An Efficient Independence Sampler for Updating Branches in Bayesian Markov chain Monte Carlo Sampling of Phylogenetic Trees

ANDRE J. ABERER1•, ALEXANDROS STAMATAKIS1•2, AND FREDRIK RONQUIST3
1Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany; 2Institute for Theoretical Informatics, Karlsruhe Institute of Technology, D-76131 Karlsruhe, Germany; and 3Department of Biodiversity Informatics, Swedish Museum of Natural History, PO Box 50007, SE-104 05 Stockholm, Sweden

Revised 19 January 2015; reviews returned 13 March 2015; accepted 9 June 2015

Abstract.—Sampling tree space is the most challenging aspect of Bayesian phylogenetic inference. The sheer number of alternative topologies is problematic by itself. In addition, the complex dependency between branch lengths and topology increases the difficulty of moving efficiently among topologies. Current tree proposals are fast but sample new trees using primitive transformations or re-mappings of old branch lengths. This reduces acceptance rates and presumably slows down convergence and mixing. Here, we explore branch proposals that do not rely on old branch lengths but instead are based on approximations of the conditional posterior. Using a diverse set of empirical data sets, we show that most conditional branch posteriors can be accurately approximated via a Γ distribution. We empirically determine the relationship between the logarithmic conditional posterior density, its derivatives, and the characteristics of the branch posterior. We use these relationships to derive an independence sampler for proposing branches with an acceptance ratio of ~90% on most data sets. This proposal samples branches between 2× and 3× more efficiently than traditional proposals with respect to the effective sample size per unit of runtime. We also compare the performance of standard topology proposals with hybrid proposals that use the new independence sampler to update those branches that are most affected by the topological change. Our results show that hybrid proposals can sometimes noticeably decrease the number of generations necessary for topological convergence. Inconsistent performance gains indicate that branch updates are not the limiting factor in improving topological convergence for the currently employed set of proposals. However, our independence sampler might be essential for the construction of novel tree proposals that apply more radical topology changes.

Bayesian phylogenetic inference using Markov chain Monte Carlo (MCMC) sampling (for reference implementations, see Lartillot et al. 2009; Drummond et al. 2012; Ronquist et al. 2012) is becoming increasingly popular. One of the most attractive features of the MCMC framework is its ability to handle complex models. The recent introduction of graph-based models (Höhna et al. 2014) will further extend the flexibility of the Bayesian phylogenetic approach.

A tree topology with branch lengths is the core of any phylogenetic model. Branch lengths correspond to the expected number of substitutions per site that occurred between two nodes in the tree. The branch length parameters account for a substantial fraction of phylogenetic model complexity. For instance, the standard GTR + Γ model (Tavaré 1986; Yang 1994) that we use here has nine free parameters (excluding tree topology and branch lengths). However, for a tree with 100 tips the number of branch length parameters is 197 (approximately twice the number of tips). Therefore, MrBayes and ExaBayes (Aberer et al. 2014) spend up to 50% of total inference time in branch length proposals. Yet, for models with branch-specific (e.g., in P4, see Foster 2004) or site-specific parameters (e.g., in PhyloBayes, see Lartillot et al. 2009) the effort that is necessary to reliably integrate over these additional parameters may be higher than the branch length sampling effort.

Branch length parameters are tightly linked to the respective topology and are essential to improving topological convergence (i.e., correctly estimating posterior probabilities of clades). Thus, MCMC topology proposals are often combined with branch length updates to expedite convergence (e.g., the LOCAL proposal or certain proposals in MrBayes that employ modest branch length updates, see Larget and Simon 1999, Lakner et al. 2008). Such hybrid proposals often combine branch multipliers with classical topological updates, such as the nearest neighbor interchange (stNNI), the extending subtree pruning and regrafting (eSPR), the extending tree bisection and reconnection (eTBR) move, or parsimony-posterior-guided SPR moves (parsSPR / lnPostSPR). Apart from the node slider move that affects two branch lengths, stand-alone branch multipliers are used to integrate over branches in MrBayes (for details on any of the aforementioned proposals, see Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.d6p3m5). Multipliers are uninformed proposals that rescale the current branch with a random factor (drawn from a sliding window on the logarithmic scale). The window size is typically tuned after a specified number of generations, such that every fourth proposal is accepted (i.e., the target acceptance rate is 25%). This is done to increase the probability of obtaining updates with similarly high posterior probabilities, but also ensures
that the proposed values are not too similar to the current value.

One inherent problem with branch changes in the context of topological updates is how branches should be transferred from the current topology to the proposed, new topology. Strictly speaking each topology comes with a unique set of branch length parameters. Thus, a topological move induces a drastic update with respect to branches: all branch length parameters are removed and replaced by new ones. In practice, one uses various mappings to reuse old branches. Typically, we assume that two branches can be mapped to each other if they define the same split (i.e., taxon bipartition) in either tree. Obviously, a newly proposed tree does not share all splits with the current tree and thus not all branches can be mapped. Therefore, it is reasonable to assume that when we propose a new topology, it has a negative impact on time-to-convergence when we map branches from bipartitions that are not contained in the proposed topology any more. Applying a multiplier to these mapped branches further decreases the acceptance probability for hybrid proposals (Lakner et al. 2008) as the proposed lengths often have a negative impact on the likelihood of the proposed topology.

Ideally, if we know the conditional posterior distribution of a branch (given all other parameters), we can construct a highly efficient proposal that is always accepted, a so-called Gibbs sampler. While the direct derivation of a Gibbs proposal for branch lengths does not appear to be feasible, the data augmentation approach allows for conjugate Gibbs sampling (Lartillot 2006). The augmentation step consists in sampling a substitution history given the current model parameters (including branch lengths) for each site. The resulting simplification of the augmented likelihood then allows direct analytical integration via a Gibbs sampler. However, construction of topological moves becomes more challenging when data augmentation is used.

Here, we explore independence sampling Tierney (1994) to propose branches in an informed way. An independence sampler proposes a parameter update for a branch (given all other parameters) that is distributed according to a density kernel \( q \) and that is independent of the current parameter value \( x \). The efficiency of an independence proposal exclusively depends on the similarity of \( q \) to the posterior of interest. Ideally, if \( q \) is approximately identical to the conditional posterior, then the proposal density ratio \( \frac{q(x)}{p(x)} \) cancels the posterior ratio and as a consequence the proposal behaves like a Gibbs sampler. As the proposed parameter updates are entirely uncorrelated (similar to the Gibbs sampler but in stark contrast to multiplier proposals), it is not necessary to tune proposal parameters to artificially low acceptance rates (e.g., 25% in case of multipliers). Instead, the observed acceptance rates directly reflect the efficiency of the independence sampler.

To find a suitable approximation to use as kernel for an independence sampler, it is necessary to characterize the conditional posterior distributions of branches. To the best of our knowledge, this problem has not been studied yet. Therefore, we initially analyze the shape of conditional posterior distributions of branches using a range of different data sets (15 data sets in total). Subsequently, we examine the relationship between the curvature of the log-likelihood and the standard deviation of conditional branch posteriors, and show that they are strongly correlated. With this we then construct an efficient independence sampler for branch lengths based on a Newton–Raphson estimation of the mode and curvature of the conditional branch posterior.

Furthermore, to explore the potential of the independence sampler for updating branch lengths in topology moves, we first examine how branch length posteriors change for newly proposed topologies. Based upon these observations, we assess various schemes for mapping old branches to new branches when updating the topology. Finally, we evaluate the efficiency of hybrid topology proposals, which use the independence sampler to update the critical branches of moves by comparing them against standard proposals.

### Conditional Posteriors of Branches

In the following, we empirically characterize the shape of conditional posterior distributions (CPDs) of branches (one at a time) for 15 data sets (see Table 1 for details). In other words, we sample the posterior of a focal branch given fixed values for all remaining branch lengths and other model parameters. To design a suitable density kernel for the independence sampler, we assess how accurately candidate density functions can be fitted to the CPDs.

#### Empirical Characterization

To obtain a representative set of branch length CPDs, we ran a chain on each data set for between 50,000 and 100,000 generations until the chain was in a state typical for the stationary distribution of a Markov chain. We then started a second round of chains from the aforementioned initial state. Here, each chain only samples a single branch (in contrast to sampling from the joint posterior). For the second round of sampling, all model parameters except the sampled branch were fixed. We extracted a sample from the chain every 10 generations and continued sampling each branch until we attained a highly accurate sample with an effective sample size (ESS) of at least 10,000.

For comparatively short branches (see Fig. 1a, where the mean \( \mu = 0.00173 \)), we observe an exponential-like distribution, while branch CPDs with a comparably high mean as in Figure 1d (\( \mu = 1.06 \)) are bell-shaped. Distributions that lie between these two extremes usually show a positive/right skew. Associated log posterior density curves show a short left tail and a long right tail (see Supplementary Figure 2 on Dryad at http://dx.doi.org/10.5061/dryad.h3pm5). However, for
long branches, CPDs sampled from this curve can be an almost symmetrical bell-shaped distribution. In case of Figure 1a, the logarithmic posterior density is linear, which from a phylogenetic point of view indicates that the most probable event is that no substitution occurred at all (i.e., the CPDs are exponential), there is also weak signal regarding the phylogenetic relationships. Finally, we observed a linear relationship between the skewness and the coefficient of variation (defined as \( \text{cv}^2 = \frac{\text{skewness}^2}{\text{standard deviation}} \)). Across all data sets, the skewness is \( \text{cv}^2 \leq 0.27 \). In contrast, for bell-shaped CPDs with a large mean (see Fig. 1d), the \( \Gamma \) and log-normal distribution perfectly match the observed posterior, whereas the normal distribution only performs 3% worse than the \( \Gamma \) fit. As expected, the log-normal distribution achieves a much lower expected acceptance of 94% and a proposal based on a fitted normal distribution only achieves 59% acceptance. In contrast, for bell-shaped CPDs with a large mean (see Fig. 1d), the \( \Gamma \) and log-normal distribution perfectly match the observed posterior, whereas the normal distribution only performs 3% worse than the \( \Gamma \) and log-normal. Here, the Weibull distribution shows a poor fit and would only be accepted in 83% of all cases. Figure 1b displays

### Table 1. Data sets used for analyses

| id  | No. of taxa | No. of characters | Type | No. of unique patterns | Gaps [%] | fit(b) | fit(c) | asdsf [%] | nucl.div. | No. of trees | No. of trees [%] |
|-----|-------------|------------------|------|------------------------|---------|--------|--------|-----------|-----------|-------------|-----------------|
| dat-24 | 24 | 141090 DNA | 4600 | 0.26 | 3.43 | -0.496 | 0.574 | 0.144 | 1 | 3 |
| dat-27 | 27 | 1949 | DNA | 934 | 20.05 | 1.08 | -0.473 | 1.25 | 0.206 | 120 | 20,037 |
| dat-36 | 36 | 1812 DNA | 1020 | 0.02 | 1.99 | -0.490 | 0.108 | 0.231 | 7 | 1,889 |
| dat-41 | 41 | 1137 DNA | 768 | 10.79 | 1.51 | -0.490 | 0.732 | 0.213 | 481 | 28,766 |
| dat-43 | 43 | 1660 DNA | 954 | 11.03 | 1.49 | -0.488 | 0.447 | 0.197 | 45 | 8,112 |
| dat-50 | 50 | 1133 DNA | 489 | 9.33 | 1.56 | -0.460 | 0.414 | 0.109 | 62,435 | 16,0440 |
| dat-59 | 59 | 1824 DNA | 1037 | 0.00 | 2.72 | -0.491 | 0.190 | 0.213 | 2,076 | 50,074 |
| dat-64 | 64 | 1008 DNA | 406 | 21.21 | 1.69 | -0.473 | 0.347 | 0.280 | 89,750 | 187,755 |
| dat-71 | 71 | 1082 DNA | 445 | 35.98 | 1.59 | -0.462 | 0.355 | 0.125 | 100,005 | 198,010 |
| dat-125 | 125 | 2919 DNA | 19436 | 32.72 | 2.02 | -0.499 | 0.0534 | 0.357 | 1 | 9 |
| dat-140 | 140 | 1104 AA | 1041 | 0.60 | 1.25 | -0.494 | 0.547 | 0.456 | 245 | 20,396 |
| dat-150 | 150 | 1269 DNA | 1130 | 4.77 | 1.61 | -0.486 | 0.99 | 0.220 | 100,005 | 198,010 |
| dat-354 | 354 | 460 DNA | 348 | 14.71 | 0.95 | -0.380 | 0.809 | 0.0871 | 100,005 | 198,010 |
| dat-404 | 404 | 13158 DNA | 7429 | 78.92 | 15.49 | -0.486 | 2.50 | 0.299 | 87,204 | 172,664 |
| dat-500 | 500 | 1398 DNA | 1193 | 2.48 | 1.35 | -0.475 | 0.829 | 0.131 | 100,005 | 198,010 |

Notes: fit(b) and fit(c) refers to the fitted parameters of the regression model fitting the negative value of the second derivative of the log-posterior to the standard deviation, asdsf denotes the ASDSF in 10 (8 in case of dat-404) reference runs; nucl.div denotes the nucleotide diversity in the alignment and #trees[%] denotes the number of trees in the X-% credible for reference runs.
a CPD for which the Weibull distribution outperforms any competing distribution with about 97% estimated acceptance rate compared to—for instance—91% for a \( \Gamma \) distribution. Also, for the distribution depicted in Figure 1c, the Weibull distribution exhibits a better proposal density than the \( \Gamma \) distribution by a small margin of 3%. Note that the latter branch is an internal branch of the hard-to-resolve data set dat-500. Despite the accurate estimate of the true CPD for this branch (ESS > 10,000), none of the fitted distributions under examination represents a perfect match, and the region around the mode of the posterior distribution is not well-shaped. Such difficult-to-fit branch length posteriors appear to be very uncommon, however.

Figure 2 confirms our observations from Figure 1: the normal distribution fits posteriors with small \( \text{cv} \) values well, but becomes inefficient (with expected acceptance rates of less than 70%) for a \( \text{cv} \) that is close to 1 (i.e., exponential-like distributions). The same holds for the log-normal distribution, which however still better fits most CPDs with intermediate skewness/\( \text{cv} \) values. For the vast majority of CPDs, the \( \Gamma \) distribution has close to optimal (>95%) expected acceptance values. However, in the \( \text{cv} \) range between 0.5 and 1.0, there are outliers for which the \( \Gamma \) distribution merely achieves 85%–95% expected acceptance. In contrast to this, the Weibull distribution shows optimal performance for strongly skewed branch length CPDs. In particular, it has a higher expected acceptance for the aforementioned outlier group that is problematic for the \( \Gamma \) distribution. However, for bell-shaped distributions (i.e., a \( \text{cv} \) close to 1.0), the performance of the Weibull distribution consistently drops to 80%–85% expected acceptance. Across all CPDs under examination, the \( \Gamma \) distribution achieves an average of 97.4% expected acceptance. If we assume that branch length posteriors are in fact
distributed according to a Γ distribution, we can also explain the linear relationship between skewness and \( \text{CV} \), since the Γ distribution has a skewness of \( \frac{1}{\sqrt{\lambda}} \) and \( \text{CV} \) of \( \sqrt{\lambda} \). Given the excellent fit of the Γ distribution to branch CPDs, we focus on this distribution in the following for the design of an independence branch sampler, yet we also consider the Weibull distribution for comparison.

**INDEPENDENCE BRANCH SAMPLER**

We can now derive an independence sampler for branches. Essentially, the proposal fits the previously discussed distributions to the conditional branch posterior using information obtained from Newton–Raphson (NR) optimization of the conditional posterior. We introduce a tuning mechanism to improve its performance (i.e., to achieve higher acceptance rates) and compare empirical acceptance rates of the tuned variant to untuned configurations of the sampler. Subsequently, we determine the sampling efficiency of the best configuration of the independence sampler and compare it to the performance of traditional branch proposals.

**Posterior Optimization**

The fitting exercise (see Fig. 2) suggests that the Γ or the Weibull distribution would be best used as proposal density kernels in an independence branch sampler, if it were possible to fit their densities easily and quickly to the conditional branch posterior. We accomplish this via NR-based optimization of the unnormalized conditional branch posterior (i.e., the product of the likelihood and the prior). We first employ the NR algorithm to obtain the mode of the conditional branch posterior, which can then be used as the mode of the fitted Γ distribution. As a by-product of the NR algorithm, when it converges to the branch length \( v_{\text{opt}} \), we obtain the value of the second derivative of the log-posterior density \( \ln P'(v_{\text{subopt}}) \) at \( v_{\text{subopt}} \), which is the branch length of the penultimate iteration. Typically, when the NR algorithm has been run for enough iterations, \( \ln P'(v_{\text{subopt}}) \) and \( \ln P'(v_{\text{opt}}) \) are almost identical. Obviously, there exists a relationship between the second derivative of the logarithmic branch posterior and the variance of the branch CPD. Specifically, for increasing absolute values of the second derivative, the variance of the CPD decreases and vice versa. Thus, we can harness the
second derivative of the logarithmic branch posterior for fitting a $\Gamma$ distribution to the branch CPD.

We determined the $\ln P^\nu (v_{\text{opt}})$ for all branch CPDs sampled previously. Indeed, there exists a power–law relationship between the standard deviation of a branch length CPD and the negative value of $\ln P^\nu (v_{\text{opt}})$ (see Fig. 3). For 11.1% of all branches, the NR method did not converge to the optimum. This exclusively affects CPDs with a $c$ close to 1 (i.e., exponential distributions) and a mode at or close to 0. The phylogenetic likelihood library (TILI) (Flouri et al. 2014) that is used for likelihood evaluations in ExaBayes does not allow for evaluation of values below a certain threshold for reasons of numerical stability. The threshold is relative to the mean substitution rate (as determined by the GTR parameters) and on most data sets allows for minimum branch lengths between $10^{-5}$ and $10^{-7}$. As a consequence, for exponential CPDs the NR method fails by design. However, in these cases we can assume that the CPD can be approximated by an exponential distribution and that the first derivative of the log-posterior can be used to estimate $\lambda$.

In order to determine the relationship between variance of the CPD and $\ln P^\nu (v_{\text{opt}})$, we fitted a power function $f(v) = b \cdot v^c$ with $\ln P^\nu (v_{\text{opt}})$ as response variable and the standard deviation of the CPD as explanatory variable to all branches on which the NR method was successful. We solved this non-linear regression model for parameters $b$ and $c$ using numerical optimization. Generalizing the regression to $f(v) = b \cdot v^c + a$ did not improve the model fit. Table 1 displays the fitted parameters for each data set as well as a consensus fit ($b = 1.61, c = -0.473$) across branches from all data sets. For the consensus fit, we randomly chose an equal number of branches from each data set in order to account for the fact that data sets with a large number of taxa have a stronger impact on the consensus than data sets with smaller numbers of taxa.

We observe that, estimates of parameter $c$ are highly similar. For all data sets, $c$ is in the range $[-0.463, -0.499]$, except for data set dat-354 (where $c = -0.38$). Data set 354 furthermore stands out as having a particularly low nucleotide diversity, which appears to be the cause for its high proportion of exponential-like branch length CPDs and thus the NR method does not converge for these distributions. However, in contrast to dat-354, for data sets with a similarly low nucleotide diversity (e.g., dat-50), parameter $c$ is within the otherwise observed range. Notice that, in a doubly logarithmic plot such as Figure 3, parameter $c$ represents the slope of the regression line. We assume that the $c$ estimate for dat-354 is not representative, since CPDs are more similar to each other than for any other data set.

Proposal Design

Given the optimal branch length $v_{\text{opt}}$ and the second derivative of the log posterior density $\ln P^\nu (v_{\text{opt}})$, we can now formulate equations to approximate the posterior via two-parameter distributions. For the $\Gamma$ distribution, we can determine $a$ and $\beta$ as follows

\[
\begin{align*}
\frac{c}{\beta} &= v_{\text{opt}} \\
\frac{a}{\beta^2} &= b \left( -\ln P^\nu (v_{\text{opt}}) \right)^{\frac{1}{c}}.
\end{align*}
\]

where $b$ and $c$ can either be fitted or we can use the empirical consensus parameters. Solving these equations to obtain the model parameters $a$ and $\beta$ is straightforward. This is not the case for the parameters of the Weibull distribution, where numerical methods for root finding are required (see Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.63pm5).

To summarize, for a branch length $v$, our NR-based independence proposal involves (i) the optimization of the underlying branch length posterior, (ii) determining the proposal density function $q$ (a Weibull or $\Gamma$ distribution), (iii) drawing a random number $\nu^*$ that is distributed according to $q$ and has a proposal density ratio of $\frac{q(\nu^*)}{q(v)}$. To ensure the reversibility of the proposal, NR is best initialized with a random value drawn from the branch length prior. In cases, where NR optimization fails, the branch length CPD is of exponential shape. This means that, the log-posterior density is linear with a slope of $\nu_{\text{opt}} = \ln P^\nu (v)$ for arbitrary branch values $v$. Since, as previously mentioned, the first derivative of the log-exponential function is $-\lambda$, we can use $q = \exp(-b \cdot \nu_{\text{opt}})$ as proposal kernel in these cases. While we can expect $d = 2$ to account for the increased expected value of the posterior mean in these cases (see Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.63pm5 for details).

Table 1 suggests that the exponent $c$ (represented as the slope in Fig. 3) generalizes well across various data sets. However, the parameter $b$ (the intercept in Fig. 3) is data set specific. Thus, we can deploy an autotuning scheme for the multiplicative factor $b$ to increase the acceptance ratio of the proposal. While for a multiplier proposal 25% is the desired acceptance rate, in case of the independence branch sampler, we intend to achieve an acceptance ratio of 100%. In contrast to the tuning of conventional proposals, we have to keep track of the current acceptance ratio and change the direction of tuning proposal parameter $b$ if we observe a lower acceptance ratio with respect to the previous tuning interval of typically 100 generations. Thus, we can tune parameter $b$ in case of NR convergence and parameter $d$ for the proposal kernel in case of non-convergence. This implies that we need to keep track of the observed acceptance ratios separately for updates based on converged and non-converged NR optimization.
Acceptance Ratio of the Independence Sampler

We implemented three flavours of NR-based $\Gamma$ proposals: (i) using individual (i.e., data set-specific) estimates for $b$ ($\Gamma_{\text{indi}}$), (ii) using the consensus over all data sets for $b$ ($\Gamma_{\text{cons}}$), and (iii) a scheme that autotunes parameter $b$ in case of convergence and parameter $d$ in case of non-convergence as described above ($\Gamma_{\text{tuned}}$).

Given the occasionally superior performance of the Weibull distribution, we also implemented a NR-based Weibull proposal ($W_{\text{tuned}}$) for comparison that employs the same tuning scheme $\Gamma_{\text{tuned}}$. For each data set we ran 10 chains for 300,000 generations using (i) one of the 4 proposals ($3 \Gamma, 1$ Weibull), (ii) a tree length multiplier, and (iii) proposals of state frequencies, substitution rates and the $\alpha$ value of a $\Gamma$ distribution for modeling rate heterogeneity across sites (fixed on a topology obtained from the reference runs, proposals weighted in a mixture of 20:1:1).

While for 10 data sets $\Gamma_{\text{cons}}$ consistently achieves an acceptance ratio $> 80\%$, the acceptance ratio is as low as $\approx 65\%$ for data sets like dat-354 (see Fig. 4). On many data sets, the acceptance ratio of $\Gamma_{\text{indi}}$ outperforms $\Gamma_{\text{cons}}$ by $10\%$, such that more than $90\%$ of proposals are accepted for 6 data sets. Naturally, since per-data set parameters are not known a priori, $\Gamma_{\text{indi}}$ is impractical for Bayesian inference; we use it to quantify performance loss by generalization compared to consensus parameters. Surprisingly, for dat-50 and dat-36, $\Gamma_{\text{indi}}$ performs worse than $\Gamma_{\text{cons}}$, although according to Figure 3 the consensus parameter values are close to the per-data set parameters. For dat-125,
acceptance ratios in chains vary strongly for $\Gamma_{\text{cons}}$. Additional experiments showed that, if we do not employ a tree length multiplier as well as GTR proposals, acceptance ratios further decrease and many chains (also for other data sets) explore sub-optimal regions of the posterior landscape. Since Bayesian inference on dat-125 is otherwise straightforward (consider the small number of trees in the 50% and 100% credible sets), we assume that tree length (which is also modified by GTR proposals in ExaBayes) impacts the performance of NR-based distribution proposals.

Finally, the tuned proposal $\Gamma_{\text{tuned}}$ consistently achieves high acceptance ratios: for 10 data sets, the average is above 90% and for none of the chains is the acceptance ratio is below 85%. While for some data sets such as dat-43 or dat-64, the tuned proposal does not outperform $\Gamma_{\text{indiv}}$, 20% more proposals are accepted compared with $\Gamma_{\text{cons}}$. As expected, the tuned Weibull proposal $W_{\text{tuned}}$ consistently performs worse than $\Gamma_{\text{tuned}}$ (e.g., dat-140 by more than 10%). Data sets with a high number of exponential posteriors (dat-71 and dat-354) underline the importance of tuning the parameter $d$ of the exponential proposal kernel: here, $W_{\text{tuned}}$ as well as $\Gamma_{\text{tuned}}$ achieve comparable acceptance ratios. We find that $\Gamma_{\text{tuned}}$ is the most efficient independence sampler of the examined variants and in the following, we only focus on $\Gamma_{\text{tuned}}$ when referring to the independence branch sampler.
Sampling Efficiency

By considering the high acceptance rates for the $\Gamma_{\text{tuned}}$ proposal, we can expect a 3–4 times higher acceptance rate compared with a branch multiplier. Since the proposal is an independence sampler, proposed branch length updates $\nu$ do not depend on the current branch length. In contrast to traditional proposals, we can expect low autocorrelation among generations in a chain that uses the independence sampler. In the following, we compare the sampling efficiency of $\Gamma_{\text{tuned}}$ with traditional proposals.

For each data set with $n$ taxa, we executed one chain and set the chain length and sampling frequency as a function of $n$ (total number of generations: 20,000 · $(2n-3)$, sample extraction every $n$ generations), because thinning obscures low acceptance ratios (i.e., high autocorrelation) among states in the chain. We used a tree sampled in the corresponding reference run for the data set, kept all parameters fixed (using equal stationary state frequencies, exchangeability rates, and data set, kept all parameters fixed (using equal stationary

Thus, with respect to runtime, $\Gamma_{\text{tuned}}$ does not require more than an additional 10–20% of runtime (see Fig. 5c). For all proposals, the factor dominating runtime appears to be the evaluation of conditional probability vectors (CPVs) for evaluating the likelihood at a randomly selected branch in the tree. The cost for NR in $\Gamma_{\text{tuned}}$ is moderate, specifically given its at least quadratic rate of convergence (see Gill et al. 1981). The ratio of the relative mean ESS (see Fig. 5b) and the relative runtime (see Fig. 5c) of $\Gamma_{\text{tuned}}$ allows us to assess the sampling efficiency per CPU time unit of $\Gamma_{\text{tuned}}$ relative to the multiplier proposal. We observe that in the same amount of time, $\Gamma_{\text{tuned}}$ produces samples with an ESS that is between 1.59× and 12.63× (median: 2.72×) higher than the ESS of the multiplier proposal samples. Thus, with respect to runtime, $\Gamma_{\text{tuned}}$ integrates over branches at least 2.41× more efficiently than the multiplier proposal in all except two data sets.

In general, Figure 5a corroborates our expectation that $\Gamma_{\text{tuned}}$ sometimes performs suboptimally for branch length CPDs with a $\nu$ between 0.5 and 1 (i.e., for posteriors on which NR does not converge). Since exponential CPDs on most data sets are rarer, parameter $d$ for approximating exponential distributions typically has been tuned less often than the parameter $c$ (for the $\Gamma$ distribution). For chains running longer than in our experiment, sampling efficiency may nonetheless improve with continued tuning of $d$. We observe a negative correlation between sampling efficiency of the two traditional proposals and $\Gamma_{\text{tuned}}$. That is, in contrast to $\Gamma_{\text{tuned}}$, the two traditional proposals perform particularly well on exponential-like CPDs and comparably bad on bell-shaped CPDs. This can be explained by considering the proposal density ratios of the multiplier and node slider that interact favorably with exponential CPDs. For instance, assume we propose new branch lengths $\nu^* = m \cdot \nu$ (where $\nu$ is the current branch length and $m$ is a multiplier) using a multiplier proposal on an exponential CPD. For $\nu^* > \nu$, the posterior ratio decreases (reducing acceptance probability). However, this means that $m > 1$ (as well as the proposal density ratio of this proposal) which increases the acceptance probability. The opposite is true for $\nu^* < \nu$. While this effect is beneficial for exponential CPDs of exponential shape (i.e., branches with a high $cE$ in Fig. 5a), it reduces the efficiency for sampling from the left tail of a bell-shaped CPDs (i.e., a high $cE$). For the node slider (with proposal density ratio $m^2$, where $m$ is the multiplier), the effect of the proposal density ratio on the sampling behavior is even more pronounced.

TOPOLOGICAL MOVES AND BRANCH LENGTHS

We now extend our analysis to the change of branch CPDs induced by topological moves. As established initially, there exist no inherently correct way of mapping branches from tree $t$ to a proposed tree $t^*$. Even if we assume that two branches are identical when corresponding to the same bipartition, a topological move will nevertheless result in at least one pair of orphan, “unmappable” branches. Consider a stNNI move (equivalent to an eSPR that stops the traversal after a single step, see Fig. 6a), where subtree $S$ is removed from adjacent nodes $A$ and $B$. Here, the bipartition induced by $v_{SB}$ is the only one that changes due to the move. Mapping the branch $v_{SB}$ to the branch inducing the new bipartition allows all other branches to stay associated with their initial bipartitions. For a bolder move, where a pruned subtree traverses three branches (thus named a SPR-3 move, Fig. 6b) we change three bipartitions. For a SPR-3, MrBayes uses a mapping where $v_{SB}$ becomes the branch length connecting the moving subtree $S$ to the last node that has been traversed by $S$ ($v_{DS}$ in Fig. 6b).

For motivating an alternative (less ad hoc) mapping, we consider the property of transitive closures. If branch lengths are ignored, for each of the three stochastic operators—stNNI, eSPR, and eTBR—a bolder proposal (with stNNI being less bold than eSPR and eSPR being less bold than eTBR) is the transitive closure of the less bold operators. For instance, each eTBR can be expressed as two eSPRs and an eSPR can be expressed as an stNNIs. With branch lengths this becomes more complicated. If we map branches as in MrBayes, we lose this transitive property. In other words, if branch lengths are considered, one eSPR cannot be decomposed into a series of stNNI moves. We can formulate an alternative NNI-transitive mapping for the eSPR proposal that
FIGURE 5. Comparison of sampling efficiency and runtime efficiency between the NR-based $\Gamma_{\text{tuned}}$ proposal, the branch length multiplier and the node slider. Panel a) depicts the effective sample size (ESS) versus the coefficient of variation for conditional posterior distributions of each branch length in a data set. Panel b) depicts the average ESS for data sets relative to the average ESS obtained by the multiplier proposal. Panel c) depicts the mean runtime of proposals relative to the mean runtime of the branch length proposal.
Impact of Topological Moves on Branch Lengths

Whether the MrBayes mapping or the transitive mapping is more efficient for a topological move depends on its impact on the branch length posteriors. We examined this empirically. First, we ran a chain using default settings for each data set for 20,000 generations to achieve a typical state of the Markov chain. Then, we enumerated all possible moves for a SPR-\( n \) move (where \( n \in \{1, 4\} \)). Furthermore, given the large number of such possible moves, we limited the number of analyzed moves to 200 randomly chosen SPR-\( n \) moves (in cases where more moves existed). For each distinct move, we sampled branch lengths from the joint branch length posterior before and after the move. In addition, because of the high computational cost, apart from the central bipartitions (class \( v_{SB} \)), branch length CPDs are decreased by three orders of magnitude, after applying the move (see Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.63pm5). If we only consider moves with a reasonably high acceptance probability (\( r > -100 \) log-likelihood units, see Fig. 7 and Supplementary Figure 10 on Dryad at http://dx.doi.org/10.5061/dryad.63pm5), we obtain less extreme ratios.

If we consider all moves (regardless of impact on likelihood), there is no correlation among inner branches (mapped with inner-default) and branches in the class close (\( r = -0.032 \)). For branches in the moves depicted in Figure 7, there is a weak correlation (\( r = -0.232 \)) and if we only focus on moves with a high acceptance probability of \( r > -10 \) log units, we obtain a moderate negative correlation of \( r = -0.403 \). In other words, moves that are accepted have a strong tendency to shorten inner branches and at the same time extend the immediately adjacent branches of involved subtrees (although the impact is weaker than for inner branches). This is because, after burnin and before applying a move, the topology is in a comparably parsimonious state (i.e., the topology pairs up taxa, such that few substitutions explain the observed alignment). Moves applied to such chain states are more likely to yield a less probable model for the alignment. Thus, when more substitutions have taken place between the common ancestor and the respective subtrees, branches adjacent to the inner branches are elongated, while inner branches are shortened.

Design of Hybrid Proposals

Our analysis of branch posterior changes induced by topological moves suggests that the acceptance ratio of topological moves may suffer from suboptimal branch lengths in the new topology. In MrBayes, we typically apply a weak branch length multiplier (multiplies by 0.95-0.99) during the burnin period and then use the default setting during the sampling period. However, for more complex trees, a stronger multiplier (multiplies by 0.9-0.8) may be necessary to ensure that the branch lengths converge to a reasonable shape. This is especially important for deep branches, which tend to be less well resolved and may be impacted more strongly by topological changes.

As expected, a topological change: (i) has a strong impact on the inner branches of the move (inner-default and inner-alt in Fig. 7); (ii) affects branches at the root of involved subtrees to a lesser degree (class close); and (iii) nearly has no effect on branches that are further away from the central bipartitions (class distant). Particularly, for bolder moves, we observe that if we apply the MrBayes mapping (inner-default), branch length CPDs change substantially less than if we apply the NNI-transitive mapping. While for the MrBayes mapping a clear trend toward shorter inner branch lengths is observable, the NNI-transitive mapping leads to branch length changes that are as pronounced as the worst case for inner-default (i.e., \( v_{SB} \), the branch that is removed and remapped) and that vary in both directions. Among branches at the root of involved subtrees, there is a general tendency to longer means after the move has been applied, specifically for the root of the moving subtree (5-0). In some cases, means of branch length CPDs are decreased by three orders of magnitude after applying the move (see Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.63pm5).

The left-hand version is typically applied in MrBayes/ExaBayes, the right-hand version represents an alternative, NNI-transitive version.
Figure 7: Ratios of the means of branch length CPDs before and after application of a SPR-n move (here, indicated as dist-n). Only moves that do not decrease the likelihood by more than -100 units are depicted. Values above the line indicate that the branch length mean is longer after the topological change. Branch names on the y-axis (following conventions explained in the main text and illustrated in Fig. 6) refer to the branch label prior to application of the topological move. For branches with vanishing bipartitions, inner-default displays the ratios according to the MrBayes-mapping and inner-alt displays the ratio according to the alternative NNI-transitive mapping. close and distant display branch categories by their depth in the respective subtrees (more distant branches not shown).

Figure 7 suggests that for the NNI, it is sufficient to either only propose a length for the remapped branch length $v_{SB}$ (see Fig. 6b). A branch length multiplier is based on the previous value and is likely to propose a less optimal value. In contrast to this, a central advantage of the independence sampler is that it can propose branch lengths de novo. That is, we can directly (without mapping) propose a branch length that most likely fits the proposed topology. Ideally, if we want to propose more than one branch at once for a topological update, we have to propose according to the joint optimum of all proposed branches. In other words, we have to iteratively optimize all branch lengths to be proposed until all have converged. We consider a branch as converged when it changes by less than 0.01 compared to the previous iteration (see Zwickl 2006, for more involved convergence criteria).
(try) or (as a bolder alternative) to also propose lengths of branches A-S, B-0, B-C, and S-0. For SPR-n moves (with \( n > 1 \)), we can construct increasingly bold moves by consecutively increasing the set of branches \( S \) to be proposed in the following order: (i) \( S-B \), (ii) last branch traversed by subtree \( S \) (C-D in SPR-3), (iii) \( S-0 \), (iv) A-S, (v) first non-traversed branch \( D-E \) in SPR-3, continuing with subtree branches (e.g., D-0) and inner branches (e.g., B-C) in the direction from the reattachment location toward the pruning location. However, since reversibility is an essential requirement for Bayesian proposals, we have to choose the branch proposal set \( S \) such that the original state can be restored (i.e., such that we obtain a non-zero proposal density ratio). For instance, if in a SPR-3 move we propose a new length for branch C-D, we have to propose B-C in the forward move as well to obtain a non-zero proposal density ratio.

Here, we examine four increasingly bold classes of hybrid proposals (topology plus branch length update) by adapting (i) only the remapped branch S-B (abbrev. switch), (ii) all inner branches (abbrev. inner, e.g., B-C, C-D, D-S in SPR-3), (iii) all inner as well as the branches adjacent to the path of branches traversed by \( S \) (A-S and D-E in SPR-3, abbrev. inner* ) and (iv) all branches in (i)–(iii) as well as the root branches of all traversed subtrees (abbrev. close, S-0, B-0, C-0 and D-0 in SPR-3). Since after determining the joint optimum of all branches \( v_i^* \in S \), each branch is proposed from a separate distribution \( v_i^* \), the proposal density ratio can be computed as the ratio of \( \prod_{i \in S} f_i(v_i^*) \) and \( \prod_{i \in S} f_i(v_i) \) (i.e., we also have to determine the joint optimum of the reverse move).

We applied all four classes to the eSPR and parsSPR moves. For the stNNI, only switch and close are relevant. Since the eTBR is equivalent to two SPR moves, we can construct analogous hybrid proposals by combining two hybrid SPR moves and including the bisected branch into the proposal set \( S \) in every case. We also extended the four update alternatives described above to design a posterior-guided SPR move (lnPostSPR). For instance, for a hybrid posterior-guided proposal of switch, we initially determine the posterior probability at each reattachment location in a user-defined radius around the pruning location after optimizing only the remapped branch length \( v_{SB} \). However, such posterior-guided hybrid proposals rapidly become computationally prohibitive.

**Evaluation**

In the following, we assess the convergence behavior of the aforementioned five topological proposals (stNNI, eSPR, eTBR, parsSPR, and lnPostSPR) either (i) in their plain form, (ii) as a hybrid proposal in combination with a branch length multiplier on all branches in \( S \) (only for modes switch and inner) or (iii) as hybrid proposals in combination with a NR-based \( \Gamma \) proposal on all branches in \( S \) (for all four modes). We ran eight independent chains starting from a random tree and employing default non-topological proposals and the proposal under examination. Furthermore, the independence branch sampler replaced the branch length multiplier and node slider. After 1,000,000 generations, we determined the average standard deviation of split frequencies (ASDSF) of sampled trees to a reference tree set created for the respective data set (see Table 1, see Supplementary Material on Dryad at http://dx.doi.org/10.5061/dryad.63pm5 for remaining details).

**Convergence to the posterior distribution**

While multiplier hybrid proposals perform equally well on most data sets, including the multiplier clearly slow convergence for data sets dat-43, dat-59 and dat-64. For data sets dat-125 and dat-27, multiplier hybrid proposals converge on average as fast as the respective plain proposals (except for the eSPR). Here, \( \Gamma \) proposals consistently achieve a higher level of convergence than the plain proposal. Similarly, \( \Gamma \) hybrid proposals overall achieve a better level of convergence on data sets dat-36 and dat-125. Both these data sets have particularly small 50% and 90% credible tree sets. We also observe advantageous convergence behavior of \( \Gamma \)-based proposals on dat-150 for the parsSPR and lnPostSPR (the three remaining classes of topological moves did not converge).

In contrast to dat-36 and dat-125, for data set dat-150 no tree was sampled twice in the reference runs. This suggests a flat posterior landscape that only allows the two guided proposals to reach the region of high posterior probability. While any kind of additional branch length proposal (also the multiplier in case of the parsSPR) improves convergence behavior, there still is a noticeable proportion of chains employing hybrid proposals that end up in local optima. We conclude that, while \( \Gamma \) hybrid proposals may decrease burnin time, they do not consistently traverse a rough topological posterior landscape more efficiently.

For data set dat-354, adding a \( \Gamma \) proposal to any set \( S \) clearly slows convergence, while adding a multiplier slows convergence to a lesser degree. It seems likely that the \( \Gamma \) proposal performs poorly in this case because there is a large number of branches for which the NR method does not converge (see Fig. 3). Nevertheless, the stand-alone independence branch sampler was highly efficient on this data set. For all chains, we observed an acceptance ratio ranging from 89% and 90% (identical to results from Fig. 4).

While multiplier hybrid proposals perform equally well on most data sets, including the multiplier clearly slow convergence for data sets dat-43, dat-59 and dat-64. For data sets dat-125 and dat-24, \( \Gamma \)-lnPostSPR hybrid proposals achieve particularly rapid and accurate convergence to the topology of the reference runs (in some cases with an ASDSF \( \approx 0.0625 \)). For both data sets, a huge amount of data (in terms of characters) is available for the comparably small number of taxa (see Table 1). Additionally, we previously observed predominantly narrow bell-shaped CPDs (i.e., with a
FIGURE 8. Performance of four classes of hybrid proposals employing a \( \Gamma \)-proposal (G-switch, G-inner, G-inner\(^*\), and G-close) compared to hybrid proposals employing a multiplier (M-inner and M-switch) contrasted by default proposals (None) for six data sets. Each box depicts the ASDSF of eight independent chains to the concatenated set of reference trees. The dashed line represents a lax convergence threshold, the solid line a strict threshold for convergence.

small \( cv \), see Fig. 5a). The performance on these two data sets suggests that analyses on phylogenomic data sets (where the conditional posterior is highly concentrated around the mode) benefit most from hybrid proposals employing the independence branch sampler.

CONCLUSION

Our results suggest that a \( \Gamma \) distribution accurately describes the CPDs of branch lengths (with some exceptions, where a Weibull distribution yields a more precise model). We determined that the skewness and \( cv \) (both closely entangled for a \( \Gamma \) distribution) are important characteristics that essentially divide branch length CPDs into two classes of distributions: bell-shaped distributions, which correspond to strongly supported clades (long branches), and exponential-like distributions, which correspond to weakly supported or unsupported clades (short branches). We notice a relationship between the number of trees in the credible set and the proportion of exponential-like distributions: the more trees in the credible set, the higher the proportion of exponential-like distributions.
While a huge number of branch CPDs were visually inspected for this study, we did not encounter a single, clearly bimodal distribution. For a small proportion of branch lengths, we observed that the mode was flat, but not beyond a point that could not be accurately matched by a $\Gamma$ distribution. We determined the $\Gamma$ distribution to describe branch CPDs most accurately, the data augmentation approach implemented in PhyloBayes ensures that the conditional posterior has the shape of a $\Gamma$ distribution. Accordingly, its proposals are drawn from a $\Gamma$ distribution.

We designed an independence sampler for branch lengths based on fitting a $\Gamma$ distribution to branch length CPDs using output from the NR branch optimization algorithm (optimum and first and second derivatives). We showed that the independence sampler is highly efficient in terms of acceptance ratios and allows us to propose branch lengths de novo without using previous branch length values. While slightly slower than a traditional branch proposal, the independence sampler yields substantially higher ESS values for sampled branch lengths. For almost all data sets the ESS-efficiency per unit of runtime for the NR-based $\Gamma$ proposal was between $2\times$ and $3\times$ higher compared to the multiplier proposal. Interestingly, the independence sampler often has a sub-optimal acceptance ratio of 50%–70% during burnin that often remains sub-optimal if a chain ends up in a local optimum. Our results suggest that the efficiency of commonly used branch length proposals (branch multiplier and node slider) depends on their respective proposal density ratios (being more efficient on exponential-like distributions). In contrast, the performance of the independence sampler degrades when NR does not converge, that is, when the branch length CPDs are exponential.

While we consider an independence sampler for branches particularly useful, analogous independence samplers could be designed for other continuous parameters. Our independence sampler relies on the fact that (i) proposal densities with two parameters are flexible enough to be fitted to the target distribution and (ii) that NR generates sufficient information (optimum and second derivative) for this fitting. The design of higher-dimensional independence samplers is challenging, since it is unlikely that only two parameters suffice for fitting a proposal density. The runtime efficiency of our independence proposal stems from the efficient posterior optimization of a single branch. For non-stationary models, the likelihood depends on the position of the root (e.g., the tree-heterogeneous model proposed in P4). Efficient realization of an independence sampler for branch-related parameters in such models does not appear feasible, since the optimization effort increases with the distance of a branch to the root.

We also examined how branch length CPDs change under topological moves. Our results show that the branch length mapping currently implemented in MrBayes is better than alternative schemes, as it minimizes the impact of the move on branch lengths. In particular, the initial branches that are traversed by a pruned subtree are impacted less by the MrBayes mapping (i.e., the mean of the distribution changes less drastically). These observations are also relevant for SPR-moves with subsequent branch length optimization in the maximum likelihood framework.

Based on our observations, we developed a strategy for proposing topology and branch length changes simultaneously with traditional topological moves. Our results show that the resulting hybrid proposals using the $\Gamma$-based independence sampler for branch length updates improve convergence behavior in some cases. However, they do not consistently and substantially outperform existing proposals. These results indicate that ill-fitting branch lengths are not a predominant limiting factor for the convergence efficiency of current topological proposals. Thus, we assume that the failure to achieve topological convergence on hard data sets is mainly associated with the fact that the transition between trees with high posterior probabilities requires more than one topological operation (e.g., two consecutive SPR moves). A new class of more aggressive moves that erase and reconstruct parts of the topology have the potential to perform well even on difficult data sets. However, for such updates to succeed, the simultaneous de novo proposal of branch lengths is essential. Thus, the independence sampler developed in this study marks an important first step toward more sophisticated topological proposal mechanisms.

In ExaBayes v1.4, we base the independence branch sampler as the major proposal for integrating over branches. Since we identified the branch length multiplier as particularly efficient for exponential CPDs, we employ it as a secondary proposal at a 9% smaller frequency compared with the independence sampler. Given the efficiency of the independence sampler, we decreased the proportion of generations used for sampling branches and instead propose changes to the topology more frequently. Thus, via increased sampling of the tree topology parameter, we can accelerate topological convergence.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.63pm5.

ACKNOWLEDGEMENTS

This research was funded by Heidelberg Institute for Theoretical Studies (A.J.A. and A.S.) and by the Swedish Research Council (grant 2011-5622 to FR). We would like to thank Nicolas Lartillot for numerous helpful suggestions and comments.

REFERENCES

Aberer A.J., Kötter K., Stamatakis A. 2014. Exabayes: massively parallel bayesian tree inference for the whole-genome era. Mole. Biol. Evol. 31:2553–2556.
