Bayesian Phylogenetic Inference

Sebastian Höhna
Bayesian Phylogenetic Inference

Sebastian Höhna

November 21, 2011

Abstract

In this thesis we consider two very different topics in Bayesian phylogenetic inference. The first paper, “Inferring speciation and extinction rates under different sampling schemes” by Sebastian Höhna, Tanja Stadler, Fredrik Ronquist and Tom Britton, focuses on estimating the rates of speciation and extinction of species when only a subsample of the present day species is available. The second paper “Burnin Estimation and Convergence Assessment” by Sebastian Höhna and Kristoffer Sahlin focuses on how to analyze the output of Markov chain Monte Carlo (MCMC) runs with respect to convergence to the stationary distribution and approximation of the posterior probability distribution.

The birth-death process is used to describe the evolution of species diversity. Previous work enabled the estimation of speciation and extinction rates under the assumption of a constant rate birth-death process and complete sampling of all extant species. We extend the complete sampled birth-death process to incomplete sampling with three different types of sampling schemes: random sampling, diversified sampling and clustered sampling. On a set of empirical phylogenies with known sampling fraction we observe that taking the sampling fraction into account gives better fitting models, either by random sampling or diversified sampling.

The current trend in Bayesian phylogenetic inference is to extend the available models by using more complex models and/or hierarchical models. This renders Bayesian inference by means of the MCMC algorithm very intricate. Performance of single or multiple MCMC runs need to be assessed. We investigate which methods are used in Bayesian phylogenetics to assess the performance of MCMC runs, which methods are available from other research areas and compile a strategy on how to assess convergence and how to estimate the burnin automatically in a statistically sound framework.

Keywords: Phylogenetics, Birth-Death Process, Bayesian inference, Markov chain Monte Carlo.

*Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden, hoehna@math.su.se.
Acknowledgements

I would like to thank my supervisor Tom Britton for teaching me efficiency and structure as a scientist. I would also like to thank my co-supervisor Fredrik Ronquist for giving me insights into his enormous knowledge and experience in evolutionary biology and phylogenetics. I’m very grateful to Fredrik Olsson, my officemate, for his invaluable help in statistical methods and the frequent discussion on more or less any subject. Finally I want to thank all the people who contributed scientifically to the content of this thesis. Tanja Stadler, for her contribution on the joint paper. Kristoffer Sahlin for his work on our joint manuscript which developed from his Master thesis. Per Alström for his collaboration on another project, providing the data for the second manuscript and pointing out the need for the work.
List of Papers

This thesis consists of two papers:


For the first paper, S. Höhna did the simulations, the inference and wrote most of the manuscript. S. Höhna, T. Stadler and T. Britton jointly developed the models. F. Ronquist contributed to the discussion of the biological data. All authors read and accepted the final manuscript. K. Sahlin initiated the project of the second manuscript with his Master thesis supervised by S. Höhna on burnin estimation. S. Höhna performed all simulations and analyses used in the manuscript and wrote the manuscript. K. Sahlin commented and agreed on the final manuscript.
1 Introduction

Phylogenetics, though a rapidly evolving subject, has a long history. Darwin introduced with his book “On the origin of species” the concept of evolution by descent and a common ancestor among all living species. His main idea focused on the concept of natural selection, that species adapt influenced by their environment by reproduction and selection of the fittest individuals, and speciation. A population of individuals from one species might separate into two or more and those new populations evolve independently and hence form new species.

In the early days the relationship between the species were estimated based on similarities of characteristics, so called traits or phenotypes, such as vertebrates, carnivores, fur, wings and teeth. The number of matching traits defines the similarity of species. In the evolutionary theory, species which separated recently will have very similar traits because they only had little time to evolve independently. Hence, species which had very similar traits where grouped together. Since the discovery of DNA (deoxyribonucleic acid) in 1953 the field of evolutionary biology has changed dramatically. The similarities of species is now not based only on the number of different traits but rather more on the genetic difference, which is the difference between genes of the species under study. Inferring evolutionary trees from gene data is called phylogenetics (Nei and Kumar, 2000). The divergence time between two species is estimated using stochastic models of gene evolution.

Many studies in phylogenetics are concerned with reconstructing the phylogenetic tree (Nei and Kumar, 2000; Huelsenbeck et al., 2001; Felsenstein et al., 2004). Other studies assume the phylogenetic tree to be given and aim to infer for instance the rates of diversification (Nee, 2006; Ricklefs, 2007). The rates of diversification (speciation and extinction rates) can be estimated from molecular phylogenies under various assumptions using a birth-death process (Nee et al., 1994). An ancestral species splits into two descendant species with rate $\lambda$ and goes extinct with rate $\mu$.

The simple models assume constant rates over time, equal rates for each species and complete sampling of the present day species. In many analyses this simple model does not fit the observed data very well. The knowledge about species diversity over time, speciation and extinction rates all obtained from the fossil record disagrees with estimates from molecular phylogenies (Quental and Marshall, 2009, 2010). Extensions which take environmental changes into account use rates which are time-dependent (Nee et al., 1994; Rabosky, 2006; Rabosky and Lovette, 2008; Morlon et al., 2010). Time-dependent rates can also approximate diversity-dependent models when the speciation rate depends on the number of species alive (Etienne et al., 2011).

In our first paper we argue that the observed phylogenies and estimated speciation
and extinction rates do not agree with paleontological studies because often incomplete phylogenies are considered (Cusimano and Renner, 2010). We extend the constant rate birth-death process for three types of incomplete sampling: diversified sampling where only representative of groups are sampled, random sampling where every species has the same sampling probability, and clustered sampling where only complete groups are sampled.

Bayesian inference has become very popular in phylogenetic studies especially due to its ability to include continuously more complex and/or hierarchical models (Ronquist and Deans, 2010). Bayesian inference in phylogenetics owes most of its popularity to its simplicity of interpretation despite being computational expensive. Studies in phylogenetics have to rely on the observed data and cannot repeat any experiments which motivates to condition on the data, as done in Bayesian inference. Furthermore, the primary interest are the parameters of a model and their uncertainty. Maximum likelihood estimates, compared with Bayesian posterior probability distribution, need the addition of bootstrapping to provide measures of uncertainty.

Commonly the MCMC (Metropolis et al., 1953; Hastings, 1970) algorithm is used to approximate the posterior probability distribution. The demands on the inference methods have increased rapidly. Since the MCMC algorithm is a stochastic approximation which only guarantees for infinitely many iterations and samples to represent the true posterior distribution some measurements, so called convergence diagnostics, are necessary to validate the approximations (Cowles and Carlin, 1996). The main requirements on a sufficient MCMC run can be summarized in three parts: (1) The chain has converged to the stationary distribution, (2) The samples are representative for the posterior distribution and (3) The approximation, e.g. for the posterior mean, has an acceptable low uncertainty. Additional to the question if the output generated by an MCMC algorithm is representative for the posterior distribution it is of great interest to estimate the burnin length $n_0$ which is the number if iterations needed until the chain has reached the stationary distribution.

In our second paper we investigate the currently used methods of convergence assessment in Bayesian phylogenetic inference. Then, we examine these methods on their statistical foundation and compare them to other methods in known in the statistical literature. We conclude the paper by providing a general framework for estimating the burnin and assessing convergence.
References


6


Inferring speciation and extinction rates under different species sampling schemes

Sebastian Höhna †, Tanja Stadler ‡, Fredrik Ronquist ‡ and Tom Britton§

November 16, 2011

Abstract

The birth-death process is widely used in phylogenetics to model speciation and extinction. Recent studies have shown that the inferred rates are sensitive to assumptions about the sampling probability of lineages. Here, we examine the effect of the method used to sample lineages. Whereas previous studies have assumed random sampling, we consider two extreme cases of biased sampling: “diversified sampling”, where tips are selected to maximize diversity, and “cluster sampling”, where sample diversity is minimized. Diversified sampling appears to be standard practice, e.g., in analyses of higher taxa, while cluster sampling may occur under special circumstances, e.g., in studies of geographically defined floras or faunas. Using both simulations and analyses of empirical data, we show that inferred rates may be heavily biased if the sampling strategy is not modeled correctly. In particular, when a diversified sample is treated as if it were a random or complete sample, the extinction rate is severely underestimated, often close to 0. Such dramatic errors may lead to serious consequences, e.g. if estimated rates are used in assessing the vulnerability of threatened species to extinction. Using Bayesian model testing across 18 empirical data sets, we show that diversified sampling is commonly a better fit to the data than complete, random or cluster sampling. Inappropriate modeling of the sampling method may at least partly explain anomalous results that have previously been attributed to variation over time in birth and death rates.

Keywords: Birth and death process, speciation, extinction, phylogenetics, species tree, sampling, inference.
Introduction

A number of important problems in the life sciences are related to the birth and death over time of genetic lineages. In macroevolutionary studies, for instance, examples include estimation of speciation and extinction rates, identification of adaptive radiations and mass extinctions, and studies of the shape of phylogenetic trees (Nee, 2006). Similar problems occur in population genetics (e.g., Wakeley (2008)) and in epidemiology (e.g., Tanaka et al. (2006)), among many other fields.

A popular stochastic model for many of these problems is the constant-rate birth-death process (BDP) (Kendall, 1948; Feller, 1968; Thompson, 1975). It is a branching process, with each lineage having a constant rate ($\mu$) of dying (extinction rate) and a constant rate ($\lambda$) of giving birth to an additional lineage (speciation rate). Trees resulting from the BDP include both extant (surviving to the current time) and extinct lineages. Molecular sequence data are only rarely available for extinct lineages, so phylogenetic trees inferred from molecular data are typically restricted to extant lineages. We call such trees extant trees. In the literature they have often been referred to as “reconstructed trees” because they result from phylogenetic inference (phylogeny “reconstruction”) (Harvey et al., 1994; Nee et al., 1994a). Fortunately, it turns out that the BDP model allows speciation and extinction rates to be estimated not only from complete trees but also from extant trees (Nee et al., 1994a).

The $\gamma$-statistic (Pybus and Harvey, 2000) is a measure of how well the extant tree fits the BDP, with values close to zero indicating support for a pure birth process (no extinction), while positive values support the BDP (with extinction), a model predicting more recent speciation times in the extant tree than expected from the pure birth process. A negative $\gamma$ value indicates that the speciation times are older than those from a pure birth process, a pattern that is not expected under the BDP. If a tree with a negative $\gamma$ value is nevertheless analyzed under the BDP, the estimated extinction rate will be close or equal to zero.

Despite abundant evidence of extinction in the fossil record, many empirical data sets (Nee, 2006; Purvis, 2008) have negative $\gamma$ values, resulting in estimated extinction rates close to zero. Rabosky and Lovette (2008) proposed an explanation for this observation, namely that speciation or extinction rates vary over time, violating the constant-rate assumption of the BDP. Specifically, if speciation rates decrease over time or extinction rates increase over time, we might expect negative $\gamma$ values. Unfortunately, Rabosky (2010) shows that non-constant speciation and extinction rates cannot be estimated without fossil data, making it difficult to test this idea in most organism groups. Furthermore, not all variation in speciation and extinction rates result in negative $\gamma$ values. For instance, Quental and Marshall (2010) simulated trees under a birth-death process
with decreasing net diversification rate (speciation rate minus extinction rate) but the resulting trees did not show significantly negative $\gamma$ values. Hence, shifts in diversification rates may not be the main reason for the apparently underestimated extinction rates in many empirical studies.

A completely different explanation for the negative $\gamma$ values in empirical data sets is given by Cusimano and Renner (2010). They suggest that the observed negative $\gamma$ may not be real but instead an artifact caused by biased taxon sampling. They show, by simulation, that if taxa are sampled such that only deep nodes in the complete tree are retained, which is equivalent to our diversified sampling, then negative $\gamma$-values are produced even under the constant-rate BDP (with positive death rate $\mu$). However, they do not explore the effects of biased taxon sampling in depth, nor do they derive likelihood functions allowing estimation of BDP parameters under biased taxon sampling. This is what we set out to do in the current paper.

Yang and Rannala (1997) and Stadler (2009) show how BDP parameters can be estimated when tips are sampled uniformly at random. However, in practice sampling is typically associated with some bias. In this paper, we explore the effects of such biased sampling. In particular, we focus on two extreme sampling methods with opposite types of bias. The first method, which we denote diversified sampling, strives to maximize diversity in the sample. A sample with maximum diversity is obtained by selecting the $n$ tips to be as distantly related as possible. In the second method, cluster sampling, the $n$ tips are selected to be as closely related as possible.

In phylogenetic studies, it is commonly an explicitly stated goal to maximize the representation of subtaxa. For instance, in an analysis of relationships within a family, biologists will often try to include exemplars of all tribes or all genera. This approach is close to our diversified sampling method. Cluster sampling is probably less frequent but may occur under special circumstances, for instance when the samples come from geographically restricted areas, such as islands. It could also occur when the sampling involves a bias linked to traits evolving on the tree. For instance, if sampling of a clade of microorganisms involves cultivation in a particular medium, the final sample might represent clusters of lineages that have independently evolved the ability to grow in this medium.

In this paper, we compare diversified and cluster sampling with random sampling. We derive the probability densities for extant trees under diversified sampling and cluster sampling. This allows maximum likelihood and Bayesian estimation of BDP parameters. Using simulations, we show that the parameters are estimated accurately if we choose the appropriate sampling scheme. However, using simulations and reanalyses of empirical data sets, we show that there may be dramatic biases in the estimated relative
death rate (the death rate divided by the birth rate) if the sampling procedure is not modeled correctly. In particular, if we treat a diversified sample as if it were a random or complete sample – a situation that appears to be common in the literature – the relative death rate will be severely underestimated, often equal to zero. Using Bayesian model choice across 18 empirical data sets, we also show that diversified sampling often fits the data significantly better than cluster, random or complete sampling. We end the paper by discussing the extent to which the failure to accommodate sampling bias might explain the negative $\gamma$ values and underestimated extinction rates associated with many empirical data sets.

**Methods**

We will begin this section with a general definition of the BDP, including parameters and properties, and briefly review known results, which will be used when deriving densities of trees under different sampling schemes.

**The BDP model**

The constant-rate birth-death process (BDP) starts at some time of origin $t_0$ with one species (we use species and lineage as synonyms in this paper). As we move from the time of origin toward the present, the number of species can increase by one or decrease by one due to the fact that a species splits into two species or that a species goes extinct. The probability that a species splits into two species (a speciation or birth event) during a short time period of length $dt$ equals $\lambda dt$ and the probability that a species goes extinct (a death event) during such a time period equals $\mu dt$. We assume that the process is supercritical ($\lambda > \mu$) because otherwise the whole group is doomed to eventually go extinct. Further, it is assumed that all species behave independently and according to the same rules. Hence, if at present there are $n$ species, the time until the next event occurs is exponentially distributed with rate $n(\lambda + \mu)$.

The tree resulting from the BDP, including both extinct and extant lineages, will be referred to as the *complete tree* (Fig. 1a). Deleting the extinct lineages from the complete tree yields the *extant tree* (Fig. 1b). For later use, let $m$ denote the number of species in the extant tree.

It is well known (Ricklefs, 2007; Purvis, 2008) that, for a fixed number $m$ of extant species, increasing the relative magnitude of the death rate (i.e. increasing $\mu/\lambda$) shifts the speciation points towards the leaves of the extant tree. On the other hand, if $\mu/\lambda$ is small, we expect the speciation points to be further away from the leaves of the extant tree. This is a fundamental property of the BDP, which we will return to later in the
Figure 1: A complete tree (left) and its extant counterpart (right) induced by the BDP together with the times $T = (t_1, \ldots, t_{m-1})$ of the reconstructed speciation events and the time of origin $t_0$. Note that only the $m-1$ speciation times from the reconstructed tree are marked since only data for the reconstructed speciation events are usually available.

paper; it also forms the basis for the $\gamma$-statistic (Pybus and Harvey, 2000). The relative death rate ($\mu/\lambda$) hence affects the location of the speciation points, a very relevant feature when considering different sampling schemes. The net diversification $\lambda - \mu$ on the other hand, mainly affects the number of speciation points (and not their location). Since we have the same number of sampled points $n$, irrespective of sampling scheme, the estimation of $\lambda - \mu$ is hardly affected by the different sampling schemes. For this reason inference is focused on estimation of the relative death rate $\mu/\lambda$ in what follows.

Often only a sample of the leaves in the extant tree is observed. Assume hence that $n$ out of the $m$ ($2 \leq n \leq m$) extant lineages are sampled, implying that $m-n$ leaves are removed from the extant tree, and let $\rho = n/m$ denote the sampling fraction. Further, let degree-two vertices in the resulting sampled tree be suppressed. Also, the edge before the first speciation event is deleted. The resulting tree is the sampled tree of the process. Note that the sampled tree is typically what we observe and on which we have to base our inference.

Throughout the paper, we set the present time to $t = 0$ and assume that the time increases going into the past. The time of the origin will be denoted by $t_0$. The time of the first split among sampled lineages, that is, the most recent common ancestor of all sampled lineages, will be denoted by $t_1$. The ordered set of the $n-1$ reconstructed speciation times is denoted by $T = (t_1, \ldots, t_{n-1})$ with $t_1 > \ldots > t_{n-1}$ (Fig. 1b).

To simplify the proofs, we will work throughout with distribution densities on oriented trees, that is, trees where we distinguish between the descendants of each bifurcation (Ford et al., 2009). An oriented sampled tree without leaf labels will be denoted by $T$. In many cases, it is more natural to consider densities on labeled trees, i.e., trees with unique leaf labels but no orientation. The two densities are related by a simple
conversion factor; specifically, by multiplying the density of an unlabeled oriented tree with $\frac{2^{n-1}}{m!}$, we obtain the density of a labeled tree (Gernhard, 2008).

**Parameters of the BDP**

The parameters of the BDP as we defined it in the previous section are $\lambda, \mu, t_0, \rho$. Given a tree, we want to estimate $\lambda$ and $\mu$. We will now explain how to deal with $\rho$ and $t_0$.

First, we consider $\rho$, which is related to the number of leaves in the tree. The number of taxa, $m$, is an observed quantity of the BDP. In practice, however, the size of the sampled tree, $n$, is often controlled by the investigator while the number of taxa, $m$, is typically unknown. Any estimates of $m$ are tied intimately to estimates of the sampling fraction, $\rho$. Therefore, it is natural to consider $\rho$ as a fixed parameter of a sampled BDP, and to condition the density of a sampled BDP on $n$.

Next we consider the time of origin of the process, $t_0$. It is quite often the case that we have no information about $t_0$, or that it is difficult to formulate our prior beliefs about $t_0$ in a Bayesian context. In this situation, it is common to assume a uniform prior on $(0, \infty)$ for the time of origin of the process. The unconditional probability of obtaining a finite tree is then 0, but by conditioning on obtaining $n$ sampled species, we obtain a proper probability density (Aldous and Popovic, 2005; Gernhard, 2008).

When we do have some information about the time of origin of the process, it is more likely to be associated with the time of the most recent common ancestor of the sampled tree, $t_1$, than with $t_0$. As an alternative to assuming a uniform prior on $t_0$, it is thus natural to condition the birth-death process on $t_1$. This is equivalent to considering two birth-death processes starting at time $t_1$, each producing some sampled extant species. When conditioning the BDP density on $t_1$, it is not strictly necessary to condition on $n$.

In the remainder of this paper, we will focus on the BDP densities conditioned on $n$, and we will assume a uniform prior on $t_0$. For convenience during the derivation, and for use in the Bayesian context, we will also provide the densities conditioned on $n$ and $t_1$. The densities conditioned on only $t_1$ are given in the Appendix for completeness.

**Properties of the birth-death process**

Under a birth-death process with complete sampling, i.e. observing the full extant tree ($n = m$), the probability that a lineage leaves exactly one descendant after time $t$ is denoted $p_1(t)$ and the probability that the lineage goes extinct before time $t$ is denoted
The probabilities are given in Kendall (1949),

\[ p_0(t) = \frac{\mu(1 - e^{-(\lambda-\mu)t})}{\lambda - \mu e^{-(\lambda-\mu)t}}, \]

\[ p_1(t) = \frac{(\lambda - \mu)^2 e^{-(\lambda-\mu)t}}{2(\lambda - \mu e^{-(\lambda-\mu)t})^2}. \]

Further, each permutation of the \( n - 1 \) speciation events is equally likely, and there is a one-to-one correspondence between a set of speciation times with a fixed permutation and an oriented tree \( T \) (Ford et al., 2009). Hence, the times of the \( n - 2 \) speciation events under the birth-death process (assuming complete sampling) conditioned on the time of the most recent common ancestor being at time \( t_1 \), are independent and identically distributed, each having density function

\[ f(s|t_1) = \mu \frac{p_1(s)}{p_0(t_1)}, \]

and distribution function

\[ F(s|t_1) = \frac{p_0(s)}{p_0(t_1)}, \]

where \( s (0 \leq s \leq t_1) \) is the time of the speciation event (Thompson, 1975). The density function and distribution function will be used extensively in the derivation of the densities of trees assuming diversified and cluster sampling.

**Sampling Schemes and Probability Densities**

In this section, we define random sampling, diversified sampling, and cluster sampling (Fig. 2). We also derive the probability density of the sampled trees for each sampling method.

For the random sampling scheme, one can distinguish two different scenarios. Under the first, called \( \rho \)-sampling, it is assumed that each lineage is sampled with some fixed probability \( (\rho_0) \), and we condition on the final sample containing exactly the number of lineages \( (n) \) included in the reconstructed tree. Thus, the actual sampling fraction, \( \hat{\rho} \), will vary from sample to sample. In the second scenario, \( n \)-sampling, it is assumed that we know the total number of lineages \( (m) \), from which \( n \) lineages are sampled with uniform probability, i.e. the actual sampling fraction \( \rho = n / m \) is known and fixed. The \( \rho \)-sampling scenario is the one typically considered under the random sampling scheme, because it is mathematically more convenient (Nee et al., 1994b; Yang and Rannala, 1997; Stadler, 2009).

For diversified and cluster sampling – as we do not sample uniformly – the \( \rho \)-sampling scenario is inapplicable. Therefore, we have to assume \( n \)-sampling for these methods.
Unfortunately, $n$-sampling leads to an awkward density for the random sampling approach (Stadler, 2009). Therefore, we will assume in the remainder of the text that we are using $\hat{\rho}$-sampling for the random sampling approach and $n$-sampling for the other methods. We will use $\rho$ loosely to refer to $\hat{\rho}$ or $\rho$ depending on context. In general, we expect the differences between $\rho$-sampling and $n$-sampling to be small, usually negligible, for the random sampling method; if for example $m = 100$ the difference in estimated speciation and extinction rates between the sampling probability $\rho$ and the sampled fraction $\hat{\rho}$ will be less than 5% with 95% probability (data not shown). The correlation of the speciation and extinction rates with different sampling probabilities is shown in Stadler (2009) and visualized in Figure 3.

**Random sampling (RS)**

Under the random sampling scheme, we assume that each extant species is sampled with a constant probability $\rho$ (Yang and Rannala, 1997; Stadler, 2009). Given a fixed complete tree of size $m$, the size of the sampled tree then follows a binomial distribution with parameters $m$ and $\rho$ and an expected sample size $\rho m$. To condition on $n$, we simply marginalize the density over different values of $m$ (Stadler, 2009).

Let $\hat{p}_0(t)$ and $\hat{p}_1(t)$ be the probability that a species at time $t$ has 0 and 1 sampled descendants at time 0, respectively, assuming random sampling. We have from Yang and
Figure 3: Two example data sets presenting the impact of $\rho$ on the estimated $\hat{\mu} / \lambda$. For the left (resp. right) figure the $\rho$ was assumed to be 0.88 (resp. 0.74).

Rannala (1997),

\[
\begin{align*}
\hat{p}_0(t) &= 1 - \frac{\rho(\lambda - \mu)}{\rho \lambda + (\lambda(1 - \rho) - \mu)e^{-(\lambda - \mu)t}}, \\
\hat{p}_1(t) &= \frac{\rho(\lambda - \mu)^2 e^{-(\lambda - \mu)t}}{(\rho \lambda + (\lambda(1 - \rho) - \mu)e^{-(\lambda - \mu)t})^2}.
\end{align*}
\]

The density of the sampled tree $T$ conditioned on $t_1$, the time of the most recent common ancestor, and $n$ can be derived from these equations using the method outlined in (Rannala and Yang, 1996). The density is (see Stadler (2010a)),

\[
f_{RS}(T|t_1, n) = \frac{1}{n-1} \prod_{i=2}^{n-1} \frac{(\lambda - \mu)\hat{p}_1(t_i)}{(1 - \hat{p}_0(t_1)) (1 - e^{-(\lambda - \mu)t_1})}. \tag{3}
\]

When conditioning only on $n$, we need to assume a prior distribution for $t_0$ the time of origin of the process. Assuming a uniform (improper) prior (Aldous and Popovic, 2005) on $(0, \infty)$ yields the density of the tree $T$ conditioned on $n$ (Stadler, 2009),

\[
f_{RS}(T|n) = n \frac{\hat{p}_1(t_1)}{1 - \hat{p}_0(t_1)} \prod_{i=1}^{n-1} \lambda \hat{p}_1(t_i). \tag{4}
\]

Note that in the RS densities, if $\lambda - \mu$ and $\rho \lambda$ are constant, the equations are
invariant (see also Stadler (2009)), meaning that only two out of the three parameters ($\lambda$, $\mu$ and $\rho$) can be estimated.

**Diversified sampling (DS)**

Under the diversified sampling scheme, we assume that we sample a fixed fraction $\rho = n/m$ of extant species. The $n$ species to sample are chosen such that the sum of edge lengths in the resulting sampled tree is maximized, i.e. the most distant species are sampled. This is equivalent to sampling species such that precisely the oldest $n - 1$ speciation times are included in the sampled tree.

**Derivation of $f_{DS}(T|t_1, n)$**

Note that $m = n/\rho$ is the total number of extant species, both sampled and non-sampled extant species. The joint density of the first $n - 1$ speciation events among all $m - 1$ speciation events yields the tree density of the sampled tree $T$. Remembering that every speciation event is independent and identical distributed conditioned on the number of species alive today $m$, we can derive the joint density from a binomial distribution with the condition that all missing $m - n$ speciation events occurred after the $n - 1$ speciation event. Using Equations 1 and 2, we obtain,

$$f_{DS}(T|t_1, n) = \frac{1}{n-1} \binom{m-2}{n-2} F(t_{n-1}|t_1)^{m-n} \prod_{i=2}^{n-1} f(t_i|t_1)$$

where $n - 1$ is the number of possibilities to insert the root in the permutation of speciation events. The binomial coefficient $\binom{m-2}{n-2}$ is the number of permutations of the $n - 2$ sampled speciation times $t_2, \ldots, t_{n-1}$ (excluding the root) with the $m - n$ later speciation events respecting the order of the $n - 2$ and the $m - n$ events. $F(t_{n-1}|t_1)$ is the probability of a speciation event being more recent than the most recent observed speciation time $t_{n-1}$ and there are $m - n$ such unobserved speciation times (see Figure 2.b).

**Derivation of $f_{DS}(T|n)$**

Now we condition only on the number of sampled species, $n$. Again, we assume a uniform prior on $(0, \infty)$ for $t_0$ the time of origin. In order to calculate $f(T|n)$, we first note that $f(T|t_0, n) = f(T|t_0, m)$ as $n$ is determined by $m$. We calculate $f(T, t_0|n) = f(T, t_0|m) = f(T|t_0, m)f(t_0|m)$ and then integrate over $t_0$. Equation 1 holds also when replacing the
most recent common ancestor by the time of origin (Gernhard, 2008). Then,

\[ f_{DS}(T|t_0, m) = \left( \frac{m-1}{n-1} \right) F(t_{n-1}|t_0)^{n-n} \prod_{i=1}^{n-1} f(t_i|t_0). \]

From Gernhard (2008) Equation 3, we further have that the distribution of the time of the origin, given the the number of extant species \( m \), is given by

\[ f(t_0|m) = m\lambda \left( \frac{\lambda}{\mu} p_0(t_0) \right)^{m-1} p_1(t_0). \]

Therefore,

\[ f_{DS}(T|n) = \int_{t_1}^\infty f_{DS}(T|t_0, m)f(t_0|m)dt_0 \]

\[ = m \left( \frac{m-1}{n-1} \right) \lambda^m \mu^{n-m} p_0(t_{n-1})^{m-n} \prod_{i=1}^{n-1} p_1(t_i) \int_{t_1}^\infty p_1(t_0)dt_0 \]

\[ = m \left( \frac{m-1}{n-1} \right) \lambda^m \mu^{n-m} (\lambda - \mu) p_0(t_{n-1})^{m-n} e^{-(\lambda - \mu)t_1} \frac{1}{(\lambda - \mu e^{-(\lambda - \mu)t_1})} \prod_{i=1}^{n-1} p_1(t_i) \]

is the desired probability density for the inference on the speciation and extinction rates used later in this paper.

**Cluster sampling (CS)**

We now analyze the complete opposite of diversified sampling, namely cluster sampling. We assume that the sampled tree includes the root of the complete tree; diversified sampling then retains the \( n - 2 \) nodes closest to the root, while cluster sampling retains the \( n - 2 \) nodes closest to the tips (the reason for assuming that the root of the complete tree is included also in the cluster sample is simply that the root of the sampled tree, irrespective of sampling scheme, is typically the tree of interest). Retaining the \( n - 2 \) most recent nodes in the tree aims to minimize the diversity (i.e. the sum of edge lengths) of the subtree given we keep the root, just as diversified sampling maximizes it. Note that some tree topologies cannot be sampled strictly according to this definition of cluster sampling. Specifically, strict cluster sampling requires that the \( n - 2 \) most recent nodes represent two separate clusters in the tree being descendants of the two branches at the root. Of course there are complete trees that cannot be sampled strictly according to this criterion. However, the subtree on \( n \) leaves of these complete trees which minimizes the diversity has greater diversity than a tree with speciation events at the times determined by cluster sampling. Thus our cluster sampling provides a lower bound on the time of
speciation events and diversity. We use our definition of cluster sampling as it is easier to analyze than more general cases of minimum-diversity sampling.

Under strict cluster sampling, we have

\[ f_{CS}(T|t_1,n) = \frac{1}{n-1} \left( \frac{m-2}{n-2} \right) (1 - F(t_2|t_1))^{m-n} \prod_{i=2}^{n-1} f(t_i|t_1), \]  

where \( f(t_i|t_1) \) and \( F(t_2|t_1) \) were defined in (1) and (2) respectively. The equation is obtained analogously to \( f_{DS}(T|t_1,n) \).

**Derivation of \( f_{CS}(T|n) \)**

The equation for \( f_{CS}(T|n) \) is derived in a similar manner to \( f_{DS}(T|n) \). We have,

\[ f_{CS}(T|t_0,m) = \left( \frac{m-1}{n-1} \right) (F(t_1|t_0) - F(t_2|t_0))^{m-n} \prod_{i=1}^{n-1} f(t_i|t_0). \]

As the density \( f(t_0|m) \) is independent of any clustering scheme, we get in the same way as in Section Derivation of \( f_{DS}(T|n) \),

\[ f_{CS}(T|n) = m \left( \frac{m-1}{n-1} \right) \lambda^{m-1} \mu^{n-m} (\lambda - \mu)(p_0(t_1) - p_0(t_2))^{m-n} \frac{e^{-(\lambda-\mu)t_1}}{(\lambda - \mu e^{-(\lambda-\mu)t_1})} \prod_{i=1}^{n-1} p_1(t_i). \]

being the probability density of the tree given that \( n \) species were sampled. This density will be used for the inference under the cluster sampling scheme.

**Inference on simulated an empirical data**

In this section we first investigate the introduced bias on speciation and extinction rates. We constructed a set of simulated data and sampled species under each of sampling schemes. Then, we estimated \( \lambda \) and \( \mu \) using the probability densities from Equations (4), (6) and (8) on the set of simulated data. A hill-climbing algorithm (Nelder and Mead, 1965) implemented in the “stats” package of R (R Development Core Team, 2009) was used to find the maximum likelihood estimates.

In the second part we analyzed the speciation and extinction rates on empirical data using maximum likelihood and Bayesian inference to verify if the extinction rate estimates are zero under all sampling schemes. We conclude our analysis by by comparing the models on the empirical data using Bayes factors.
Table 1: Mean estimate of $\hat{\lambda}$ from 1000 simulated trees and the 95% confidence interval, conditioning on $n$. True parameters were: $\lambda = 2$, $\mu = 1$ (so $\frac{\mu}{\lambda} = 0.5$ is the true value), $m = 200$, $n = 100$.

<table>
<thead>
<tr>
<th>Actual sample \ Inference assumption</th>
<th>Diversified sample</th>
<th>Random sample</th>
<th>Cluster sample</th>
<th>Complete sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversified Sample</td>
<td>0.504 [0.136,0.723]</td>
<td>0.000 [0.000,0.000]</td>
<td>0.000 [0.000,0.000]</td>
<td>0.000 [0.000,0.000]</td>
</tr>
<tr>
<td>Random Sample</td>
<td>0.999 [0.987,1.000]</td>
<td>0.506 [0.056,0.745]</td>
<td>0.000 [0.000,0.000]</td>
<td>0.020 [0.000,0.511]</td>
</tr>
<tr>
<td>Cluster Sample</td>
<td>1.000 [0.999,1.000]</td>
<td>0.937 [0.766,0.985]</td>
<td>0.531 [0.151,0.766]</td>
<td>0.873 [0.533,0.970]</td>
</tr>
</tbody>
</table>

Simulated Data

First, we simulated 1000 random trees with the software package TreeSim (Stadler, 2010b). Each tree had $m = 200$ extant species, and birth and death rates equal to $\lambda = 2.0$ and $\mu = 1.0$, respectively. Next, we took one subsample of size $n = 100$ from each of the extant trees (hence $\rho = n/m = 0.5$) using the cluster sampling and diversified sampling methods. For the random sampling method, we obtained a subsample of size $n$ close to 100 by randomly deciding on the inclusion of each tip with a constant probability $\rho_0 = 0.5$, and subsequently determined $\rho$ as the resulting ratio $n/m$.

This gave three subsampled trees from each simulated tree. For each of these trees, we then estimated the speciation rate $\lambda$ and extinction rate $\mu$ under four different assumptions, namely that the subsampled tree was produced by: (1) a cluster sample; (2) a random sample; (3) a diversified sample; or (4) a complete sample, i.e., erroneously that $m = 100$ and $\rho = \rho_0 = 1$.

Inference on the Simulated Data

In the simulations, we examined three types of data — diversified sample, random sample, and cluster sample — and for each data set, we examined four types of inference procedures — assuming data were a diversified sample, a random sample, a cluster sample or a complete sample. Clearly, inference should work satisfactorily when the correct assumption is made about the sampling method but what systematic biases, if any, occur when this is not the case? For example, what happens if data are treated as a random sample of the extant tree when data were in fact collected so as to have species as distant as possible in the sampled tree, i.e. we have a diversified sample?

The diversification rate was correctly estimated when the actual sample was produced under any of the three sampling schemes, and the analysis was done assuming the correct sampling scheme (see Tables 1 and 2). Furthermore, if the analysis was done assuming the random sampling scheme or the complete sampling scheme, then the net diversification rate can be estimated correctly or with only a minor bias. Until now
Table 2: Mean estimate of $\hat{\lambda} - \mu$ from 1000 simulated trees and the 95% confidence interval, conditioning on $n$. True parameters were: $\lambda = 2$, $\mu = 1$ (so $\lambda - \mu = 1.0$ is the true value), $m = 200$, $n = 100$.

<table>
<thead>
<tr>
<th>Actual sample \ Inference assumption</th>
<th>Diversified sample</th>
<th>Random sample</th>
<th>Cluster sample</th>
<th>Complete sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversified Sample</td>
<td>1.014 [0.688,1.417]</td>
<td>1.221 [1.025,1.425]</td>
<td>0.394 [0.300,0.506]</td>
<td>0.839 [0.608,0.996]</td>
</tr>
<tr>
<td>Random Sample</td>
<td>0.003 [0.000,0.178]</td>
<td>1.006 [0.672,1.465]</td>
<td>0.415 [0.310,0.547]</td>
<td>0.959 [0.671,1.217]</td>
</tr>
<tr>
<td>Cluster Sample</td>
<td>0.000 [0.000,0.000]</td>
<td>0.953 [0.280,2.582]</td>
<td>0.976 [0.501,1.502]</td>
<td>0.953 [0.280,2.582]</td>
</tr>
</tbody>
</table>

the complete or random sampling scheme were the only available assumptions on how the data was obtained and thus this had a negligible affect on the diversification rate estimate. However, it had a large effect on the relative extinction rate as we show in the remainder of the paper. We summarize the results in Table 3. Note that due to the relatively small size of the trees ($m = 200$) and the small sampling fraction ($\rho = 0.5$) the variance in the trees is large and thus the confidence intervals too.

Table 3: General pattern for $\hat{\mu}$ (or $\hat{\frac{\lambda}{\mu}}$) for different types of sampling and different inference assumptions.

<table>
<thead>
<tr>
<th>Actual sample \ Inference assumption</th>
<th>Diversified sample</th>
<th>Random sample</th>
<th>Cluster sample</th>
<th>Complete sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversified Sample</td>
<td>good</td>
<td>$\hat{\mu} &lt; \mu_0$</td>
<td>$\hat{\mu} &lt;&lt; \mu_0$</td>
<td>$\hat{\mu} &lt;&lt; \mu_0$</td>
</tr>
<tr>
<td>Random Sample</td>
<td>$\hat{\mu} &gt; \mu_0$</td>
<td>good</td>
<td>$\hat{\mu} &lt; \mu_0$</td>
<td>$\hat{\mu} &lt; \mu_0$</td>
</tr>
<tr>
<td>Cluster Sample</td>
<td>$\hat{\mu} &gt;&gt; \mu_0$</td>
<td>$\hat{\mu} &gt; \mu_0$</td>
<td>good</td>
<td>$\hat{\mu} &gt; \mu_0$</td>
</tr>
</tbody>
</table>

In a further analysis we studied the effect of the sampling fraction $\rho$, the tree size $m$ and the true extinction rate $\frac{\lambda}{\mu}$, on the bias. For these additional simulations, we used the diversified sampling scheme to create the subsamples and the random sampling scheme for inference. We simulated again 100 trees with $\lambda = 2.0$ and $\mu = 1.0$. First, we fixed the size of the extant trees $m$ to 200 and subsampled the trees with $\rho$ varying between 0.05 and 0.95 (Fig. 4.a). The results show that, even when subsampling is acknowledged but the wrong sampling scheme is assumed, we can observe $\hat{\mu}$ estimates of 0 when $\rho < 0.75$.

Second, we fixed the sampling size $n$ to 100 and varied $m$ between 110 and 200. The same effect as in the previous set of simulations is observed (see Figure 4.b). Note that fixing the sample size $n$ to 100 and varying $m$ between 110 and 200 has the same effect as fixing $n$ and varying $\rho = [0.5, 0.91]$.

Third, we fixed $\rho$ to 0.8 and let $m = [10, 200]$ change, which results in $n$ varying between 8 and 160. There is hardly any effect observable on the mean estimates depending on the tree size (see Figure 4.c).
Figure 4: Illustration of how estimation bias (of $\mu$ and $\lambda$) depends on parameter values. The trees for a-c were created with $\lambda = 2.0$ and $\mu = 1.0$. All analyses were performed under diversified sampling but inference assuming random sampling. Top Left: Fixed $m = 200$ with varying $\rho = n/m$. Top Right: Fixed $n = 100$ with varying $m$. Bottom Left: Fixed $\rho = 0.8$ with varying $m$ and $n$. Bottom Right: Fixed $m = 200$, $n = 160$ (so $\rho = 0.8$) and $\lambda = 2.0$ and varying $\mu = [0.05, 1.95]$. 
In the last set of simulations, we fixed all sampling parameters \( m = 200, n = 160 \) and \( \rho = 0.8 \) but this time had a varying \( \frac{\mu}{\lambda} = [0.025, 0.975] \) with \( \lambda = 2.0 \). The results show that the value of \( \mu \) has little effect on the bias of the estimates.

Using the same simulated data, we also tried inference under the complete sampling assumption. The results were very similar to those described above, except shifted towards lower absolute \( \lambda \) and \( \mu \) estimates (data not shown).

From this investigation, we conclude that the sampling fraction \( \rho \) has the strongest impact on the introduced bias. The tree size \( m \) and the relative extinction rate \( \frac{\mu}{\lambda} \) have hardly any effect when \( \rho \) remains the same.

**Empirical Data**

Phillimore and Price (2008) studied shifts over time in speciation and extinction rates in a large set of bird phylogenies. Their set of trees showed a large amount of support for rate shifts and rejected the constant birth death process in more than 50% of the trees. For our study we selected the trees which had a known sampling fraction of \( \rho \leq 0.90 \), which reduced their number of data sets from originally 45 to 18 trees (see Table 4).

Table 4: Data sets from Phillimore and Price (2008), removing trees with \( \rho > 0.9 \)

<table>
<thead>
<tr>
<th>Nr</th>
<th>Phylogeny</th>
<th>Species (n)</th>
<th>Missing m</th>
<th>( \rho )</th>
<th>age (in MYA)</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aegotheles</td>
<td>8</td>
<td>1 9</td>
<td>0.89</td>
<td>9.67</td>
<td>-0.434</td>
</tr>
<tr>
<td>2</td>
<td>Amazona</td>
<td>28</td>
<td>3 31</td>
<td>0.90</td>
<td>6.72</td>
<td>-1.856</td>
</tr>
<tr>
<td>3</td>
<td>Anas</td>
<td>45</td>
<td>6 51</td>
<td>0.88</td>
<td>8.35</td>
<td>-1.377</td>
</tr>
<tr>
<td>4</td>
<td>Anthus</td>
<td>37</td>
<td>9 46</td>
<td>0.80</td>
<td>12.65</td>
<td>-2.855</td>
</tr>
<tr>
<td>5</td>
<td>Caciques and oropendolas</td>
<td>17</td>
<td>2 19</td>
<td>0.89</td>
<td>7.86</td>
<td>-1.765</td>
</tr>
<tr>
<td>6</td>
<td>Dendroica, Parula, Seinurus, Vermivora</td>
<td>40</td>
<td>5 45</td>
<td>0.89</td>
<td>9.09</td>
<td>-2.224</td>
</tr>
<tr>
<td>7</td>
<td>Grackles and allies</td>
<td>36</td>
<td>4 40</td>
<td>0.90</td>
<td>8.42</td>
<td>-2.828</td>
</tr>
<tr>
<td>8</td>
<td>Hemispingus</td>
<td>12</td>
<td>2 14</td>
<td>0.86</td>
<td>15.69</td>
<td>-1.635</td>
</tr>
<tr>
<td>9</td>
<td>Myiarchus</td>
<td>19</td>
<td>3 22</td>
<td>0.86</td>
<td>9.59</td>
<td>1.854</td>
</tr>
<tr>
<td>10</td>
<td>Phylloscopus and Seicercus</td>
<td>59</td>
<td>11 70</td>
<td>0.84</td>
<td>12.33</td>
<td>-2.991</td>
</tr>
<tr>
<td>11</td>
<td>Puffinus</td>
<td>24</td>
<td>3 27</td>
<td>0.89</td>
<td>7.84</td>
<td>1.490</td>
</tr>
<tr>
<td>12</td>
<td>Ramphastos</td>
<td>8</td>
<td>3 11</td>
<td>0.73</td>
<td>8.11</td>
<td>-0.483</td>
</tr>
<tr>
<td>13</td>
<td>Sterna</td>
<td>34</td>
<td>10 44</td>
<td>0.77</td>
<td>21.66</td>
<td>1.365</td>
</tr>
<tr>
<td>14</td>
<td>storks</td>
<td>16</td>
<td>3 19</td>
<td>0.84</td>
<td>11.21</td>
<td>-1.254</td>
</tr>
<tr>
<td>15</td>
<td>Tangara</td>
<td>42</td>
<td>7 49</td>
<td>0.86</td>
<td>10.10</td>
<td>-2.465</td>
</tr>
<tr>
<td>16</td>
<td>Trogons</td>
<td>29</td>
<td>10 39</td>
<td>0.74</td>
<td>24.88</td>
<td>-0.910</td>
</tr>
<tr>
<td>17</td>
<td>Turdus and allies</td>
<td>60</td>
<td>10 70</td>
<td>0.86</td>
<td>14.29</td>
<td>-2.278</td>
</tr>
<tr>
<td>18</td>
<td>Wrens</td>
<td>50</td>
<td>24 74</td>
<td>0.68</td>
<td>12.10</td>
<td>-3.628</td>
</tr>
</tbody>
</table>

**Maximum likelihood inference**

The results of reanalyzing the 18 data sets from Phillimore and Price (2008) under different assumptions on the sampling procedure are shown in Table 5. As expected, the
analyses of some of the empirical data sets show that the estimated extinction rates are significantly higher when inference is based on the presumably more realistic assumption that we have a diversified sample, rather than on the less realistic or erroneous assumptions that we have a random or complete sample (see Table 5). Nevertheless, more than three quarters of the estimates remain equal to 0 under the assumption of diversified sampling.

Table 5: Data from Phillimore and Price (2008), conditioned on \( n \) (and given \( \rho \)). Estimates for \( \frac{\hat{\mu}}{\hat{\lambda}} \) are specified.

<table>
<thead>
<tr>
<th>Tree \ inf model</th>
<th>Diversified Sample</th>
<th>Random Sample</th>
<th>Cluster Sample</th>
<th>Complete Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.767</td>
<td>0.506</td>
<td>0.307</td>
<td>0.440</td>
</tr>
<tr>
<td>4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>6</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>7</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>8</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>9</td>
<td>0.251</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>10</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>11</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>12</td>
<td>0.134</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>13</td>
<td>0.656</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>14</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>15</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>16</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>17</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>18</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Bayesian inference

We also analyzed the 18 trees from Phillimore and Price (2008) using Bayesian inference to obtain estimates of the relative extinction rate \( \frac{\hat{\mu}}{\hat{\lambda}} \) and the net-diversification rate \( \lambda - \mu \). We assumed a uniform prior on the relative extinction rate \( \frac{\hat{\mu}}{\hat{\lambda}} \sim U[0, 1] \) and a uniform improper prior on the net-diversification rate \( \lambda - \mu \sim U[0, \infty] \). The posterior probability was approximated using the Markov chain Monte Carlo (MCMC) method (Metropolis et al., 1953; Hastings, 1970) with a chain length of 10000. The chains were sampled without thinning, using a burn-in of 1000. We assessed the convergence of the chains using the within chain convergence diagnostic proposed by Geweke (1992) which is implemented in CODA (Plummer et al., 2006). Geweke’s test estimates and compares the sample means and variances of two independent parts, so-called windows,
of the chain. The chains appeared well sampled and converged according to this MCMC convergence diagnostic. Table 6 shows the mean posterior estimate and the 95% credible interval (alternatively the maximum a posteriori, MAP, and highest posterior density interval could have been shown).

Table 6: Bayesian estimates of the relative extinction rate based on data from Phillimore and Price (2008), conditioned on \( n \) (and given \( \rho \)). The marginal posterior distribution of \( \lambda \) is summarized using the posterior mean and the 95% credible interval.

<table>
<thead>
<tr>
<th>Tree \ inf model</th>
<th>Diversified Sample</th>
<th>Random Sample</th>
<th>Cluster Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.349 [0.070, 0.758]</td>
<td>0.241 [0.009, 0.702]</td>
<td>0.313 [0.027, 0.720]</td>
</tr>
<tr>
<td>2</td>
<td>0.159 [0.012, 0.474]</td>
<td>0.128 [0.016, 0.325]</td>
<td>0.240 [0.071, 0.513]</td>
</tr>
<tr>
<td>3</td>
<td>0.565 [0.223, 0.842]</td>
<td>0.525 [0.200, 0.936]</td>
<td>0.303 [0.049, 0.700]</td>
</tr>
<tr>
<td>4</td>
<td>0.135 [0.004, 0.377]</td>
<td>0.126 [0.012, 0.364]</td>
<td>0.118 [0.015, 0.299]</td>
</tr>
<tr>
<td>5</td>
<td>0.204 [0.022, 0.500]</td>
<td>0.256 [0.046, 0.535]</td>
<td>0.138 [0.008, 0.365]</td>
</tr>
<tr>
<td>6</td>
<td>0.144 [0.002, 0.400]</td>
<td>0.171 [0.005, 0.439]</td>
<td>0.089 [0.020, 0.216]</td>
</tr>
<tr>
<td>7</td>
<td>0.161 [0.019, 0.413]</td>
<td>0.134 [0.015, 0.405]</td>
<td>0.115 [0.010, 0.296]</td>
</tr>
<tr>
<td>8</td>
<td>0.344 [0.083, 0.658]</td>
<td>0.358 [0.041, 0.719]</td>
<td>0.241 [0.019, 0.555]</td>
</tr>
<tr>
<td>9</td>
<td>0.383 [0.079, 0.759]</td>
<td>0.302 [0.033, 0.759]</td>
<td>0.238 [0.012, 0.627]</td>
</tr>
<tr>
<td>10</td>
<td>0.125 [0.004, 0.318]</td>
<td>0.087 [0.006, 0.221]</td>
<td>0.053 [0.003, 0.163]</td>
</tr>
<tr>
<td>11</td>
<td>0.281 [0.020, 0.672]</td>
<td>0.298 [0.048, 0.696]</td>
<td>0.205 [0.021, 0.491]</td>
</tr>
<tr>
<td>12</td>
<td>0.321 [0.062, 0.693]</td>
<td>0.334 [0.058, 0.727]</td>
<td>0.285 [0.066, 0.588]</td>
</tr>
<tr>
<td>13</td>
<td>0.384 [0.053, 0.750]</td>
<td>0.216 [0.014, 0.638]</td>
<td>0.315 [0.046, 0.667]</td>
</tr>
<tr>
<td>14</td>
<td>0.365 [0.011, 0.851]</td>
<td>0.196 [0.014, 0.635]</td>
<td>0.168 [0.017, 0.453]</td>
</tr>
<tr>
<td>15</td>
<td>0.126 [0.005, 0.352]</td>
<td>0.101 [0.002, 0.355]</td>
<td>0.125 [0.017, 0.305]</td>
</tr>
<tr>
<td>16</td>
<td>0.255 [0.005, 0.606]</td>
<td>0.203 [0.016, 0.514]</td>
<td>0.157 [0.019, 0.410]</td>
</tr>
<tr>
<td>17</td>
<td>0.136 [0.026, 0.339]</td>
<td>0.086 [0.006, 0.225]</td>
<td>0.078 [0.006, 0.212]</td>
</tr>
<tr>
<td>18</td>
<td>0.136 [0.007, 0.409]</td>
<td>0.082 [0.013, 0.234]</td>
<td>0.063 [0.008, 0.186]</td>
</tr>
</tbody>
</table>

In contrast to the ML estimates, none of the mean posterior estimates equals zero. This is due to the different use of the likelihood in the two methods. In Bayesian inference it is common practice to report the marginal distribution of the parameter of interest, integrating out the uncertainty of other parameters (e.g. Pawitan (2001), pp 278-279). If a point estimate is desired, the mean or median of the marginal distribution is often used. In ML inference, in contrast, the estimate is based on the peak of the joint probability distribution, i.e. analyzing the likelihood as a function of the parameter of interest keeping other parameter(s) at the value maximizing the joint probability distribution, a quantity known as the profile likelihood (e.g. Pawitan (2001), pp 61-64).

As we have shown in the Tables 5 and 6 the mean (or median) of the marginal distribution and the ML estimates differ strongly and need to be interpreted differently.
An example is shown in Figure 5 where the marginal and joint distribution are given. The mean and the 95% credible interval of the marginal distribution are 0.241 and [0.009, 0.702], respectively, although the MAP and ML estimate both have $\hat{\mu}/\hat{\lambda}$ equal zero.

Figure 5: The two plots are extracted from Tracer v1.5 (Rambaut and Drummond, 2009) of samples from an MCMC run. We assigned a uniform prior probability between 0 and 1 to the relative extinction rate and an improper uniform prior probability between 0 and infinity to the net diversification rate. The tree analyzed here is Tree 1 (see Table 4) and assumed random sampling with a sampling probability of $\rho = 0.89$. The plot to the left shows the marginal distribution of the relative extinction rate with a mean estimate of 0.241 and the 95% credible interval between 0.009 and 0.702. The plot to the right shows the joint distribution of the net diversification rate and the relative extinction rate and the ML-estimates are found where the density of points is highest: $\lambda - \mu \approx 0.137$ and $\mu/\lambda \approx 0.0$.

It is also worth noting that the Bayesian marginal means no longer have the same ordering between the different sampling assumptions as was the case with the ML estimates. For tree 2, for instance, cluster sampling has the highest mean marginal estimate of the relative extinction rate (see Table 6), whereas ML estimates of the extinction rate were always highest for the diversified sampling method in our simulations. The explanation to this difference between inference procedures also lies in the fact that Bayesian inference uses the marginal likelihood whereas maximum likelihood uses the profile likelihood.

**Model Selection**

To examine whether biased sampling occurs in practice, we investigated the fit of diversified, cluster, random, and complete sampling models to all 18 trees under Bayesian inference. The evaluation of model fit was performed by computing Bayes factors by
thermodynamic integration (Lartillot and Philippe, 2006),

\[ BF = \frac{p(D|M_1)}{p(D|M_0)} = \int_0^1 \int_0^\infty p(D|\lambda, \mu, M_1) p(\lambda, \mu|M_1) d(\lambda - \mu) d\mu \]

\[ \int_0^1 \int_0^\infty p(D|\lambda, \mu, M_0) p(\lambda, \mu|M_0) d(\lambda - \mu) d\mu \],

with \( \rho = \frac{n}{m} \). We used the model switch scheme described in Lartillot and Philippe (2006) with \( C = 100 \) (\( \delta \beta = 0.01 \)) and \( Q = 10000 \). The full sampling model (\( m = n \)) was used as the reference model (\( M_0 \)) and all Bayes factors are reported on a log-scale. Hence, Bayes factors greater than 1 show strong support for the proposed model over the full sampling model, whereas Bayes factors of smaller than -1 show strong support against the proposed model. The results are given in Table 7.

Table 7: Bayes factors compared with complete sampling (log-scale – so positive values indicate better fit than complete sampling). The best model per data set is shown in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS</td>
<td>0.49</td>
<td>0.53</td>
<td>-3.04</td>
<td>3.88</td>
<td>0.79</td>
<td>1.05</td>
<td>-0.75</td>
<td>-0.77</td>
<td>0.35</td>
<td>-1.34</td>
<td>0.87</td>
<td>0.03</td>
<td>-0.09</td>
<td>-0.24</td>
<td>-1.04</td>
<td>2.10</td>
<td>2.05</td>
<td>5.92</td>
</tr>
<tr>
<td>RS</td>
<td>0.08</td>
<td>0.14</td>
<td>0.00</td>
<td>0.70</td>
<td>0.15</td>
<td>0.27</td>
<td>0.19</td>
<td>0.08</td>
<td>0.73</td>
<td>0.12</td>
<td>0.17</td>
<td>0.16</td>
<td>0.17</td>
<td>0.47</td>
<td>0.51</td>
<td>0.60</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>-0.08</td>
<td>-0.25</td>
<td>-2.02</td>
<td>-5.67</td>
<td>-0.78</td>
<td>-1.44</td>
<td>-1.55</td>
<td>-0.29</td>
<td>-0.57</td>
<td>-4.99</td>
<td>-1.58</td>
<td>-1.06</td>
<td>-1.00</td>
<td>-4.35</td>
<td>-4.71</td>
<td>-10.04</td>
<td>-1.03</td>
<td>-27.98</td>
</tr>
</tbody>
</table>

Diversified sampling is the best model in ten out of the 18 cases. In five of these cases, there is a significant increase of the Bayes factor compared to the complete sampling model and therefore a strong support for the diversified sampling as the true model. Although one might expect diversified sampling to be the best fitting model when the \( \gamma \) value is a large negative value, this is actually not always the case (see Table 4). One reason might be that the model assumes that all non-sampled speciation events occurred after the most recent speciation event. Even if there is just one very recent speciation event sampled (this is the case for tree 10, for instance), then the model may be rejected although the sample might otherwise have been very close to a diversified sample.

It is noteworthy that the random sampling model always performs better than the complete sampling model (all log Bayes factors are positive). This is a strong motivation to always consider the sampling fraction when speciation and extinction rate inference is performed.

Discussion

Empirical studies using the BDP often show an estimated death rate close to zero even though it is well known that species go extinct during evolution (Nee, 2006; Purvis, 2008). It thus appears that the BDP model often underestimates real extinction rates,
an observation that inspired the current paper.

It is widely acknowledged that incomplete phylogenies may cause underestimated extinction rates in the BDP model (Yang and Rannala, 1997; Stadler, 2009). Specifically, if the sampling fraction of species, \( \rho \), is assigned a larger value than really is the case, then the estimated extinction rate \( \mu \) will be underestimated (Stadler, 2009). This might easily happen when the number of species in the complete tree is poorly known, or when the data set is analyzed as if it represented the full extant tree even though it really represents a sample (i.e. \( \rho \) is incorrectly set to 1).

In this paper, we point to another possible explanation of the tendency of the BDP to underestimate extinction rates, namely systematic biases introduced by the sampling protocol. Often, species are sampled to be more distantly related to each other than they would be in a random sample, a protocol we refer to as diversified sampling. Diversified sampling has the effect that the sampled tree typically contains long external branches. If this is neglected in the statistical analysis and data are treated as if they came from a random sample, then the death rate will be underestimated. A data set that is analyzed as if it were the complete tree, but in fact was a diversified sample of the tree, will hence have an underestimated death rate \( \mu \) both because the sampling fraction is overestimated and because the analysis is not accounting for the sampling bias.

One possible explanation for still having \( \mu/\lambda \) estimated to 0 (when the true value is positive) could be that the sampling fraction \( \rho \) is over-estimated, i.e. that the number of species in the extant tree is larger than the \( m \) used in the analysis. We chose two data sets to illustrate this. One of these data sets had an inferred extinction rate of zero while the other had a positive inferred rate, given the estimate of the sampling fraction quoted in the original study (\( \rho \)) and under all sampling method assumptions (diversified, random and cluster sampling). We now varied the assumed sampling fraction, \( \rho \), over the interval \((0, 1]\), see Figure 3. The results show that, for both diversified and random sampling, the estimated extinction rate increases with decreased sampling fraction. For diversified sampling in particular, the rate estimate increases quite dramatically when the assumed sampling fraction falls below a certain critical value, which varies between the trees, and the estimate \( \hat{\mu}/\hat{\lambda} \) converges to 1 as \( \rho \to 0 \). Contrary to this behavior, the relative extinction rate estimate \( \hat{\mu}/\hat{\lambda} \) under the cluster sampling assumption increases with increasing \( \rho \). As expected, the estimate under any of the three sampling schemes converges to the estimate \( \hat{\mu}/\hat{\lambda} \) of the complete sampling when \( \rho \to 1 \). We conclude from these illustrations that if \( \rho \) is over-estimated (i.e. the number \( m \) of extant species are under-estimated) under the random sampling and diversified sampling this might lead to \( \hat{\mu}/\hat{\lambda} \) being close or equal to 0 whereas it would be distinctly larger if the correct value of \( \rho \) would be used.
As mentioned earlier, a low death rate produces speciation points more centered in the tree, while a high death rate pushes the speciation points towards the tips. For this reason, one might expect that if we had a diversified sample and thought it was a random sample, we would underestimate the death rate. This is exactly the effect we see; in fact, the estimated death rate in the simulations effectively becomes zero. A similar effect occurs if we have a random sample but assume it is a cluster sample. The opposite effect, an overestimate of the extinction rate, occurs if we think we have a diversified sample but actually have a random sample or cluster sample.

Conclusions

Our simulations illustrate the importance of accommodating the sampling bias in the inference procedure. For normal-sized data sets (10 to 100 taxa) and moderate values of the sampling fraction (50 to 90 percent), the estimates of the extinction rate often differ in order of magnitude. Assuming diversified sampling results in the highest estimates, followed by random sampling (intermediate estimates) and cluster sampling (lowest estimates). However, if the actual sampling strategy is accounted for in the model, unbiased estimates of the extinction rate can be obtained.

If the sampling fraction is known but only little is known about the sampling scheme, or none of the available sampling schemes truly represents the sample, then the two sampling schemes DS and CS will produce estimates which may be interpreted as bounds for the true estimate irrespective of the sampling scheme.

Accounting for both the sampling fraction $\rho$ and the effect of diversified sampling, we show that the BDP estimates positive extinction rates for some empirical data sets (3 out of 17), for which the BDP produces zero extinction rate estimates under the random sampling assumption. Nevertheless, extinction rate estimates remain effectively zero for a large number of empirical data sets. It was illustrated that one possible explanation for this could be that the sampling fraction has been over-estimated (i.e. the number of extant species under-estimated) for some of the data sets.

Purvis (2008) gives a number of possible explanations for underestimated extinction rates, both sampling issues: incomplete phylogenies and unrepresentative clades due to non-random selection of groups for study, as well as other reasons: overestimated rates of recent splits, unacknowledged recent splits, and lastly non-constant rates of speciation and extinction (see also Rabosky and Lovette (2008)).

The effect of underestimating the extinction rate is probably a mixture of all above mentioned reasons, a different mixture for different trees. The contribution of the present paper has been to identify and illustrate the consequences of sampling species in a non-uniform manner. As we showed by Bayes factor computation, our proposed method of
acknowledging a sampling bias is in over 50% of the empirical trees the best fitting model.

Acknowledgments

We would like to thank Albert Phillimore for kindly providing us with his data. We would also like to thank Dan Rabosky for helpful discussions on the subject. The research was supported by the Swedish Research Council (TB and FR).

References


Appendix

A Derivations of various probability densities

A.1 Derivation of $f_{DS}(T|t_1)$

Recall that $m = n/\hat{\rho}$ is the total number of extant species, both sampled and non-sampled extant species, and that conditioning on $n$ also determines $m$, if $\rho$ is fixed. We
have,

\[ f_{DS}(T|t_1) = f_{DS}(T|t_1, n) f(n|t_1) \]

\[ = f_{DS}(T|t_1, n) f(m|t_1) \]

as \( m = n/\rho \). We have with Equation (1),

\[ f(m|t_1 = t_{\text{mrca}}) = \sum_{i=1}^{m-1} \frac{p_i(t_1)p_{m-i}(t_1)}{(1 - p_0(t_1))^2} \]

\[ = (m - 1) \left( \frac{\lambda}{\mu} \right)^{m-2} \frac{p_1(t_1)^2p_0(t_1)^{m-2}}{(1 - p_0(t_1))^2}. \]

Further, with \( f_{DS}(T|t_1) \) from Equation (5), we have,

\[ f_{DS}(T|t_1) = \left( \frac{m - 1}{n - 1} \right) \left( \frac{\lambda p_0(t_1)}{\mu} \right)^{m-2} \frac{p_1(t_1)^2}{(1 - p_0(t_1))^2} \]

\[ F(t_{n-1}|t_1)^{m-n} \prod_{i=2}^{n-1} f(t_i|t_1). \]  

\[ (9) \]

### A.2 Derivation of \( f_{CS}(T|t_1) \)

Analogously to the diversified sampling, we derive the probability density conditioned on \( t_1 \) for cluster sampling:

\[ f_{CS}(T|t_1) = \left( \frac{m - 1}{n - 1} \right) \left( \frac{\lambda p_0(t_1)}{\mu} \right)^{m-2} \frac{p_1(t_1)^2}{(1 - p_0(t_1))^2} (1 - F(t_2|t_1))^{m-n} \prod_{i=2}^{n-1} f(t_i|t_1). \]  

\[ (10) \]

26
Burnin Estimation and Convergence Assessment

Sebastian Höhna∗ and Kristoffer Sahlin†

November 17, 2011

Abstract

Estimating the burnin length and assessing convergence purely from the output of an MCMC run is increasingly important in Bayesian phylogenetic inference. Previously, methods for estimating the burnin and assessing convergence have been ad-hoc, such as the minimum number of effective samples or the deviation in split frequencies. In this paper we compare the currently used methods to convergence assessment methods from the mathematical literature, namely the Geweke test and the Heidelberger-Welch test. The latter two show strong advantages in being statistically consistent and unbiased. Statistical consistency and unbiasedness was verified on simulated data with known posterior distributions. Both methods consider convergence as the Null hypothesis. The Null hypothesis is rejected based on standard p-values, which are easier to interpret than a threshold as used by the effective sample size. We extend these convergence assessment methods for single and multiple chains. Furthermore, we test the performance of the convergence assessment methods on an empirical dataset and conclude that tests for convergence to the same stationary distribution from independent runs are most adequate. Additionally, we developed an automatic procedure that finds the optimal burnin in the cases we studied. All methods we tested are implemented in the open source software RevBayes (http://www.revbayes.net/).

Keywords: Convergence Diagnostics, Burnin Estimation, Bayesian inference, Markov chain Monte Carlo, Phylogenetics.

1 Introduction

Bayesian inference in phylogenetics has become very popular in recent years (Huelsenbeck et al., 2001). The seminal contributions by Yang and Rannala (1997), Mau et al.∗

∗Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden, hoehna@math.su.se.
†School of Computer Science and Communication, Royal Institute of Technology (KTH), Karolinska Institute Science Park, SE-171 65 Solna, Sweden
(1999), and Li et al. (2000) enabled approximation of the posterior distribution of the parameters of interest — e.g., the tree topology, the substitution parameters, the clock rates and the divergence times — via the Markov chain Monte Carlo (MCMC) algorithm. Adopting these ideas many software packages, such as MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), BEAST (Drummond and Rambaut, 2007), PhyloBayes (Lartillot et al., 2009) and MCMCTree in PAML (Yang, 2007), have implemented several variants of the MCMC algorithm. The availability of ready-to-use and easy-to-use Bayesian phylogenetic inference software packages has increased the number of applications for Bayesian inference, for example, approximately every second article published in 2010 in Systematic Biology uses the MCMC algorithm (see Table 1). However, with the great potential of the MCMC algorithm originated two new challenges: designing efficient MCMC algorithms (see Lakner et al. (2008) and Höhna and Drummond (2011) for recent advances) and assessing convergence to the stationarity distribution of the MCMC runs (see e.g. Nylander et al. (2008)). The challenges become more prevalent in recent studies, as can be seen by the number of reported failures to converge. These studies use more complex and demanding models, such as hierarchical models (with several layers of hyperpriors), or integrative models (e.g., simultaneous estimation of several parameters such as alignment, substitution rates, clock rates, divergence times, diversification rates, gene trees and species tree).

In the present paper we will focus on the second problem, the assessment of convergence. We define a chain having converged when the sample frequencies are a sufficiently close approximation of the stationary distribution of the chain. However, the MCMC algorithm is a stochastic approximation method for the posterior probability distribution and hence will always produce some uncertainty in the approximated estimates. Nevertheless, there are some necessary requirements on the MCMC algorithm for conducting scientific analysis, which we will define as:

1. Precision: The uncertainty of the estimator must be smaller than a given tolerance value.
2. Stability: Longer chains or more samples will not lead to significantly different estimates, given the tolerated uncertainty.
3. Reproducibility: Repeated chains from random starting values will give the same estimates, given the tolerated uncertainty.

A mathematical description of the requirements follows in the section Statistically Motivated Convergence Assessment Methods. Although the mathematical and statistical literature is filled with many attempts of assessing convergence of MCMC runs (for reviews and comparison, see Cowles and Carlin (1996); Brooks and Roberts (1998a);
Table 1: Review of the recent literature using MCMC.

<table>
<thead>
<tr>
<th>Journal</th>
<th>Articles</th>
<th>Used MCMC</th>
<th>Assessed convergence</th>
<th>Multiple runs</th>
<th>Manual/visual</th>
<th>Split frequencies</th>
<th>PSRF</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syst Biol</td>
<td>79</td>
<td>41</td>
<td>37</td>
<td>37</td>
<td>25</td>
<td>19</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>MBE</td>
<td>16</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>JEB</td>
<td>19</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BMC Evol Biol</td>
<td>35</td>
<td>16</td>
<td>16</td>
<td>13</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Evolution</td>
<td>22</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>171</td>
<td>72</td>
<td>65</td>
<td>59</td>
<td>36</td>
<td>30</td>
<td>9</td>
<td>13</td>
</tr>
</tbody>
</table>

Notes: Research articles published in the July 2011 issue of Molecular Biology and Evolution (MBE), Evolution and the Journal of Evolutionary Biology (JEB), published in June 2011 in BMC Evolutionary Biology (BMC Evol Biol) and between February 2010 and July 2011 in Systematic Biology (Syst Biol) were studied if they used MCMC methods, multiple runs and convergence assessment methods. Observed convergence assessment methods were: Manual/Visual assessment, deviation in split frequencies, the potential scale reduction factor (PSRF) and the effective sample size (ESS).

El Adlouni et al. (2006)), the amount of methods used in phylogenetic studies is rather small and the methods themselves are rather primitive. We searched five journals (Systematic Biology, Evolution, Molecular Biology and Evolution, Journal of Evolutionary Biology and BMC Evolutionary Biology) to assess the frequency of published articles using MCMC and furthermore which convergence assessment methods were used. For Evolution, Molecular Biology and Evolution, and the Journal of Evolutionary Biology we searched all research articles in the July 2011 issue; for BMC Evolutionary Biology all research articles published in June 2011; and for Systematic Biology all research articles published in issues between February 2010 and July 2011. The results are summarized in Table 1. Most studies used some convergence assessment method, though some method lack the mathematical foundation (see Section Convergence Assessment Methods currently used in Bayesian Phylogenetic Inference for a discussion). The methods used were: Manual or visual inspection of the chains and trace plots (e.g. with Tracer (Rambaut and Drummond, 2011)); Comparing split frequencies between independent runs (e.g. using AWTY (Nylander et al., 2008)); Calculating the within and between chain variance by means of the potential scale reduction factor (Gelman and Rubin, 1992b); and checking the effective sample size (ESS). The most striking observation of the literature review is that biologists use the convergence assessment methods available in the software packages they know, but unfortunately the methods available are not satisfying as shown below. This shows the lack of easy-to-use convergence assessment software packages that contain mathematically sound and consistent methods. With this paper we aim to provide such a software package.

In the published articles which we evaluated for this paper, we observed the following
tools: Tracer (Rambaut and Drummond, 2011), AWTY (Nylander et al., 2008), MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) and PhyloBayes (Lartillot et al., 2009). More sophisticated convergence assessment software packages exist in other research areas, such as CODA (Plummer et al., 2006) and BOA (Smith, 2007), which are freely available but more cumbersome in their usage. Note, BOA uses the same source code internally as CODA for the convergence assessment methods, thus, we restrict all further evaluation on CODA. We did not see a single research article of the 171 we looked at using either one, CODA or BOA. A summary of the convergence methods implemented in these software packages is given in Table 2.

The aim of the present paper is to evaluate the available convergence assessment methods, compare them with the currently applied methods and suggest new improved methods. We start the remainder of this paper with an empirical study from the perspective of a user of the MCMC algorithm who wants to perform a Bayesian phylogenetic analysis on some model and data. We analyze an empirical dataset and try to assess convergence with the methods which we have found in our literature review. Then we discuss, evaluate and propose convergence assessment methods from the statistical literature. Additionally, we provide an algorithm that automatically finds the optimal burnin. We conclude this paper with a short discussion and a recommendation on how the burnin should be estimated and the convergence should be assessed.

2 Convergence Assessment Methods currently used in Bayesian Phylogenetic Inference

In our evaluation of the currently used convergence assessment methods we found the following methods: Manual/visual inspection, split frequencies, potential scale reduction factor and the ESS. In this section we apply and discuss all four methods on MCMC output from an empirical dataset. The dataset we use is the Cettiidae dataset with one mitochondrial gene and three nuclear introns analyzed with the multispecies coalescent

<table>
<thead>
<tr>
<th>Software</th>
<th>Manual/visual</th>
<th>Split frequencies</th>
<th>PSRF</th>
<th>ESS</th>
<th>Geweke test</th>
<th>H-W test</th>
<th>S-Stationarity</th>
<th>M-Stationarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWTY</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BOA</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CODA</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MrBayes</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PhyloBayes</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RevBayes</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Tracer</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
method implemented in *BEAST (Heled and Drummond, 2010) (for the original analysis, model and data description see Alström et al. (2011)). We ran 10 independent chains, each 100,000,000 iterations with samples taken every 10,000 with the default settings of *BEAST. The likelihood traces are shown in Figure 1.

2.1 Manual or visual trace inspection as a convergence assessment method

Interestingly, half of the studies running MCMC algorithms reported that convergence was assessed (not exclusively) by manual or visual inspection. However, the description of manual or visual inspection of the traces is rather vague. In this section we elaborate what we understand one can do with visual or manual inspection of the traces. We will start with a visual inspection of the chains, first the single chains independently and then all chains together.

A trace plot of a well mixing MCMC run, which converged already to its stationarity distribution, is often described as looking like a caterpillar with a horizontal mean (no notable slope of the mean)(Gilks et al., 1996). We discarded manually a burnin until the runs seemed to have stabilized by looking at the traces of the 10 runs we performed on the Cettiidae dataset (see Figure 1). Next, we observe that all runs except run 6 fulfill the description of a horizontal caterpillar. However, some runs, e.g. run 3, have a big jump in the likelihood at the second half of the run, which might lead to the assumption that there could be another jump or plateau which the run has not reached yet. Comparing the likelihood levels of the plateaus of the different runs, e.g. run 1 and run 5, we see a significant discrepancy, although each run in itself seems to have reached the stationarity distribution. Consequently, a single chain which has reached a plateau and stabilized might be misleading.

In the second part of the manual inspection we look at the numerical output provided by Tracer. We have summarized the most important information in Table 3. All but run 6 have “high” ESS’s which indicates convergence (for more details see the section below). Most studies using manual inspection report that estimates of independent runs are “close enough”. A statistically sound procedure would be to compute the confidence intervals around the parameters of interest using the standard error of the mean (SEM) and testing if the intervals overlap. A 95% confidence interval is approximated by estimate ± 2 * SEM. Although the chains appeared to mix well, we can clearly identify different estimates from the different chains where the confidence intervals do not overlap (e.g., runs 9 and 10).

In conclusion, visual and manual inspection of a single chain might be misleading, as in the situation of our dataset. Single runs reported high ESS’s and appeared to
Figure 1: Trace plots of the log-likelihood of the 10 independent runs on the Cettiidae dataset. The visually estimated burnin is shown in light-gray. The runs were performed with *BEAST for 100,000,000 iteration with samples taken every 10,000 iterations. Run 6 fails to stabilize and several runs, e.g., run 1 and run 2 stabilize at different mean estimates which indicates non-convergence.
have stabilized at the the stationarity distribution. Nevertheless, using multiple runs — preferably more than 4 — can help to identify non-convergence. Instead actual convergence cannot be assessed, as for example run 3 might find another not yet discovered plateau. The judgement of visual inspections are subjective and it is often not clear which methods have been used if the authors state that convergence was assessed manually. Furthermore, manual assessment should not be necessary because the calculation can be implemented more efficiently in software packages.

2.2 The effective sample size as a convergence assessment method

Often the ESS is used as a convergence assessment method, although there exists no direct mathematical foundation explaining why or how this can be achieved. The software Tracer suggests, that runs with an ESS of over 200 have certainly converged, runs with an ESS between 100 and 200 should be treated carefully and runs with an ESS less than 100 failed to converged. Again, the mathematical literature is missing an explanation for these thresholds. Even if an MCMC output with only few effective samples is available, these samples can possibly be from the stationarity distribution and thus the chain could have converged. On the other hand, a high ESS can be misleading if the chain was stuck in an area of the parameter space which represents a local optimum. If we look at the ESS for the 10 runs presented in Table 3, we see that only run 6 clearly failed to converge. Although we already noticed in the previous section that run 1 and run 5 converged to a different plateau and give different estimates, their ESS’s give acceptable values. Nevertheless, a justification for the ESS as a convergence assessment method is that the ESS is directly related to the precision of an estimator and with a higher ESS the precision also becomes higher. Finally, the ESS can be used as a neces-

<table>
<thead>
<tr>
<th>Run</th>
<th>mean( ln(Posterior))</th>
<th>SEM( Posterior )</th>
<th>ESS( Posterior )</th>
<th>mean( Root height )</th>
<th>SEM( Root height )</th>
<th>ESS( Root height )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14270.0004</td>
<td>3.3917</td>
<td>312.9408</td>
<td>0.0087</td>
<td>0.0001</td>
<td>181.1252</td>
</tr>
<tr>
<td>2</td>
<td>14545.8295</td>
<td>4.8138</td>
<td>161.8855</td>
<td>0.0104</td>
<td>0.0001</td>
<td>84.6038</td>
</tr>
<tr>
<td>3</td>
<td>13920.0921</td>
<td>4.0108</td>
<td>157.9417</td>
<td>0.1383</td>
<td>0.0006</td>
<td>268.6075</td>
</tr>
<tr>
<td>4</td>
<td>14552.6048</td>
<td>4.2733</td>
<td>177.6672</td>
<td>0.1299</td>
<td>0.0004</td>
<td>461.009</td>
</tr>
<tr>
<td>5</td>
<td>13916.6144</td>
<td>2.6729</td>
<td>444.8419</td>
<td>0.1379</td>
<td>0.0004</td>
<td>647.3908</td>
</tr>
<tr>
<td>6</td>
<td>14331.9155</td>
<td>18.0599</td>
<td>12.9575</td>
<td>0.1776</td>
<td>0.0015</td>
<td>41.2940</td>
</tr>
<tr>
<td>7</td>
<td>13909.1763</td>
<td>2.8348</td>
<td>451.6379</td>
<td>0.1377</td>
<td>0.0004</td>
<td>878.6416</td>
</tr>
<tr>
<td>8</td>
<td>13911.2336</td>
<td>3.916</td>
<td>265.3323</td>
<td>0.1387</td>
<td>0.0008</td>
<td>175.6784</td>
</tr>
<tr>
<td>9</td>
<td>14549.7631</td>
<td>3.0386</td>
<td>278.8076</td>
<td>0.0005</td>
<td>0.0001</td>
<td>178.7145</td>
</tr>
<tr>
<td>10</td>
<td>13908.9141</td>
<td>2.7004</td>
<td>450.1918</td>
<td>0.1375</td>
<td>0.0004</td>
<td>499.6338</td>
</tr>
</tbody>
</table>

Notes: ESS smaller than 100 are considered to indicate non-convergence and highlighted in red; ESS between 100 and 200 are highlighted in orange and might have converged but should be handled with care.
sary statistic, i.e. that a certain ESS is required to get a sufficient precision, but is not a sufficient statistic which can detect non convergence. Another weakness, the biased behavior, is shown in Figure 4 and discussed in a later Section Estimating the stability of the estimator.

2.3 The potential scale reduction factor as a convergence assessment method

The potential scale reduction factor (PSRF) was introduced by Gelman and Rubin (1992b) to compare the within and between chain variance. We assume that its popularity in Bayesian phylogenetic inference using MCMC stems from the implementation in MrBayes. The necessity of comparison between multiple independent runs was highlighted in the previous sections by showing that single run convergence assessment methods are potentially misleading (see also Gelman and Rubin (1992a)). The PSRF is simple to compute but hard to interpret because the definition says: The independent runs have converged to the same stationarity distribution if the PSRF is close to 1. How close it should be and what the frequentist properties for repeated multiple runs are is however unknown. For the Cettiidae dataset we obtained PSRF of 4.5229 for the posterior probabilities and 8.0778 for the root-height. The comparably high PSRF indicates non-convergence between the independent runs. This failure to converge of the independent runs to the same posterior distribution agrees with our conclusions from the previous sections.

2.4 The difference in split frequencies as a convergence assessment method

Split frequencies, which correspond directly to the posterior probabilities of that split, provide the advantage over tree topologies that split frequencies are sampled more frequently and hence more precise estimates can be obtained. However, method such as AWTY (Nylander et al., 2008), only compare the split frequencies of independent runs instead of performing a statistical test for convergence. The problem arising with such a quantitative comparison is that no information of the stochastic properties of the MCMC algorithm is used and no interpretation for repeated runs is given. For example, our requirement 2, that more samples lead to the same estimates cannot be given with any degree of certainty with a quantitative method. A better procedure is to treat each split frequency as an independent parameter which is either 0 (absent) or 1 (present). Then, each split parameter can be analyzed with the same methods as all other parameters. All advantages of the statistically motivated methods apply to the split frequencies too.
Only comparing the split frequencies, as done in AWTY, can be misleading. We will show this with a simple example. We ran two pairs of MCMC with the same data and model, but with different tree proposals (see Figure 2). The first pair produced slow mixing chains and the second pair faster mixing chains. If we only looked at the split frequencies, which were 0.9503 and 0.9318 for the first pair and 0.8921 and 0.9124 for the second pair, we would think that each pair has converged equally well. However, by plotting the traces of the split with states being either present or absent, we can calculate the ESS for each split and obtained 42.02 and 25.84 for the slow mixing chains and 235.23 and 225.4 for the fast mixing chains. Further calculations lead to standard errors of the split frequencies which were approximately 0.025 even for the fast mixing chains. This indicates that there is even in the fast mixing chain a considerably high error and exact estimates should be considered carefully. Nevertheless, using these calculations one can obtain much more precise information about convergence than by only comparing the split frequencies.

Nevertheless, we compared the 10 independent runs using AWTY. Convergence was

![Figure 2: Trace plot of a split in the Cettiidae dataset. If the split is present we label it 1, if the split is absent we label it 0. We run twice two chains, each pair with different MCMC settings. The data and model was the same and thus should be the posterior probabilities. Top: Two slow mixing chains with 0.9503 and 0.9318 as the estimates of the posterior probability of the split. The left chain has an ESS of 42.02 and the right chain an ESS of 25.84. Bottom: Two faster mixing chains with posterior probability estimates of 0.8921 and 0.9124 and ESS’s of 235.23 and 225.4.](image)
rejected based on the large deviation in split frequencies between several runs, which
agrees with the previous results. Furthermore, by looking at the trace plots of the split
parameters we observed problems in mixing. Slow mixing is commonly the cause for
slow convergence, as we believe in this situation.

3 Statistically Motivated Convergence Assessment Meth-
ods
After we looked at convergence assessment methods from the point of view of a phyloge-
neticist, we develop in this section a mathematical framework for assessing convergence
from MCMC output. Convergence assessment methods, often also called convergence
diagnostics, can be divided into two classes: First, a theoretical analysis of the transi-
tion kernel, i.e. the proposal mechanism from one state in the chain to another, and
second, analyzing the output of one or multiple MCMC runs (Cowles and Carlin, 1996).
Though analyzing the theoretical properties of a specific transition kernel helps to de-
sign new and more efficient MCMC algorithms, it is of lesser use when the convergence
of MCMC runs from any available software package needs to be assessed. Therefore we
restrict our attention to methods analyzing the output of MCMC runs. We consider the
MCMC algorithm itself as a “black-box” of which we do not know any details of the
implementation.

In order to be able to develop convergence assessment methods, we need a few basic
mathematical definitions on how the MCMC output was generated. Let us assume that
the software packages, from which we obtained the MCMC output, used the Metropolis-
Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). All other variants of MCMC
algorithm — e.g., the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith,
1990) — can be considered as special cases of the Metropolis-Hastings algorithm and
considering only the Metropolis-Hastings algorithm is sufficient for our purpose of treat-
ing the MCMC algorithm as a “black-box”. Therefore the methods used in this study
can still be used if the software or implementation of the MCMC algorithm has changed.

Following standard notation, let us define the state of the MCMC at iteration \( t \) by \( X_t \). A new state \( X_{t+1} \) is proposed from the proposal distribution \( P(X_{t+1}|X_t) \) and
accepted with probability \( \alpha = \min(1, \frac{P(X_{t+1}) \times P(X_t|X_{t+1})}{P(X_t) \times P(X_{t+1}|X_t)}) \). Depending on the proposal
distribution chosen, we can obtain three types of MCMC output, see Figure 3. Figure
3.a shows a trace of independent samples, Figure 3.b few jumps which are commonly a
sign of too broad proposal distributions and therefore too few accepted proposals, and
Figure 3.c shows a “sticky” trace which results from too narrow proposal distributions
and hence too many accepted proposals. We will use these three types of MCMC output
Let us consider for the remainder of this paper the situation that one is interested in estimating the mean parameter $\mu$ of a posterior distribution. Note that the accuracy of other quantiles can be assessed in a similar fashion. Although the MCMC method produces correlated samples, the posterior mean can be estimated by the common sample mean:

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} X_t$$  \hspace{1cm} (1)

where $n$ is the number of samples from the chain and $X_t$ is the state of the chain at iteration $t$. The sample mean estimator $\hat{\mu}$ will converge to the true population mean by the Law of Large Numbers (Geyer, 1992; Gamerman and Lopes, 2006, p. 125). Applying the Central Limit Theorem and using standard notation, $n$ is the number of samples, $\mu$ the distribution mean, $\sigma$ the standard deviation of the distribution (commonly estimated by the square root of the sample variance), we get

$$\sqrt{\frac{\tau}{n}} \times \frac{\hat{\mu} - \mu}{\sigma} \sim \text{Norm}(0, 1),$$  \hspace{1cm} (2)

where we defined $\tau$ as the autocorrelation time (ACT), see e.g. (Robert, 1998, p. 103)). The ACT and the number of samples directly constitute to the effective sample size (ESS):

$$\text{ESS} = \frac{n}{\tau}. \hspace{1cm} (3)$$

We obtain the standard deviation of the mean estimator $\hat{\mu}$, which is defined by the standard error of the mean (SEM) for correlated samples, by transformation of equation 2:

$$\text{SEM} = \sqrt{\frac{\tau \times \sigma^2}{n}}. \hspace{1cm} (4)$$

Figure 3: Three common traces plots from MCMC runs. Left: Independent samples from the posterior distribution. Middle: Samples from a chain with high correlation because of low acceptance probability. Right: “Sticky” trace due to high correlation and high acceptance probability.
After having defined the necessary quantities ($\hat{\mu}$, ACT, SEM and ESS), we will proceed by estimating the precision of an estimator, then the stability and reproducibility and finally how to automatically remove the burnin phase.

### 3.1 Estimating the precision of the estimator

The precision of an estimator is not only important for knowing how good the estimate is but also for comparing if the difference of two estimates is statistically significant. Because of this importance we will carefully study if and how good the precision of an estimator can be assessed. As we showed in the previous section, the precision of the estimator is defined by the SEM. More samples lead to a smaller SEM and thus to a higher precision. Testing for the precision of the estimator is a straightforward computation, if the SEM is known. Unfortunately, a good estimate of the SEM is only available for uncorrelated samples or if the autocorrelation time is known, a requirement that the MCMC output does not fulfill. Instead, several approximations have been proposed: batch means (Geyer, 1992), fitting a linear regression to the log spectrum (Heidelberger and Welch, 1983), initial sequence estimators (Straatsma et al., 1986) (implemented in Tracer) and the autoregressive process (Hamilton, 1994) (implemented in CODA). In this paper we omit the details of the SEM approximation methods and refer the reader to Thompson (2010) for a review. Thompson (2010) studies the behavior of the different SEM approximations with some artificially constructed traces of which the autocorrelation is known. She concludes, that the autoregressive process is the most accurate approximation and the batch means provide the best trade-off between calculation time and accuracy.

We evaluated the different SEM approximation methods on common MCMC output with unknown autocorrelation. To be able to assess the performance of the SEM approximations we needed repeated simulations with known “true” mean $\mu$ to construct coverage probabilities on how often the true mean fell inside the confidence interval. Therefore, we first constructed MCMC’s with posterior distributions being either the normal distribution, the exponential distribution or the gamma distribution and modified the proposal function so that we obtained the three different types of MCMC traces shown in Figure 3. One could use slightly more relevant distributions for phylogenetics by simulating data under a phylogenetic model and then using the same model for inference. However, simulating data and performing inference under a realistic phylogenetic is very time consuming, error prone and might introduce a bias in the estimates so that the “true” mean does not equal the inferred mean. Furthermore, most posterior distribution resemble some type of standard statistical distribution (e.g. normal, lognormal or gamma distribution) and therefore we chose to simulate our data under these simplistic
distributions. We computed the confidence intervals on a $\alpha = 0.5$ level and assessed the frequency of how often the true mean fell inside the confidence interval. The expected frequency was the confidence level of 50%. A higher frequency indicates too large SEM’s which result into too wide confidence interval and a lower frequency indicates too small SEM’s. An example for the standard normal distribution for different run length of the MCMC is shown in Figure 4. We conclude that the autoregressive function as an SEM approximation slightly outperforms the other three methods, which confirms the conclusions in Thompson (2010).

![Figure 4: Left: Coverage probabilities of the true mean in the 50% confidence interval for different ACT approximation methods. Middle: Coverage probabilities of the true mean in the 50% confidence interval for different convergence assessment methods. Coverage probabilities smaller than 0.5 indicate too many rejected runs and coverage probabilities larger than 0.5 indicate too few rejected runs. Right: Instead of dividing a single run into $k$ batches the Multiple Stationarity test cross-validates the mean of $k$ runs. We started 10 MCMC runs from independent starting values uniformly distributed between -3 and 3 for a standard normal posterior distribution. The MCMC was run from 2000 to 10000 iterations and this was procedure repeated 10000 times. We compared the coverage probability (expected 0.5) for the Multiple Stationarity test to the average potential scale reduction factor.](image)

We summarize this section by pointing out that the SEM can be reliable approximated if the ESS is sufficiently large ($> 20$, data not shown). Furthermore, the SEM can be used to test for convergence to a specified precision threshold in two ways: absolute precision (the so called Monte Carlo Standard Errors Flegal et al. (2008)) and relative precision (Jones et al., 2006). Let us define the absolute precision test as

$$\text{SEM} < k \quad (5)$$

and the relative precision test as

$$\frac{\text{SEM}}{\hat{\mu}} < k \quad (6)$$

where $k$ is the accepted threshold for the precision. Note, finding an appropriate threshold $k$ is problem and parameter specific. A parameter which ranges between 0 and
1 should have a much higher precision than a parameter which has a variance much larger. Nevertheless, thresholds such as the fourth significant digit are often useful in applications (Flegal et al., 2008).

3.2 Estimating the stability of the estimator

We consider an estimator to be stable, if additional samples — i.e. longer runs — result in the same estimate disregarding small variations due to the stochastic error. Samples of an MCMC run will only fulfill the requirement if the run has converged to the stationarity distribution of the Markov chain. However, if the run is stuck in a peak of a local optimum the MCMC output might seem as being at the stationarity distribution too. Due to the unidentifiability of the two cases, one cannot design a test of stationarity that is able to attest convergence if the posterior distribution is not known a priori. Instead, tests identifying non-convergence are the only feasible alternative (Cowles and Carlin, 1996). In this section, we will describe two of the most used stationarity tests from the mathematical literature (the Heidelberger-Welch test and the Geweke test), develop a new test for stationarity, define the convergence test based on the ESS-threshold and evaluate all four tests on simulated data with a known stationarity distribution. The Geweke test, Heidelberger-Welch test and our third test, the “Single Stationarity test” are statistically motivated and test if the run is located at the stationarity distribution and reject if not.

Heidelberger and Welch (1983) proposed a test that calculates the evolution of the mean estimator \( \hat{\mu}(i) \) at a given iteration \( i \) of the chain. Following the notation of Brooks and Roberts (1998b) we denote \( Y_0 = 0, Y_n = \sum_{t=1}^{n} X_t, \bar{X} = \frac{1}{n} Y_n, \hat{S}(0) \) as the estimate of the spectral density, \( \lfloor a \rfloor \) the integer part of \( a \), then

\[
\hat{B}_n(s) = \frac{Y_{\lfloor ns \rfloor} - \lfloor ns \rfloor \bar{X}}{(n\hat{S}(0))^{1/2}} \quad \text{with } 0 \leq s \leq 1
\]  

is distributed as a Brownian Bridge. If \( \hat{B}_n \) indeed follows a Brownian bridge can be tested with the Cramér-von Mises test. The stationarity of the Markov chain is rejected based on the p-value of the Cramér-von Mises test.

Geweke (1992) suggested to compare a first interval \( A \) of the run, say 10%, with a second interval \( B \), say the second half of the run. We denote the number of samples in interval \( A \) and interval \( B \) by \( n_A \) and \( n_B \), and the autocorrelation times of the samples in interval \( A \) and \( B \) by \( \tau_A \) and \( \tau_B \). Under the assumption of stationarity, the normalized difference of the sample means of the two intervals (\( \hat{\mu}_A \) and \( \hat{\mu}_B \)) are standard normally
distributed,

\[ z = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\tau_A}{n_A} + \frac{\tau_B}{n_B}}}. \]  

(8)

The p-value of the standard normal distribution for \( z \) is then used to reject convergence.

Following up on the idea of Geweke (1992), we introduce a new stationarity test, which is denoted simply by “Single-Stationarity test” in the remainder. Let us divide the samples of the MCMC into \( k \) equally sized batches (see Figure 5). Then we cross-compare the different batches and calculate the p-value using equation 8. Because this procedure uses multiple tests, we need to apply a correction for the p-values. We suggest two corrections, first the deviation of the p-values from the uniform distribution, and second, adjust the significance level by the Sidak-correction and reject if:

\[ \max_i \left\{ P_{N(0,1)}\left(\left|\frac{\hat{\mu}_i - \hat{\mu}_{\text{SEM}}}{\text{SEM}}\right|\right) \right\} < 1 - (1 - \alpha)\frac{1}{k}. \]  

(9)

Figure 5: The Single Stationarity test divides the MCMC output into \( k \) equally sized batches (here \( k = 5 \)) and cross-validates if the normalized batch means are standard normally distributed.

Additionally to the three mathematically consistent stationarity tests, we define a “ESS-test” which assesses convergence if an MCMC output reports an ESS > 200. This test is an ad-hoc procedure assessing convergence purely based on if enough samples are available. Nevertheless, the ESS test is directly correlated to the precision tests suggested by Jones et al. (2006) and Flegal et al. (2008).

We evaluated the four stationarity tests following the procedure of the previous section on evaluating the SEM approximations. We performed repeatedly 10000 MCMC runs with a known posterior distribution and studied the frequency of assessing convergence on an \( \alpha = 0.5 \) level. With a 0.5 confidence level we expect that an unbiased test will reject 50% of the chains because just by chance 50% of the runs should have more extreme samples than allowed in the 0.5 confidence interval. All three statistically mo-
tivated tests, the Single-Stationarity test, the Geweke-test, and the Heidelberger-Welch test clearly outperformed the ESS-test, as seen in Figure 4. The Single-Stationarity test and the Geweke test use the same underlying foundation and give unbiased convergence assessments. Both tests provide a simple statistic, the p-value, which is interpreted as the probability of an MCMC producing an output at least as extreme as the observed output, if the chain is at its stationarity distribution. The same is true for the Heidelberger-Welch test. The ESS-test instead is biased in the way that with too few samples or too short runs it will reject all runs and with long runs it will accept all runs. Furthermore, the ESS-test does not take the stochastic error into account and has no interpretation for repeated runs.

3.3 Estimating the reproducibility of the estimator

In the previous section we have shown that failure of convergence to the stationarity distribution can be assessed if the chain is still moving towards the stationarity distribution (e.g., see Figure 3). However, the stationarity tests cannot distinguish if the chain is stuck in a local optimum or if it has reached the stationarity distribution. All methods based on single run analysis have this disadvantage. It can only be overcome using multiple runs from random starting values (Gelman and Rubin, 1992a). Convergence can be assessed with similar tests as the test for stationarity of single runs, for instance instead of breaking the trace into \( k \) batches each run can be considered as a batch and the “Single Stationarity test” applied over the independent runs, which we call “Multiple Stationarity test”. This procedure has the advantage over the potential scale reduction factor by Gelman and Rubin (1992b) that it can be interpreted in a statistical manner, namely with p-values. As an example we run 10 MCMC’s repeatedly 10000 times for 2000 to 10000 iterations with posterior distribution being \( \text{N}(0,1) \) and starting values uniformly drawn between -3 and 3. The results are shown in Figure 4. The PSRF converged quickly to 1.0 and the Multiple Stationarity test to 0.5, which was the defined significance level. Both methods work well in our example, though we believe that the p-value of the Multiple Stationarity test is easier to interpret compared to the PSRF.

3.4 Estimating the burnin

Above we have considered all samples of the MCMC algorithm contributing equally to the estimator \( \hat{\mu} \). However, the Markov chain spends the first \( n_0 \) iteration in the transient phase before having converged to the stationarity distribution. Figure 6 shows a common MCMC run with an initial transient phase, the so-called burnin. Although the mathematical literature gives no clear explanation on why to remove all samples
Figure 6: Evaluation of the burnin estimation methods. Top left: A common MCMC output with a burnin phase until iteration $n_0 \approx 1000$. Right: The ESS-Max function for the MCMC output shown on the left. Bottom left: The coverage probability of the true mean using a 50% confidence interval after removing the burnin suggested by the Geweke test, Heidelberger-Welch test (H-W), Single Stationarity test and the ESS-Max function. Bottom right: The average estimates burnin $n_0$. 
until iteration $n_0$ and heavily debates if it should be done, it is common practice to do so (see e.g. Brooks and Roberts (1998a) and Cowles et al. (1999)). We advocate to remove the burnin phase because it gives higher precisions of the estimator with shorter runs. The common practice in phylogenetics varies between removing 10%, 25%, 50% or a manually specified fraction of the samples. Instead of giving a specific value for all runs, which expectedly will not be the optimal choice for most applications, we provide in the following an algorithm that finds automatically the optimal burnin:

1. Obtain an MCMC output and set the burnin $n_0 = 0$.
2. Remove $n_0$ samples from the beginning of the run.
3. Test if the run has converged to its stationarity distribution.
4. If the test was successful, use $n_0$ as the burnin; otherwise proceed.
5. Add $b$ to $n_0$.
6. If $n_0$ has not exceeded a pre-specified limit (e.g. 50% or 75% of the total run length), go back to step 2.

We used $b = 10$ and the Heidelberger-Welch test, the Geweke test and the Single Stationarity test as test criteria. Additionally, we define the function “ESS-Max” by

$$ n_0 = \max_i (ESS(X\{X_0, ..., X_i\})) $$

where $ESS(X\{X_0, ..., X_i\})$ is the effective sample size for the MCMC output $X$ after removing the first $i$ samples.

We aim to identify which of the burnin estimation methods is unbiased with respect to the stationarity distribution and has the highest power by finding the shortest burnin phase. Therefore we designed an MCMC with posterior distribution being N(0,1), a starting value far outside the posterior mode ($X_0 = 20$) and a narrow proposal distribution which leads to slow mixing. We run this chain for 10000 repetitions over 2000 to 10000 iterations. For each run we estimated the convergence under the four methods: three using iterative removal of samples and testing for convergence (Heidelberger-Welch test, Geweke test and Single Stationarity test) and the ESS-Max function. After the burnin has been found we assessed convergence if the known true mean fell into the 50% confidence interval of the sample mean.

The Heidelberger-Welch test underestimated the burnin phase (see Figure 6). The Geweke-test and the Stationarity-test performed better but were also biased towards slightly too short burnin lengths. The too short burnin phase results into biased sample
means and therefore in too low coverage probabilities. The ESS-Max instead is not influenced by the number of samples and always give the best estimate for the burnin phase.

4 Discussion

4.1 Results on the Cettiidae MCMC output

In section Convergence Assessment Methods currently used in Bayesian Phylogenetic Inference we analyzed the output of 10 runs of the Cettiidae dataset using methods other research have used in Bayesian phylogenetic inference. In section Statistically Motivated Convergence Assessment Methods we discussed various methods from the statistical literature for assessing convergence. In this section we will apply the methods from the statistical literature and compare the results to our previous analysis. First, we estimated the optimal burnin by the ESS-Max function and removed the burnin phase from all runs. The optimal burnin, see Table 4 and 5, is very similar to the manually set convergence from Figure 1. Next, we calculated the ESS and assessed convergence using the Heidelberger-Welch test, Geweke test and Single Stationarity test all with a confidence level of $\alpha = 0.01$. Most of the runs passed at least the stationarity test. The Heidelberger-Welch test and the Single Stationarity test reported failure of convergence for run 1, which we did not detect by manual or visual inspection. Run 6 was rejected only by the Single Stationarity test, however, the estimated burnin was very high and thus the ESS was considerably low. Finally, combined convergence of all chains to the same posterior estimates was rejected by both the PSRF and our Multiple Stationarity test.

4.2 Interpretation of p-values

The p-values we introduced to the phylogenetic literature with this paper for rejecting convergence have a simple frequentist interpretation. The p-value gives the probability of observing an MCMC output deviating as much as the present samples from random, though correlated, samples under the condition of the run being at the stationarity distribution. A low p-value indicates extreme fluctuation in the trace which is either due to chance or not being at stationarity yet. We recommend a relatively high significance level, e.g., $\alpha = 0.05$, for rejecting convergence and being conservative. On the other hand, high p-values do not guarantee convergence. The tests should be understood as hypothesis tests where the Null hypothesis is that the chain is at its stationarity distribution.
Table 4: Convergence results of the 10 MCMC runs on the Cettiidae dataset considering the posterior traces.

<table>
<thead>
<tr>
<th>Run</th>
<th>Burnin</th>
<th>ESS</th>
<th>H-W</th>
<th>Geweke</th>
<th>S-Stationarity</th>
<th>PSRF</th>
<th>M-Stationarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1420</td>
<td>555.2578</td>
<td>failed</td>
<td>passed</td>
<td>failed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>330</td>
<td>349.0286</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5770</td>
<td>309.2341</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1750</td>
<td>350.0877</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>370</td>
<td>745.9279</td>
<td>failed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8310</td>
<td>187.7583</td>
<td>passed</td>
<td>failed</td>
<td></td>
<td>4.5229</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>560</td>
<td>642.2175</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3600</td>
<td>439.0881</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>370</td>
<td>490.2233</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1680</td>
<td>651.9614</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Convergence results of the 10 MCMC runs on the Cettiidae dataset considering the root-height traces.

<table>
<thead>
<tr>
<th>Run</th>
<th>Burnin</th>
<th>ESS</th>
<th>H-W</th>
<th>Geweke</th>
<th>S-Stationarity</th>
<th>PSRF</th>
<th>M-Stationarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>210</td>
<td>240.5276</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>650</td>
<td>263.1986</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6090</td>
<td>415.5855</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1570</td>
<td>1332.1170</td>
<td>passed</td>
<td>failed</td>
<td></td>
<td>8.0778</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1520</td>
<td>1206.4580</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4850</td>
<td>360.3083</td>
<td>failed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>260</td>
<td>1076.2700</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2480</td>
<td>555.3910</td>
<td>passed</td>
<td>passed</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>810</td>
<td>316.0205</td>
<td>passed</td>
<td>passes</td>
<td>passed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1700</td>
<td>848.4104</td>
<td>passed</td>
<td>failed</td>
<td>failed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.3 Causes of failure to converge

True failure to converge, that is not due to the stochastic error, is caused by inefficient implementations of the MCMC algorithm for the specific dataset and model. A more thorough investigation for the Cettiidae dataset of all parameters provided insight into very slow mixing of the tree topologies which caused the failure to converge. If all parameters of the model are available, the slowest mixing parameter can be identified. Only if all parameters have converged to their stationarity distribution can the estimates for any parameter be trusted. Low ESS despite passing stationarity test for some less important parameters do not hinder the correctness of the estimates of the parameters of interest.

We want to point out, that convergence assessment methods and therefore also this paper are not focused on detecting the cause for the non-convergence. The cause is specific to the software used for generating the MCMC output and therefore cannot be assessed independently. Nevertheless, the convergence assessment methods can indicate which parameter mixed slowly by reporting failed test of stationarity, high uncertainty in the estimates, low ESS and long burnin phases.

4.4 How convergence should be assessed in practice

Our results show that an optimal burnin can be found using the ESS-Max function. This burnin can be automatically computed and maximizes the information in the available MCMC output. The burnin should not exceed 50% of the chain length. A large burnin introduces unreliable convergence assessments and longer runs should be preferred. Next the precision of the estimator should be assessed, either by the relative error using Equation 6, the absolute error using Equation 5, the ESS or a combination of the three. A low precision or low ESS does not indicate non-converge but shows high uncertainty in the approximated estimator and more samples should overcome the low precision. Then, convergence of single parameters can be rejected with methods such as the Single Stationarity test, the Geweke test and the Heidelberger-Welch test. Applying several tests is commonly recommended (Cowles and Carlin, 1996). After the single chains have passed a precision test and a stationarity test, there is still a high probability that the runs got stuck in different parts of the parameter space, as reported by our analysis. Convergence to the same estimates can be rejected using the Multiple Stationarity test and p-values. If convergence was rejected by one of the tests, more or longer runs are needed to obtain better approximates.
4.5 Conclusions

In this paper we showed that there is a great need of convergence assessment methods, not only in phylogenetics (Brooks and Roberts, 1998b). We compared commonly applied methods with more sophisticated methods from the mathematical literature. The results show that by manual or visual analysis one can quickly overlook non-convergence. In the same way one can be overconfident of convergence if only the split frequencies are compared instead of assessing the uncertainty in the split frequency estimate. Furthermore, we show that the ESS should not be used as a stand-alone convergence assessment method. Instead we suggest using the Single Stationarity test to reject if the chain has converged to the stationary distribution. Most importantly, with the complexity of the used models multiple runs are unavoidable for rejecting convergence. Our Multiple Stationarity test provides a simple method for rejecting convergence of multiple runs to the same posterior distribution based on standard p-values. Additionally, we introduce an algorithm and a function, the ESS-Max, which can automatically find the optimal burnin for an MCMC output. The methods described in this paper are implemented in the open source software package RevBayes (http://www.revbayes.net/).

5 Acknowledgements

The authors would like to thank Tom Britton and Fredrik Ronquist for commenting on the manuscript.

References


