform priors on substitution rate ratios of the GTR model. For this reason, we used exponential priors on both of these parameters in the final runs, such that more prior probability was put on moderate parameter values.

Thus, it is important to avoid extreme priors and to monitor convergence and mixing for all model parameters to discover potential problems with weak data when analyzing complex models using the Bayesian MCMC approach.

Model Sensitivity and Topological Incongruence

Recent computer simulations have shown that with certain types of extreme model deviations, Bayesian inference can result in excessively high posterior probabilities for clades (Suzuki et al., 2002). The incongruence between the trees obtained under different models in our study (Figs. 3, 4) also illustrates the model sensitivity of the Bayesian approach. However, the incongruence only affected a small number of taxa; about two thirds of the clades remained constant under all models. Furthermore, model sensitivity of the topology decreased, very high posterior clade probabilities became less common, and conflict between morphology and molecules was weakened as model complexity increased. It is a matter of speculation whether even more complex models, for instance allowing process heterogeneity across the tree, might continue this trend, but the Bayesian MCMC approach certainly seems to encourage fruitful exploration

of this question in the future. In any case, preliminary simulation results and observations from other data sets suggest that there is indeed a general tendency for oversimplified models to be associated with excessive credibility in topologies that may not be correct.

A much-debated issue is whether different types of data should be combined in the first place (e.g., de Queiroz et al., 1995), and a number of tests have been used to address this question (Huelsenbeck and Bull, 1996; Cunningham, 1997; Barker and Lutzoni, 2002; Buckley et al., 2002; Dowton and Austin, 2002). A Bayes factor test of combinability is a straightforward extension of the partitioning techniques used here. The integrated likelihood of a model in which the morphological and molecular partition had unlinked topologies is compared with that of a model in which these partitions had linked topologies but unlinked branch lengths. The only new component needed for this type of analysis is an MCMC sampler that will change a topology with two independent sets of branch lengths.

As long as we do not know whether current models are missing important components, however, it is difficult to provide an appropriate answer to the combinability question. In the context of the models explored here, it seems likely that a formal Bayes factor comparison would suggest that our morphological and molecular data partitions evolved on different topologies. Nevertheless, we would be inclined to interpret this suggestion as the result of model imperfection rather than true mismatch between morphological and molecular trees. The evolutionary explanations for topological incongruence seem relatively far-fetched in comparison with the possible model components that might be missing. Examples of the latter include morphological convergence, which might explain why morphology groups the two clades of gall inducers on woody hosts (the oak gallers and the woody non-oak gallers) but the molecules do not, and molecular process heterogeneity across the tree, which might explain why the three longest terminal branches (the woody non-oak gallers) appear basally in the molecular tree but not in the morphological tree. Potentially, both of these hypotheses can be addressed with appropriate stochastic models. Until this is done, it seems premature to consider explanations involving processes causing topological incongruence.

Bayesian Model Selection

Thus far, model selection in statistical phylogenetics has usually been based on criteria comparing ML estimates (for exceptions, see Suchard et al., 2001; Aris-Brosou and Yang, 2002; Bollback, 2002). Our results illustrate several advantages of adopting a Bayesian approach to model comparison instead. The comparison is based on the model likelihood, which can be conveniently estimated using the harmonic mean of the likelihood values from the MCMC samples. Despite the warnings issued in the literature concerning the sensitivity of the harmonic estimator, we observed the estimate to be reasonably stable among independent MCMC runs. The maximum range of the estimate among the four independent runs for any of the 18 different models analyzed was seven log likelihood units, which was negligible in comparison with the differences among models.

Unlike standard likelihood ratio tests, Bayesian model comparison integrates over uncertainty in model parameters. This integration makes the result more generally valid because it does not depend on a single topology or a particular set of ML point estimates of model parameters. An additional advantage of the Bayesian approach is that it allows comparison of nonnested models, unlike likelihood ratio tests.

Our results show that Bayesian model comparison can favor simple models over much more complicated ones. In the most extreme example, an 8-parameter model (2-MkG-HKYIG) was preferred over a 36-parameter model (5-Mk-GTR). Preference for a simpler model only occurred in comparisons between nonnested models among the results reported here (Table 8).

It is difficult to answer the general question of whether Bayesian model selection strikes a reasonable balance between model complexity and estimation error. Among the models we tried, the best one was also the most parameter rich. It was difficult to find components in this model (5-MkΓ-GTRIΓ) that were not well justified. A couple of parameters had diffuse or strongly skewed marginal posterior distributions, i.e., the gamma shape and the proportion of invariants for the two small nuclear protein-coding gene fragments (EF1α and LWRh), and some parameter sets had posterior distributions largely coinciding with a submodel, i.e., the substitution rates of the GTR model for the same gene fragments. However, the model likelihood decreased when any of these parameters were excluded from the model, indicating that they were important. Albeit the decrease seemed small in relation to the total model likelihood, it was still sufficient to indicate strong support for the inclusion of these parameters (cf. Table 1). We experienced problems with convergence and mixing in models that appeared to be radically overparameterized and are therefore not in a position to say whether Bayesian model selection criteria would have favored exclusion of some parameters from these models. Practical problems of obtaining convergence and mixing may turn out to be a more severe constraint on model complexity than are Bayesian model selection criteria per se.

We are confident that likelihood ratio tests would have given results similar to those of the Bayesian selection criteria for the nested model comparisons. Standard hLRTs, the AIC, and the BIC applied to the four individual molecular partitions all suggested models for each partition similar or identical to the ones in the 5-MkΓ-GTRIΓ model (unpubl. data). Adding ML values across partitions and estimating the morphology partition ML using extremely long MCMC runs indicated that applying any of these selection criteria to the combined data and the range of models examined here would have resulted in the same preferred model as did the Bayesian approach. It seems likely that this will be the general pattern: Bayesian and likelihood approaches will give

similar results. Thus, in practice the main advantage of the Bayesian model selection approach is its computational convenience and its applicability to all types of model comparisons.

Even though Bayes factors provide good orientation tools in the selection among models, many aspects of model choice still will have to depend on the judgment of the investigator. The range of models to be tested must be determined by the investigator, but it is difficult to avoid some arbitrariness in this step. For instance, we decided to test models in which the partitions corresponded to genes, but other ways of partitioning the data might have been considered, such as lumping the three protein-coding genes and then dividing these sites into two or three partitions based on codon position. Alternatively, we could have considered codon models instead of single-nucleotide models for the protein-coding genes. One of the most exciting features of the Bayesian approach is that it allows investigators to examine a larger set of models than has been possible previously, and one can foresee future Bayesian implementations that systematically walk through a large space of predefined models either in sequences of runs on individual models or in runs simultaneously averaging across all available models (Suchard et al., 2001).

Phylogeny and Evolution of Gall Wasps

In many ways, the combined analyses presented here support previous conclusions concerning the phylogeny and evolution of gall wasps based purely on morphological data (Ronquist, 1994; Liljeblad and Ronquist, 1998; Ronquist and Liljeblad, 2001). For instance, these data confirm the rarity of host-plant shifts, the presence of three unrelated Asteraceae galling clades (*Aulacidea* + *Isocolus*, *Antistrophus*, and *Phanacis* + *Timaspis*), the distinctness of the *Diastrophus* lineage from other gallers of woody host plants, and the close relationship between the Aylacini Rosaceae gallers and the inquilines (the IR clade).

However, there are also two major conclusions from previous work that are not upheld in this analysis. First, the inquilines do not form a monophyletic group; instead, the current results suggest that the IR clade has the inquilines as basal lineages and the gallers (*Diastrophus* + *Gonaspis* and *Xestophanes*) nested deeply inside. Although this set of relationships is certainly possible, it would imply a rather complicated history of shifts between gall induction and inquilinism within a small cynipid subclade, which is unlikely given that there has not been a single shift between these life strategies in other parts of the gall-wasp phylogeny. The result is also surprising in view of the fact that the inquilines form one of the most well-supported clades in the morphological analysis (Figs. 2a, 2b).

The second surprising result is the grouping of three gallers of woody non-oak hosts (*Pediaspis*, *Diplolepis*, and *Eschatocerus*) outside the rest of the Cynipidae (Figs. 2–4). These genera have traditionally been regarded as being close to the oak gallers (e.g., Kinsey, 1920; Weidner, 1968), and this conclusion has been supported by morphological data, albeit not as strongly as inquiline monophyly. It is of course possible that the morphological result is due to convergence among gall inducers attacking woody host plants and that the woody non-oak gallers are indeed basal in the Cynipidae phylogeny, as suggested in the present analysis. However these three terminal taxa also have the longest terminal branches in the molecular phylogeny (Fig. 2c), which might indicate problems in the analysis of the molecular data.

These and other questions pertaining to gall-wasp phylogeny and evolution will be examined in more detail elsewhere in the context of a phylogenetic analysis based on a larger taxon sample (Nylander et al., in prep.).

ACKNOWLEDGMENTS

We are grateful to the following people for supplying and identifying specimens and for assistance in field work: James Cook, Johan Liljeblad, Zhiwei Liu, Göran Nordlander, Juli Pujade, Antonis Rokas, Kathy Schick, Mikael Sporrong, Graham Stone, Zoe Tsekoura, and Hege Vårdal. Afsaneh Ahmadzadeh gave valuable help with laboratory work. Computational facilities were provided by Mikael Thollesson and the Linnaean Center of Bioinformatics, Uppsala. Chris Simon, Thomas Buckley, David Posada, and an anonymous reviewer gave valuable comments that improved the manuscript. This research was supported by a grant to J.A.A.N from the Helge Ax:on Johnson Foundation, a Swedish Research Council grant to F.R. (621-2001-2963), NSF grants to J.P.H. (DEB-0075406 and MCB-0075404), and funding from the Spanish Ministry of Science and Technology, research project REN2002-03518, to J.L.N-A.

REFERENCES

Aitkin, M. 1991. Posterior Bayes factors. J. R. Stat. Soc. B 53:111–142.

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* Second International Symposium on Information Theory (B. N. Petrov and F. Csaki, eds.). Akademiai Kiado, Budapest.

Aris-Brosou, S., and Z. Yang. 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Syst. Biol. 51:703–714.

Askew, R. R. 1984.The biology of gallwasps. Pages 223–271 *in* Biology of gall insects (T. N. Anantakrishnan, ed.). Edward Arnold, London.

Barker, F. K., and F. M. Lutzoni. 2002. The utility of the incongruence length difference test. Syst. Biol. 51:625–637.

Bartlett, M. S. 1957. A comment on D. V. Lindley's statistical paradox. Biometrika 44:187–192.

Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.

Buckley, T. R., P. Arensburger, C. Simon, and G. K. Chambers. 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. Syst. Biol. 51:4–18.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference, a practical information-theoretic approach, 2nd edition. Springer, New York.

Cunningham, C. W. 1997. Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. 14:733–740.

DeBry, R. W. 1999. Maximum likelihood analysis of gene-based and structure-based process partitions, using mammalian mitochondrial genomes. Syst. Biol. 48:286–299.

de Queiroz, A., M. J. Donoghue, and J. Kim. 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. 26:657–681.

Dowton, M., and A. D. Austin. 2002. Increased congruence does not necessarily indicate increased phylogenetic accuracy—The behavior of the incongruence length difference test in mixed-model analyses. Syst. Biol. 51:19–31.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.
Gamerman, D. 1997. Markov chain Monte Carlo: Stochastic simulation for Bayesian inference. Chapman and Hall, London.
Gelfand, A. E., and D. K. Dey. 1994. Bayesian model choice: Asymptotics and exact calculations. J. R. Stat. Soc. B 56:501–514.
Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood. Pages 156–163 *in* Computing science and statistics: Proceedings of the 23rd Symposium on the Interface (E. M. Keramidas, ed.). Interface Foundation, Fairfax Station.
Goloboff, P. A. 1993. Estimating character weights during tree search. Cladistics 9:83–91.
Gu, X., Y.-X. Fu, and W.-H. Li. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546–557.
Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21:160–174.
Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.
Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: Traditional and Bayesian approaches. Nat. Genet. 4:275–284.
Huelsenbeck, J. P., and J. J. Bull. 1996. A likelihood-ratio test to detect conflicting phylogenetic signal. Syst. Biol. 45:92–98.
Huelsenbeck, J. P., and K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28:437–466.
Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.
Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. Biometrics 17:754–755.
Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.
Jeffreys, H. 1961. Theory of probability, 3rd edition. Oxford Univ. Press, Oxford, U.K.
Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. M. Munro, ed.). Academic Press, New York.
Kass, R. E., and A. E. Raftery. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795.
Kinsey, A. C. 1920. Phylogeny of cynipid genera and biological characteristics. Bull. Am. Mus. Nat. Hist. 42:357–402.
Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20: 86–93.
Larget, B., and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.
Lavine, M., and M. J. Schervish. 1999. Bayes factors: What they are and what they are not. Am. Stat. 53:119–122.
Lewis, P. O. 2001a. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.
Lewis, P. O. 2001b. Phylogenetic systematics turns over a new leaf. Trends Ecol. Evol. 16:30–37.
Liljeblad, J., and F. Ronquist. 1998. A phylogenetic analysis of higher-level gall wasp relationships (Hymenoptera: Cynipidae). Syst. Entomol. 23:229–252.
Lindley, D. V. 1957. A statistical paradox. Biometrika 44:187–192.
Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.
Newton, M. A., and A. E. Raftery. 1994. Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). J. R. Stat. Soc. B Met. 56:3–48.
Nieves-Aldrey, J. L. 2001. Hymenoptera, Cynipidae. Pages 1–636 *in* Fauna Ibérica, Volume 16 (M. A. Ramos, ed.). Museo Nacional de Ciencias Naturales, CSIC, Madrid.
Posada, D., and K. A. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.
Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002. Combining multiple data sets in a likelihood analysis: Which models are the best? Mol. Biol. Evol. 19:2294–2307.
Rodríguez, J., L. Oliver, A. Marín, and R. Medina. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142:485–501.
Rokas, A., J. A. A. Nylander, F. Ronquist, and G. N. Stone. 2002. A maximum-likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera: Cynipidae): Implications for insect phylogenetic studies. Mol. Phylogenet. Evol. 22:206–219.
Ronquist, F. 1994. Evolution of parasitism among closely related species: Phylogenetic relationships and the origin of inquilinism in gall wasps (Hymenoptera, Cynipidae). Evolution 48:241–261.
Ronquist, F. 1999. Phylogeny, classification and evolution of the Cynipoidea. Zool. Scr. 28:139–164.
Ronquist, F., and J. P. Huelsenbeck. 2003. MrBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Ronquist, F., and J. Liljeblad. 2001. Evolution of the gall wasp–host plant association. Evolution 55:2503–2522.
Sanderson, M. J., and J. Kim. 2000. Parametric phylogenetics? Syst. Biol. 49:817-829.
Schwartz, G. 1978. Estimating the dimensions of a model. Ann. Stat. 6:461–464.
Stone, G. N., K. Schönrogge, R. Atkinson, D. Bellido, and J. Pujade-Villar. 2002. The population biology of oak gall wasps (Hymenoptera, Cynipidae). Annu. Rev. Entomol. 47:633–668.
Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.
Sullivan, J., K. E. Holsinger, and C. Simon. 1995. The effect of topology on estimates of among-site rate variation. J. Mol. Evol. 42:308–312.
Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99:16138–16143.
Swofford, D. L. 1998. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer, Sunderland, Massachusetts.
Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.
Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic. Acids. Res. 22:4673–4680.
Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol 51:689–702.
Tierney, L. 1994. Markov chains for exploring posterior distributions (with discussion). Ann. Stat. 22:1701–1762.
Verdinelli, I., and L. Wasserman. 1995. Computing Bayes factors using a generalization of the Savage–Dickey density ratio. J. Am. Stat. Assoc. 90:614–618.
Wakely, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. 11:436–442.
Wasserman, L. 2000. Bayesian model selection and model averaging. J. Math. Psychol. 44:92–107.
Weidner, H. 1968. Zur kenntnis der Gallwespentribus Aulacini (Hymenoptera, Cynipidae). Entomol. Z. (Frankf. Main) 78:105–120.
Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.
Yang, Z. 1996a. Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol. Evol. 11:367–372.
Yang, Z. 1996b. Maximum likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

First submitted 30 April 2003; reviews returned 6 August 2003; final acceptance 9 October 2003
Associate Editor: Thomas Buckley