Blackwell
Publishing

Cladistics 22 (2006) 171–185

Cladistics

www.blackwell-synergy.com

The relative performance of Bayesian and parsimony approaches when sampling characters evolving under homogeneous and heterogeneous sets of parameters

Mark P. Simmons[1,*,†], Li-Bing Zhang[1,†], Colleen T. Webb[1], Aaron Reeves[1] and Jeremy A. Miller[2]

[1]*Department of Biology, Colorado State University, Fort Collins, CO 80523, USA;* [2]*Department of Entomology, California Academy of Sciences, 875 Howard Street, San Francisco, CA 94103, USA*

Accepted 15 December 2005

Abstract

We tested whether it is beneficial for the accuracy of phylogenetic inference to sample characters that are evolving under different sets of parameters, using both Bayesian MCMC (Markov chain Monte Carlo) and parsimony approaches. We examined differential rates of evolution among characters, differential character-state frequencies and character-state space, and differential relative branch lengths among characters. We also compared the relative performance of parsimony and Bayesian analyses by progressively incorporating more of these heterogeneous parameters and progressively increasing the severity of this heterogeneity. Bayesian analyses performed better than parsimony when heterogeneous simulation parameters were incorporated into the substitution model. However, parsimony outperformed Bayesian MCMC when heterogeneous simulation parameters were not incorporated into the Bayesian substitution model. The higher the rate of evolution simulated, the better parsimony performed relative to Bayesian analyses. Bayesian and parsimony analyses converged in their performance as the number of simulated heterogeneous model parameters increased. Up to a point, rate heterogeneity among sites was generally advantageous for phylogenetic inference using both approaches. In contrast, branch-length heterogeneity was generally disadvantageous for phylogenetic inference using both parsimony and Bayesian approaches. Parsimony was found to be more conservative than Bayesian analyses, in that it resolved fewer incorrect clades.
© The Willi Hennig Society 2006.

Investigators often favor the sampling of characters that are evolving at different rates from one another, allowing slower evolving characters to resolve early derived (or "basal") clades and faster evolving characters to resolve more recently derived clades (e.g., Graybeal, 1994; Pennington, 1996; Baker et al., 2001; Giribet et al., 2001; Kjer et al., 2001). This widely accepted approach traces back to Hillis (1987, p. 25). Despite their potential benefits for resolving the more recently derived clades, it is possible for rapidly evolving characters to cause errors (e.g., long-branch attraction; Felsenstein, 1978) among early derived clades, overwhelming the phylogenetic signal retained in the slower evolving characters. As Reed and Sperling (1999, p. 286) noted, "A character may be both 'good' and 'bad', depending on what level of divergence it is being used to resolve."

Rate of evolution is only one of several parameters (e.g., character-state space [number of states available for each character to change among], character-state frequencies, relative rates of change among character states) that affect the informativeness of characters for phylogenetic inference. One may apply different likelihood models to different characters when sampling multiple process partitions (i.e., groups of characters with identical histories and evolving under similar parameters; e.g., Bull et al., 1993; Reed and Sperling,

*Corresponding author: Mark P. Simmons, Department of Biology, Colorado State University, Fort Collins, CO 80523-1878, USA. *E-mail address:* psimmons@lamar.colostate.edu

†These two authors contributed equally to the paper.

1999; Caterino et al., 2001), though delimiting natural process partitions will often be difficult (Siddall, 1997). Sampling different process partitions may also be advantageous to parsimony-based phylogenetic inference (Wheeler et al., 1993; Cummings et al., 1995). As Wheeler et al. (1993, p. 16) asserted: "It is highly unlikely that some factor would be canalizing protein coding, ribosomal and morphological features in such a way that they [all process partitions] are all fooled identically."

An alternative approach when applying model-based methods of phylogenetic inference is to analyze characters that are all evolving under the same parameters (including rate of evolution). For example, Jow et al. (2002, p. 1596) stated that "We believe that taking a well-defined set of sites within the RNA helices and using an evolutionary model appropriate to these sites is likely to increase the reliability of results." Similarly, Kelchner (2002, p. 1664) noted that sequence data from group II introns, "which evolve under nearly the same selective constraints... suggests that sequences from multiple G2 introns can be readily combined" under the same model for model-based phylogenetic inference.

These two alternative approaches are illustrated by comparing Barkman et al.'s (2000) and Graham and Olmstead's (2000) character-sampling strategies for resolving the early derived lineages of flowering plants. Although Barkman et al. (2000, p. 13166) expected congruent histories among the nine loci that they sampled from the mitochondrial, plastid and nuclear genomes, they asserted that "well-supported congruent phylogenetic estimates from all three genomic compartments would result in the highest confidence of angiosperm relationships." In contrast, Graham and Olmstead (2000) sampled 17 genes from the plastid genome, thereby sampling many slowly evolving characters that were expected to be robust to any problems caused by long-branch attraction. In effect, Barkman et al. (2000) undertook the extra effort and expense of sampling characters that evolved under different sets of parameters, whereas Graham and Olmstead (2000) chose to sample many, easier-to-amplify sequences that evolved under a similar set of parameters.

In this study, we tested the hypothesis that despite the extra effort, it is beneficial for the accuracy of phylogenetic inference to sample characters that are evolving under different sets of parameters for phylogenetic inference. We did so using both Bayesian MCMC (with the likelihood criterion) and parsimony approaches. Within the context of nucleotide characters we examined differential rates of evolution among characters, differential character-state frequencies and character-state space and differential relative branch lengths among characters. The results of this study are relevant for determining gene-sampling strategies. If sampling multiple different genes that are evolving under different sets of parameters (i.e., rates of spontaneous mutations and/or selective constraints) is advantageous, then this helps to justify the time and expense (e.g., cloning nuclear genes) necessary to sample such loci (e.g., Wheeler et al., 1993; Stanger-Hall and Cunningham, 1998; Barkman et al., 2000; Simmons et al., 2001). Alternatively, if it is not advantageous, then we should simply sequence the genes that are easiest to amplify and that are evolving at an "appropriate" rate (i.e., slow enough to avoid potential for long-branch attraction, yet fast enough to provide sufficient variation; e.g., Graham and Olmstead, 2000; Rai et al., 2003).

The results of this study also bear on the choice between parsimony and model-based approaches for analyzing empirical data. Several simulation-based studies have shown maximum likelihood (Felsenstein, 1973) to be robust to violations of assumptions regarding rate heterogeneity among characters, and between character states (e.g., Huelsenbeck, 1995; Sullivan and Swofford, 2001). Recent studies (Kolaczkowski and Thornton, 2004; Gadagkar and Kumar, 2005; Gaucher and Miyamoto, 2005; Spencer et al., 2005) have compared the relative performance of parsimony and model-based approaches with respect to differential relative branch lengths among characters. However, to our knowledge, no simulation studies have compared the relative performance of parsimony and model-based approaches by progressively incorporating more different sets of heterogeneous parameters (differential rates of evolution among characters, differential character-state frequencies and character-state space, and differential relative branch lengths among characters) and progressively increased the severity of this heterogeneity, as we do in this study.

Our *a priori* hypotheses were that parsimony would perform relatively better than model-based approaches as: (1) more different sets of heterogeneous parameters were incorporated in the simulations, and (2) the severity of the heterogeneity increased. The first hypothesis is *contra* Yang's (1996a, p. 297) hypothesis that "parsimony may be expected to perform worse when rate variation among sites exists than when rates are constant." Yang (1996a, p. 294) concluded that, "As the complexity of the process of nucleotide substitution in real sequences is well recognized, the likelihood method appears preferable to parsimony." Our hypotheses were constructed on two related bases. The basis for our first hypothesis was that as the likelihood model being used becomes increasingly underparameterized relative to the simulated data, its performance would decrease relative to parsimony.

The basis for our second hypothesis was that model-based approaches, for which the model and/or model parameters are estimated rather than given (*contra* Huelsenbeck, 1995; Yang, 1996a; Sullivan and Swofford, 2001) are less likely to accurately fit the pattern of evolution when the characters are evolving under more different, and increasingly heterogeneous, parameters (as is the case with empirical data). That is, one may more accurately fit a model to data that are evolving under a uniform set of parameters. The exception to this is the heterogeneity that the models incorporate in a similar (or identical) manner to which the heterogeneity was simulated (e.g., if gamma-based rate heterogeneity among characters, in which $\alpha = \beta$, was simulated and used by a model-based method of phylogenetic inference; see Sullivan and Swofford, 2001).

Materials and methods

Simulations

Matrices were simulated using the Evolver program within the PAML suite (Yang, 1997). The "MCaa.dat" parameter file was used to simulate the nucleotide characters with a proportional model of evolution (i.e., the probability of change from one character state to another is proportional to their frequencies; Felsenstein, 1981). Ten replicate matrices were simulated for each set of model parameters. Two thousand characters were simulated for each matrix to reduce stochastic errors caused by the use of fewer characters, and to emulate the number of parsimony-informative characters that are often available in contemporary multigene analyses. This relatively high number of characters was expected to alleviate any problems with contemporary, commonly used likelihood models (i.e., GTR + I + Γ or one of its more restrictive variants) trying to estimate too many parameters from too few characters.

Characters were simulated on to a tree of 36 terminals, with branch lengths following a molecular clock (Fig. 1). Thirty-six terminals were selected to be representative of many empirical studies and to allow thousands of separate tree searches to be computationally tractable. The tree topology was selected so as to be intermediate between a totally symmetrical tree (which is generally very easy to reconstruct given a molecular clock) and a totally asymmetrical tree (which is often very difficult to reconstruct given a molecular clock, due to short internal branches and many long-terminal branches). The tree consisted of 12 terminal clades, each of which was composed of three terminals. All internal branches that connected these 12 clades were relatively short (one branch segment; see below), making phylogenetic inference more difficult. The branch lengths leading to, and within, the terminal clades towards the "base" of the tree (terminals 7–30) were varied in length creating four cases in which long-branch attraction could occur between long internal branches that are not sister groups (clades [7, 8, 9], [13, 14, 15], [19, 20, 21] and [25, 26, 27]), and four cases in which long-branch repulsion could occur between long internal branches that are sister groups (Siddall, 1998; clades [10, 11, 12], [16, 17, 18], [22, 23, 24] and [28, 29, 30]). As a result, the tree topology with its relative branch lengths was not expected to be strongly biased in favor of either the parsimony or likelihood approaches.

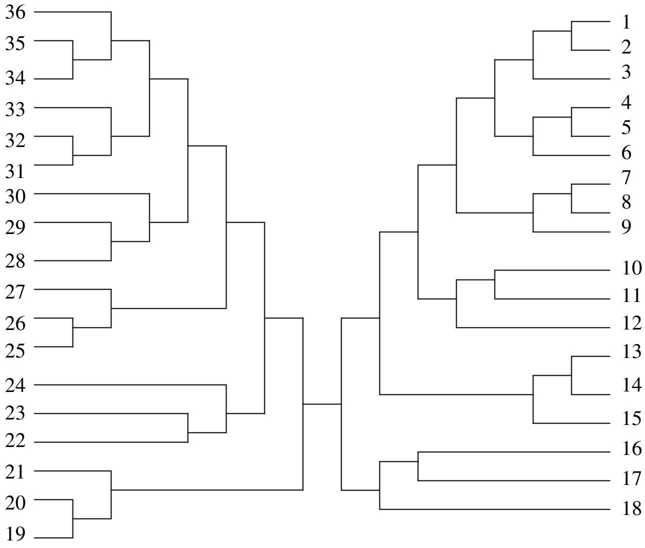

Fig. 1. Tree topology and relative branch lengths (drawn proportionally along the horizontal axis) used for simulations.

The simulated tree consisted of 69 total branches, including 36 terminal branches and 33 internal branches. Of the 69 branches, 45 were one branch segment long, 11 were two branch segments long, four were three branch segments long, four were four branch segments long, four were five branch segments long and one was six branch segments long (Fig. 1). Four branch lengths per branch segment (hereafter "rates of evolution") were examined: 0.2, 0.3, 0.4 and 0.5. These resulted in an average of 24.2, 36.3, 48.4 and 60.5 expected substitutions per character, respectively, across the entire tree. Note that these branch lengths represent highly divergent sequences for which phylogenetic inference by any method is expected to be difficult. These rates of evolution, which average from 35% to 88% internodal change, roughly correspond with the higher end of rates typically examined in four-taxon simulations (from 2% to 74% internodal change; e.g., Gaut and Lewis, 1995; Huelsenbeck, 1995). Shorter branch lengths (e.g., 0.05 per branch segment) were not examined because the correct tree would not have been challenging to infer (as long as a sufficient number of characters were sampled) and would have led to trivial results. Extremely long branch lengths (e.g., 1.0 per branch segment) were excluded because this does not represent a realistic level of variability for genes that are typically used for phylogenetic inference, and the genes would probably be unalignable across all 36 terminals.

Table 1
Combinations of model parameters for which heterogeneity was simulated

Simulation	Rate heterogeneity among characters: Gamma	Rate heterogeneity among characters: Dual	Nucleotide frequencies	Relative branch lengths	Number of partitions per run
A	+	–	–	–	1
B	–	+	–	–	2
C	–	–	+	–	2 (2–6)*, 4 (7–10)
D	–	–	–	+	2
E	+	–	+	–	2 (2–6), 4 (7–10)
F	–	+	+	–	4 (2–6), 8 (7–10)
G	+	–	–	+	2
H	–	+	–	+	4
I	–	–	+	+	4 (2–6), 8 (7–10)
J	+	–	+	+	4 (2–6), 8 (7–10)
K	–	+	+	+	8 (2–6), 16 (7–10)

*Numbers in parentheses refer to particular simulation runs.

Table 2
Increasing heterogeneity in substitution rate among characters, nucleotide frequencies and relative branch lengths per branch segment for simulation runs 1–10

	Rate heterogeneity among characters			Nucleotide frequencies				Relative branch lengths	
	Gamma	Dual							
Simulation run	alpha	char. 1	char. 2	A	G	C	T	char. 1	char. 2
1	infinite	100%	100%	0.25	0.25	0.25	0.25	100%	100%
2	100	110%	90%	0.30	0.30	0.20	0.20	110%	90%
3	50	120%	80%	0.35	0.35	0.15	0.15	120%	80%
4	10	130%	70%	0.40	0.40	0.10	0.10	130%	70%
5	5	140%	60%	0.45	0.45	0.05	0.05	140%	60%
6	1	150%	50%	0.50	0.50	0	0	150%	50%
7	0.75	160%	40%	0.60	0.40	0	0	160%	40%
8	0.5	170%	30%	0.70	0.30	0	0	170%	30%
9	0.25	180%	20%	0.80	0.20	0	0	180%	20%
10	0.1	190%	10%	0.90	0.10	0	0	190%	10%

Three model parameters were varied in the simulations: rate heterogeneity among sites, nucleotide frequencies and rate of evolution. Increasing heterogeneity in each of these parameters was simulated individually, and in all possible combinations (Table 1). For each of the three model parameters that were varied, heterogeneity was progressively increased from the first run (no heterogeneity) to the tenth run (extreme heterogeneity; Table 2). For each of the four rates of evolution simulated, the total expected number of substitutions was identical for each of the 10 runs across all combinations of parameters simulated, making the results for each of these simulations directly comparable with the others.

Rate heterogeneity among sites was simulated using either a discrete gamma distribution (Yang, 1993) with 400 separate rate categories (to more finely partition heterogeneity) or two discrete rates (dual rates) that were proportionally adjusted. Note that the heterogeneities simulated in runs 2–10 (Table 2) for these two approaches are not directly comparable with one another with respect to the degree of heterogeneity, other than the fact that the rate heterogeneity becomes progressively more extreme within each approach. For the dual rates approach, runs 2–10 each required two simulations, with 50% of the characters simulated in each (Table 1).

Heterogeneity in nucleotide frequencies was simulated using all four nucleotides potentially represented at each character in simulations 1–5, and two nucleotides potentially represented at each character in simulations 6–10 (Table 2). Considering all 2000 characters together, each nucleotide was equally represented (25%) for all 10 simulations. Runs 2–6 each required two simulations, with 50% of the characters simulated in each, whereas runs 7–10 each required four simulations, with 25% of the characters simulated in each (Table 1). For example, run seven consisted of:

60% A, 40% G (500 characters); 40% A, 60% G (500 characters); 60% C, 40% T (500 characters), and 40% C, 60% T (500 characters).

Heterogeneity in relative branch lengths (as in a covarion model and heterotachy; Fitch and Markowitz, 1970; Lopez et al., 2002) was simulated such that all branches that were shortened for half of the characters were lengthened for the other half of the characters, and vice versa. These relative increases and decreases were applied to 68 of the 69 branches in the tree; the one exception being the "basal-most" branch of the tree connecting terminals 1–18 with terminals 19–36 (Fig. 1). For example, the six-segment-long branch was 11.4 segments long for half of the characters and 0.6 segments long for the other half of the characters in run 10. Starting at the base of the tree, sister branches were varied such that one was shortened and one was lengthened for each group of characters. This was performed in a consistent manner such that branches on the right side of the tree (for terminals 1–18) drawn below their sister branch were lengthened for the first group of characters, and shortened for the second group of characters. In contrast, branches on the left side of the tree (for terminals 19–36) drawn below their sister branch were shortened for the first group of characters and lengthened for the second group of characters.

The matrices were re-simulated in order to vary two or three of the model parameters simultaneously. This required the product of the number of partitions for each of the parameters involved when simulated individually (Table 1). For example, varying dual rate heterogeneity and relative branch lengths simultaneously (simulation H; Table 1) required four partitions of 500 characters each. So, for the fifth run, characters 1–500 were simulated such that they evolved at 140% of the initial rate, and the first set of sister branches was lengthened by 140%, whereas the second set of sister branches was shortened by 60%. Characters 501–1000 were simulated such that they evolved at 140% of the initial rate, and the first set of branches was shortened by 60%, whereas the second set of sister branches was lengthened by 140%. Characters 1001–1500 were simulated such that they evolved at 60% of the initial rate, and the first set of sister branches was lengthened by 140%, whereas the second set of sister branches was shortened by 60%. Characters 1501–2000 were simulated such that they evolved at 60% of the initial rate, and the first set of branches was shortened by 60%, whereas the second set of sister branches was lengthened by 140%. The blocks of simulated characters were then concatenated together to form a matrix of all 36 terminals for 2000 characters using CONCAT (available at: http://www.biology.colostate.edu/Research/). A total of 4000 simulated matrices were generated: 11 sets of simulations × 9 runs per set = 99 + 1 initial baseline simulation = 100 times 10 replicates per run = 1000 × 4 rates of evolution = 4000.

Tree searches

Each of these 4000 simulated matrices was analyzed using both equally weighted parsimony and Bayesian MCMC analyses (Yang and Rannala, 1997). Parsimony jackknife analyses (Farris et al., 1996) were conducted using PAUP* 4.0b10 (Swofford, 2001) with the removal probability set to approximately e^{-1} (37%), and "jac" resampling emulated. One thousand replicates were performed with 10 tree-bisection-reconnection searches per replicate and a maximum of 10 trees held per search.

Modeltest 3.06 (Posada and Crandall, 1998) was used to select the single best-fit likelihood model among the 56 examined for Bayesian analyses, using the first replicate from each of the 400 runs (a run is composed of 10 replicate matrices simulated under a given set of model parameters). The likelihood ratio test (Huelsenbeck and Crandall, 1997) was used to select among models, given that it has been found to outperform the Akaike Information Criterion (Akaike, 1974) in selecting the correct model from simulated data (Posada and Crandall, 2001; but see Pol (2004) for concerns regarding alternative model-selection starting points). The selected model, using the default priors for nucleotide frequencies, shape of the gamma distribution (if applicable), and/or proportion of invariant sites (if applicable), was then used for all replicates within the corresponding run by MrBayes 3.0b4 (Huelsenbeck and Ronquist, 2001). The model selected for each set of parameters is available as supplemental data at: http://www.biology.colostate.edu/Research/. A single Bayesian analysis was performed for each matrix, consisting of four chains with default temperature settings, 500 000 generations, and trees sampled every 100 generations. The 2001 trees sampled from the last 200 100 generations were used to construct majority rule consensus trees in PAUP*, which served to estimate the posterior probability for each of the resolved clades. A representative sample of matrices was examined to ensure that stationarity in log-likelihood had been reached within the first 300 000 generations. Maximum likelihood analyses were not performed because they are not currently computationally tractable for conducting a sufficient number of conventional bootstrap (Felsenstein, 1985) or jackknife replicates for each of the 4000 matrices (Sanderson and Kim, 2000; but see Waddell et al., 2002).

Statistical analyses

Both 50% and 95% majority rule consensus trees were calculated for the parsimony jackknife and Bayesian analyses of each matrix. These two cutoffs represent the range of support values for clades that are generally considered to have sufficient support to test phylogenetic hypotheses. PEST version 2.2 (Zujko-Miller and Miller,

2003) was used to determine the number of clades correctly and incorrectly resolved between the parsimony jackknife and Bayesian trees relative to the reference tree (i.e., the tree topology on which the characters were simulated) for each matrix. This was used to calculate the overall success of resolution (Simmons and Miya, 2004), for which the maximum score was 33, for all clades correctly resolved, and the worst possible score was −33, in which all clades from the reference tree would be contradicted. The overall success of resolution scales linearly to the Robinson–Foulds distance (Robinson and Foulds, 1981; Penny and Hendy, 1985) for fully resolved trees, which would range from 0 to 66 for our trees. The average congruence for each run of 10 replicates was determined using CONDENSE (Simmons et al., 2004b; available at: http://www.biology.colostate.edu/Research/), but the statistical analyses incorporated the replicates individually using standard multiple regression techniques as opposed to using average replicate values.

Two groups of statistical analyses using multiple regression were performed in JMP IN (SAS Institute, Inc.) to address our hypotheses. Multiple regression is generally used to learn more about the relationship between several independent or predictor variables (e.g., the method of phylogenetic analysis, the rate of evolution, the degree of heterogeneity) and a dependent or response variable (e.g., the number of correctly resolved clades). Multiple regression partitions the variability seen in the dependent variable into proportions that can be explained by each of the independent variables (main effects) as well as by interactions among the independent variables. Our simulations were done using several parameter sets, and we were often interested in questions involving the relationship between a specific independent variable and the dependent variable across parameter sets. The multiple-regression approach allows us to address these types of questions across the entire dataset instead of using subsections of the data. The main advantages to this approach are a greater statistical power, a larger scope of inference, and the ability to deal with any correlations among the independent variables of interest. Multiple regression can also be used to obtain least squares means estimates (sometimes called "adjusted mean"). The least squares means for a specific independent variable are means that have been corrected for the effects of other independent variables (such as those generated by correlations). Independent contrasts of the least squares means are then used to compare if they are significantly different from one another.

Because of the inherent correlation between data for 50% and 95% majority-rule consensus trees, we analyzed the data separately for the 50% and 95% cutoffs. It is not possible to appropriately analyze the data from both the 50% and 95% cutoff values in a standard regression model without running into problems with over-specification of the model and ill-conditioned regression.

"Group 1" of the multiple regressions were performed separately using the number of correctly resolved clades, the number of incorrectly resolved clades and the overall success of resolution (the number of clades correctly resolved minus the number of clades incorrectly resolved) for each replicate as response variables. Eleven such sets of regressions were performed, corresponding to each set of simulations (Table 1). Each individual regression model included the following main effects: the method of phylogenetic analysis (categorical: Bayesian or parsimony), the rate of evolution (continuous: 0.2, 0.3, 0.4 or 0.5 branch lengths per branch segment), and the degree of heterogeneity (continuous: 1–10). The interactions, "method of analysis × rate of evolution" and "method of analysis × degree of heterogeneity", were also included in the model. All variables were treated as fixed effects. Residuals were normally distributed and no high-leverage points or outliers were observed. The residuals met the assumptions of multiple regression, indicating that multiple regression on the untransformed data was appropriate. Least-squares mean estimates of the categorical independent variables were obtained in addition to the parameter estimates.

"Group 2" of the multiple regressions were performed using overall success of resolution as the response variable. The model included as fixed main effects the partition treatment (i.e., set of simulations) as a categorical variable (Table 1; 11 categories), the method of phylogenetic analysis (categorical), and the rate of evolution (continuous). The degree of heterogeneity was nested within the partition treatment as the interpretation of heterogeneity differed across treatments. The interactions included were "method of analysis × partition treatment", "method of analysis × rate of evolution", and "method of analysis × partition treatment (heterogeneity)". Residuals were normally distributed and no high-leverage points or outliers were observed. Again, this indicates that the assumptions of the multiple-regression approach were met. Least-squares mean estimates of the categorical independent variables were obtained in addition to the parameter estimates. In order to address the hypothesis regarding the inclusion of single versus multiple partitions, we performed independent contrasts on the estimated means for the "method of analysis × partition treatment" interaction. Because two methods were used to simulate rate heterogeneity among characters, but only one method each for branch-length and character-state heterogeneity, the contrasts had to be appropriately weighted. So as not to over-weight the effect of rate heterogeneity among sites relative to nucleotide-frequency and branch-length heterogeneity, results from the two methods for simulating heterogeneity among sites were each given one-half the weight of the results from each of the other two parameters that were simulated individually (i.e., gamma 0.167, dual 0.167, nucleotide 0.333, branch 0.333) A similar procedure was followed for heterogeneous parameters simulated in pairs (i.e., gamma + nucleotide 0.167, dual + nucleotide 0.167, gamma + branch 0.167, dual + branch 0.167, nucleotide + branch 0.333).

Groups of analyses with multiple tests were Bonferroni-corrected in order to control for spurious significant results that can be the result of large numbers of comparisons. Results from models that explained less than 30% of the variability in the response variable (i.e., $R^2 < 0.30$) were considered to have an inadequate explanatory power and are indicated as such in the results. Highly significant models generally had a P-value of less than 0.0001 unless otherwise reported. All of the regressions in Groups 1 and 2 showed a significant lack-of-fit, but many of them also had high R^2 values ($R^2 > 0.70$). The significant lack-of-fit is due to our inclusion only of variables that were hypothesized to be of importance, but indicates that more of the variance in the data could be explained if additional interactions were included.

Results and discussion

The results were generally not qualitatively different when using either the 50% or 95% consensus trees for our analyses (but see Table 3). Unless otherwise noted, we assessed the relative performance of the parsimony and Bayesian analyses using the overall success of resolution (the number of clades correctly resolved minus the number of clades incorrectly resolved in the jackknife or posterior-consensus tree). The overall success of resolution for the parsimony and Bayesian analyses using the 95% cutoff are presented in Figs 2–4. Figures for the number of clades correctly resolved, the number of clades incorrectly resolved, and the overall success of resolution using both the 50% and 95% cutoffs are available as supplementary material in Figs S1–S15 at: http://www.biology.colostate.edu/Research/ and http://www.blackwell-synergy.com.

Parsimony versus Bayesian analyses overall

We assessed the relative performance of the parsimony and Bayesian analyses using Group 1 analyses, with the overall success of resolution as the response variable. The relative performance was determined by the parameter estimate for the method of analysis main effect. A significant positive value indicated that Bayesian analyses performed relatively better, and a significant negative value indicated that parsimony performed relatively better (Table 3). Of the 11 sets of simulations, Bayesian analyses significantly outperformed parsimony for two sets of simulations with respect to the overall success of resolution using the 50% cutoff, and five sets of simulations using the 95% cutoff. On the other hand, parsimony significantly outperformed Bayesian analyses for six sets of simulations using the 50% cutoff, and four sets using the 95% cutoff (Table 3).

When the simulation parameters (rate heterogeneity among sites, differential nucleotide frequencies and differential branch lengths) were examined individually,

Table 3
The relative performance of Bayesian and parsimony methods, as well as the relative performance under increasing rates of evolution, based on the overall success of resolution, for each of the simulations using 50% and 95% cutoffs

	Relative performance				Relative performance under increasing rates of evolution‡					
	50% cutoff		95% cutoff		50% cutoff			95% cutoff		
Simulation	Bayes/ pars est.	Signif.	Bayes/ pars est.	Signif.	Bayes slope	Pars. slope	Signif. of int.	Bayes slope	Pars. slope	Signif. of int.
Gamma rate	0.32	0.03*	2.22	<0.0001	−3.23	−2.73	0.07	−3.42	−3.37	0.88
Dual rate	−0.36	0.03*	1.07	<0.0001	−4.51	−3.04	<0.0001	−4.61	−3.29	<0.0001
Nucleotide freq.	−3.58	<0.0001	−1.29	<0.0001	−7.47	−4.68	<0.0001	−5.30	−3.69	<0.0001
Branch lengths	−5.00	<0.0001	−3.39	<0.0001	−5.91	−3.46	<0.0001	−5.03	−4.03	0.003*
Gamma + freq.	1.35	<0.0001	3.02	<0.0001	−4.45	−3.77	0.09	−3.35	−2.91	0.12
Dual + freq.	−1.29	<0.0001	0.46	0.006*	−7.87	−5.22	<0.0001	−5.71	−3.59	<0.0001
Gamma + branch	0.34	0.09	1.91	<0.0001	−3.80	−2.85	0.009*	−3.67	−3.28	0.21
Dual + branch	−1.53	<0.0001	0.39	0.009*	−6.17	−3.29	<0.0001	−5.37	−3.15	<0.0001
Freq. + branch	−3.93	<0.0001	−2.34	<0.0001	−7.81	−5.31	<0.0001	−5.66	−3.62	<0.0001
Gamma + freq. + branch	0.96	<0.0001	2.18	<0.0001	−4.83	−3.91	0.04*	−3.95	−2.95	0.002*
Dual + freq. + branch	−2.03	<0.0001	−1.40	<0.0001	−8.09	−5.15	<0.0001	−6.32	−3.52	<0.0001

*Not significant at the 0.05 level after Bonferroni correction.

‡Slopes were calculated as the parameter estimate for rate of evolution ± the parameter estimate for the "method of phylogenetic analysis × rate of evolution interaction" (+ for Bayes, – for parsimony). All rate-of-evolution main effects were significant ($P < 0.0001$). Reported significance values are for the phylogenetic analysis × rate of evolution interaction.

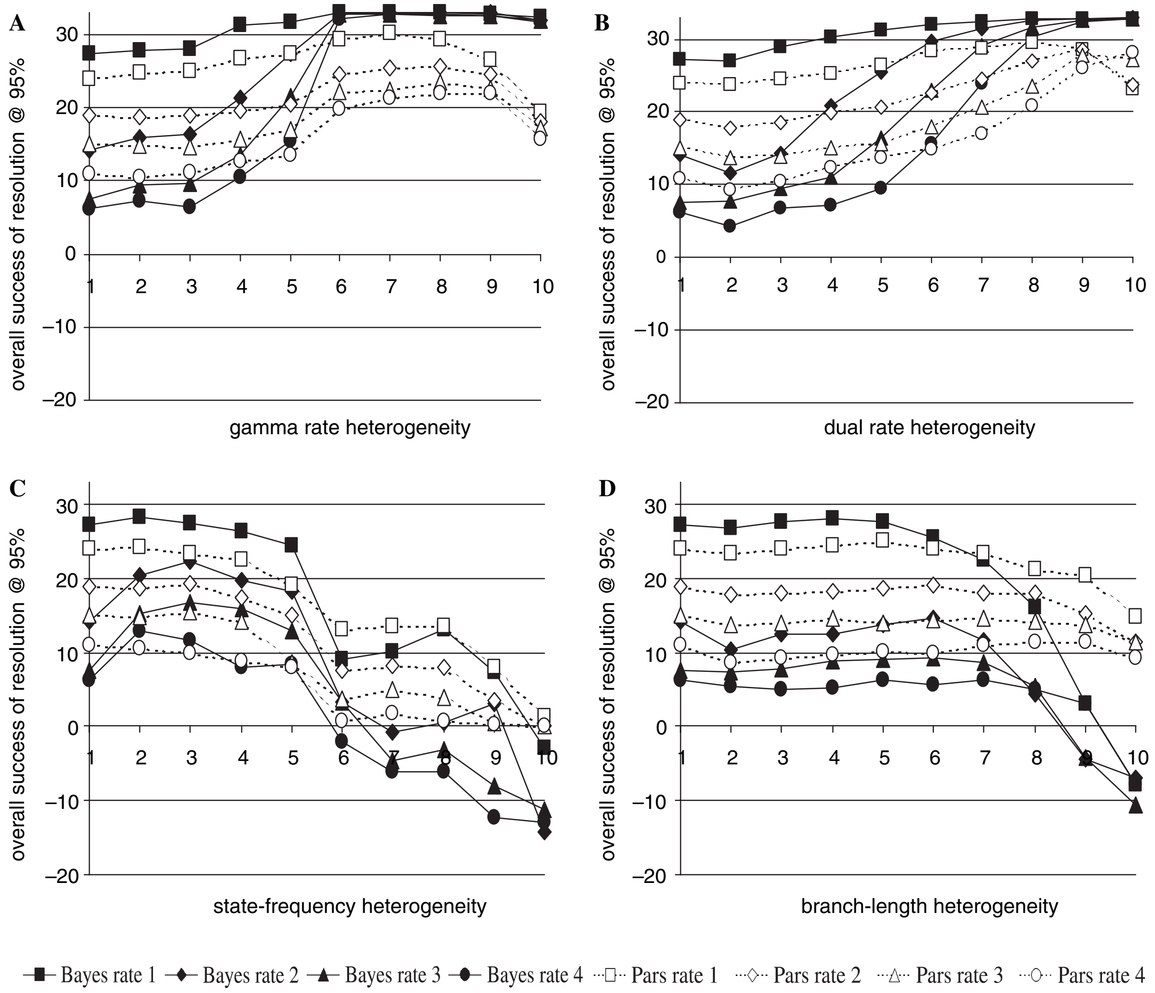

Fig. 2. Average overall success of resolution (the number of clades correctly resolved minus the number of clades incorrectly resolved) by parsimony and Bayesian analyses at each of the four rates of evolution (0.2, 0.3, 0.4 and 0.5 branch lengths per branch segment) for the 10 simulation runs in which heterogeneity was simulated in only a single model parameter. (A) Gamma-rate heterogeneity. (B) Dual-rate heterogeneity. (C) State-frequency heterogeneity. (D) Branch-length heterogeneity.

Bayesian analyses outperformed parsimony for rate heterogeneity among sites (except for dual rate heterogeneity at the 50% cutoff), yet was outperformed by parsimony under differential nucleotide frequencies and branch lengths. Of the 72 simulations that invoked gamma or dual rate heterogeneity among sites (i.e., runs 2–10 for each of the four rates), 70 of the models selected by the hierarchial likelihood ratio test incorporated gamma-distributed rate heterogeneity. In contrast, of the 36 simulation runs that invoked differential nucleotide frequencies, only four of the selected models incorporated unequal nucleotide composition. The four that did were estimated by Modeltest to have all four nucleotide frequencies between 24% and 26%. Because covarion-type models (Lockhart et al., 1998; Tuffley and Steel, 1998; Galtier, 2001) were not incorporated in Modeltest, the likelihood models selected were unable to effectively account for differential branch lengths among characters. Taken together, these results indicate that Bayesian analyses performed better than parsimony when heterogeneous simulation parameters were incorporated into the substitution model. However, parsimony outperformed Bayesian MCMC when heterogeneous simulation parameters were not incorporated into the Bayesian substitution model.

Effect of an increased rate of evolution

We determined the relative performance of parsimony to Bayesian analyses as the rate of evolution increased. This was done by examining the significant interaction plots for the "method of analysis × branch length per branch segment" interaction in the Group 1 analyses with overall success of resolution as the

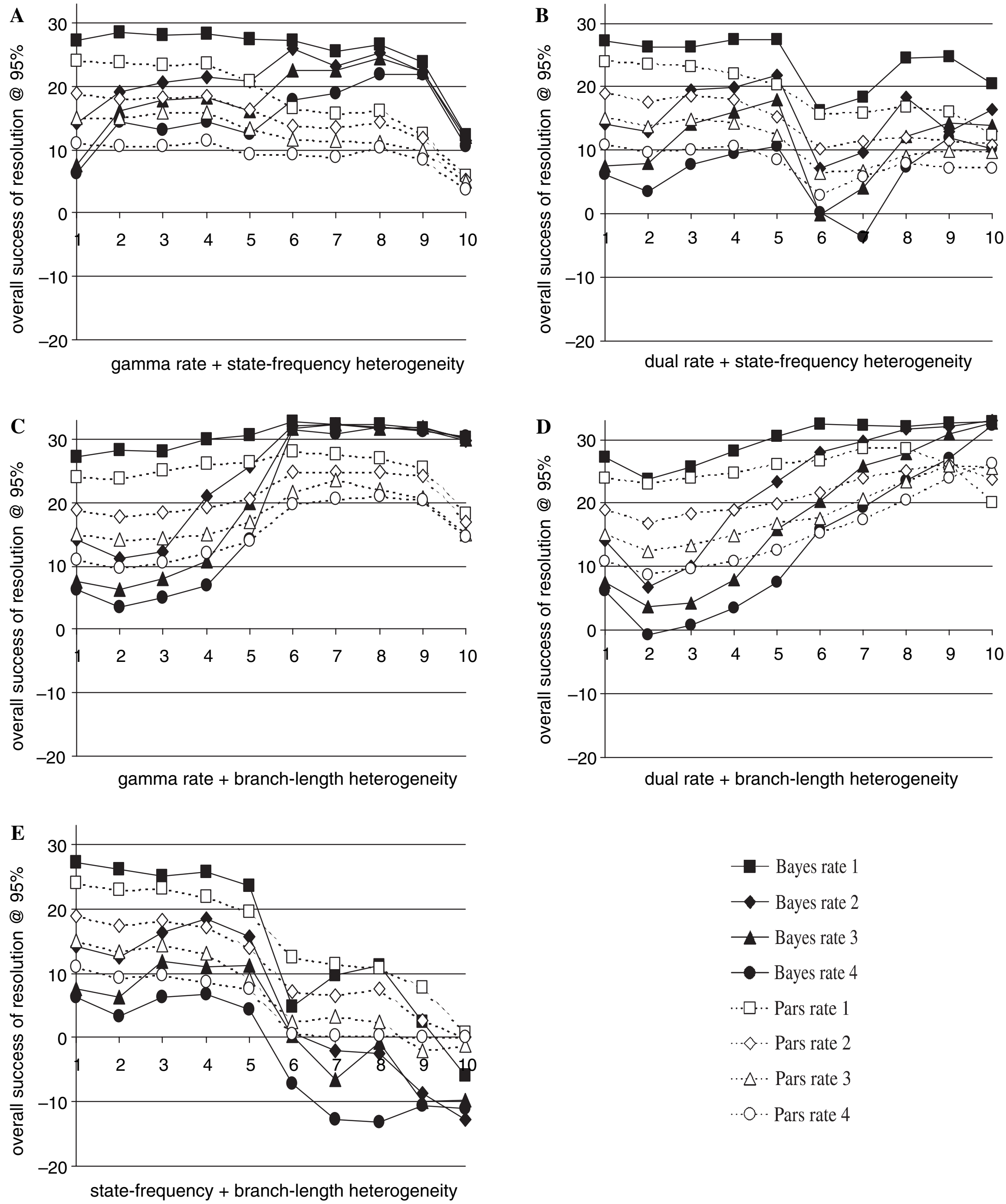

Fig. 3. Average overall success of resolution by parsimony and Bayesian analyses at each of the four rates of evolution for the 10 simulation runs in which heterogeneity was simulated in pairs of model parameters simultaneously. (A) Gamma-rate + state-frequency heterogeneity. (B) Dual-rate + state-frequency heterogeneity. (C) Gamma-rate + branch-length heterogeneity. (D) Dual-rate + branch-length heterogeneity. (E) State-frequency + branch-length heterogeneity.

response variable. The higher the rate of evolution simulated, the better that parsimony performed relative to Bayesian analyses with respect to the overall success of resolution (i.e., parsimony had a more positive [or less negative] slope than Bayesian analyses). This result was evident in all 11 sets of simulations, using both the 50% and 95% cutoffs, for which six were significant at both cutoffs (Table 3). In some cases (e.g., dual rate heterogeneity at the 95% cutoff) parsimony performed relatively better at higher rates of evolution, in spite of

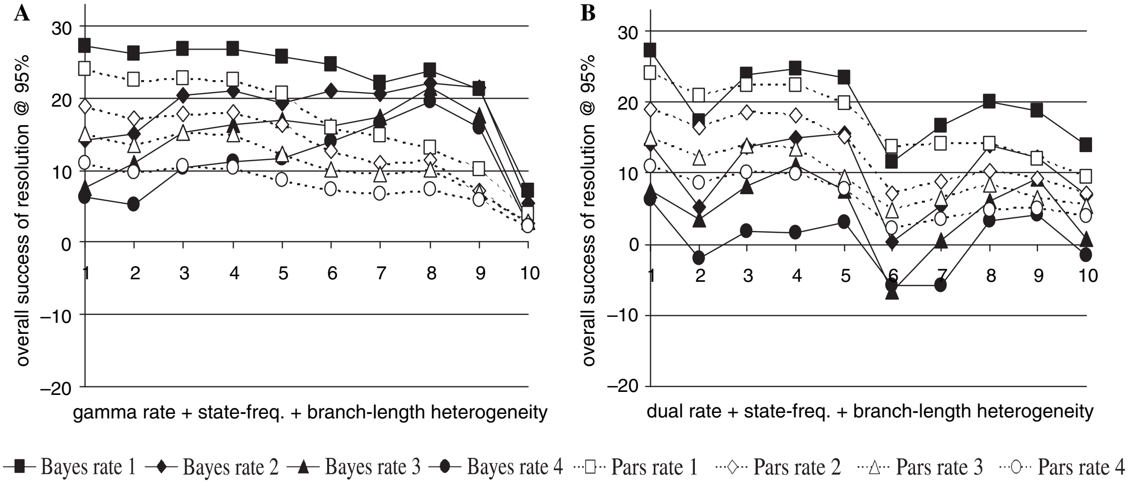

Fig. 4. Average overall success of resolution by parsimony and Bayesian analyses at each of the four rates of evolution for the 10 simulation runs in which heterogeneity was simulated in all three model parameters simultaneously. (A) Gamma rate + state-frequency + branch-length heterogeneity. (B) Dual rate + state-frequency + branch-length heterogeneity.

the Bayesian analyses performing better overall (Table 3). Note that this set of statistical analyses was not performed to test an *a priori* hypothesis regarding the relative performance of Bayesian and parsimony methods at different rates of evolution. Rather, the analyses were performed after observing this trend in the results (Figs 2–4, S1–S3). The better relative performance of parsimony at higher rates of evolution may be explained by the increasing difficulty of accurately estimating the model using the likelihood ratio test and/or accurately estimating model parameters in Bayesian analyses.

Gamma versus dual rate heterogeneity

The Bayesian analyses were found to perform significantly better than parsimony with respect to the overall success of resolution when characters with rate heterogeneity among sites were simulated using the gamma distribution rather than dual rates (independent contrasts, $t = 2.20$, $P = 0.03$ at the 50% cutoff, and $t = 4.75$, $P < 0.0001$ at the 95% cutoff). The least squares mean estimate for Bayesian analyses minus that for parsimony using gamma-distributed heterogeneity was 0.65 at the 50% cutoff, and 4.44 at the 95% cutoff. This is for least squares means from the Group 2 analyses of overall success of resolution. In contrast, the average for Bayesian analyses minus that for parsimony using dual rate heterogeneity was −0.71 at the 50% cutoff, and 2.14 at the 95% cutoff. A model that incorporated gamma-distributed rate heterogeneity was selected by Modeltest and used in the Bayesian analyses for 70 of the 72 simulations that incorporated gamma-distributed or dual rate heterogeneity.

This result suggests that simulation studies comparing the relative performance of likelihood and parsimony methods under rate heterogeneity among sites are biased in favor of likelihood. This bias occurs when the rate heterogeneity used to simulate the data is incorporated in the same way in which the rate heterogeneity is accounted for by the likelihood model used to infer the tree. We know of no reason why simulating dual rate heterogeneity would be more amenable for phylogenetic inference using parsimony rather than likelihood approaches. Note that dual- and gamma-distributed rate heterogeneity are not directly comparable measures, and we cannot rule out the possibility that inherent differences between these two types of rate heterogeneity (or the differences in scaling used here) explain these patterns. To be conservative regarding the possibility of bias in favor of likelihood, we suggest that future studies should simulate rate heterogeneity in the data differently from the way it is accounted for by the likelihood model used to infer the tree. This may also be done by using empirical data from known phylogenies (e.g., Buckley and Cunningham, 2002).

Effect of increasing the number of heterogeneous model parameters

We tested our hypothesis that parsimony would outperform likelihood-based approaches using independent contrasts in our Group 2 analyses. Parsimony was not found to perform relatively better than the Bayesian analyses, as more different sets of heterogeneous parameters were incorporated in the simulations. Rather, the general pattern observed was that Bayesian and parsimony analyses converged in their performance as the number of heterogeneous model parameters simulated increased.

The difference in the average least squares mean estimates between Bayesian and parsimony analyses was −5.72 (i.e., on average, parsimony resolved 5.72 more clades than Bayesian MCMC with respect to overall success of resolution) when the heterogeneous parameters were simulated individually, −3.00 when simulated in pairs and −1.06 when all three heterogeneous parameters were simulated together, at the 50% cutoff. At the 95% cutoff, the difference was −2.02 for singles, 0.368 for pairs, and 0.775 for triples. Three of the four differences in the average least squares mean estimates of overall success of resolution were found to be significant (independent contrasts, 50% singles versus pairs, $t = -8.83$, $P < 0.0001$; 50% pairs versus triples, $t = -5.21$, $P < 0.0001$; 95% singles versus pairs, $t = -9.87$, $P < 0.0001$) at both the 50% and 95% cutoffs when comparing the pairs with the singles and the triples with the pairs. The exception was the insignificant difference in performance between parsimony and Bayesian analyses when comparing the triples with the pairs at the 95% cutoff (independent contrast, $t = -1.42$, $P = 0.16$).

Sampling process partitions

To test whether increasing the number of process partitions sampled enhanced the performance of parsimony and Bayesian approaches, we performed a second set of independent contrasts on the Group 2 analyses. Increasing the number of process partitions sampled was found to be disadvantageous, on average, for parsimony-based phylogenetic inference. The average least squares mean estimates of overall success of resolution for parsimony was 17.74 when the heterogeneous parameters were simulated individually, 15.21 when simulated in pairs and 11.98 when all three heterogeneous parameters were simulated together at the 50% cutoff. Using the 95% cutoff, the averages were 15.70 for singles, 14.34 for pairs and 12.00 for triples. All four of the decreases in the average least squares mean estimates of overall success of resolution were found to be significant (independent contrasts, 50% singles versus pairs, $t = 11.57$, $P < 0.0001$, 50% pairs versus triples, $t = 12.28$, $P < 0.0001$; 95% singles versus pairs, $t = 7.96$, $P < 0.0001$; 95% pairs versus triples, $t = 11.39$, $P < 0.0001$) at both the 50% and 95% cutoffs when comparing the doubles with the singles and the triples with the doubles.

Increasing the number of process partitions sampled had a less uniform effect on Bayesian-based phylogenetic inference. The average least squares mean estimate of the overall success of resolution for Bayesian analyses increased when the heterogeneous parameters were simulated in pairs rather than individually at the 95% cutoff (not significant at the 50% cutoff), yet decreased when all three heterogeneous parameters were simulated together, at both the 50% and 95% cutoffs. The averages were 12.01 for singles, 12.21 for pairs and 10.92 for triples at the 50% cutoff, and 13.68 for singles, 14.71 for pairs and 12.78 for triples at the 95% cutoff. The two differences were both significant at the 95% cutoff (independent contrasts, singles versus pairs, $t = -6.00$, $P < 0.0001$; pairs versus triples, $t = 9.38$, $P < 0.0001$), as was the difference between pairs and triples at the 50% cutoff (independent contrasts, $t = 4.92$, $P < 0.0001$). However, the difference between the singles and doubles at the 50% cutoff was not significant (independent contrasts, $t = -0.91$, $P = 0.36$). Note that these analyses assumed that the unit of measurement for the overall success of resolution was uniform across its entire range.

These results suggest that, on average, it is not preferable to deliberately increase the number of process partitions sampled for parsimony-based phylogenetic inference (when the process partitions differ from one another based on the three heterogeneous parameters simulated here). As such, when possible, we should simply sequence loci that are easy to amplify and are evolving at an "appropriate" rate and manner, following Graham and Olmstead (2000), rather than expending extra resources on sampling loci that are evolving under different sets of parameters. However, this suggestion is predicated by two critical assumptions that in many cases will obviate its practical application. First, this suggestion is based on the notion that genes that are evolving at the "appropriate" rate and manner (e.g., not having overly constrained character-state space; Naylor et al., 1995; Simmons et al., 2004b) may be accurately identified. One way to hedge one's bets would be to sample multiple process partitions. Second, this suggestion does not take into account the potential for introgression, lineage sorting or horizontal transfer (Doyle, 1992). When these issues are liable to be of concern (e.g., introgression through hybridization among closely related plant species (Rieseberg et al., 2000); lineage sorting in lineages with short internal branches and large effective population sizes (Moore, 1995); or horizontal transfer of mitochondrial genes in flowering plants (Bergthorsson et al., 2003; Won and Renner, 2003)), we suggest that multiple coalescent genes be sampled, following Doyle (1992).

Effects of increasing heterogeneity

To test the relative sensitivity of parsimony and Bayesian analyses to increasing heterogeneity, we examined the interaction plots for models from the Group 1 analyses with a significant "method of analysis × degree of heterogeneity" interaction. The method with the shallower slope was deemed to be less sensitive. Of the 11 sets of simulations, Bayesian MCMC was found to be significantly less sensitive than parsimony to increasing

Table 4
The relative effect of increasing heterogeneity on Bayesian and parsimony methods, based on the overall success of resolution

	50% cutoff			95% cutoff		
Simulation	Bayes slope	Pars. slope	Signif. of int.	Bayes slope	Pars. slope	Signif. of int.
Gamma rate	1.89	1.26	< 0.0001	2.58	0.68	< 0.0001
Dual rate	2.31	1.45	< 0.0001	2.65	1.29	< 0.0001
Nucleotide freq.	−5.23	−3.93	< 0.0001	−3.20	−2.02	< 0.0001
Branch lengths	−2.98	−0.99	< 0.0001	−1.96	−0.33	< 0.0001
Gamma + freq.	0.32	−1.21	< 0.0001	0.21	−1.16	< 0.0001
Dual + freq.	−0.39	−1.00	< 0.0001	−0.10	−0.91	< 0.0001
Gamma + branch	2.21	0.81	< 0.0001	2.60	0.55	< 0.0001
Dual + branch	2.78	1.45	< 0.0001	2.79	1.15	< 0.0001
Freq. + branch	−4.91	−3.89	< 0.0001	−2.99	−2.07	< 0.0001
Gamma + freq. + branch	−0.36	−1.85	< 0.0001	−0.19	−1.44	< 0.0001
Dual + freq. + branch	−0.94	−1.85	< 0.0001	−0.53	−1.24	< 0.0001

Slopes were calculated as the parameter estimate for degree of heterogeneity ± the parameter estimate for the "method of phylogenetic analysis × degree of heterogeneity" interaction (+ for Bayes, – for parsimony). All increasing-heterogeneity main effects were significant ($P < 0.0001$). Reported significance values are for the "method of phylogenetic analysis × rate of heterogeneity" interaction.

heterogeneity for four sets of simulations with respect to the overall success of resolution using both the 50% and 95% cutoffs. On the other hand, parsimony was found to be significantly less sensitive than Bayesian analyses for seven sets of simulations using both the 50% and 95% cutoffs (Table 4). Note that this set of statistical analyses was not performed to test an *a priori* hypothesis regarding the relative performance of Bayesian and parsimony methods at different degrees of heterogeneity. Rather, the analyses were performed after observing these trends in the results (Figs 2–4, S1–S3).

In order to test the relative performance of the two methods as heterogeneity increased, we again examined the interaction plots for models from the Group 1 analyses with a significant "method of phylogenetic analysis × degree of heterogeneity" interaction. The method with the higher (if positive; lower if both were negative) slope was determined to have the better performance. Parsimony was not found to perform uniformly better than Bayesian MCMC as the severity of the heterogeneity increased. Only in those three simulations in which rate heterogeneity among sites was not incorporated did parsimony perform significantly better than Bayesian analyses (Table 4). In contrast, Bayesian MCMC significantly outperformed parsimony in all eight of the simulations in which rate heterogeneity among sites was incorporated. All of these results were significantly supported at both the 50% and 95% cutoffs (Table 4).

We examined the parameter estimates for rate heterogeneity in the Group 1 analyses with the overall success of resolution as the response variable in order to test whether or not rate heterogeneity among sites was advantageous overall. Analyses where the parameter estimate for the rate heterogeneity main effect was significant and positive were considered to indicate an overall advantageous effect of increasing heterogeneity. Up to a point, rate heterogeneity among sites was found to be generally advantageous for phylogenetic inference using both the parsimony and Bayesian approaches (Fig. 2A,B). This result was significant for both parsimony and Bayesian analyses based on the overall success of resolution using both the 50% and 95% cutoffs (Table 5). The exception to this was the nearly universal decrease in overall success of resolution for both parsimony and Bayesian analyses when going from homogeneous rates to slight (110%/90%) rate heterogeneity using dual rates (Fig. 2B). Taken as a whole, the overall success of resolution significantly increased as both gamma rate heterogeneity and dual rate heterogeneity progressively increased. Once the rate heterogeneity became too extreme (gamma distributed with $\alpha = 0.1$), however, both parsimony and Bayesian analyses showed decreases in the overall success of resolution for all four rates simulated (Fig. 2A). This result significantly supports the widely used strategy of deliberately sampling characters that are

Table 5
The overall effect of increasing heterogeneity based on overall success of resolution

	50% cutoff		95% cutoff	
Simulation	Het. est.	Signif.	Het. est.	Signif.
Gamma rate	1.57	< 0.0001	1.63	< 0.0001
Dual rate	1.88	< 0.0001	1.97	< 0.0001
Nucleotide freq.	−4.58	< 0.0001	−2.61	< 0.0001
Branch lengths	−1.98	< 0.0001	−1.14	< 0.0001
Gamma + freq.	−0.45	< 0.0001	−0.47	< 0.0001
Dual + freq.	−0.70	< 0.0001	−0.50	< 0.0001
Gamma + branch	1.51	< 0.0001	1.58	< 0.0001
Dual + branch	2.12	< 0.0001	1.97	< 0.0001
Freq. + branch	−4.40	< 0.0001	−2.53	< 0.0001
Gamma + freq. + branch	−1.10	< 0.0001	−0.81	< 0.0001
Dual + freq. + branch	−1.40	< 0.0001	−0.88	< 0.0001

Table 6
The relative performance of Bayesian and parsimony methods based on the number of clades incorrectly resolved

	50% cutoff				95% cutoff			
Simulation	Bayes mean	Pars. Mean	t	Signif.	Bayes mean	Pars. mean	t	Signif.
Gamma rate	1.36	0.50	7.44	< 0.0001	0.02	0.00	2.49	0.01*
Dual rate	2.66	1.08	9.58	< 0.0001	0.16	0.01	3.90	< 0.0001*
Nucleotide freq.	13.79	5.87	25.12	< 0.0001	4.31	0.01	17.61	< 0.0001
Branch lengths	9.68	2.69	24.72	< 0.0001	3.43	0.16	13.13	< 0.0001
Gamma + freq.	5.28	3.56	7.02	< 0.0001*	1.11	0.07	11.22	< 0.0001*
Dual + freq.	9.69	5.24	15.38	< 0.0001	3.48	0.17	17.08	< 0.0001
Gamma + branch	2.66	1.66	5.35	< 0.0001	0.35	0.06	5.15	< 0.0001*
Dual + branch	4.86	1.64	13.40	< 0.0001	1.17	0.05	9.79	< 0.0001*
Freq. + branch	16.19	7.68	27.08	< 0.0001	6.52	0.20	23.07	< 0.0001
Gamma + freq. + branch	7.62	5.45	7.70	< 0.0001	2.62	0.39	14.09	< 0.0001
Dual + freq. + branch	12.51	7.34	16.70	< 0.0001	6.70	0.58	25.19	< 0.0001

*The model had low explanatory power (R^2 < 0.30).

evolving at different rates to improve phylogenetic inference when analyzing one's data using parsimony and/or Bayesian analyses.

In contrast to rate heterogeneity among sites, branch-length heterogeneity was found to be generally disadvantageous for phylogenetic inference using both parsimony and Bayesian approaches (Fig. 2C). This result was significant for both analyses based on the overall success of resolution using both the 50% and 95% cutoffs (Table 5). Note that these statistical analyses were not performed to test an *a priori* hypothesis regarding the relative performance of the Bayesian and parsimony methods to increasing branch-length heterogeneity. Rather, the analyses were performed after observing these trends in the results (Figs 1, S1). A probable explanation for the poor performance of Bayesian MCMC to increasing branch-length heterogeneity is that the likelihood models used in this study (and in the vast majority of empirical likelihood-based analyses) all assume proportional branch lengths among characters. Our result is consistent with Chang's (1996) examples, showing how likelihood methods can be inconsistent under a general heterogeneous model, and Kolaczkowski and Thornton's (2004) four-taxon simulations.

Incorrect resolution by parsimony and Bayesian analyses

By using independent contrasts to compare the least mean squares estimates for parsimony and Bayesian analyses from the Group 1 analyses with number of incorrect clades resolved as the response variable, we could test how well each method of analysis performed with respect to the number of clades incorrectly resolved. Parsimony was found to be more conservative than Bayesian MCMC in that it resolved fewer incorrect clades. This result was found to be significant for all 11 sets of simulations at the 50% cutoff, and 10 of the 11 sets at the 95% cutoff (Table 6). This set of statistical analyses was used to test the *a priori* hypothesis that Bayesian analyses would resolve significantly more incorrect clades than would the parsimony jackknife analyses. These results are consistent with previous studies that found Bayesian posterior probabilities to be inflated relative to ideal support values, in contrast to both the bootstrap (Suzuki et al., 2002; Cummings et al., 2003; Taylor and Piel, 2004) and the jackknife (Simmons et al., 2004a). These studies were the bases for our *a priori* hypothesis.

Conclusions

Our two *a priori* hypotheses were refuted. First, parsimony was not found to perform relatively better than Bayesian MCMC as more different heterogeneous parameters were incorporated in the simulations. Rather, the general pattern observed was that the parsimony and Bayesian analyses converged in their performance as the number of simulated heterogeneous model parameters increased. Second, parsimony was not found to perform uniformly better than Bayesian MCMC as the severity of the heterogeneity increased. Only in those three simulations in which rate heterogeneity among sites was not incorporated did parsimony perform better than Bayesian MCMC. In contrast, Bayesian MCMC outperformed parsimony in all eight of the simulations in which rate heterogeneity among sites was incorporated.

The convergence in performance of parsimony and Bayesian analyses as the number of heterogeneous model parameters increased is an important result to consider when extrapolating the results of simplistic simulation studies (including this one) to determine how empirical data should be analyzed. Although likelihood-based methods generally perform well when their assumptions are met in simulations (e.g., Huelsenbeck, 1995; Yang, 1996a; Sullivan and Swofford, 2001), their performance may not be significantly better than parsimony when applied to empirical characters (Anderson and Swofford, 2004), which are generally heterogeneous.

One could perform Bayesian analyses using different models for each of the process partitions that we simulated in each of the 11 sets. However, for the results to be empirically relevant, the boundaries between process partitions must be identifiable in the empirical data. An initial way of testing approaches that attempt to delimit process partitions (Lartillot and Philippe, 2004; Pagel and Meade, 2004) would be to randomize our characters and then apply the approaches, wherein partitions are discovered rather than set *a priori*.

Mixed likelihood models (e.g., Yang, 1996b; DeBry, 1999; Nylander et al., 2004) are a step forward in addressing the heterogeneity we simulated, though delimiting natural process partitions remains problematic (Siddall, 1997). Furthermore, mixed models using the GTR + I + Γ model or one of its more restrictive variants would still not accommodate differential branch lengths (wherein a branch that is long for one partition is short for another partition, and vice versa), as was simulated here. If process partitions were to be accurately delimited, and mixed, covarion-type models were to be accurately estimated and applied, each of the heterogeneous parameters examined in this study could be addressed using likelihood approaches. However, a question would remain: could all of the necessary parameters be effectively estimated for the models from each process partition or would this result in overparameterization (see Cunningham et al. (1998) and Lemmon and Moriarty (2004))?

Acknowledgments

We thank two anonymous reviewers, Thomas Buckley, John Freudenstein, Melissa Islam, Scot Kelchner, Kurt Pickett, Chris Randle, Pat Reeves and Susanne Renner for helpful discussions and/or suggestions that improved the manuscript.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control. 19, 716–723.

Anderson, F.E., Swofford, D.L., 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol. Phylogenet. Evol. 33, 440–451.

Baker, R.H., Wilkinson, G.S., DeSalle, R., 2001. Phylogenetic utility of different types of molecular data used to infer evolutionary relationships among stalk-eyed flies (Diopsidae). Syst. Biol. 50, 87–105.

Barkman, T.J., Chenery, G., McNeal, J.R., Lyons-Weiler, J., Elisens, W.J., Moore, G., Wolfe, A.D., dePamphilis, C.W., 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc. Natl Acad. Sci. USA, 97, 13166–13171.

Bergthorsson, U., Adams, K.L., Thomason, B., Palmer, J.D., 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature, 424, 197–201.

Buckley, T.R., Cunningham, C.W., 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol. Biol. Evol. 19, 394–405.

Bull, J.J., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D.L., Waddell, P.J., 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42, 384–397.

Caterino, M.S., Reed, R.D., Kuo, M.M., Sperling, F.A.H., 2001. A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera: Papilionidae). Syst. Biol. 50, 106–127.

Chang, J.T., 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Biosci. 134, 189–215.

Cummings, M.P., Otto, S.P., Wakeley, J., 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12, 814–822.

Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., Winka, K., 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52, 477–487.

Cunningham, C.W., Zhu, H., Hillis, D.M., 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution, 52, 978–987.

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

Doyle, J.J., 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. Syst. Bot. 17, 144–163.

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics, 12, 99–124.

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22, 240–249.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.

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

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39, 783–791.

Fitch, W.M., Markowitz, E., 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4, 579–593.

Gadagkar, S.R., Kumar, S., 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol. Biol. Evol. 22, 2139–2141.

Galtier, N., 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18, 866–873.

Gaucher, E.A., Miyamoto, M.M., 2005. A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. Mol. Phylogenet. Evol. 37, 928–931.

Gaut, B.S., Lewis, P.O., 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12, 152–162.

Giribet, G., Edgecombe, G.D., Wheeler, W.C., 2001. Arthropod phylogeny based on eight molecular loci and morphology. Nature, 413, 157–161.

Graham, S.W., Olmstead, R.G., 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am. J. Bot. 87, 1712–1730.

Graybeal, A., 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. Syst. Biol. 43, 174–193.

Hillis, D.M., 1987. Molecular versus morphological approaches. Annu. Rev. Ecol. Syst. 18, 23–42.

Huelsenbeck, J.P., 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44, 17–48.

Huelsenbeck, J.P., Crandall, K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28, 437–466.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics, 17, 754–755.

Jow, H., Hudelot, C., Rattray, M., Higgs, P.G., 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol. Biol. Evol. 19, 1591–1601.

Kelchner, S.A., 2002. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. Am. J. Bot. 89, 1651–1669.

Kjer, K.M., Blahnik, R.J., Holzenthal, R.W., 2001. Phylogeny of the Trichoptera (caddisflies): characterization of signal and noise within multiple datasets. Syst. Biol. 50, 781–816.

Kolaczkowski, B., Thornton, J.W., 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature, 431, 980–984.

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109.

Lemmon, A.R., Moriarty, E.C., 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53, 265–277.

Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D.H., Charleston, M.A., Howe, C.J., 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15, 1183–1188.

Lopez, P., Casane, D., Philippe, H., 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19, 1–7.

Moore, W.S., 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. Evolution, 49, 718–726.

Naylor, G.J.P., Collins, T.M., Brown, W.M., 1995. Hydrophobicity and phylogeny. Nature, 373, 565–566.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47–67.

Pagel, M., Meade, A., 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53, 571–581.

Pennington, R.T., 1996. Molecular and morphological data provide phylogenetic resolution at different hierarchial levels in *Andira*. Syst. Biol. 45, 496–515.

Penny, D., Hendy, M.D., 1985. The use of tree comparison metrics. Syst. Zool. 34, 75–82.

Pol, D., 2004. Empirical problems of the hierarchial likelihood ratio test for model selection. Syst. Biol. 53, 949–962.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics, 14, 817–818.

Posada, D., Crandall, K.A., 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50, 580–601.

Rai, H.S., O'Brien, H.E., Reeves, P.A., Olmstead, R.G., Graham, S.W., 2003. Inference of higher-order relationships in the cycads from a large chloroplast data set. Mol. Phylogenet. Evol. 29, 350–359.

Reed, R.D., Sperling, F.A.H., 1999. Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus *Papilio*. Mol. Biol. Evol. 16, 286–297.

Rieseberg, L.H., Baird, S.J.E., Gardner, K.A., 2000. Hybridization, introgression, and linkage evolution. In: Doyle, J.J., Gaut, B.S. (Eds.), Plant Molecular Evolution. Kluwer Academic Publishers, Dordrecht, pp. 205–224.

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

Sanderson, M.J., Kim, J., 2000. Parametric phylogenetics? Syst. Biol. 49, 817–829.

Siddall, M.E., 1997. Prior agreement: arbitration or arbitrary? Syst. Biol. 46, 765–769.

Siddall, M.E., 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris Zone. Cladistics, 14, 209–220.

Simmons, M.P., Miya, M., 2004. Efficiently resolving the basal clades of a phylogenetic tree using Bayesian and parsimony approaches: a case study using mitogenomic data from 100 higher teleost fishes. Mol. Phylogenet. Evol. 31, 351–362.

Simmons, M.P., Pickett, K.M., Miya, M., 2004a. How meaningful are Bayesian posterior probabilities? Mol. Biol. Evol. 21, 188–199.

Simmons, M.P., Reeves, A., Davis, J.I., 2004b. Character-state space versus rate of evolution for phylogenetic inference. Cladistics, 20, 191–204.

Simmons, M.P., Savolainen, V., Clevinger, C.C., Archer, R.H., Davis, J.I., 2001. Phylogeny of the Celastraceae inferred from 26S nrDNA, phytochrome B, *atpB, rbcL, and* morphology. Mol. Phylogenet. Evol. 19, 353–366.

Spencer, M., Susko, E., Roger, A.J., 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. 22, 1161–1164.

Stanger-Hall, K., Cunningham, C.W., 1998. Support for a monophyletic Lemuriformes: overcoming incongruence between data partitions. Mol. Biol. Evol. 15, 1572–1577.

Sullivan, J., Swofford, D.L., 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50, 723–729.

Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl Acad. Sci. USA, 99, 16138–16143.

Swofford, D.L., 2001. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer, Sunderland, MA.

Taylor, D.J., Piel, W.H., 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol. Biol. Evol. 21, 1534–1537.

Tuffley, C., Steel, M., 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147, 63–91.

Waddell, P.J., Kishino, H., Ota, R., 2002. Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. Genome Informatics, 13, 82–92.

Wheeler, W.C., Cartwright, P., Hayashi, C.Y., 1993. Arthropod phylogeny: a combined approach. Cladistics, 9, 1–39.

Won, H., Renner, S.S., 2003. Horizontal gene transfer from flowering plants to *Gnetum*. Proc. Natl Acad. Sci. USA, 100, 10824–10829.

Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10, 1396–1401.

Yang, Z., 1996a. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42, 294–307.

Yang, Z., 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42, 587–596.

Yang, Z., 1997. PAML: a Program Package for Phylogenetic Analysis by Maximum Likelihood. CABIOS. 13, 555–556.

Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Mol. Biol. Evol. 14, 717–724.

Zujko-Miller, C., Miller, J.A., 2003. PEST: Precision estimated by sampling traits. http://www.gwu.edu/~clade/spiders/pestDocs.htm, Program distributed by the authors.

Supplementary material

The authors have provided supplementary data and figures, which can be viewed online with the article at http://www.blackwell-synergy.com