7-1-2018

The behavior of metropolis-coupled Markov chains when sampling rugged phylogenetic distributions

Jeremy M. Brown
Louisiana State University

Robert C. Thomson
University of Hawai'i at Mānoa

Recommended Citation

Brown, J., & Thomson, R. (2018). The behavior of metropolis-coupled Markov chains when sampling rugged phylogenetic distributions. *Systematic Biology, 67* (4), 729-734. https://doi.org/10.1093/sysbio/syy008

Syst. Biol. 67(4):729–734, 2018

DOI:10.1093/sysbio/syy008
Advance Access publication February 15, 2018

The Behavior of Metropolis-Coupled Markov Chains When Sampling Rugged Phylogenetic Distributions

JEREMY M. BROWN[1,*] AND ROBERT C. THOMSON[2]

[1]*Department of Biological Sciences and Museum of Natural Science, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA;*
[2]*Department of Biology, University of Hawai'i 2538 McCarthy Mall, Edmondson Hall Room 216, Honolulu, HI 96822, USA*
**Correspondence to be sent to: Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA; E-mail: jembrown@lsu.edu.*

Received 3 November 2017; reviews returned 5 February 2018; accepted 6 February 2018
Associate Editor: Mark Holder

Abstract.—Bayesian phylogenetic inference relies on the use of Markov chain Monte Carlo (MCMC) to provide numerical approximations of high-dimensional integrals and estimate posterior probabilities. However, MCMC performs poorly when posteriors are very rugged (i.e., regions of high posterior density are separated by regions of low posterior density). One technique that has become popular for improving numerical estimates from MCMC when distributions are rugged is Metropolis coupling (MC^3). In MC^3, additional chains are employed to sample flattened transformations of the posterior and improve mixing. Here, we highlight several underappreciated behaviors of MC3. Notably, estimated posterior probabilities may be incorrect but appear to converge, when individual chains do not mix well, despite different chains sampling trees from all relevant areas in tree space. Counterintuitively, such behavior can be more difficult to diagnose with increased numbers of chains. We illustrate these surprising behaviors of MC^3 using a simple, non-phylogenetic example and phylogenetic examples involving both constrained and unconstrained analyses. To detect and mitigate the effects of these behaviors, we recommend increasing the number of independent analyses and varying the temperature of the hottest chain in current versions of Bayesian phylogenetic software. Convergence diagnostics based on the behavior of the hottest chain may also help detect these behaviors and could form a useful addition to future software releases. [Metropolis coupling; Markov chain Monte Carlo; Bayesian phylogenetic inference.]

Bayesian phylogenetic inference involves sampling from posterior distributions of trees, which sometimes exhibit local optima, or peaks, separated by regions of low posterior density. Markov chain Monte Carlo (MCMC) algorithms are the most widely used numerical method for generating samples from these posterior distributions, but they are susceptible to entrapment on individual optima in rugged distributions when they are unable to easily cross through or jump across regions of low posterior density. Ruggedness of posterior distributions can result from a variety of factors, including unmodeled variation in evolutionary processes and unrecognized variation in the true topology across sites or genes. Ruggedness can also become exaggerated when constraints are placed on topologies that require the presence or absence of particular bipartitions (often referred to as positive or negative constraints, respectively). These types of constraints are frequently employed when conducting tests of topological hypotheses (Bergsten et al. 2013; Brown and Thomson 2017). Negative constraints can lead to particularly rugged distributions when the data strongly support a forbidden clade, because monophyly of the clade can be disrupted by inserting outgroup taxa in many different ways. However, topological moves between the alternative disruptions are very difficult, because they require swaps between the inserted outgroup taxa while the data constrain taxa from the forbidden clade to remain close together on the tree. While this precise form of ruggedness is particular to negative constraints, trees with high posterior density can be separated by similarly complicated topological rearrangements, even in the absence of constraints.

Metropolis coupling, also called parallel tempering, is one strategy for avoiding entrapment in local optima (Geyer 1991; Altekar et al. 2004; Yang 2014, p. 245–247) and has been implemented by default for many years in the popular Bayesian phylogenetics software package, MrBayes (Ronquist and Huelsenbeck 2003; Ronquist et al. 2012). In Metropolis coupling (often denoted MC^3 for Metropolis-coupled MCMC), m chains are run in parallel with each assigned its own stationary distribution. The first, or cold, chain samples directly from the targeted posterior distribution, while each of the other heated chains samples from a flattened transformation of the posterior. By flattening the posterior, heated chains are able to traverse between peaks more easily. However, because the heated chains are not targeting the posterior itself, samples are recorded only from the cold chain. Periodically, the chains may swap positions, which allows the heated chains to act as scouts for the cold chain.

To achieve the flattening effect, heated chains sample points in proportion to the posterior density raised to an exponent <1. Specifically, each chain's distribution

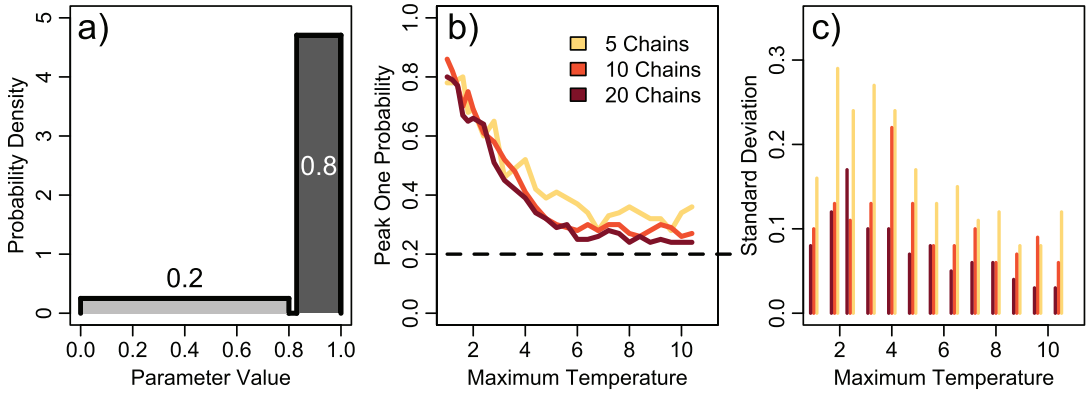

FIGURE 1. a) A probability distribution used to illustrate the behavior of Metropolis-coupled Markov chain Monte Carlo (MC^3) when sampling highly structured distributions. Peak One, depicted in light gray, has uniformly low density. Peak Two, depicted in dark gray, has uniformly high density. A valley of very low probability density separates the two peaks. The total probability of Peak One is 0.2 and the total probability of Peak Two is ~0.8. b) Estimated probabilities of Peak One when sampling the probability distribution with MC^3. Lines of varying color and brightness correspond to different numbers of coupled chains. The dashed line indicates the true probability. c) Standard deviations of estimated Peak One probabilities. Twenty replicate analyses were run for each combination of chain number and temperature.

has density $p(\tau,\theta|D)^{1/T}$, where $p(\tau,\theta|D)$ is the posterior distribution of topologies, τ, and model parameters, θ, conditioned on the data, D. T is a chain-specific temperature and a commonly used incremental heating scheme sets $T=1+\lambda(j-1)$, where λ defines the heating increment and j gives a chain's index between 1, the cold chain, and m, the hottest chain. Proposals for swaps between the positions of chains i and j are accepted with probability

$$\min\left(1,\left[\frac{p(\tau_j,\theta_j|D)}{p(\tau_i,\theta_i|D)}\right]^{\frac{1}{T_i}-\frac{1}{T_j}}\right). \quad (1)$$

For a more thorough explanation of Metropolis coupling, we direct interested readers to Yang (2014, pp. 245–247).

Despite the enormous number of empirical phylogenetic studies that employ Metropolis coupling (at a minimum, all those that use MrBayes with its default MC^3 setting), little effort has been devoted to understanding the behavior of this technique when sampling from very rugged phylogenetic distributions, which are the cases in which MC^3 is generally viewed as being most useful. In addition, little theoretical or empirical research exists to demonstrate the efficacy of this strategy. While attempting to improve sampling and convergence for a set of phylogenetic analyses with rugged distributions (e. g., Brown and Thomson 2017), we encountered several behaviors of MC^3 that we believe are not widely appreciated. Notably, estimated posterior probabilities may be incorrect but appear to converge, when individual chains do not mix well, despite different chains sampling trees from all relevant areas in tree space. Counterintuitively, such behavior can be more difficult to diagnose with increased numbers of chains.

The goal of this article is to illustrate these behaviors using both a simple one-parameter example and two empirical phylogenetic examples (one constrained and one unconstrained), while explaining their causes and potential consequences. To detect and mitigate the effects of these behaviors, we recommend increasing the number of independent analyses and varying the temperature of the hottest chain in current versions of Bayesian phylogenetic software. Convergence diagnostics based on the behavior of the hottest chain may also help detect these behaviors and could form a useful addition to future software releases.

The first, and most fundamental, consideration is that different local peaks in the posterior distribution may not be sampled in proportion to their posterior probability, even if swaps between chains on these peaks are regularly accepted. To understand why this occurs, we can examine the acceptance probability for chain swaps starting at the extreme where $\lambda=0$ and our MC^3 analysis employs only two chains. In this situation, each temperature (T) equals one and both chains target the posterior distribution. If the two chains become entrapped on different peaks, any proposed swap between them will be accepted with probability 1 even when the posterior densities of the two chain's current positions are very different. Looking at the acceptance probability, if $T_i=T_j$, the ratio of the posterior densities will be raised to the 0th power, and the outcome will always be 1. As a consequence, both peaks would be sampled with equal frequency. If λ is greater than 0, this ratio will no longer be 1 and not all proposed swaps will be accepted, but they may still be accepted at frequencies that differ markedly from the difference in their posterior density. To demonstrate this phenomenon, we constructed a one-parameter probability distribution with two distinct peaks, and then attempted to estimate the probability of each peak using MC^3 with varying numbers of chains and temperatures (Fig. 1). When temperatures were low, both peaks were sampled regularly but at frequencies that differed substantially from their true probabilities (Fig. 1b). Similar effects occur in two

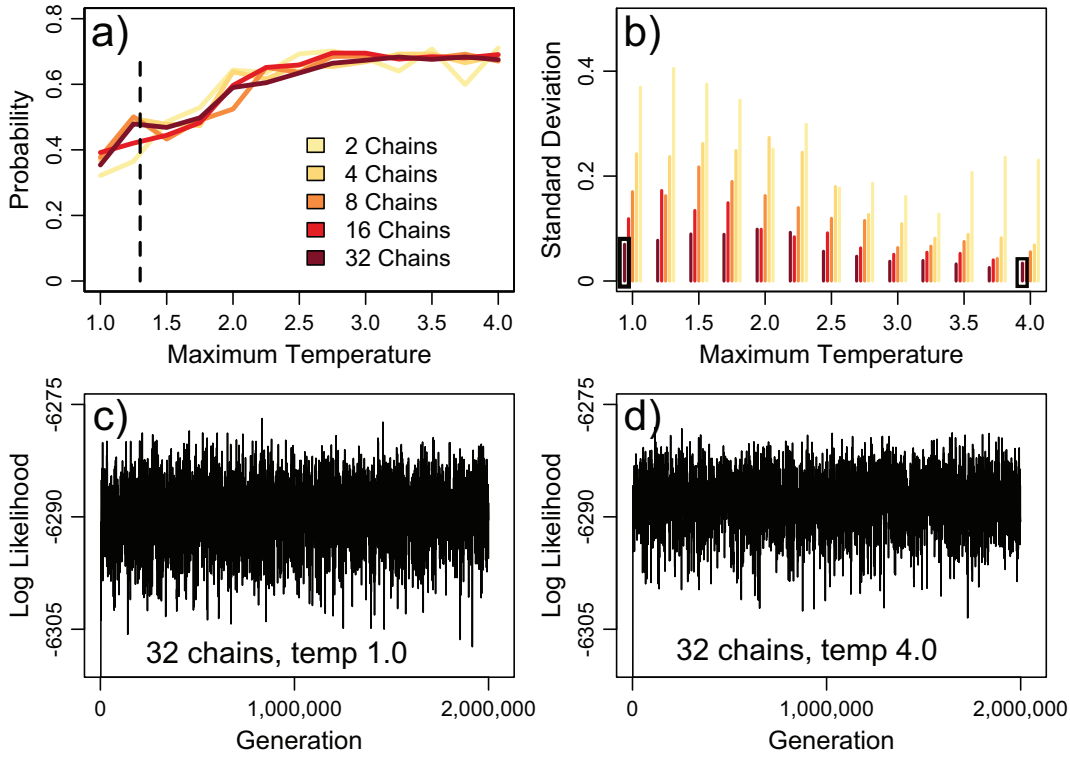

FIGURE 2. a) Estimated probabilities of monophyly for human (*Homo sapiens*) and either of two birds (*Gallus gallus* or *Taeniopygia guttata*) when analyzing the first empirical example data set (data from Crawford et al. 2012, see Methods section for details) with a negative constraint on bird monophyly. Topologies that broke up bird monophyly by grouping one of the two birds with human consistently had the highest estimated posterior probability. The vertical dashed line shows the default maximum temperature (i. e., temperature of the hottest chain) in MrBayes, which uses four chains by default. b) Standard deviations in estimated probabilities across 24 replicate analyses for each combination of chain number and temperature. Panels c) and d) show trace plots of likelihoods for two example analyses (indicated with black boxes in panel b). Both appear to mix well and sample similar likelihoods, despite substantial differences in estimated posterior probabilities for topological hypotheses.

exemplar phylogenetic analyses with known rugged distributions, but true probabilities are unknown in these cases (Figs. 2 and 3).

There are two important downstream consequences of this sampling phenomenon. First, when employing small numbers of chains, multiple peaks may be sampled at nearly identical frequencies across runs, even if those frequencies bear little resemblance to the true posterior probabilities. For instance, if we run two replicate analyses in the extreme case outlined above ($m = 2, \lambda = 0$), a lack of convergence might be immediately obvious if one analysis samples only Peak One and the other only Peak Two, or if one analysis samples only one peak and the other samples both. However, if both analyses end up with one chain on each peak, they both may appear to mix well in trace plots, and they will have nearly identical frequencies of samples across peaks (0.5, in this case). The probability of this spurious, but seemingly strong, evidence of convergence between analyses will depend on the number of runs, the number of chains, and the probability of any individual chain becoming entrapped on each peak. For the constrained empirical example that we use here, the probability of apparent convergence is roughly 25% for two independent runs that employ two chains when the temperature is low. This value was calculated based on the observed similarity in estimated probabilities across runs and should serve as strong motivation to run more than two independent analyses. Estimated probabilities and marginal distributions are so similar in these cases that convergence would seem to have occurred using any reasonable diagnostic that relies on comparing values between runs.

The second consequence is that increasing the number of chains can cause *all* independent runs to sample peaks at similar frequencies, despite these frequencies still differing strongly from their true probabilities. This phenomenon occurs because the proportion of samples collected from a peak is related to the proportion of chains entrapped on that peak. Essentially, increasing the number of chains gives a false sense of convergence. This phenomenon is clearly seen in both the one-parameter (Fig. 1) and phylogenetic (Figs. 2 and 3) examples, when examining the standard deviation in estimated probabilities across runs (Figs. 1c, 2b, and 3b). As the number of chains increases, the standard deviation drops for all temperatures, despite very different estimates at different temperatures. If the probability of a chain becoming entrapped on Peak One is P_1 and on Peak Two is P_2, the number of chains on each peak will follow a binomial distribution. As more chains are utilized, the proportion entrapped on each peak will approximate P_1 and P_2 with increasing precision. For our one-parameter

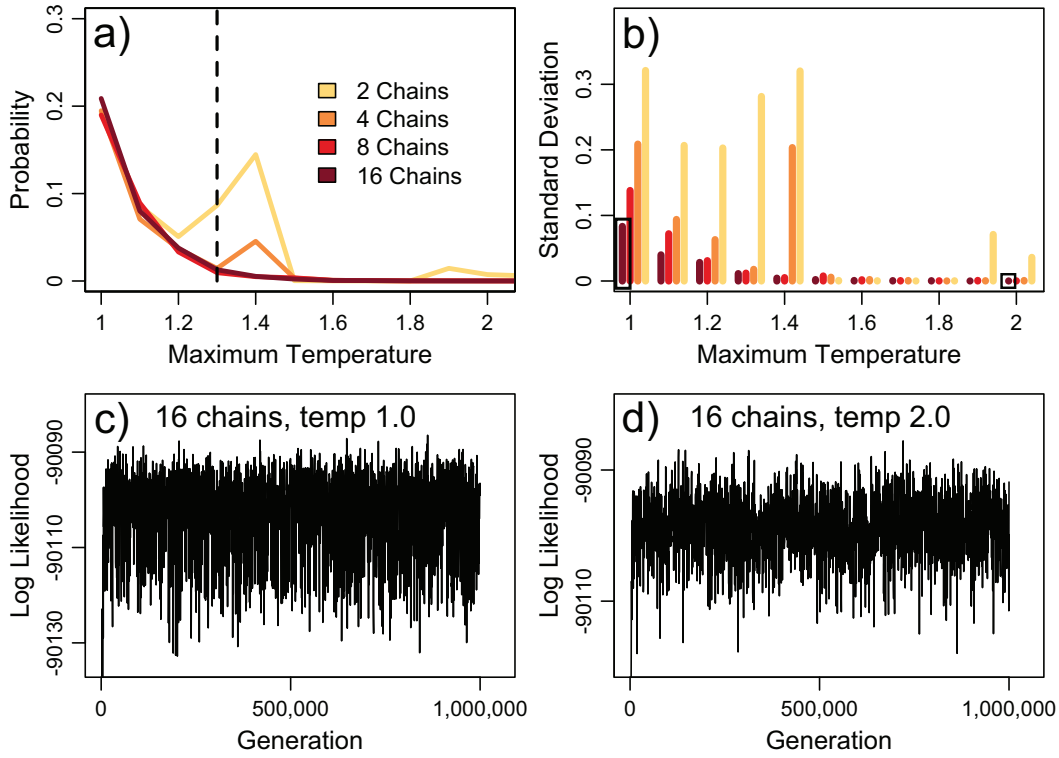

FIGURE 3. a) Estimated probabilities of monophyly for human (*Homo sapiens*) and mouse (*Mus musculus*) when analyzing the second empirical example data set (data from Green et al. 2014, see Methods section for details). The vertical dashed line again shows the default maximum temperature in MrBayes. b) Standard deviations in estimated probabilities across 24 replicate analyses for each combination of chain number and heating. Panels c) and d) show trace plots of likelihoods for two example analyses. As in Figure 2, both appear to mix well and sample similar likelihoods, despite substantial differences in estimated posterior probabilities.

example (Fig. 1), the probability of entrapment is much higher for Peak One given its width, even though it has a much lower total probability than Peak Two. Consequently, when the maximum temperature is too cool, the estimated probability of Peak One is much higher than it should be, but all the independent analyses swap well across peaks and return similar estimates. Whether these estimates are sufficiently similar to suggest convergence will depend on the number of runs, the number of chains, and the rigor of any particular convergence diagnostic. Because of this effect, we recommend that attempts to improve mixing for rugged distributions do not simply employ an increase in the number of chains while keeping the maximum temperature constant. Ideally, a series of analyses would be run with different maximum temperatures to look for sensitivity in the estimated probabilities and new diagnostics would be implemented to ensure that at least some of the hot chains are able to freely move between peaks. In addition, users can monitor the maximum standard deviation in bipartition frequencies across runs. If a sufficiently small threshold for this value is used as a convergence diagnostic, it may be possible to detect differences in the number of chains entrapped on different peaks as long as the required threshold is greater than the expected stochastic variation across runs.

Rugged phylogenetic distributions may become more common, as data sets now frequently include many genes with different underlying histories. Topological hypothesis tests, requiring the use of negative constraints during inference, may exacerbate this ruggedness because they create deep valleys (regions of zero posterior probability) in tree space that standard tree proposals are unable to easily cross. The collective effects outlined above are important to consider in these cases, particularly when considering different strategies for improving mixing and assessing convergence in troublesome analyses. Our results suggest two practical steps for those using current software to perform analyses employing MC^3 with rugged distributions. First, we recommend using more than two independent runs. In general, the probability of spurious, but seemingly precise, convergence drops off quickly as more independent runs are employed. Second, mixing across rugged landscapes is improved much more by adjusting the temperature of the hottest chain, rather than the number of chains. In fact, increasing the number of chains while keeping the maximum temperature constant can result in apparent improvements in convergence, driven not by improved mixing but instead by having a larger sample from a binomial (or multinomial) distribution.

Additions to Bayesian phylogenetic inference software may also help users identify and mitigate these problems. For example, convergence diagnostics based on the behavior of only the hottest chain could indicate if the maximum temperature is sufficiently high. If the hottest chain does not pass standard convergence diagnostics (including comparing the hottest chains from multiple independent runs), this would provide evidence that the temperature is too cool. Convergence diagnostics based on heated chains are not currently included in standard Bayesian phylogenetic inference software but would be conceptually straightforward to add. A second useful extension to current phylogenetic MCMC software could be the addition of new topology moves. The empirical examples that we explore here both involve cases where high probability trees are separated by two standard tree moves, whether nearest-neighbor interchange (NNI) or subtree-prune-regraft (SPR). However, a move that simply swapped the labels of two taxa in a tree should easily be able to traverse these valleys.

Metropolis coupling clearly improves convergence to the true posterior for many analyses, and we do not suggest that it be avoided. Rather, care and thought is warranted when setting up Metropolis-coupled analyses and interpreting their output. The same care that should be applied to all MCMC analyses.

Methods

One-Parameter Metropolis-Coupled Markov Chain Monte Carlo

To explore the behavior of MC^3, we defined a target distribution with two peaks (optima) that each had uniform density (Fig. 1a). Peak One was wider, [0.00,0.80], and shorter (probability density = 0.25) than Peak Two, which was relatively narrow, [0.83,1.00], and tall (probability density = 4.71). The two peaks were separated by a Valley, (0.80,0.83), of much lower density (probability density $=1\times10^{-6}$). The total probability of Peak One was 0.2 and that of Peak Two was ~0.8. More precisely, the total probability of the Valley and Peak Two summed to 0.8, but the Valley probability was very low (3×10^{-8}). Each chain employed a symmetric, uniform proposal distribution of width 0.01, so that the Valley could not be jumped with a single move. Twenty replicate analyses were run for 500,000 generations, sampling every 10th generation, and discarding the first 10% of samples as burn-in. In each generation, each chain performed a standard Metropolis–Hastings update with probability 0.5 or two chains attempted to swap positions with probability 0.5. Definitions of λ, T, and proposal ratios for Metropolis coupling follow Yang (2014). Python code to conduct these analyses is available from https://github.com/jembrown/toyMC3/.

Empirical Phylogenetic Data

Empirical phylogenetic data for the constrained example were taken from Crawford et al. (2012), who studied amniote phylogeny using ultraconserved elements (UCEs) by sampling 10 species broadly distributed across the major amniote groups. Brown and Thomson (2017) also analyzed these data and used Bayes factors to quantify the strength with which each UCE locus supported the monophyly of each major amniote group. Using their results, we selected the five UCE loci that most strongly supported the monophyly of birds and concatenated them. In order of support, these loci are chr9_6291 (399 sites), chr8_3325 (360 sites), chr4_7064 (334 sites), chr7_10489 (382 sites), and chr7_10865 (536 sites).

Empirical data for the unconstrained example were taken from Green et al. (2014), who studied amniote phylogeny and genomic evolution. One data set they used to infer phylogeny was comprised of UCEs sampled across 21 species comprising all major amniote groups. In the context of an ongoing project studying UCE evolution, we haphazardly selected 20 loci from their set of 633 UCE loci that comprise "taxon group 1" and assembled them into a concatenated alignment totaling 17,499 base pairs. The concatenated alignment and individual locus names are available in the Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.584m3.

These sets of loci were chosen because the behavior of Bayesian phylogenetic analyses using these data clearly illustrates the features of MC^3 that we wish to highlight. The same general principles apply to analyses of other data sets (e.g., the full concatenated data set of Crawford et al. 2012).

Bayesian Phylogenetic Analyses

Empirical Bayesian phylogenetic analyses were conducted in a modified version of MrBayes v3.2.5, with a small change to keep the software from turning off topology moves under some patterns of topological constraints (described further in Brown and Thomson 2017). The modified code is provided in the supplementary material for that article and on GitHub (https://github.com/jembrown/mrbayes_3.2.5_topoMoveFix). Empirical analyses for the constrained example assumed a general time-reversible (GTR)+I+Γ model of sequence evolution, while those for the unconstrained example assumed a GTR model with no rate variation. For both sets of analyses, in addition to modifying the number of Metropolis-coupled chains and the heating parameter ("Temp"), we also adjusted the frequency of proposed chain swaps ("Swapfreq") to every 4th generation. We adjusted the swap frequency, because we noticed that if swaps are accepted every time they are proposed (for instance, when $T=1$) and exactly two chains are used, the frequency with which different peaks are written to file (every 500th generation) does not accurately reflect how often they are sampled by the cold chain. This problem occurs because writing to file, in effect, applies a secondary filter to the sampling procedure. If swapping between two peaks is very regular, this secondary filter can result

in large differences between the actual and reported sampling frequencies.

For the constrained example, amniote phylogeny was inferred using the UCE data described above, but a negative constraint was imposed on the monophyly of birds (*Gallus gallus* and *Taeniopygia guttata*). All analyses were run for 2,000,000 generations, with samples taken every 500 generations, for a total of 4001 samples. The first 1000 samples (25%) were discarded as burn-in. For the unconstrained example, analyses were run for 1,000,000 generations, with samples taken every 500 generations, for a total of 2001 samples. The first 200 samples (10%) were discarded as burn-in.

Funding

This work was supported by the National Science Foundation [DEB-1355071 to J. M. B. and DEB-1354506 to R. C. T.]. Portions of this research were conducted with high-performance computing resources provided by Louisiana State University (http://www.hpc.lsu.edu).

Acknowledgments

The authors thank Michael Landis, Karen Cranston, Thomas Buckley, Mark Holder, and two anonymous reviewers for helpful comments that improved this manuscript. Bret Larget suggested the use of a convergence diagnostic that focuses on the hottest chain.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.584m3.

References

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

Bergsten J., Nilsson A.N., Ronquist F. 2013. Bayesian tests of topology hypotheses with an example from diving beetles. Syst. Biol. 62:660–673.

Brown J.M., Thomson R.C. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.

Crawford N.G., Faircloth B.C., McCormack J.E., Brumfield R.T., Winker K., Glenn T.C. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8:783–786.

Geyer C.J. 1991. Markov chain Monte Carlo maximum likelihood. In: Keramidas E.M., editor. Computing science and statistics: Proceedings of 23rd Symposium Interface. Fairfax Station: Interface Foundation. p. 153–163.

Green R.E., Braun E.L., Armstrong J., Earl D., Nguyen N., Hickey G., Vandewege M.W., St. John J.A., Capella-Gutierrez S., Castoe T.A., Kern C., Fujita M.K., Opazo J.C., Jurka J., Kojima K.K., Caballero J., Hubley R.M., Smit A.F., Platt R.N., Lavoie C.A., Ramakodi M.P., Finger J.W., Suh A., Isberg S.R., Miles L., Chong A.Y., Jaratlerdsiri W., Gongora J., Moran C., Iriarte A., McCormack J., Burgess S.C., Edwards S.V., Lyons E., Williams C., Breen M., Howard J.T., Gresham C.R., Peterson D.G., Schmitz J., Pollock D.D., Haussler D., Triplett E.W., Zhang G., Irie N., Jarvis E.D., Brochu C.A., Schmidt C.J., McCarthy F.M., Faircloth B.C., Hoffmann F.G., Glenn T.C., Gabaldon T., Paten B., Ray D.A. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science 346:1254449.

Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Ronquist F., Teslenko M., Van Der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Yang Z. 2014. Molecular evolution: a statistical approach. Oxford (UK): Oxford University Press.