Syst. Biol. 55(4):553–565, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150600812544

Searching for Convergence in Phylogenetic Markov Chain Monte Carlo

ROBERT G. BEIKO,[1] JONATHAN M. KEITH,[2] TIMOTHY J. HARLOW,[1] AND MARK A. RAGAN[1]

[1]*ARC Centre in Bioinformatics and Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia; E-mail: r.beiko@gmail.com (R.G.B.) and ARC Centre in Bioinformatics*
[2]*Department of Mathematics, The University of Queensland, Brisbane, Australia*

Abstract.— Markov chain Monte Carlo (MCMC) is a methodology that is gaining widespread use in the phylogenetics community and is central to phylogenetic software packages such as MrBayes. An important issue for users of MCMC methods is how to select appropriate values for adjustable parameters such as the length of the Markov chain or chains, the sampling density, the proposal mechanism, and, if Metropolis-coupled MCMC is being used, the number of heated chains and their temperatures. Although some parameter settings have been examined in detail in the literature, others are frequently chosen with more regard to computational time or personal experience with other data sets. Such choices may lead to inadequate sampling of tree space or an inefficient use of computational resources. We performed a detailed study of convergence and mixing for 70 randomly selected, putatively orthologous protein sets with different sizes and taxonomic compositions. Replicated runs from multiple random starting points permit a more rigorous assessment of convergence, and we developed two novel statistics, δ and ε, for this purpose. Although likelihood values invariably stabilized quickly, adequate sampling of the posterior distribution of tree topologies took considerably longer. Our results suggest that multimodality is common for data sets with 30 or more taxa and that this results in slow convergence and mixing. However, we also found that the pragmatic approach of combining data from several short, replicated runs into a "metachain" to estimate bipartition posterior probabilities provided good approximations, and that such estimates were no worse in approximating a reference posterior distribution than those obtained using a single long run of the same length as the metachain. Precision appears to be best when heated Markov chains have low temperatures, whereas chains with high temperatures appear to sample trees with high posterior probabilities only rarely. [Bayesian phylogenetic inference; heating parameter; Markov chain Monte Carlo; replicated chains.]

Bayesian phylogenetic analysis implemented via Markov chain Monte Carlo (MCMC) is gaining widespread use in the phylogenetics and systematics communities (Huelsenbeck et al., 2001; Larget and Simon, 1999; Li et al., 2000; Mau et al., 1999; Rannala and Yang, 1996; Yang and Rannala, 1997). The core of Bayesian methodology, as it applies to a broad range of applications including phylogenetic inference, can be summarized in the following steps (see also text by Gelman et al., 2004):

- construction of a *sampling distribution*, $p(D|\theta)$, which defines the putative distribution of possible observations D given unknown parameters θ,
- construction of a *prior distribution*, $p(\theta)$, which expresses what is known or believed about the parameters θ prior to observing the data,
- analytic determination of the *posterior distribution*, $p(\theta|D)$, using Bayes' rule and given actual observations D,
- sampling of the posterior distribution (possibly using MCMC) and estimation of certain desired probabilities (such as bipartition probabilities in phylogenetics) based on this sample. This step is known as Monte Carlo estimation and it is performed only if analytic determination of the required probability is difficult or impossible.

The sampling distribution and the prior distribution together constitute a Bayesian model. In fact, the above methodology can be generalized to allow for several different models, by conditioning on the model M in the sampling and prior distributions, $p(D|\theta, M)$ and $p(\theta|M)$, and by assigning prior probabilities p(M) to the various models (Green, 1995). However, a tree topology can equally well be regarded as a model or a parameter, and hence Bayesian phylogenetics can be implemented using conventional Metropolis-Hastings algorithms (Metropolis et al., 1953; Hastings, 1970).

Often the major practical issue in implementing Bayesian methodology is to ensure adequate sampling of the posterior distribution. This can be hindered by poor convergence and mixing properties of the MCMC algorithm used for sampling. In theory, properly constructed MCMC methods are ergodic; they must eventually converge to the intended distribution (Roberts, 1996; Roberts and Rosenthal, 1996; Tierney, 1994, 1996). (The term "convergence" is used here in the distributional sense appropriate when discussing Markov chains. A sequence of values generated by MCMC does not converge to a point, but rather the distribution of *potential* values converges to a limiting distribution.) In practice, however, MCMC methods can spend many iterations trapped in the vicinity of a single mode of a multimodal distribution or in a limited region of the space. Slow mixing may necessitate a very long sampling phase to obtain low variance Monte Carlo estimates (Mossel and Vigoda, 2005). Multimodality, and consequent problems with convergence and mixing, is known to occur in phylogenetic applications (Chor et al., 2000; Hillis et al., 2005; Maddison, 1991; Salter and Pearl, 2001; Steel, 1994). Indeed, evidence we present in this paper suggests that multimodality is typical when the number of taxa is large (>30).

An important factor influencing the efficiency of Metropolis-Hastings samplers is the choice of proposal mechanism (Al-Awadhi et al., 2004; Brooks et al., 2003b; Gilks et al., 1996; Rotondi, 2002). Ideally, proposals should facilitate transitions between alternative modes,

and such proposals are the subject of current research (see, for example, Chauveau and Vanderkhove, 2002; Tjelmeland and Hegstad, 2001). The question of tuning proposal mechanisms for phylogenetic MCMC deserves explicit treatment in the literature, but is beyond the scope of this paper.

An alternative method of facilitating transitions between alternate modes, employed by the MrBayes program (Huelsenbeck and Ronquist, 2001; Altekar et al., 2004) is Metropolis-coupled MCMC (Geyer, 1991; Geyer and Thompson, 1996; Gilks and Roberts, 1996), in which several Markov chains are run simultaneously. One chain, called the "cold" chain, is designed to have the required posterior distribution as its limiting distribution, whereas the other chains, referred to as "heated" chains, have distributions proportional to the posterior distribution raised to a power less than one as their limiting distribution. Heated chains thus converge to a distribution that is closer to uniform than the posterior distribution, and so are less likely to generate long sequences of values in the neighborhood of a single mode. These heated chains are allowed to swap elements with the cold chain (or with other heated chains), thus enabling the cold chain to converge and mix more rapidly. Users of MrBayes are required to select the number of chains and the "heating parameter," which determines the "temperatures" (inverses of the exponents) of the heated chains. Only the cold chain is used to generate samples.

Another important practical issue in MCMC is determining at what point the burn-in period should be terminated; that is, at what point convergence can be said to have occurred. If the starting point for an MCMC analysis is a random point in the sampled space, there will typically be a steady increase in the likelihood score of elements generated by the chain. At some point, however, likelihood scores will stabilize (they do not converge, but decreases become more frequent until they balance increases). Such stabilization is typically assessed through graphical inspection of serial log-likelihood values and may be regarded as a necessary but not sufficient criterion for convergence of the Markov chain. Methods for assessing MCMC convergence are reviewed by Cowles and Carlin (1996) and Brooks and Roberts (1998), but there are currently no fail-safe methods: in practice, convergence can be easy to reject but impossible to accept, unless the chains are long enough to sample from every point in parameter space. In this paper, we propose a simple and pragmatic criterion for assessing convergence.

MCMC practitioners are divided over whether it is more efficient to base inferences and convergence diagnostics on one long chain or on two or more replicate runs started from independent, over-dispersed starting points (Gelman and Rubin, 1992; Geyer, 1992; Huelsenbeck et al., 2002). Geyer (1992) made a strong case for the use of a single long chain, arguing that some statistical distributions (such as the "witch's hat") can show similar results across many chains that had not yet converged. Another disadvantage of replicate chains is that each chain has its own burn-in phase, which must be discarded. When convergence is slow, a substantial amount of computational effort is thus expended in generating samples that must be discarded. Gilks et al. (1996) claim that "It is now generally agreed that running many short chains, motivated by a desire to obtain independent samples from [the target distribution], is misguided unless there is some special reason for needing independent samples." However, in the multiple-chain approach, the precision of estimated posterior probabilities across replicates can be used to support the accuracy of the mean of these estimates. Many MCMC convergence diagnostics used in the statistical community (Brooks and Gelman, 1998; Brooks and Giudici, 2000; Brooks et al., 2003a; Brooks and Roberts, 1998; Cowles and Carlin, 1996; Gelman, 1996; Gelman and Rubin, 1992; Li et al., 2000) rely on stable posterior probability estimates from multiple replicated chains. It is generally recognized that when multiple processors are available, it is worthwhile to run multiple chains simultaneously (Gilks et al., 1996). This strategy has been implemented in version 3.1.1 of MrBayes, with convergence diagnostics based on the sampling frequencies of various topological features. Here we present empirical results that show the practical value of working with multiple short chains.

As with other phylogenetic methods, most studies incorporating MCMC and based on biological sequences have used nucleotide characters for phylogenetic reconstruction. Amino acid characters have been examined in a minority of cases, such as Ragan et al. (2003) and Xiong and Bauer (2002). Analyses performed mainly to test the properties of Bayesian MCMC have also tended to focus on nucleotide rather than amino acid characters (Erixon et al., 2003; Huelsenbeck et al., 2001; Suzuki et al., 2002), with a few recent exceptions (Douady et al., 2003; Mar et al., 2005). Although some of the properties derived from these analyses are likely to be consistent across different types of characters, it is worth investigating amino acids as well, due to the different number of character states, rates of evolution, and substitution matrices that need to be considered. Amino acid characters are also most appropriate to analyse the deep relationships among prokaryotic data sets used below, many of which cover multiple phyla or domains.

We present here a replication-based study of the stability of MCMC runs performed on microbial protein sequence data sets with 5 to 140 members. Because we dealt with empirical (as opposed to simulated) protein data, we do not know the phylogenetic history of these sequences and cannot assume that all sets of genes from a given set of taxa will have the same history, due to the potential influence of factors such as lateral genetic transfer which contradict the classic organismal phylogeny suggested by trees of 16S ribosomal DNA (Brown, 2003; Gogarten et al., 2002; Raymond et al., 2002). Thus, in this study we were not concerned with assigning high posterior probabilities to the "true" phylogenetic relationships of each data set whatever they may be, but rather with getting statistically equivalent estimates of bipartition probabilities from many replicated runs of the same data set.

METHODS

Data Sets

Annotated proteins from 144 sequenced microbial genomes were clustered according to Harlow et al. (2004) to yield 22,437 sets of putatively orthologous protein data sets, each containing between 4 and 144 sequences. Each of these data sets was then aligned using the default settings of T-COFFEE (Notredame et al., 2000). Thirty small (5 taxa), 10 medium (10 taxa), 10 large (30 taxa), and 20 very large (40 to 144 taxa) aligned data sets were drawn from this set, with two conditions: each locus had to contain at least one protein that was not annotated as "putative" or "hypothetical," and each could not have a trivial phylogenetic resolution (which would occur if fewer than four nonidentical sequences were present in a given data set). The result was 70 data sets with different taxonomic distributions and levels of sequence divergence (Appendix 1; available at http://systematicbiology.org). These sets represent a cross section of the orthologous sets used to assess lateral genetic transfer among microbial genomes (Beiko et al., 2005), the results of which needed to be interpreted in light of possible violations of substitution matrix models and of sampling instability. Prior to phylogenetic analysis, the least reliable regions of each alignment were removed using GBLOCKS (Castresana, 2000). Conservative settings (described in Beiko et al., 2005) were used such that most of each alignment was retained, with the elimination of only the "ragged" ends and other regions with large numbers of alignment gaps. All 70 alignments can be accessed via the Systematic Biology website at http://systematicbiology.org.

Protein Analysis with MrBayes

The first phase of the analysis consisted of replicated MCMC runs (with Metropolis coupling) using MrBayes 3.04 beta (Altekar et al., 2004; Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). The main purpose of this phase was to investigate whether estimates of bipartition posterior probabilities, which are more stable than posterior probabilities of entire trees (see Evaluation of Likelihood Stabilization and Chain Convergence below), differed among replicated runs that had apparently converged, inasmuch as their likelihoods appeared to have stabilized. Each of the 70 data sets was analyzed, with among-site rate variation modelled using a four-category discrete approximation to the gamma distribution (Yang, 1994). Instead of choosing a single amino acid substitution matrix, the command "prset aamodelpr=mixed(0,0.25,0,0.25,0,0.25,0,0,0.25,0)" was used, allowing the Markov chain to move among four empirically derived matrices: JTT (Jones et al., 1992), mtREV (Adachi and Hasegawa, 1996), VT (Muller and Vingron, 2000), and WAG (Whelan and Goldman, 2001). These empirical matrices all have the same number of free parameters, so higher likelihood values associated with one substitution matrix must be due to better model fit, and not to the addition of potentially superfluous parameters. Ten replicate runs, each consisting of a single MCMC analysis with a randomly chosen starting tree, were performed for each data set. The total Markov chain length was 500,000 generations for the 5- and 10-taxon data sets, 1 million generations for the 30-taxon data sets, and 2 million generations for the 20 largest data sets. All MrBayes runs in this analysis consisted of four Markov chains (one cold and three heated), with a uniform prior between 0.1 and 50 for the gamma-shape parameter, and an exponential prior with an expectation of 0.1 (brlenspr=Unconstrained:Exp(10.0) in MrBayes) for the distribution of branch lengths. We retained the default proposal mechanism in MrBayes 3.04 beta, which defines the relative probabilities of changing a particular parameter of the model: at each iteration of the MCMC sampler, there was a probability of 0.5769 associated with a nearest-neighbor interchange to the current tree topology; of 0.1154 to a tree bisection and reconnection (TBR) move; of 0.2308 to a change in the length of a branch in the tree; and 0.0385 each to changes in the gamma-shape parameter or the choice of empirical amino acid substitution matrix. Although the choice of proposal mechanism can have a substantial impact on the sampling effectiveness of a Markov chain, we kept these default ratios constant within this work to focus on the issue of short versus long chains. Except for variation in chain length and temperature as described below, all of the MrBayes settings used in phases 2 and 3 were identical to those used in phase 1.

In the second phase, five 30-taxon data sets (30-2, 30-3, 30-6, 30-8, and 30-9) were selected at random and five replicates of each were run for 101 million generations. The burn-in phase was identified and discarded using the method described in the following section, and the first 100 million generations after the end of burn-in summarized to yield "reference" estimates of bipartition posterior probability. The posterior probability estimates from these runs were treated as benchmarks for comparison with estimates based on shorter runs if all five reference runs yielded essentially the same mean bipartition posterior probability estimates (based on values of the ε statistic; see following section for further details). Shorter replicated runs (between 10,000 and 1 million generations) were combined and compared to single runs of equivalent length to assess both the precision of the bipartition posterior probability estimates across replicates of the same length, and the similarity of both estimates to the corresponding reference runs.

The effect of chain temperature (as determined by the heating parameter, T) on apparent convergence and parameter mixing was examined in the third phase of the analysis. Three data sets (30-8, VL-4, and VL-19) that had produced particularly variable bipartition posterior probability estimates (given their size) in the second phase of analysis were subjected to 10 replicated runs of 2 million generations each, with a total of 13 different settings of T: 0.01, 0.05, 0.1, 0.2 (the default), 0.5, 0.75, 1, 1.33, 2, 5, 10, 20, and 100. Convergence was assessed by comparing the bipartition posterior probability estimates derived from the post-likelihood stabilization phase of each Markov chain for each temperature setting.

Evaluation of Likelihood Stabilization and Chain Convergence

Because all MCMC runs in this analysis began with random tree topologies, the first portion of each run contained a set of low-probability trees that preceded sampling of a stationary distribution. A common way to assess convergence is through visual examination of a plot of likelihood versus generation: if stabilization of likelihood values is observed, then the samples that precede this stable condition are discarded. We implemented an automatic assessment of this effect that is more conservative than visual inspection. For each run, we searched for the end of the burn-in phase by comparing successive log-likelihood values along the Markov chain and marked as the end the first tree in the chain that had a log-likelihood value greater than the mean of the last 100,000 generations in the entire run. A very early point thus identified would signify a quick rise to stationarity, while a point near the end of the run (especially within the last 100,000 generations) could indicate that stationarity was never reached. In phase 1, the largest burn-in value obtained for each protein data set was rounded up to the next multiple of 50,000 generations, and this number of generations was discarded from the analysis. Because no burn-in value exceeded 500,000 generations in phase 2, this number of generations was discarded from the beginning of each run prior to analysis of results.

To assess whether apparent likelihood stabilization (as defined above) is a reliable indicator of convergence with respect to topology, we compared summary statistics for the replicated runs. Overall estimates of the posterior probability of trees were not used as a comparative criterion for practical reasons: in the 5-taxon cases, support levels of the 15 possible unrooted trees could easily be compared based on their posteriors, but in the 30-taxon cases, in which 30,000 trees (3 million generations/100 generations per sample) were sampled, there were often more than 25,000 distinct trees, with few or no trees appearing in more than 1% of samples. Because the maximum a posteriori (MAP) tree (Huelsenbeck et al., 2002; Rannala, 2002; Yang and Rannala, 1997) is not strongly supported, we chose to examine bipartition posteriors summarized across the entire MCMC run. The number of distinct bipartitions sampled in these cases was much lower than the number of distinct trees, and bipartition posteriors from different runs (or within different fragments of a single run) could easily be compared. These summaries were generated by running the "sumt" command in MrBayes on the generated tree files. Perl scripts (available on request) were used to summarize the bipartition data.

We define two quantities that are useful in comparing the posterior probabilities of multiple Markov chains, which monitor the convergence of bipartition or clade probabilities as proposed by Huelsenbeck et al. (2002). The quantity δ is the aggregated difference between estimated posterior probabilities for a set of bipartitions sampled from a pair of MCMC runs. For a data set with n sequences, a strictly bifurcating tree will have a total of (n − 3) bipartitions, so δ will be bounded by zero (all sampled bipartitions have equal posterior probabilities in both runs) and 2(n − 3) (all bipartitions sampled in one chain are completely absent from the other). This absolute aggregated difference is useful in describing the sampling variability between a pair of experiments, each of which could be a single MCMC run or a summary of multiple runs that reports the mean estimated posterior probability for each bipartition. The quantity ε is used to summarize the variability within a set of replicated MCMC runs. The replicated MCMC runs from each set of experiments were summarized by computing the mean estimated posterior probability and standard deviation of each observed bipartition across the set of replicates. Nodes with unstable support across replicates could thus be easily identified, and the sum of all bipartition standard deviations was computed to produce ε, an overall picture of the stability across replicates.

RESULTS

First Phase: Convergence of Small and Large Data Sets

The burn-in values, calculated in the manner described above for each replicate of the 70 data sets examined in phase 1, are shown in Figure 1. The burn-in phase is apparently very short for the small data sets: the maximum burn-in value of the 300 runs performed on the 5-taxon data sets was only 1900 generations. Calculated burn-in values >50,000 generations are rarely observed for data sets with fewer than 80 sequences, but a rapid rise in the number of required generations is seen above this level. Most data sets with >80 sequences had at least one calculated burn-in value >100,000 generations: data set VL-16 (102 proteins) was an extreme case, requiring in one case over 400,000 generations in the burn-in phase. Perhaps not coincidentally, this data set was by far the shortest (only 67 amino acids) after being trimmed with GBLOCKS.

The estimated posterior probabilities associated with each of the four substitution matrices considered are shown in Figure 2. In a majority of cases (60 out of 70), one of the four permitted amino acid substitution matrices had an estimated posterior probability greater than 0.99. Although the WAG matrix was preferred in all size categories, larger data sets showed the strongest preference for WAG, with 9 out of ten 30-sequence data sets and 20 out of 20 very large data sets assigning a PP $\geq$ 0.99 to this matrix. The mtREV matrix was a considerably worse fit to the data than the other matrices, and never had an estimated posterior probability >0.01 in any of the 50 data sets. Because a mitochondrial substitution regime was not expected in the microbial data sets, a non-negligible posterior probability for the mtREV matrix in any of the data sets would have most likely suggested a poor fit of the data to any substitution matrix, rather than a specific preference for mtREV.

The log-likelihood appeared to stabilize in all 700 runs. This is consistent with convergence, but does not guarantee it. We examined more powerful ways of

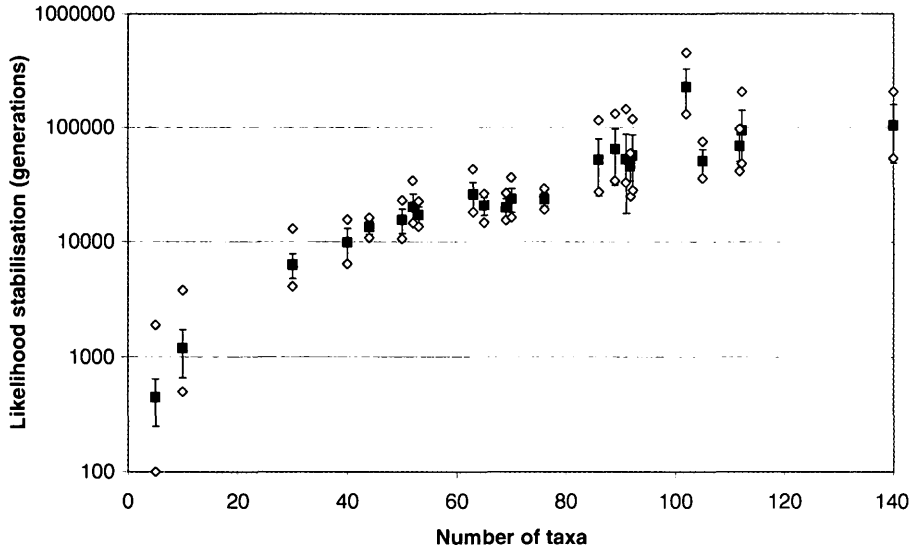

FIGURE 1. Likelihood stabilization points recorded for each replicate run among 70 protein data sets (10 replicates each of 30 data sets of size 5, 10 data sets of size 10, 10 data sets of size 30, and 20 data sets of sizes 40 to 140). Mean values ± standard deviation are shown, and empty diamonds represent the maximum and minimum values.

rejecting convergence by comparing the average log-likelihoods and the bipartition posterior probabilities among the ten replicated runs of each protein data set. Significant differences in these values would suggest either that post-convergence sampling was inadequate or that convergence had not occurred. The results are summarized in Table 1. The number of generations summarized depends on the size of the protein data set: 250,000 generations after apparent burn-in for the 5-taxon data sets, 300,000 for the 10-taxon data sets, 500,000 for the 30-taxon data sets, and 1.5 million for the data sets of sizes 40 to 140.

The standard deviations of the mean log-likelihoods across 10 replicate runs are shown in Table 1 for each protein data set. The values appear at first glance to be small. Although the range of log-likelihood scores was frequently greater than 30 within the sampling phase of most replicates, the standard deviation of replicate means was typically less than 1.0, with the exception of the seven largest data sets. The deterioration of apparent convergence is illustrated in Figure 3: though some bipartitions in both sets appear to mix well, the ranges of bipartition posterior probabilities are much larger in data set VL-19 than in set VL-4.

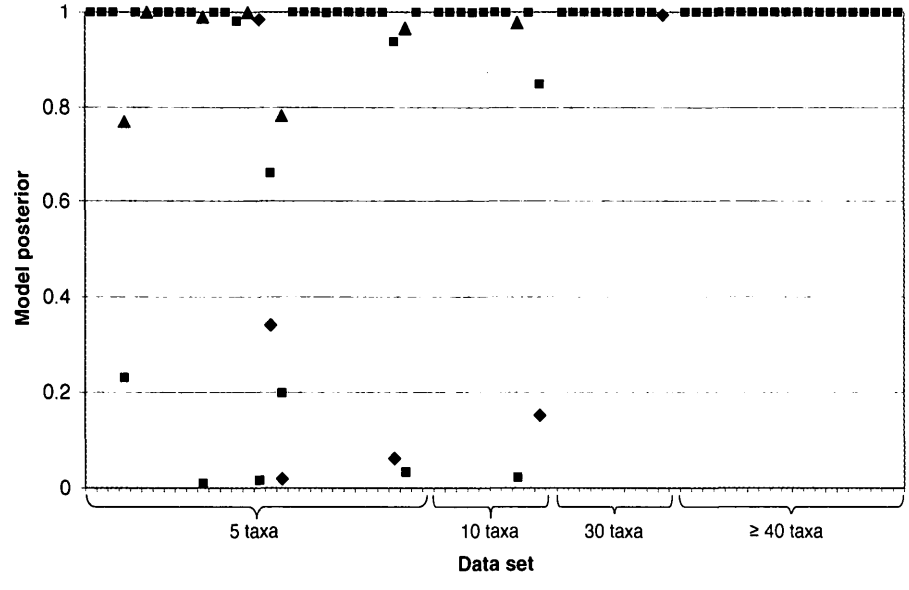

FIGURE 2. Posterior probabilities of amino acid substitution models across 10 replicates of 50 protein data sets. Models that achieved a posterior probability greater than 0.01 are shown: WAG (squares), VT (triangles), and JTT (diamonds). The mtREV model, which was permitted but never achieved a posterior above the minimum threshold, is not shown.

TABLE 1. Markov chain Monte Carlo sampling stability across replicates for thirty 5-taxon protein data sets, ten 10-taxon protein data sets, ten 30-taxon protein data sets, and 20 data sets of sizes 40 to 140. For each data set, the mean log-likelihood of all sampled trees is shown, followed by the standard deviation of this mean, computed from the mean value of each replicated run ($n = 10$). The last three columns summarize the support of bipartitions sampled during the MCMC run. The first of these columns shows the number of bipartitions with a mean posterior probability greater than 0.01 from the set of 10 replicated runs. The second bipartition column indicates the maximum difference in posterior observed for any single bipartition between a pair of replicates, which is bounded by 0 (posterior probabilities are identical across all replicates) and 1 (at least one bipartition has a posterior of 1.0 in one replicate, and 0.0 in another). The last column is the mean of all standard deviations of bipartition posterior probabilities for the set of replicates.

Data set	Mean lnL	Standard deviation of lnL	No. of nodes PP > 0.01	Maximum PP difference	Mean SD of PP	Data set	Mean lnL	Standard deviation of lnL	No. of nodes PP > 0.01	Maximum PP difference	Mean SD of PP
5-1	−1913.31	0.06	4	0.042	0.0069	10-6	−1223.42	0.09	11	0.024	0.0023
5-2	−2929.87	0.04	4	0.037	0.0072	10-7	−1338.34	0.09	12	0.037	0.0004
5-3	−8388.76	0.06	2	0.003	0.0004	10-8	−4143.75	0.13	16	0.039	0.0064
5-4	−2521.29	0.10	4	0.027	0.0059	10-9	−2405.62	0.10	13	0.054	0.0063
5-5	−1613.14	0.12	2	0.006	0.0008	10-10	−1618.22	0.14	29	0.030	0.0043
5-6	−3103.65	0.05	4	0.019	0.0040	30-1	−5934.39	0.53	133	0.458	0.0277
5-7	−3004.77	0.04	2	0	0	30-2	−11,232.6	0.42	62	0.211	0.0113
5-8	−5632.91	0.04	2	0	0	30-3	−9043.14	0.32	50	0.114	0.0090
5-9	−3282.97	0.06	4	0.016	0.0027	30-4	−15,890.4	0.49	63	0.389	0.0295
5-10	−5750.92	0.09	2	0.011	0.0015	30-5	−14,807.9	0.32	42	0.130	0.0080
5-11	−1134.32	0.06	2	0.001	0.0001	30-6	−8546.73	0.30	42	0.073	0.0067
5-12	−2933.18	0.04	9	0.032	0.0065	30-7	−11,020.8	0.26	43	0.093	0.0060
5-13	−3261.93	0.05	2	<0.001	<0.0001	30-8	−4974.37	0.29	45	0.074	0.0072
5-14	−1383.08	0.12	2	0	0	30-9	−4510.52	0.35	96	0.094	0.0068
5-15	−4846.30	0.05	4	0.022	0.0040	30-10	−3896.52	0.40	83	0.064	0.0072
5-16	−2199.59	0.06	4	0.002	0.0058	VL-1	−10,449.8	0.36	96	0.379	0.0265
5-17	−2941.04	0.04	2	0.002	0.0004	VL-2	−17,117.1	0.32	75	0.086	0.0061
5-18	−3093.99	0.05	2	0.001	0.0002	VL-3	−16,943.8	0.55	75	0.139	0.0098
5-19	−2482.52	0.06	4	0.020	0.0043	VL-4	−6367.03	0.53	173	0.441	0.0167
5-20	−6847.96	0.04	2	0	0	VL-5	−15,636.1	0.52	122	0.070	0.0055
5-21	−2879.45	0.05	2	0	0	VL-6	−14,630.7	0.69	140	0.473	0.0237
5-22	−1752.12	0.09	4	0.021	0.0039	VL-7	−12,130.0	0.59	150	0.143	0.0084
5-23	−1958.37	0.14	4	0.043	0.0072	VL-8	−11,458.5	0.28	128	0.223	0.0105
5-24	−4938.42	0.06	2	0.003	0.0004	VL-9	−13,548.2	0.48	194	0.148	0.0091
5-25	−4702.04	0.05	4	0.021	0.0040	VL-10	−19,998.7	0.45	129	0.280	0.0118
5-26	−3381.38	0.06	2	<0.001	0.0001	VL-11	−25,466.2	0.58	213	0.380	0.0189
5-27	−1686.42	0.12	2	0.027	0.0056	VL-12	−27,703.1	0.48	165	1.000	0.0228
5-28	−1501.86	0.09	4	0.023	0.0046	VL-13	−25,849.9	0.90	159	0.308	0.0174
5-29	−3031.08	0.05	2	0	0	VL-14	−20,094.1	1.30	240	0.556	0.0323
5-30	−1642.59	0.13	2	0	0	VL-15	−26,418.2	1.35	189	0.576	0.0281
10-1	−11,294.7	0.13	11	0.076	0.0092	VL-16	−6436.22	4.89	651	1.000	0.0671
10-2	−3492.61	0.06	8	0.026	0.0022	VL-17	−24,958.1	1.39	235	1.000	0.0607
10-3	−3771.54	0.10	7	0.003	0.0003	VL-18	−16,804.0	4.46	326	1.000	0.0605
10-4	−3744.68	0.14	12	0.048	0.0047	VL-19	−17,193.5	3.20	429	1.000	0.0892
10-5	−6372.96	0.14	7	<0.001	<0.0001	VL-20	−36,394.6	7.30	363	1.000	0.0769

This increased difference between replicate Markov chains is also evident when the number and frequency of sampled bipartitions are considered. Within the smallest protein data sets, the extreme and average differences between estimated bipartition posterior probabilities are small (Table 1). The largest difference in estimated bipartition posterior probability between any pair of 5-taxon replicates is 0.043, seen in data sets 5 to 23. This difference results from an estimated posterior probability of 0.686 in replicate 9 for a certain bipartition, with a corresponding figure of 0.643 in replicate 1. The average of all standard deviations of bipartition posterior probability within each protein data set was always less than 0.01, again showing very stable estimates. The maximum difference between bipartitions increased with larger protein data sets, with a difference of 0.076 in data set 10-1, and 0.458 in data set 30-1. The largest differences were seen in the largest protein data sets: the data sets of size 89, 102, 105, 112, and 140 all exhibited bipartitions that had a posterior probability of 1.0 in one or more out of 10 replicates, but were completely absent (posterior = 0.0) in others. Thus, taxonomic groupings with the highest possible level of support within one Markov chain were completely absent from another. For instance, one of a pair of incompatible bipartitions is favored in 9 out of 10 replicate runs of data set VL-12, but is completely absent from the tenth. Both bipartitions are present to some extent in four of the replicates, so it is possible (though rare) for the Markov chain to move from one configuration to the other. The switch from one bipartition to the other occurs in both directions in some replicate runs, so it appears that both are significant components of the stationary distribution.

It is clear from these observations that either convergence had not occurred or individual chains had not been run long enough after convergence to generate an adequate sample. The former possibility is usually assumed when parallel runs differ significantly, but the

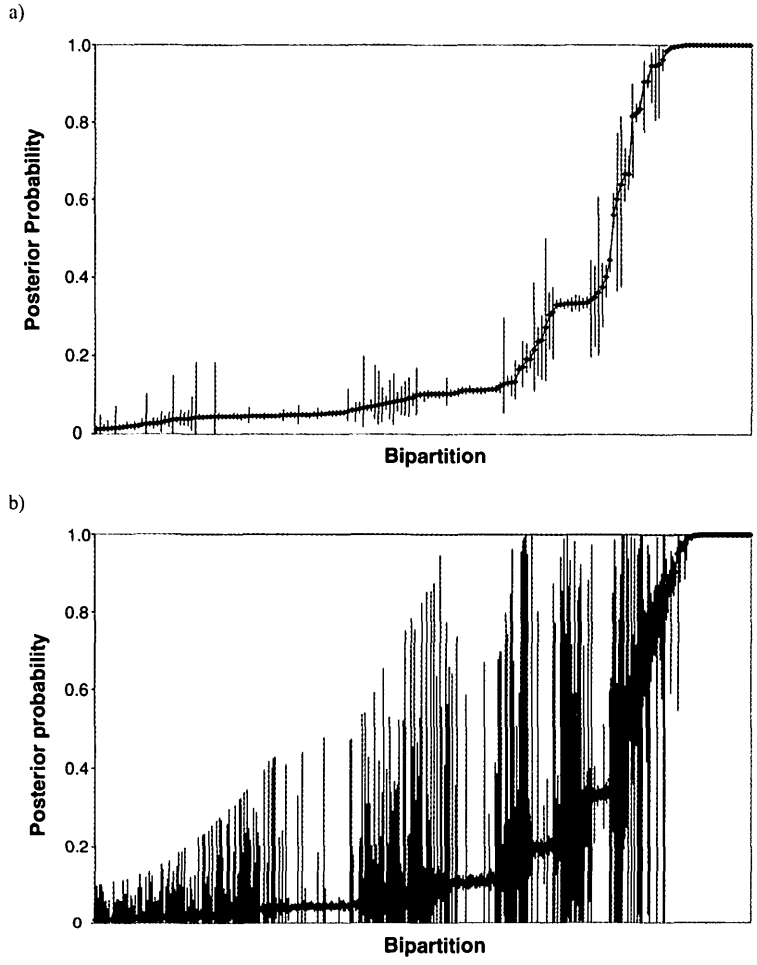

FIGURE 3. Mean posterior probability of bipartitions sampled in 10 replicate MCMC runs of data sets (a) VL-4 (52 sequences) and (b) VL-19 (112 sequences). Bipartitions are sorted in increasing order of posterior probability; vertical bars indicate the range of estimated posterior probabilities across replicates.

latter explanation is also plausible. Even after convergence, slow-mixing chains may move only infrequently between modes or distant regions, so that small samples from individual chains may have significant differences even if collectively they represent an adequate sample.

Second Phase: Accuracy of Single Long Versus Several Short Runs

If the variation in estimated bipartition posterior probabilities seen in the larger data sets is due to inadequate sampling rather than lack of convergence, then there are two possible ways to obtain less variable estimates. One is to use longer MCMC runs. However, MCMC is computationally expensive, and the sharp rise in bipartition instability observed in the largest protein data sets suggests that runs of 10 million generations or greater may be required (unless improved mixing can be achieved by other means, such as changes to the proposal mechanism of the sampler). An alternative strategy that may be preferable when multiple processors are available is to run a large number of chains in parallel from different starting points and pool the samples. However, if the variation is due to a lack of convergence, only the former strategy (running longer chains) is appropriate.

To test whether multiple short runs could be used to estimate bipartition probabilities, we examined five 30-taxon data sets in detail. Five replicate runs of each of these data sets were performed from different random starting points to yield a more reliable estimate of the bipartition posterior probabilities, against which shorter runs of the same data sets could be judged. MCMC runs of 101 million generations were performed, with the first 100 million generations after burn-in used to compute bipartition posterior probabilities. When the five long replicates of each data set were compared, the largest standard deviation of posterior probability for any bipartition was 0.004, and 98.4% of all bipartitions had associated standard deviations less than or equal to 0.001 (Table 2).

For comparative purposes, the short, replicated runs were combined into a single metachain and summarized to yield a set of mean bipartition posterior probabilities. Figure 4 shows that for 13 out of 25 combinations of run length and data set, the metachain summaries

TABLE 2. Markov chain Monte Carlo sampling stability across replicates for 5 30-taxon protein data sets. For each data set, the mean log-likelihood of all sampled trees is shown, followed by the standard deviation of this mean, computed from the mean value of each replicated run ($n = 5$). The last three columns summarize the support of bipartitions sampled during the MCMC run. The first of these columns shows the number of bipartitions with a mean posterior probability greater than 0.01 from the set of ten replicated runs. The second bipartition column indicates the maximum difference in posterior observed for any single bipartition between a pair of replicates, which is bounded by 0 (posterior probabilities are identical across all replicates) and 1 (at least one bipartition has a posterior of 1.0 in one replicate, and 0.0 in another). The last column is the mean of all standard deviations of bipartition posterior probabilities for the set of replicates.

Data set	Mean lnL	Standard deviation of lnL	No. of nodes PP > 0.01	Maximum PP difference	Mean SD of PP
30-2	−11,232.70	0.07	129	0.004	<0.001
30-3	−9043.20	0.05	111	0.001	<0.001
30-6	−8546.68	0.03	81	0.002	<0.001
30-8	−4974.31	0.03	120	0.002	<0.001
30-9	−4510.70	0.01	196	0.001	<0.001

were more similar to the target posterior probability than summaries of a single Markov chain of the same length, whereas in the remaining 12 cases the single Markov chain was more similar to the reference. The ratio of metachain δ value to single run δ value varied between 0.42 (30-3, 10,000 generations) and 1.51 (30-2, 50,000 generations). Nonconvergence as expressed by δ values was also influenced by the data set under consideration, and there was a strong trend of increasing accuracy with increasing length of the Markov chain. The accuracy of individual MCMC runs may be data set and start point–specific, but these results give no clear advantage to either single long or multiple short chains in terms of accurately estimating the reference distribution of posterior probabilities. This observation is consistent with the possibility that variation among parallel chains is due to slow mixing rather than lack of convergence.

There was a weak tendency for metachain and long chain δ values to become more similar to one another as runs increased in length. For each run length, we pooled the ratio of the larger δ value to the smaller across all five data sets to obtain a mean ratio for that run length. The largest such ratio was 1.54, obtained from the pooled samples corresponding to the shortest run length, and the smallest was 1.16, obtained from pooled δ ratios of the longest runs. Although there is some instability in these estimates due to small sample sizes, it appears that the ratio of single long chain δ to metachains δ will approach 1.0 as run length increases, which would be expected where run length is infinite. For the metachains considered here, there was also a strong relationship between the metachain/reference chain δ value and the computed ε value for the metachain ($R^2 = 0.753$, $df = 23$, $P = 1.16 \times 10^{-8}$).

Third Phase: Mixing and the Effect of Heating on Convergence

If a Markov chain is mixing well, then it should provide an adequate sample of the posterior distribution in a relatively small number of generations. Mixing is affected by the complexity of the model under consideration,

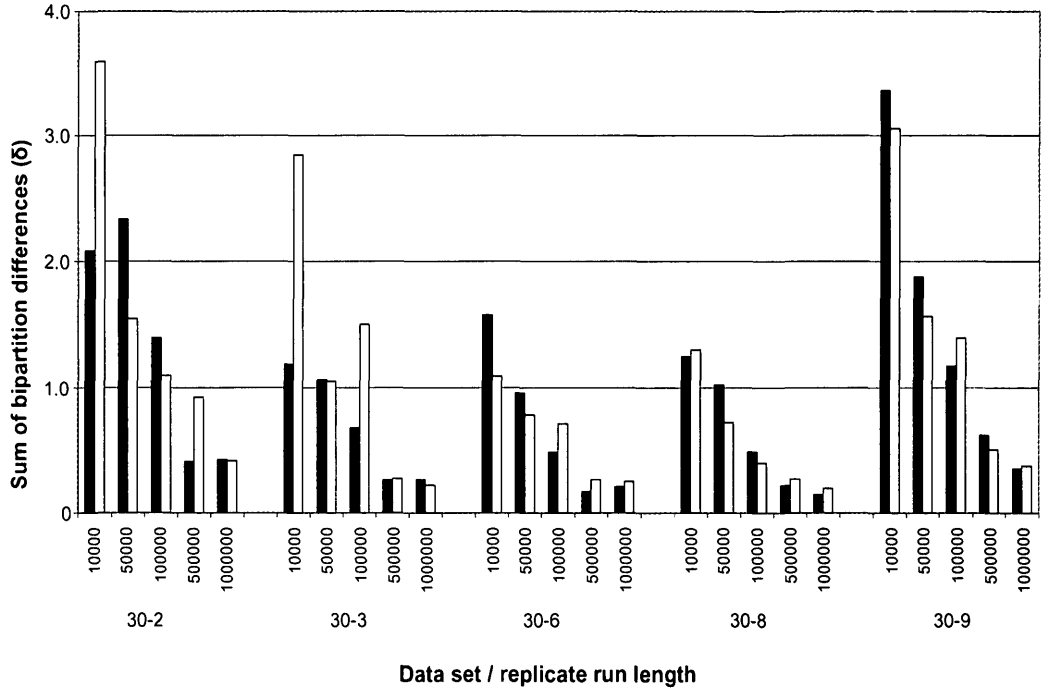

FIGURE 4. Comparison of bipartition summaries for 10 replicate short runs of data sets 30-2, 30-3, 30-6, 30-8, and 30-9 versus a single long run. The height of each bar is equal to the δ value between the reference runs (of length 5 × 100 million generations) and either a metachain of the length indicated on the x-axis assembled from 10 short runs (filled bars), or a single run of length equal to the corresponding metachain (empty bars). Larger values of δ indicate a greater difference between the bipartition posterior probabilities estimated in a given run and the corresponding reference posterior estimate, implying a less accurate estimate of bipartition posterior probabilities from that run.

the operations used to change the model parameters at each generation, and (in Metropolis-coupled MCMC) the number and temperatures of the heated Markov chains. The temperature of a heated chain affects the probability that it will accept a proposed move that yields a drop in posterior probability. Temperature settings that are too low may produce heated chains that have difficulty escaping local regions of high posterior probability, while temperature settings that are too high will produce chains that accept many proposed moves and sample almost randomly from tree space. The convergence and mixing properties for three data sets (30-8, VL-4, and VL-19) of different sizes were assessed by comparing the bipartition posterior probabilities across replicates for 13 different values of the heating parameter T, by investigating the frequency of chain swaps.

The purpose of Metropolis-coupled MCMC is to improve the mixing properties of the cold chain and thus to decrease the variance of Monte Carlo estimates such as the bipartition probability estimates. The quantity ε, which is based on the standard deviation of bipartition probability estimates across replicated runs, thus provides a measure of the success of the Metropolis-coupled strategy and a means of identifying the optimal heating parameter T for a given data set. Figure 5 summarizes the relationship between ε and T for the above-mentioned data sets. Uncorrected values of ε varied widely between the sets, with values for 30-8 in the range 0.25 to 0.5, values for VL-4 between 1.5 and 6.0, and values for VL-19 between 28 and 61. To allow comparisons across sets, values of ε computed for each of the three data sets were divided by the maximum ε obtained from that data set. All three data sets showed a similar pattern of normalized ε variation versus T, with lower ε for T values between 0.01 and 0.2, followed by an increase across T of 0.5, 0.75 and 1.0, then little variation up to 100. Thus it appears that for these data sets, lower T values yield more stable estimates of bipartition posterior probabilities. This increased precision does not necessarily imply better accuracy: it is possible that very low temperature values prevent effective sampling, such that the Markov chain is unlikely to escape regions of high likelihood that are locally good but globally suboptimal. To assess whether the chains from runs with low T values did in fact converge on the target distribution, we computed the mean posterior probability across ten replicates for each T setting of data set 30-8, and computed the δ value between each of these and the mean posterior probability of the five sets of reference runs from the previous experiment. Regression analysis showed no relationship ($R^2 < 1.0 \times 10^{-5}$) between the T value of each set of 10 replicated runs and the corresponding δ value, suggesting that low values of T did not lead to poor convergence relative to higher T values. Instead, it appears that low T values yielded the estimators with the lowest variance, indicated by low ε values across replicates.

It is clear that the Metropolis-coupled strategy is not working here; heated chains are apparently not producing good proposals for swaps. To test whether this is the case, we examined the frequency with which proposed chain swaps occur. Because they do not contribute directly to the posterior distribution of model parameters,

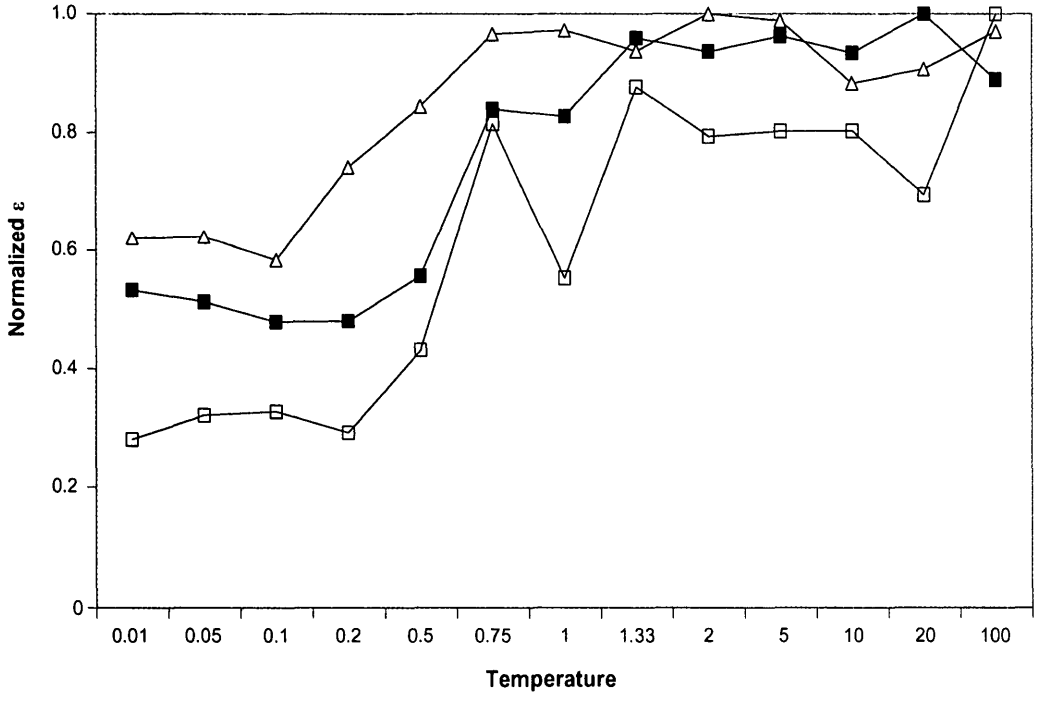

FIGURE 5. Normalized ε values computed from 10 replicates of 13 different temperatures for data sets 30-8 (triangles), VL-4 (empty squares), and VL-19 (filled squares). Higher values of ε indicate a wider spread of bipartition posterior probability estimates across a given set of replicate runs, and therefore lower precision of those individual estimates.

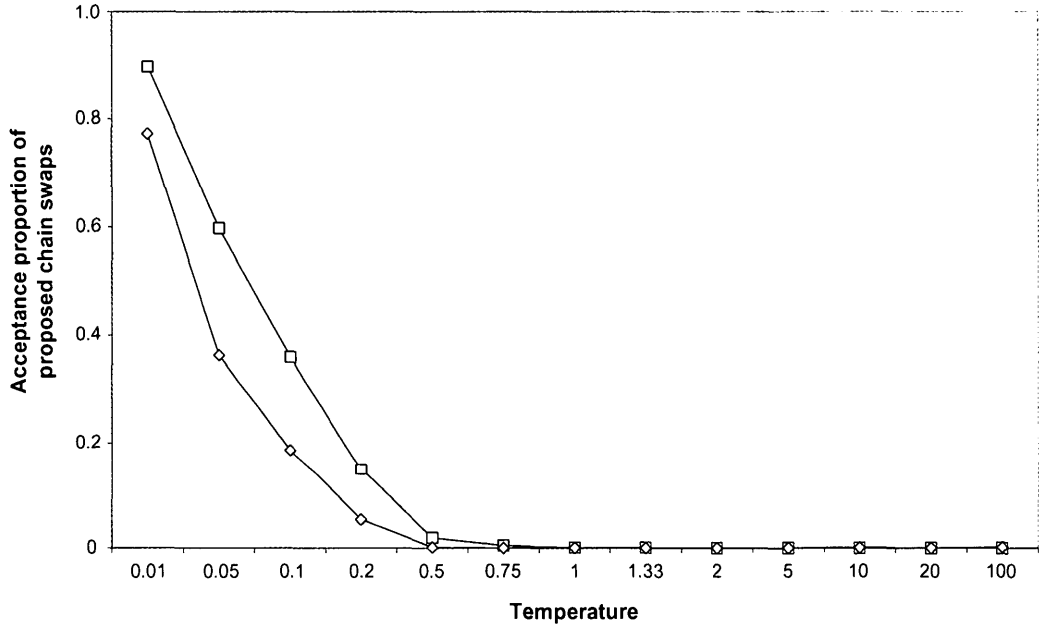

FIGURE 6. Acceptance proportions of proposed chain swaps in Metropolis-coupled MCMC runs for data sets VL-4 (squares) and VL-19 (diamonds). Each point represents the mean acceptance proportion across 10 replicated runs, each 2 million generations in length.

heated chains are useful only when they are sampling regions of high posterior probability. High-temperature chains have a greater probability of accepting moves that yield a decrease in posterior probability, and therefore have a higher capacity to sample widely through parameter space. However, if this wider sampling never produces models that are as good as those that are being sampled by the cold chain, then swaps between the two chains will be rare. Figure 6 shows the relationship between T value and chain swap acceptance for data sets VL-4 and VL-19: in both data sets, there was a monotonic decrease of chain-swap acceptance with increasing temperature. T values of 1 or greater had an acceptance proportion that was invariably less than 0.01, implying essentially no benefit from the use of heated chains. Conversely, relatively low T values led to a high frequency of chain-swap acceptance.

DISCUSSION

Data Set Size and Replication

The first phase of our analysis showed the deterioration of sampling stability that tends to occur with increasing data set size. Though the largest protein data sets were run for more generations and generally had burn-in phases that were shorter than 200,000 generations according to our criteria, wide variation in posterior support was seen for some bipartitions in the data sets containing more than 30 sequences. The worst possible case was observed for some bipartitions in protein data sets of size 89 or larger, with bipartitions universally present in some replicate runs (posterior probability = 1.00) and absent from others (posterior probability = 0.00). Thus, as has been suggested previously (e.g., Huelsenbeck, 2001), likelihood stabilization is not analogous to topological convergence: if a large number of trees must be sampled to obtain an accurate estimate of the posterior distribution of bipartitions, then even several million generations after burn-in may not be sufficient to capture this distribution.

Why do these large differences in bipartition posterior probability occur? They are not universal within a protein data set, as even the largest protein data set had some bipartitions that were sampled with equal frequency across all 10 replicates. Nor were they seen in all of the large data sets: none of the data sets of size 91 or 92 showed this effect among their replicated runs. These partitions do not appear to come from disjoint regions of non-zero posterior probability within tree space, because some replicates were able to switch from one state to another, yielding a bipartition posterior greater than 0 and less than 1 that is probably still a poor estimate of the posterior probability. One possible explanation for the behavior of these bipartitions is that they are plausible only in combination with other specific bipartitions, and that topological switching between two plausible alternatives can occur only if other nodes within the tree are suitably arranged. If the conditions for switching of certain bipartitions are rarely met, and the tree is so large that the appropriate change is rarely proposed, then the switch may never occur within the lifetime of any reasonable Markov chain, and which alternative bipartition is completely favored will depend on the sampling path during the burn-in phase. By performing replicated runs, we were able to identify these instabilities, which show that for the poorly mixing feature at least, much longer runs (of potentially one or two orders of magnitude greater) will be necessary to obtain an accurate posterior probability estimate of these poorly mixing bipartitions.

If a sufficient number of iterations is run, then the Markov chain will sample every point in tree space in proportion to its posterior probability. Not surprisingly, we found that 5-taxon data sets, which can produce only 15 tree topologies, required very few iterations prior to convergence on a stable topology distribution. An advantage of small data sets is that the topology space can be traversed in a small number of iterations: for 5-taxon data sets, no more than two TBR operations are necessary to convert one possible tree into another. In contrast, larger trees from larger data sets can be separated by many interconversion steps. The complexity of this problem creates the potential for regions of high posterior probability that can trap a chain of finite length, leading to outcomes that depend on the (typically random) choice of starting tree topology. This dependency is the primary motivation for replicated runs.

Our investigation of single long versus several short runs of 30-taxon data sets in phase 2 showed that several short replicates merged into a metachain yielded estimates of the reference posterior probability distribution that were not consistently better or worse than those estimated from single runs of length equal to the metachain. Metachains can be treated in the same way as a single, long chain, but the key advantage of metachains is that they contain discontinuities in sampling space at the breakpoints between the replicated chains. These discontinuities increase the overall independence of samples within the chain, thus yielding the same spread of samples as a single, highly autocorrelated chain of a greater length and reducing the variance of Monte Carlo estimates. Though we quantified this difference in detail only for data sets containing 30 proteins each, it is likely that the sampling advantages inherent in replicated runs increase as data set size increases.

Effect of Temperature on Sampling Efficiency

The results of the third phase of our analysis suggest that changing the heating parameter T of an MCMC run has an effect on the precision of the estimate of the posterior distribution, and on the mixing of the Markov chain. T settings below 0.5 yielded ε values that were 30% to 60% smaller than the ε values obtained from runs with T settings greater than 1.0. Our results suggest that low heating parameter values, and consequently heated chains of lower temperatures, improve the precision of posterior probability estimates, without a loss of accuracy.

At first glance, this conclusion (that lower temperatures yield more precise estimates of posterior probability) might appear counterintuitive: chains with higher temperatures can sample more widely from tree space, theoretically increasing the chance that the algorithm can move between disjoint regions of tree space that all have high posterior probabilities. However, a potential problem with heated chains is that they can spend much of their time sampling trees with low posterior probability, rarely (if ever) swapping with the cold chain, and ultimately making little or no contribution to the final set of samples. Thus the heating parameter (for a fixed number of chains) must strike a balance between reducing the number of chain swaps and enhancing exploration of the space. In the examples under consideration, it appears that the optimal balance is achieved with a heating parameter of zero or close to zero. In other words, tree space is best explored by parallel cold chains, with or without Metropolis coupling. If this is generally true, then it will be more efficient to run individual cold chains without Metropolis coupling and pool the results, because then all chains can be sampled instead of just the cold chain.

Suggestions for Phylogenetic MCMC

An important practical recommendation resulting from this study is to use multiple parallel chains for phylogenetic MCMC. The reasons for this are several. Firstly, large samples can be generated efficiently by running independent chains on parallel computers, although multiple burn-ins have to be discarded, decreasing gross computational efficiency. This is not possible for a single long chain. Secondly, several mainstream techniques for assessing MCMC convergence can be applied only if there are multiple chains. Thirdly, experimental results presented in this paper demonstrate that bipartition probabilities estimated by pooling the results of replicate runs were no worse on average than those estimated using a single run of equivalent length. Although not examined here, replicate chains that are significantly different may also show this pattern if the differences between chains are due to slow mixing. Slow-mixing chains initialized from an overdispersed distribution of starting points may converge in a distributional sense even if post-burn-in samples from individual short runs are restricted to single modes or limited regions. This claim may seem counterintuitive, but is easily defended by noting that even after convergence a slow-mixing chain may move only infrequently between modes or distant regions, so that samples from individual chains may have significant differences even though collectively they represent an adequate sample. Nevertheless, we do not advocate pooling results from significantly different chains, because in such cases one does not have evidence that convergence has in fact occurred. If one could devise a test for convergence that allows for the possibility of significant differences between small samples due to slow mixing, then such a test would enable estimates based on smaller sample sizes. One possibility is to subdivide the parallel chains into groups containing equal numbers of chains, then look for significant differences between groups, rather than between individual chains.

Whether single long or several short chains are used to estimate the posterior distribution of trees, convergence of a Markov chain should not be assumed without performing a series of tests. The metachain approach did not yield worse estimates of the reference posterior distribution than did single, long chains, and the multiple random starts allowed the use of comparative test statistics to assess convergence. We offer δ and ε, which are similar to statistics proposed by Huelsenbeck et al. (2001), as

useful tools to estimate convergence either within a run (by comparing adjacent and non-adjacent fragments) or among a set of runs. The emergence of parallel processing as a cheap and accessible alternative to shared memory multiprocessor machines produces another advantage for multiple short chains: because Markov chains are obligately serial, long runs cannot currently be spread out across large numbers of processors. Replicated short runs are inherently parallel, and make better use of cluster computing resources.

Although a small number of Markov chains appears to be sufficient for runs involving small data sets, large data sets may benefit from the increased parallelism of more chains, or more interaction between chains (Liang and Wong, 2001; Liu et al., 2000). If our hypotheses about the effects of temperature are correct, then an ideal combination of temperature and number of chains may not exist for all data sets but would instead depend on the topography of the likelihood surface (Altekar et al., 2004). Little is known about how to choose the number of Metropolis-coupled chains and their temperatures, although a general observation is that the number of chains needs to increase as the dimension or size of the search space if rapid mixing is to be maintained (Geyer, 1991; Marinari, 1998; Madras 2003; Zheng, 2003). Geyer and Thompson (1995) suggest setting the number of chains and heating parameter to achieve an acceptance rate of 20% to 40% in the context of the closely related simulated tempering algorithm. Our results indicate that using heated chains may not be beneficial in some cases. It may be more efficient to run nonheated, noninteracting chains in parallel, so that subsamples can be drawn from all chains instead of only the cold chain. Finally, future studies should compare the sampling efficiency of the strategy implemented in MrBayes with those of other algorithms such as the generalized Gibbs sampler (Keith et al., 2005), and examine the convergence properties under different proposal mechanisms, with particular emphasis on different types of moves through tree space (in addition to TBR and NNI). One move type that has been found to improve efficiency when multiple chains are implemented in parallel is to allow recombination between elements in different chains. This technique is known as evolutionary Monte Carlo (Liang and Wong, 2000, 2001). In the current context, two trees that share a common bipartition could be recombined to produce two new trees by exchanging the subtrees attached to the edges separating the two parts of the bipartition.

ACKNOWLEDGMENTS

We thank Peter Gogarten and Olga Zhaxybayeva for ideas about bipartition instability, and two anonymous referees for their comments on earlier versions of the manuscript. This work was supported by Australian Research Council Grants CE0348221, DP0342987, and DP0452412.

REFERENCES

Adachi, J., and M. Hasegawa. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42:459–468.

Al-Awadhi, F., M. Hurn, and C. Jennison. 2004. Improving the acceptance rate of reversible jump MCMC proposals. Stat. Prob. Lett. 69:189–198.

Allen, B. L., and M. Steel. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. Ann. Comb. 5:1–15.

Altekar, G., S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20:407–415.

Beiko, R. G., T. J. Harlow, and M. A. Ragan. 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. USA 102:14332–14337.

Brooks, S. P., and A. Gelman. 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. 7:434–455.

Brooks, S. P., and P. Giudici. 2000. MCMC convergence assessment via two-way ANOVA. J. Comput. Graph. Stat. 9:266–285.

Brooks, S. P., P. Giudici, and A. Philippe. 2003a. Nonparametric convergence assessment for MCMC model selection. J. Comput. Graph. Stat. 12:1–22.

Brooks, S. P., P. Giudici, and G. O. Roberts. 2003b. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. J. R. Stat. Soc. B Stat. Methodol. 65:3–39.

Brooks, S. P., and G. O. Roberts. 1998. Convergence assessment techniques for Markov chain Monte Carlo. Stat. Comput. 8:319–335.

Brown, J. R. 2003. Ancient horizontal gene transfer. Nat. Rev. Genet. 4:121–132.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Chauveau, D., and P. Vanderkhove. 2002. Improving Convergence of the Hastings-Metropolis algorithm with an adaptive proposal. Scand. J. Stat. 29:13–29.

Chor, B., M. D. Hendy, B. R. Holland, and D. Penny. 2000. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. Mol. Biol. Evol. 17:1529–1541.

Cowles, M. K., and B. P. Carlin. 1996. Markov chain Monte Carlo convergence diagnostics: A comparative review. J. Am. Stat. Assoc. 91:883–904.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248–254.

Erixon, P., B. Svennblad, T. Britton, and B. Oxelman. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst. Biol. 52:665–673.

Gelman, A. 1996. Inference and monitoring convergence. Pages 131–143 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). London: Chapman & Hall/CRC.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis. London: Chapman & Hall/CRC.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–472.

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. Fairfax Station, VA: Interface Foundation.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo. Stat. Sci. 7:473–483.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Introducing Markov chain Monte Carlo. Pages 1–19 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). London: Chapman & Hall/CRC.

Gilks, W. R., and G. O. Roberts. 1996. Strategies for improving MCMC. Pages 89–114 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). London: Chapman & Hall/CRC.

Gogarten, J. P., W. F. Doolittle, and J. G. Lawrence. 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19:2226–2238.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.

Harlow, T. J., J. P. Gogarten, and M. A. Ragan. 2004. A hybrid clustering approach to recognition of protein families in 114 microbial genomes. BMC Bioinformatics 5:45.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Hillis, D. M., T. A. Heath, and K. St. John. 2005. Analysis and visualization of tree space. Syst. Biol. 54:471–482.
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 phylogenetic trees. Bioinformatics 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.
Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.
Keith, J. M., P. Adams, M. A. Ragan, and D. Bryant. 2005. Sampling phylogenetic tree space with the generalized Gibbs sampler. Mol. Phylogenet. Evol. 34:459–468.
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.
Li, S. Y., D. K. Pearl, and H. Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. J. Am. Stat. Assoc. 95:493–508.
Liang, F. M., and W. H. Wong. 2000. Evolutionary Monte Carlo: Applications to c_p model sampling and change point problem. Stat. Sinica 10:317–342.
Liang, F. M., and W. H. Wong. 2001. Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. J. Am. Stat. Assoc. 96:653–666.
Liu, J. S., F. M. Liang, and W. H. Wong. 2000. The multiple-try method and local optimization in metropolis sampling. J. Am. Stat. Assoc. 95:121–134.
Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. 40:315–328.
Madras, N., and Z. Zheng. 2003. On the swapping algorithm. Random Structures Algorithms 22:66–97.
Mar, J. C., T. J. Harlow, and M. A. Ragan. 2005. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. BMC Evol. Biol. 5:8.
Marinari, E. 1998. Optimized Monte Carlo methods. Pages 50–81 *in* Advances in computer simulation (J. Kertész, and I. Kondor, eds.). Springer, New York.
Mau, B., M. A. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1–12.
Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1091.
Mossel, E., and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309:2207–2209.
Muller, T., and M. Vingron. 2000. Modeling amino acid replacement. J. Comput. Biol. 7:761–776.
Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.
Ragan, M. A., C. A. Murphy, and T. G. Rand. 2003. Are Ichthyosporea animals or fungi? Bayesian phylogenetic analysis of elongation factor 1 alpha of Ichthyophonus irregularis. Mol. Phylogenet. Evol. 29:550–562.
Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Syst. Biol. 51:754–760.
Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.
Raymond, J., O. Zhaxybayeva, J. P. Gogarten, S. Y. Gerdes, and R. E. Blankenship. 2002. Whole-genome analysis of photosynthetic prokaryotes. Science 298:1616–1620.
Roberts, G. O. 1996. Markov chain concepts related to sampling algorithms. Pages 45–57 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). London: Chapman & Hall/CRC.
Roberts, G. O., and J. S. Rosenthal. 2004. General state space Markov chains and MCMC algorithms. Prob. Surv. 1:20–71.
Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Rotondi, R. 2002. On the influence of the proposal distribution on a reversible jump MCMC algorithm applied to the detection of multiple change-points. Comput. Statist. Data Anal. 40:633–653.
Salter, L. A., and D. K. Pearl. 2001. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst. Biol. 50:7–17.
Steel, M. 1994. The maximum likelihood point for a phylogenetic tree is not unique. Syst. Biol. 43:560–564.
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.
Tierney, L. 1994. Markov chains for exploring posterior distributions. Ann. Stat. 22:1701–1728.
Tierney, L. 1996. Introduction to general state-space Markov chain theory. Pages 59–74 *in* Markov chain Monte Carlo in practice (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). London: Chapman & Hall/CRC.
Tjelmeland, H., and B. K. Hegstad. 2001. Mode jumping proposals in MCMC. Scand. J. Stat. 28:205–223.
Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.
Xiong, J., and C. E. Bauer. 2002. A cytochrome b origin of photosynthetic reaction centers: An evolutionary link between respiration and photosynthesis. J. Mol. Biol. 322:1025–1037.
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., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.
Zheng, Z. 2003. On swapping and simulated tempering algorithms. Stochastic Process Appl. 104:131–154.

First submitted 25 April 2005; reviews returned 10 August 2005; final acceptance 19 April 2006
Associate Editor: Jack Sullivan