Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty

MBE Advance Access published March 7, 2012

Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty

Guy Baele,[1] *Philippe Lemey,*[1] *Trevor Bedford,*[2] *Andrew Rambaut,*[2] *Marc A. Suchard,*[3,4,5] *and Alexander V. Alekseyenko*[6]

[1]Department of Microbiology and Immunology, Katholieke Universiteit Leuven, Leuven, Belgium
[2]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom
[3]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA
[4]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA
[5]Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095, USA
[6]Department of Medicine, Center for Health Informatics and Bioinformatics, New York University School of Medicine, NY 10016, USA

Recent developments in marginal likelihood estimation for model selection in the field of Bayesian phylogenetics and molecular evolution have emphasized the poor performance of the harmonic mean estimator (HME). Although these studies have shown the merits of new approaches applied to standard normally distributed examples and small real-world data sets, not much is currently known concerning the performance and computational issues of these methods when fitting complex evolutionary and population genetic models to empirical real-world data sets. Further, these approaches have not yet seen widespread application in the field, due to the lack of implementations of these computationally demanding techniques in commonly-used phylogenetic packages. We here investigate the performance of some of these new marginal likelihood estimators, specifically, path sampling and stepping-stone sampling for comparing models of demographic change and relaxed molecular clocks, using synthetic data and real-world examples for which unexpected inferences were made using the HME. Given the drastically increased computational demands of path sampling and stepping-stone sampling, we also investigate a posterior simulation-based analogue of Akaike's information criterion (AICM) through Markov chain Monte Carlo (MCMC), a model comparison approach which shares with the HME the appealing feature of having a low computational overhead over the original MCMC analysis. We confirm that the HME systematically overestimates the marginal likelihood and fails to yield reliable model classification and show that the AICM performs better and may be a useful initial evaluation of model choice but that it is also, to a lesser degree, unreliable. We show that path sampling and stepping-stone sampling substantially outperform these estimators and adjust the conclusions made concerning previous analyses for the three real-world data sets that we reanalyzed. The methods used in this paper are now available in BEAST, a powerful user-friendly software package to perform Bayesian evolutionary analyses.

Introduction

Bayesian inference has become increasingly popular in molecular phylogenetics over the past decades, with Markov chain Monte Carlo (MCMC) integration revolutionizing the field (Yang and Rannala 1997). While MCMC has provided the opportunity to infer posterior distributions under complex phylogenetic models, the computational demands associated with increasing model complexity and the amount of data available has considerably hampered assessing the performance of such models. Comparing alternative models according to objective criteria in a formal model selection procedure is becoming an essential approach to phylogenetic hypothesis testing (Suchard, Weiss and Sinsheimer 2001; Huelsenbeck et al. 2001). Here, the aim of model selection is not necessarily to find the true model that generated the data, but to select a model that best balances simplicity with flexibility and biological realism in capturing the key features of the data (Steel 2005).

A standard approach to perform model selection in a Bayesian phylogenetic framework operates through the evaluation of Bayes factors (Sinsheimer, Lake and Little 1996; Suchard, Weiss and Sinsheimer 2001). The Bayes factor is a ratio of two marginal likelihoods (i.e. two normalizing constants of the form $p(Y \mid M)$, with Y the observed data and M an evolutionary model under evaluation) obtained for the two models, M_0 and M_1, under comparison (Jeffreys 1935):

$$B_{10} = \frac{p(Y \mid M_1)}{p(Y \mid M_0)}. \quad (1)$$

In order to evaluate model fit and calculate Bayes factors, the normalization constant or marginal likelihood $p(Y \mid M)$, which measures the average fit of a model to the data, is of primary importance. Calculation of the marginal likelihood of model M requires integration of its likelihood across parameter values, weighted by the model's prior distribution

$$p(Y \mid M) = \int_{\theta \in \Theta} p(Y \mid \theta, M)\, p(\theta \mid M)\, \mathrm{d}\theta. \quad (2)$$

Among several models, one is led to choose the one of greatest marginal likelihood. The Bayes factor offers advantages over likelihood-ratio-tests comparing nested models in which one garners evidence only in favor of rejecting less complex models. Instead, the Bayes factor evaluates the relative merits of both competing models. Consequentially, models need not be nested and the marginal likelihood naturally penalizes for model complexity. Values of the Bayes factor greater than 1 are considered as evidence in favor of M_1. Given that modeling assumptions may have orders-of-magnitude effects on model fit, the log Bayes factor is often calculated. Kass and Raftery (1995) introduce different gradations to assess the log Bayes factor as evidence against M_0. A value between 0 and 1 is not

Key words: model comparison, marginal likelihood, Bayes factors, path sampling, stepping-stone sampling, demographic models, molecular clock, Bayesian inference, phylogeny, BEAST.

E-mail: guy.baele@rega.kuleuven.be

Mol. Biol. Evol. X(Y):1–9. XXXX
doi:10.1093/molbev/msl161
Advance Access publication XXXX

worth more than a bare mention, whereas a value between 1 and 3 is considered to give positive evidence against M_0. Values larger than 3 and 5 are considered to respectively give strong and very strong evidence against M_0.

Although researchers have proposed several useful methods to evaluate Bayes factors in phylogenetics, they are often limited to specific model selection situations (Lartillot and Philippe 2006). For example, Suchard, Weiss and Sinsheimer (2001) develop the Savage-Dickey ratio (Verdinelli and Wasserman 1995) as a Bayes factor estimator for nested evolutionary models in phylogenetics. Additional approaches include reversible jump MCMC to evaluate the relative merits of tree topologies (Suchard, Weiss and Sinsheimer 2005) and nesting alternative models together into a single mixture model via model averaging in phylogenetics (Lemey et al. 2009; Li and Drummond 2011). Outside of phylogenetics, one often employs approximations to the Bayes factor, such as Bayesian information criterion (Schwartz 1978) and Laplace estimators (Kass and Raftery 1995). However, these approximations often make large sample assumptions that are rarely valid in phylogenetics and break down when considering the discrete nature of tree topologies.

Among the few methods of potentially general applicability, phylogenetics has readily adopted (i) importance sampling (IS) estimators (Newton and Raftery 1994) and (ii) path sampling (PS) estimators (Ogata 1989; Gelman and Meng 1998) to compute marginal likelihoods of competing models. Occasionally, phylogeneticists refer to PS as 'thermodynamic integration' (Lartillot and Philippe 2006) in deference to the physics over statistics literature. PS methods represent very general estimators; they can be applied to any model for which MCMC samples can be obtained. These approaches allow for an overall ranking of competing models to be constructed, from which the top-performing model can easily be determined.

Lartillot and Philippe (2006) discuss and evaluate several approaches to calculate marginal likelihoods and Bayes factors in the context of phylogenetics. They examine three variants of IS, the prior arithmetic mean estimator, the posterior harmonic mean estimator (HME), the stabilized HME, and PS. Of these approaches, the HME (Newton and Raftery 1994) is by far the simplest method, only requiring samples from the posterior distribution, and has been used extensively in the field of phylogenetics (see e.g. Nylander et al. (2004)). The HME is often severly biased, overestimating the true marginal likelihood (Xie et al. 2011). Because HME estimator variance may be infinite, a modified, stabilized version has been proposed (Newton and Raftery 1994) with extensions to quantify its Monte Carlo error in phylogenetics (Redelings and Suchard 2005). Lartillot and Philippe (2006) compare the various approaches using a Gaussian model with different dimensions and an evolutionary model on a fixed tree for which exact calculation of the marginal likelihood is available. Results indicate that PS outperforms the IS variants across all scenarios, remaining well-behaved in cases with high dimensions where all three IS methods fail, even when using a huge numbers of costly posterior samples.

Recently, Xie et al. (2011) introduced a new method, called stepping-stone sampling (SS) that employs ideas from both IS and PS to estimate the marginal likelihood in a series (the stepping stones) that bridges the posterior and prior distribution of a model. Again using a Gaussian model example, the authors show that SS yields a substantially less biased estimator than PS. Further, for realistic phylogenetic models, SS importantly requires significantly fewer path steps than PS to accurately estimate the marginal likelihood with acceptably small discretization bias.

Because PS and SS offer increased model selection accuracy, in particular relative to the HME, Bayesian inference software that incorporates an array of evolutionary models would greatly benefit from the implementation of these methods. BEAST (Drummond et al. 2012) is a cross-platform program for Bayesian MCMC analysis of molecular sequences that offers a multitude of different models, such as autocorrelated and uncorrelated relaxed clock models, substitution models including heterogeneity across sites, coalescent models of population size and growth and phylogeographic models, with support for a flexible choice of prior specifications on model parameters. BEAST presents a flexible framework for testing evolutionary hypotheses without conditioning on a single tree topology. However, the rich choice in models has not been matched by state-of-the-art methods for calculating marginal likelihoods; only the HME is readily available when integrating over the uncertainty in the phylogenetic tree.

Here, we implement PS and SS approaches to test models while accommodating phylogenetic uncertainty in BEAST. We also implement a posterior simulation-based analogue of Akaike's information criterion through MCMC (AICM) (Raftery et al. 2007), which is computationally efficient as it only requires samples from the posterior, and compare the performance of PS, SS and AICM to that of the HME. Using a simulation study, we show that PS and SS consistently outperform the AICM and HME, and that the AICM outperforms the HME in four out of five simulation scenarios, when performing demographic model selection. Our results for demographic and molecular clock selection on empirical data sets indicate that PS and SS yield the most consistent results across two runs with different starting values and systematically yield more realistic model classifications. Further, the AICM yields more consistent results across runs than the HME, but like the HME, fails to consistently select the appropriate model.

Methods

Path sampling and stepping-stone sampling in BEAST

Most implementations of PS rely on drawing MCMC samples from a series of distributions, each of which is a power posterior differing only in its power, along the path going from the prior to the unnormalized posterior defined by the model M. Both Lartillot and Philippe (2006) and Xie et al. (2011) define this path to be:

$$q_\beta(\theta) = p(Y \mid \theta, M)^\beta p(\theta \mid M), \tag{3}$$

where $p(Y \mid \theta, M)$ is the likelihood function and $p(\theta \mid M)$ the prior). Hence, the power posterior is equivalent to the posterior distribution when $\beta = 1.0$ and is equivalent to the prior distribution when $\beta = 0.0$.

Lartillot and Philippe (2006) propose to evenly spread the different values of that power β between 0.0 to 1.0 and use Simpson's triangulation method to derive an expression for the marginal likelihood. The authors propose to collect one sample from each power posterior, before β is updated. Assuming $K+1$ path steps, this yields a collection of samples $(\beta_k, \theta_k)_{k=0\ldots K}$, with $\beta_0 = 0$ and $\beta_K = 1$, which are used to calculate the estimate for the marginal likelihood:

$$\ln p(Y \mid M) = \frac{1}{2K} \sum_{k=0}^{K-1} (\ln p(Y \mid \theta_k, M) + \ln p(Y \mid \theta_{k+1}, M)). \tag{4}$$

In our implementation of path sampling in BEAST (Drummond et al. 2012), we have however chosen to use multiple samples per β, requiring a small adaptation of equation 4 in that each loglikelihood is replaced by the mean loglikelihood of the samples taken at each β.

Lepage et al. (2007) advocate for the use of a sigmoidal function that places most power values near the extremes of the unit interval in their model-switch PS analysis and Friel and Petitt (2008) use equally spaced points in the interval [0,1] elevated to the fourth or fifth power. Hence, the approaches of Lepage et al. (2007) and Friel and Petitt (2008) both place most of the power values at points where the power posterior is changing rapidly. Xie et al. (2011) find that the efficiency of PS could dramatically improve by choosing β values according to evenly spaced quantiles of a Beta$(\alpha, 1.0)$ distribution rather than spacing β values evenly from 0.0 to 1.0; this is a generalization of the approach by Friel and Petitt (2008).

Xie et al. (2011) propose to calculate the marginal likelihood using n samples from a series of $K+1$ power posteriors as follows

$$p(Y \mid M) = \prod_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} p(Y \mid \theta_i, M)^{\beta_k - \beta_{k-1}}. \tag{5}$$

The authors show that numerical stability can be improved by factoring out the largest sampled likelihood for each power posterior. While the estimator for the marginal likelihood shown in equation 5 is unbiased, a bias is introduced by transforming to the log scale, which can be alleviated by increasing K.

Xie et al. (2011) show that a value of $\alpha = 0.3$ is close to optimal for their Gaussian model example, suggesting that values close to 0.3 are perhaps generally optimal. The choice $\alpha = 0.3$ results in half of the β values evaluated being less than 0.1. The authors state that the positive skewness of this distribution is useful because (with sufficient and informative data) the likelihood only begins losing control over the power posterior for β values near 0, and at that point the target distribution changes rapidly from something resembling the posterior to something resembling the prior. Conditioning on the total number of β values evaluated, placing most of the computational effort on β values near zero results in increased accuracy. In BEAST, we provide these different possibilities for spreading the power values. However, in the results of this paper, we follow the Xie et al. (2011) recommendation.

Estimation of HME and AICM

The harmonic mean estimate of the marginal likelihood only requires samples from the posterior, i.e. for $\beta = 1$ in equation 3, and can hence be calculated from an MCMC sample that is obtained by a standard Bayesian phylogenetic analyses under a particular model. If one collects n samples from the posterior, the HME is estimated as follows

$$p(Y \mid M) = \frac{n}{\sum_{i=1}^{n} \frac{1}{p(Y \mid \theta_i, M)}}. \tag{6}$$

Raftery et al. (2007) introduce the AICM as a posterior-simulation based analog of the AIC model selection criterion. AICM has the advantage that, like the harmonic mean estimator of marginal likelihood, one may estimate the AICM directly from posterior samples generated by MCMC with little additional work. Raftery et al. (2007) show that asymptotically with large amounts of data, the posterior distribution of a model's log likelihood ℓ follows

$$\ell_{\max} - \ell \sim \text{Gamma}(\gamma, 1), \tag{7}$$

where $\ell_{\max}$ represents the maximum possible log likelihood, $\gamma = k/2$ and k represents the effective number of parameters in the model. The density function of a Gamma$(\gamma, 1)$ distribution is

$$f(x) = \frac{x^{\gamma-1}\, e^{-x}}{\Gamma(\gamma)},$$

and thus the density function of the log likelihood becomes

$$f(\ell) = \frac{e^{\ell - \ell_{\max}}\, (\ell_{\max} - \ell)^{\gamma-1}}{\Gamma(\gamma)}. \tag{8}$$

Alternatively, the posterior distribution of log likelihoods may be described in terms of a deviance $\mathscr{D} = -2\ell$, such that the posterior deviance is distributed according to a shifted chi-squared distribution

$$\mathscr{D} - \mathscr{D}_{\min} \sim \chi^2(2\gamma),$$

with density function

$$f(\mathscr{D}) = \frac{2^{-\gamma} e^{(\mathscr{D}_{\min} - \mathscr{D})/2}\, (\mathscr{D} - \mathscr{D}_{\min})^{\gamma-1}}{\Gamma(\gamma)}. \tag{9}$$

Equation 7 suggests a method-of-moments estimate of γ as $\hat{\gamma} = s_\ell^2$ and $\hat{\ell}_{\max} = \bar{\ell} + s_\ell^2$, where $\bar{\ell}$ and s_ℓ^2 are the sample mean and variance of the posterior log likelihoods (Raftery et al. 2007). Thus, an estimate of the effective number of parameters k equals $2s_\ell^2$.

AIC (Akaike 1973) is commonly used for model comparison in a maximum-likelihood context, and is defined as

$$\text{AIC} = 2k - 2\ell_{\max}.$$

Models with lower values of AIC are preferred over models with higher values. An increase in the number of parameters k penalizes more complex models. Here, we follow Raftery et al. (2007) in estimating AICM as

$$\begin{aligned} \text{AICM} &= 2\hat{k} - 2\hat{\ell}_{\max} \\ &= 2(2s_\ell^2) - 2(\bar{\ell} + s_\ell^2) \\ &= 2s_\ell^2 - 2\bar{\ell}, \end{aligned} \tag{10}$$

a function of just the posterior sample mean and variance of the log likelihood. The AICM is similar in spirit to the deviance information criterion (Gelman et al. 2004).

In addition to the simple method-of-moments estimator of AICM (10), we consider estimating AICM by fitting the sampled log likelihood values to the their asymptotic density function (8) via maximum likelihood to estimate to $\hat{\ell}_{\max}$ and $\hat{\gamma}$. However, this procedure does not result in a marked improvement over the moment estimator, while suffering from a much higher computational burden. Consequently, throughout the manuscript, we use (10) as our estimate of AICM.

Performance analysis through simulation

To assess the performance of the HME, AICM, PS and SS in population genetics model comparison in which it becomes necessary to integrate over all possible trees, we perform a simulation study inspired by the coalescent analysis of Worobey et al. (2008), see below for details. We consider the sampling dates of 60 sequences that represent the diversity in the original HIV-1 group M data set and simulate dated-tip genealogies under two simple demographic models: a constant population size and an exponentially growing population size through time. The simulations under the exponential growth model include increasing growth rates: 0.01, 0.025, 0.05 and 0.10 per year, respectively. We simulate 100 genealogies under each scenario. Because variance in coalescent simulations yields much wider TMRCA distributions than the empirically observed TMRCA posterior distribution for HIV-1 group M (Worobey et al. 2008), we rescale the resulting trees by drawing the TMRCA from a normal distribution with mean 1910 and standard deviation 10.

Along each genealogy, we simulate sequences encompassing 1000 sites using GTR parameter values and a substitution rate that reflects the estimates for the real data. For each simulated data set under each demographic model, we estimate marginal likelihoods using the HME, AICM, PS and SS of both the constant population and exponential population model. For all marginal likelihood estimators, 10^8 MCMC iterations were run in BEAST, with each estimator taking no more than 3 days to complete.

Results

HIV epidemic history

We revisit a Bayesian evolutionary reconstruction of the HIV-1 group M epidemic history originally performed by Worobey et al. (2008). This study examines sequence data from a 1960 specimen from the Belgian Congo (now Kinshasa, Democratic Republic of the Congo) that show considerable divergence from the 1959 (ZR59) sequence (Zhu et al. 1998), the oldest and only known sequence sampled before 1976 at that time. Because sequences predating the recognition of AIDS are critical to defining the time of origin and the timescale of virus evolution, the authors include these in a relaxed molecular clock analysis and estimated an origin of group M near the beginning of the twentieth century (Worobey et al. 2008).

[FIG. 1 about here.]

Worobey et al. (2008) consider several different coalescent models that serve to provide a prior distribution for time-measured trees and offer a glimpse into the population dynamics of the epidemic. These models include the constant population size, exponential growth (assuming a constant growth rate through time), expansion growth (assuming an increasing growth rate through time), logistic growth (assuming a decreasing growth rate through time) and the Bayesian skyline plot demographic model (a general, non-parametric prior that enforces no particular demographic history; Drummond et al. (2006a)). The authors show that the inclusion of the 1959 and 1960 sequences seemed to improve estimation of the TMRCA of the M group, limiting the influence of the coalescent tree prior on the posterior TMRCA distributions compared with the data set that excluded these earliest cases of HIV-1. However, scientific interest also lies in characterizing through model comparison changes in the population dynamics captured by the different coalescent models rather than the direct ancestors of the sampled sequences. From the Worobey et al. (2008) paper, the HME suggests that a constant population size model provided the best fit to the data. This appears to be at odds with a model for population expansion and the Bayesian skyline plot reconstruction that suggest a more complex (and biologically plausible) demographic history of increasing HIV population size through time. The authors state that the inability to reject the constant population size model is counterintuitive because it is clear that the HIV-1 population size has increased notably and speculate that this finding might be due to the simplest model providing a good fit to a relatively short, information-poor alignment, in comparison to more parameter-rich models.

We reanalyze this HIV-1 dataset by performing two independent fittings to each possible prior model and apply the HME, AICM, PS and SS to perform model selection. Figure 1 shows the log marginal likelihoods for each model using each estimator (see Table S1 in the Supplementary Material for the actual values). Depending on which independent fitting we examine, the constant population model is either the best or the worst model according to the HME, highlighting the poor reliability of this approach. Indeed, the poor repeatability of the HME relative to PS and SS has been demonstrated before (Fan et al. 2011). Moreover, the marginal likelihoods of all five demographic models lie within a 6 and 9 log unit range, for the first and second fitting respectively. This indicates that the overall difference between the five models according to the HME is quite small, making it difficult to reliably select an appropriate demographic model. This range increases to respectively 51 and 25 log units for the AICM, indicating that this approach too suffers from poor repeatability for this data set, even though the overall ranking of the models stays the same. The AICM prefers a constant population size in both runs, which has been stated to be counterintuitive (Worobey et al. 2008). Using PS and SS, however, a drastically different situation emerges. For both fittings, the Bayesian skyline model outperforms all other models considered, whereas the constant population model performs considerably worse compared to the other demographic models. We refer to the original publication (Worobey et al. 2008) for a graphical representation of the Bayesian sky-

line model for the HIV-1 group M. This suggests that the constant population model was not originally preferred because of an information-poor alignment (Worobey et al. 2008), but because the HME fails to provide an adequate classification of the demographic models.

Marginal likelihood estimator performance

Although PS and SS arrive at a biologically more plausible outcome for HIV population size change through time, it remains difficult to ascertain that these estimators select a model closer to the truth for real-life data sets compared to the HME and the AICM. To address performance more formally, we next present a series of simulations to test the ability of marginal likelihood estimators and the AICM to correctly identify the underlying demographic model in cases where the true model is known. The simulations include constant population size and exponential growth dynamics with increasing growth rates and were modeled after the real data set (see Methods).

[Table 1 about here.]

When we simulate data under a constant population size coalescent process (Table 1), the HME is unable to distinguish between a constant population model and an exponential growth rate model, performing no better than a fair coin toss. This is also reflected in the average log Bayes factor across all 100 replicates (roughly centered around 0), indicating that on average the HME considers these two models to perform equally well. Here, the AICM outperforms the HME, correctly classifying 60 simulation replicates and yielding a positive overall difference in AICM of 0.57, in favor of the constant population model. PS and SS outperform both the HME and AICM, correctly classifying 72 out of 100 simulation replicates, yielding an average log Bayes factor of 1.76 in favor of the constant population model. This average log Bayes factor can be interpreted as the average penalty that the exponential growth rate model receives for including one additional parameter.

A simulation scenario close to the constant population size model is that of an exponentially increasing population size with a very low growth rate, 0.01 in our simulation study. In this scenario, the HME fails again to outperform a fair coin toss, again yielding an average log Bayes factor close to 0. The AICM performs slightly worse in this case and only correctly selects the exponential growth model in 45 cases, as reflected in an average AICM difference that is slightly positive. The difficulty to distinguish between an exponential population growth model with a very small growth rate and the constant population model is also shared by PS and SS, although they classify 57 out of 100 simulation replicates correctly and yield a relatively low average log Bayes factor of 0.81 in favor of the exponential growth rate model. Increasing the growth rate to 0.025 reveals that, while the performance of the HME only increases slightly (a correct classification for 59 out of 100 simulation replicate), the performance of PS and SS increases drastically to a proportion of 0.92 correct decisions. The performance of the AICM lies in between that of the HME and PS/SS, with the average difference in AICM returning a negative value for this growth rate.

Further increasing the growth rate in the simulations yields perfect performance for both PS and SS, while the AICM performs almost equally well. Although the HME performance also improves for growth rates of 0.05 and higher, it only attains a proportion of 0.80 correct classifications and the average log Bayes factor increases only slightly. With an increasing growth rate, the AICM furnishes significantly better performance than the HME and achieves perfect performance at a growth rate of 0.10. We can therefore conclude that both PS and SS significantly outperform the HME. While the AICM's performance lies in between that of the HME and PS/SS in cases where it remains difficult to distinguish between the models, AICM performs well in the face of modest to strong evidence.

In the simulation results above we use log Bayes Factor of 0 as cut off for binary classification of models. To assess the discriminatory power of the HME, AICM and PS/SS across a range of cut-offs, we plot the true positive rate as a function of the false positive rate in Figure 2. These receiver operating characteristic (ROC) curves evaluate BF distributions that compare the fit of both coalescent models on data simulated under constant population size and a particular growth rate. In every comparison, PS (and SS) exhibits a stronger discriminatory behaviour than the AICM and the HME. Hence, no matter the cutoff used when performing model comparison, PS (and SS) consistently outperforms AICM and HME. The AICM outperforms the HME in most cases and presents therefore a better alternative for the HME to get a first glimpse of the outcome of a model selection approach while maintaining computational efficiency. However, the best performing methods are clearly PS and SS, justifying the increased computational demands of these methods.

[FIG. 2 about here.]

DNA virus evolutionary rates

Firth et al. (2010) explore the use of temporally structured sequence data within a Bayesian framework to estimate the evolutionary rates for seven human double-stranded DNA (dsDNA) viruses. The authors set out to examine the ability of current inference tools to estimate relatively low evolutionary rates such as those thought to commonly characterize dsDNA viruses (Duffy, Shackleton and Holmes 2008). Of the data sets the authors analyze, we here focus on the herpes simplex virus-1 data set. Herpes viruses are large dsDNA viruses, with genomes that range from 125 to 240 kbp, that infect both vertebrates and invertebrates. Firth et al. (2010) report that the Bayesian skyline plot model outperforms the constant population size model for this data set, irrespective of whether a strict clock or an uncorrelated relaxed clock is assumed, a plausible result that we will not further discuss here. The analyses of Firth et al. (2010) also show that the performance of the strict clock (SC) is virtually identical to that of the uncorrelated relaxed clock lognormal distribution (UCLD), with both being outperformed by the uncorrelated relaxed clock exponential distribution (UCED). It remains however unclear why a more restrictive exponential function would

provide a better underlying distribution to model rate variation among lineages compared to a lognormal distribution.

Aside from using a strict molecular clock (SC), which is often deemed unrealistic due to rate variation among lineages, we have used uncorrelated relaxed clocks, for which we assume two underlying distributions: the exponential rate distribution (UCED) and lognormal rate distribution (UCLD) (Drummond et al. 2006b). Offering an alternative to the autocorrelated relaxed-clock models, these clock models assume a priori no correlation of the rates on adjacent branches of the tree. Instead, the rate on each branch of the tree is drawn independently and identically from an underlying rate distribution. We reanalyze the HSV-1 dataset to compare these models using different marginal likelihood estimators as well as using AICM. We also compare the strict and relaxed clock models in the presence and absence of the sampling dates to test for 'temporal signal'. This provides the Bayesian alternative to the likelihood ratio test that conditions on a single tree topology to test whether including the sampling dates in a dated-tip model significantly improves the fit of the clock models (Rambaut 2000; Suchard, Weiss and Sinsheimer 2003). For consistency, we perform the same number of MCMC iterations as in the original study.

[FIG. 3 about here.]

Figure 3 demonstrates that the HME, once again, returns inconsistent model rankings across independent fittings (see Table S2 in the Supplementary Material for the actual values). In the first fitting, a UCLD clock with sampling dates is the top-performer via the HME, while this model combination is the worst-performer in the second fitting. The AICM estimates are very consistent across both runs for all the models compared and seem to show that the UCED is clearly the worst-performing model, both with and without sampling dates used. PS and SS also yield consistent results across both fittings, significantly preferring the SC and UCLD over the UCED when sampling dates are used. The difference between SC and UCLD in this case is too small to conclude significance, which may not be surprising as the temporal signal might be insufficient to inform a relaxed clock model in this case. However, when the sampling dates are not used, PS and SS indicate that the SC is by far the worst-performing model which is picked up by neither the HME nor the AICM. The temporal signal appears to be significant however, because - except for the UCLD HME fit in one run - incorporating the sampling dates consistently provides a better clock fit to the HSV data.

Spread of methicillin-resistant *Staphylococcus aureus*

Staphylococcus aureus is a common cause of infections that has undergone rapid global spread over recent decades. Gray et al. (2011) are the first to apply formal phylogeographic methods to study the molecular epidemiology of bacterial pathogens, which has long been hampered by the limited genetic diversity of data sets based on individual genes. The authors investigate a whole-genome single nucleotide polymorphism (SNP) data set of health care-associated methicillin-resistant *S. aureus* sequence type 239 (HA-MRSA ST239) strains using Markov models that consider discrete diffusion among the geographical locations of sampling. Gray et al. (2011) employ the HME to perform model selection, which generally prefers complex evolutionary and population dynamic models: an uncorrelated relaxed clock and the Bayesian skyline plot model provide a better fit than a strict clock and a constant population size assumption respectively.

[Table 2 about here.]

Here, we revisit the subset of analyses that use an ascertainment bias correction (ABC) model to take into account that only variable sites are being used. Indeed, in many alignments of closely related sequences, a large number of sites are invariant and are often excluded because they are phylogenetically uninformative. However, when these sites are excluded, a correction is needed to renormalize the site probabilities to account for the difference between unobserved and excluded site patterns. Following the original analysis we consider data sets with full, intergenic and synonymous SNP inclusion (Gray et al. 2011).

Gray et al. (2011) note that for the analyses assuming a relaxed clock, three independent fittings were combined to obtain sufficient independent samples from the posterior. Since exploratory analyses using PS and SS also indicated inconsistent results in some cases, we reran the original analyses to diagnose potential issues. For the full and synonymous data sets, we encountered inadequate mixing for the parameters of the general time-reversible (GTR) nucleotide substitution model, equilibrium nucleotide frequencies and the parameters of the UCLD clock model. To ameliorate these issues, we simplified the GTR model to an HKY model, fixed the base frequencies to the empirical base frequencies and most importantly replaced the improper uniform prior on the mean rate in the UCLD model with a diffuse gamma prior. For matters of consistency, we apply the same models and priors to the intergenic data set. This resolved the apparent mixing issues, yielding proper posterior and prior distributions and consistent model ordering according to PS and SS (Table 2). It therefore remains a crucial part of any MCMC analysis to check the MCMC chain for adequate mixing and provide proper priors for all the model parameters if one wishes to estimate marginal likelihoods. Given these changes, we have also recalculated the HME estimates reported in the original paper using the new settings.

Only after providing proper priors and using PS and SS, we arrive at consistent conclusions across all three data partitions. For each data partition, an uncorrelated relaxed clock outperforms a strict clock and a Bayesian skyline plot model outperforms a constant population size assumption (Table 2).

Discussion

Recent developments in marginal likelihood estimation demonstrate the potential for more accurate Bayesian model selection based on simple Gaussian model examples and small real-world phylogenetic data sets. Here, we implement such estimators, including PS and SS, as well as the AICM, in a Bayesian inference framework for

evolutionary hypotheses testing when uncertainty remains about the underlying time-measured genealogy. Such genealogies require molecular clock assumptions and dedicated tree priors, such as coalescent models (Drummond et al. 2002), that frequently need to be scrutinized. Our simulations and analyses of empirical data sets indicate that PS and SS remain feasible without conditioning on a known phylogeny and, although computationally more demanding, consistently outperform the AICM and the HME. These latter approaches are less computationally demanding because they only require samples from the posterior distribution to perform model selection and can be calculated from a standard MCMC run. PS and SS, on the other hand, require MCMC sampling from a series of power posteriors in order to be able to calculate the marginal likelihood. Given that the accuracy of the estimator depends on the number of power posteriors that are traversed, a large number of iterations may be required to yield reliable results for large data sets and complex evolutionary models.

All of the methods mentioned are now available in BEAST (Drummond et al. 2012) through XML specification, with the HME and the AICM accessible directly in the graphical interface driven Tracer program. We provide two BEAST XML files as Supplementary Material to this paper, one illustrating the usage of the HME and AICM estimators and one illustrating the usage of the PS and SS estimators. In these two examples, the intergenic data set of Gray et al. (2011) is analyzed using a Bayesian skyline plot model and an uncorrelated relaxed clock with a log-normal distribution (UCLD). The implementations allow for an easy comparison between different models while incorporating phylogenetic uncertainty. In the current study, we focus on comparing demographic and clock models, but the general implementation allows to calculate marginal likelihoods for any model that can be fitted in BEAST, such as sequence evolution, trait evolution and phylogeographic models (see e.g. (Lemey et al. 2009, 2010)). We refer to Drummond et al. (2012) for an overview of available models. Further, our implementations allow marginal likelihoods of a series of models to be calculated independently, after which these can be compared through their Bayes factors to decide which model yields the best fit to the data and should therefore be used for parameter estimation.

As mentioned earlier, Worobey et al. (2008) show that the inclusion of the 1959 and 1960 sequences seemed to improve estimation of the TMRCA of the M group. We have shown, using path sampling and stepping-stone sampling, that the Bayesian skyline plot model is the optimal choice among the demographic models that we tested for this data set. With respect to the conclusions put forward in the work of Worobey et al. (2008), this means that the time of the most recent common ancestor obtained under the Bayesian skyline plot (TMRCA 1908, 95% HPD 1884-1924) can be selected over that of the constant population model (TMRCA 1921, 95% HPD 1908-1933), when the 1959 and 1960 sequences are included. Hence, in this scenario, the estimate of the TMRCA of the M group is relatively insensitive to the coalescent tree prior. However, should our conclusions still hold when the 1959 and 1960 sequences are excluded, the difference between the TMRCA estimates would drastically increase, with a TMRCA under the Bayesian skyline plot of 1882 (95% HPD 1831-1916) and a TMRCA under the constant population model of 1933 (95% HPD 1919-1945).

Lartillot and Philippe (2006) note that the difference between the logarithm of the marginal likelihoods of two phylogenetic models can be small compared to the two log marginal likelihoods themselves; this can lead to a poor estimate of the Bayes factor, unless the precision on each marginal likelihood estimate is very high. To counter this effect, researchers suggest constructing a single path connecting the two competing models in the space of unnormalized densities and then calculating the Bayes factor directly along this single path (Gelman and Meng 1998). By construction, this approach often results in lower estimate error for the Bayes factor in phylogenetics (Rodrigue, Philippe and Lartillot 2006). However, estimator efficiency depends on the path construction and hence, other paths between two arbitrary models may be devised. For highly structured models, such as those we find in phylogenetics, finding an efficient path between two arbitrary models is not a generic exercise and requires expert knowledge, e.g. when the models have mismatching or extra parameters. In up-coming work, we aim to provide the ability to construct such Bayes factor estimators in BEAST. The main challenge in accomplishing this is to develop a user-friendly interface for users to link common parameters between the competing models to construct effective paths. Indeed, while marginal likelihood estimation for a particular models already requires various adaptations in software, Bayes factor estimation between two arbitrary models requires much more drastic changes.

One way to circumvent the path construction difficulty is to shorten the path from posterior to prior whilst still calculating the marginal likelihood for each model separately. Recently, Fan et al. (2011) propose a more general version of SS that introduces an arbitrary "working" prior distribution that, in practice, one selects as a product of independent probability densities parameterized using MCMC samples from the posterior distribution. The authors show that if this reference distribution exactly equals the posterior distribution, the marginal likelihood can be estimated exactly. The generalized SS is considerably more efficient and does not require sampling from distributions close to the true prior that is problematic for vague choices. However, at the moment this method is restricted to evaluations on a fixed phylogenetic tree topology. Integrating over plausible tree topologies complicates generalized SS because of the need to define a reference distribution for topologies that provides a good approximation to the posterior. Future work will focus on tackling these technical hurdles and further improving marginal likelihood estimation for model selection.

Bayesian phylogenetics requires a sensible balance between parameter-richness and biological realism. A good model captures the essential features of the hypothesis being tested without introducing unnecessary error, bias and over-fitting. Accurate model comparisons are therefore a crucial part of any phylogenetic study, even though in this field of research the model will always be misspecified in the sense that all evolutionary models are severe simplifi-

cations of reality. Based on the results we presented in this paper, we advocate against the use of the HME and provide an alternative measure, the AICM, as an initial posterior-based investigation to be used with caution. While PS/SS both come with increased computational demands, they clearly provide the most accurate and consistent results and we recommend them for performing model selection.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under Grant Agreement n278433 and ERC Grant agreement no. 260864, from the National Institutes of Health (U54 RR024386-01A2, R01 GM086887, R01 HG006139 and R01 NS063897) and from The Wellcome Trust (WT092807MA). TB is supported by the European Molecular Biology Organization. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by Ghent University, the Hercules Foundation and the Flemish Government department EWI. We acknowledge the support of the National Evolutionary Synthesis Center (NESCent) through a working group (Software for Bayesian Evolutionary Analysis).

Literature Cited

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Second international symposium on information theory., Springer Verlag, volume 1, pp. 267–281.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006a. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics. 161:1307–1320.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2006b. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22:1185–1192.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution. In press.

Duffy S, Shackleton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9:267–276.

Fan Y, Wu R, Chen MH, Kuo L, Lewis PO. 2011. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. 28:523–532.

Firth C, Kitchen A, Shapiro B, Suchard MA, Holmes EC, Rambaut A. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Mol. Biol. Evol. 27:2038–2051.

Friel N, Petitt AN. 2008. Marginal likelihood estimation via power posteriors. J. R. Stat. Soc. B. 70:589–607.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. Bayesian Data Analysis. Chapman & Hall/CRC.

Gelman A, Meng XL. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Stat. Sci. 13:163–185.

Gray RR, Tatem AJ, Johnson JA, Alekseyenko AV, Pybus OG, Suchard MA, Salemi M. 2011. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant staphylococcus aureus st239 genome-wide data within a Bayesian framework. Mol. Biol. Evol. 28:1593–1603.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294:2310–2314.

Jeffreys H. 1935. Some tests of significance treated by theory of probability. In: Proceedings of the Cambridge Philosophical Society. volume 31, pp. 203–222.

Kass RE, Raftery AE. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795.

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finding its roots. PLoS Comp. Biol. 5:e1000520.

Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010. Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. 27:1877–1885.

Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24:2669–2680.

Li WLS, Drummond AJ. 2011. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol. p. (in press).

Newton MA, Raftery AE. 1994. Approximating Bayesian inference with the weigthed likelihood bootstrap. J. R. Stat. Soc. B. 56:3–48.

Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.

Ogata Y. 1989. A Monte Carlo method for high dimensional integration. Num. Math. 55:137–157.

Raftery A, Newton M, Satagopan J, Krivitsky P. 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: MBernardo J, Bayarri MJ, Berger JO, editors, Bayesian Statistics, Oxford University Press, pp. 1–45.

Rambaut A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics. 16:395–399.

Redelings BD, Suchard MA. 2005. Joint Bayesian estimation of alignment and phylogeny. Syst. Biol. 54:401–418.

Rodrigue N, Philippe H, Lartillot N. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol. Biol. Evol. 23:1762–1775.

Schwartz G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.

Sinsheimer JS, Lake JA, Little RJ. 1996. Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. Biometrics. 52:193–210.

Steel MA. 2005. Should phylogenetic models be trying to fit an elephant? Trends Genet. 21:307–309.

Suchard M, Weiss R, Sinsheimer J. 2003. Testing a molecular clock without an outgroup: derivations of induced priors on branch-length restrictions in a Bayesian framework. Systematic biology. 52:48–54.

Suchard MA, Weiss RE, Sinsheimer JS. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.

Suchard MA, Weiss RE, Sinsheimer JS. 2005. Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. Biometrics. 61:665–673.

Verdinelli I, Wasserman L. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. J. Am. Stat. Assoc. 90:614–618.

Worobey M, Gemmel M, Teuwen DE, et al. (13 co-authors). 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature. 455:661–665.

Xie W, Lewis PO, Fan Y, Kuo L, Chen MH. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.

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

Zhu T, Korber BT, Nahmias AJ, Hooper E, Shaper PM, Ho DD. 1998. An african HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature. 391:594–597.

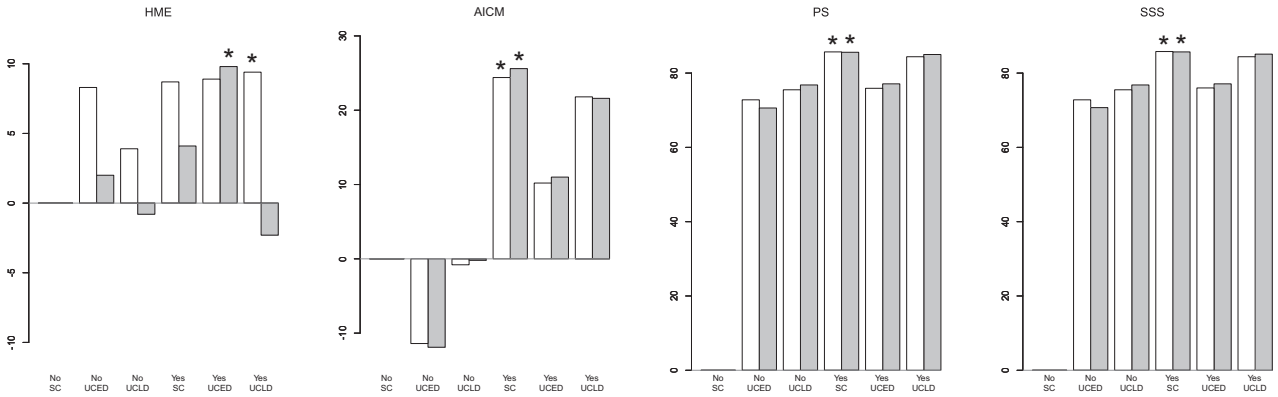

FIG. 1.—Differences in log marginal likelihood estimates and AICM for two independent fittings (first fitting shown in white, second in gray) of the HIV dataset using the harmonic mean estimator (HME), posterior-simulation Akaike information content (AICM), path sampling (PS) and stepping-stone sampling (SS). For each estimator, the constant population size model (Con) was used as the reference model and the top performing model for each fitting is indicated with a star (*). For all estimators, we employ equal amounts of computational work (MCMC iterations), as well as an equal numbers of samples from which to estimate the marginal likelihood. The HME shows drastic differences in the overall ranking of the demographic models and, depending on the fitting, may very well select a constant population size as the preferred coalescent prior. The AICM is consistent across both fittings but selects a constant population size above all other coalescent priors. PS and SS consistently select the Bayesian skyline plot (BSP) coalescent prior as the optimal choice and put the constant population size far behind the other coalescent priors. PS and SS indicate that the expansion growth model (Expan) yields the second highest fit, while the exponential (Expo) and logistic (Log) growth models yield similar performance.

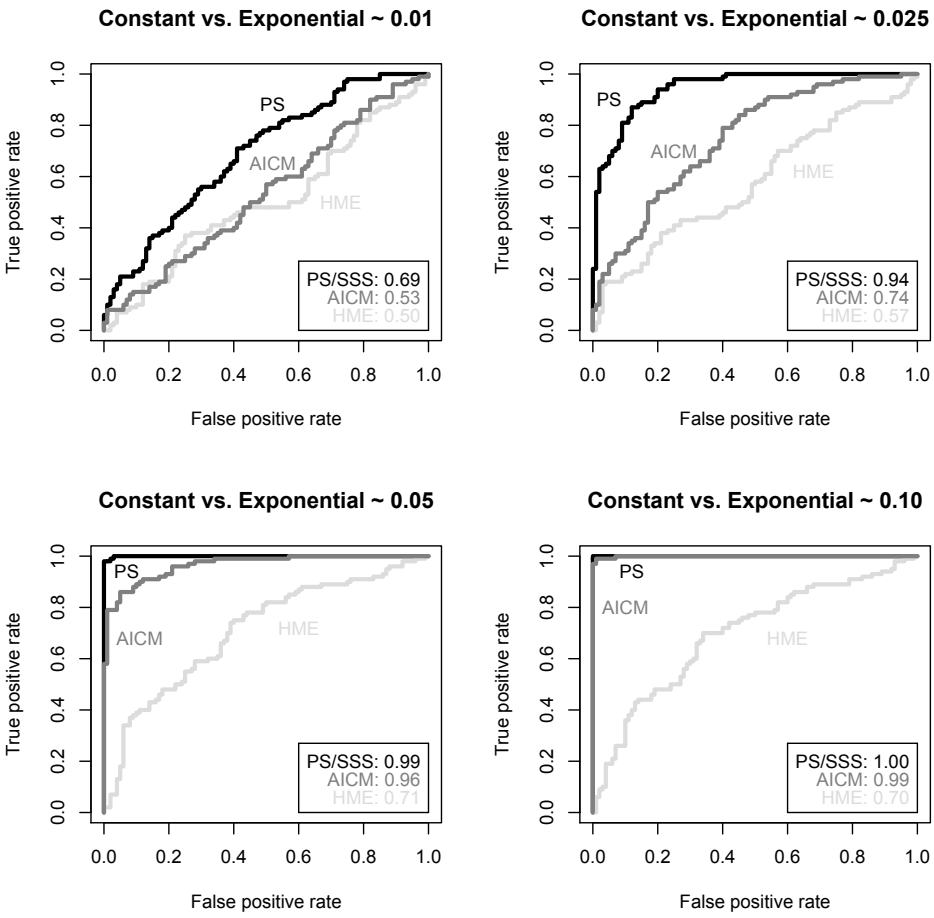

FIG. 2.—Evaluation of log Bayes factor estimates using PS (SS yields an undistinguishable plot), AICM and the HME to compare model fit, with four pairwise comparisons being shown: a constant population size versus an exponential population size with growth rates of 0.01, 0.025, 0.05 and 0.10. An increasingly strong discriminatory behaviour (low false positive rates and high true positive rates) can be seen for PS (and SS) up to a growth rate of 0.10, whereas the HME retains questionable performance. AICM performance lies in between that of the HME and PS/SS. Color-coded area under the curve (AUC) values are given at the bottom right of each plot.

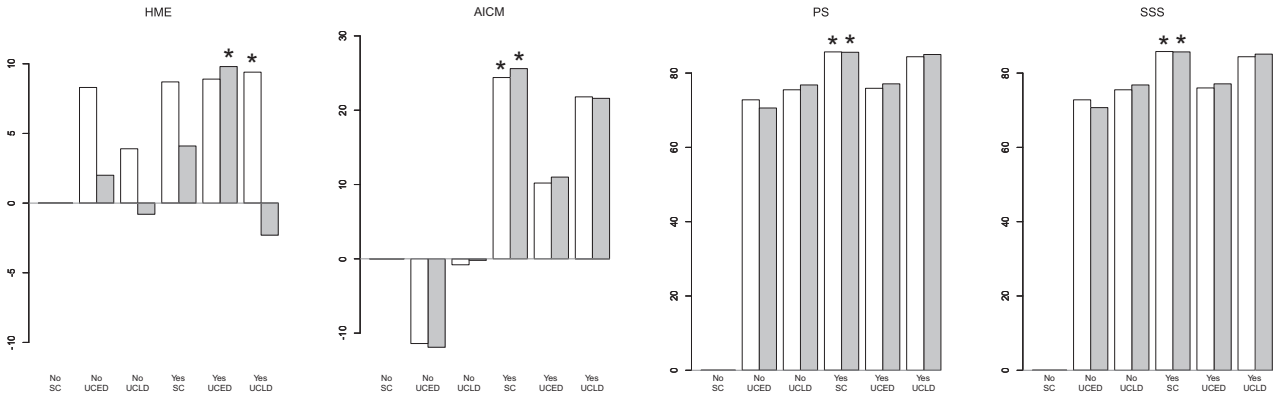

FIG. 3.—Differences in log marginal likelihood estimates for two independent fittings (first fitting shown in white, second in gray) for the HSV dataset (Firth et al. 2010) using HME, AICM, PS and SS using a strict clock (SC), an uncorrelated relaxed clock with an exponential distribution (UCED) and an uncorrelated relaxed clock with a lognormal distribution (UCLD). The data was analyzed excluding the sampling dates (No) and including the sampling dates (Yes). We used the strict clock model excluding the sampling dates as the reference model and the top performing model for each fitting is indicated with a star (*). Equal amounts of computational work (MCMC iterations) were run for all estimators, as well as an equal number of posterior samples being used to estimate the marginal likelihood. While the HME shows drastic differences in the overall ranking of the (clock) models, the AICM as well as PS and SS exhibit consistent behaviour, although disagreeing on the performance of a strict clock when the sampling dates are omitted.

Table 1 Marginal likelihood estimator performance for 100 simulated datasets under various coalescent priors using the HME, AICM, PS and SS. We employed equal amounts of computational work (MCMC iterations) for all estimators, as well as an equal number of posterior samples being used to estimate the marginal likelihood. The HME, PS and SS columns report the number of correct classifications obtained out of 100 simulations. The log BF HME, log BF PS and log BF SS report the mean log Bayes factor over all replicates between the constant population size and exponential growth coalescent priors (a positive number indicates a preference for the constant population size), while ΔAICM reports the mean difference of the AICM values across all replicates.

Coalescent prior	Growth rate	HME	AICM	PS	SS	log BF HME	ΔAICM	log BF PS	log BF SS
Constant	-	48	59	72	72	0.61	0.57	1.76	1.76
Exponential	0.010	50	45	57	57	0.28	0.20	-0.81	-0.80
Exponential	0.025	59	73	92	92	-1.33	-1.36	-6.81	-6.81
Exponential	0.050	80	99	100	100	-4.43	-4.34	-12.54	-12.54
Exponential	0.100	78	100	100	100	-7.75	-7.66	-18.24	-18.24

Table 2 Marginal likelihood estimates for two independent fittings for the HA-MRSA ST239 dataset using the HME, AICM, PS and SS (with the overall ranking of the models shown in parentheses for each estimator) after specifying proper priors. As in the original publication of Gray et al. (2011), we compare the constant population size and Bayesian skyline plot (BSP) demographic models under both a strict clock (SC) and an uncorrelated relaxed clock with a lognormal distribution (UCLD) for three data sets: a full, intergenic and synonymous data set (we refer to Gray et al. (2011) for more details on these data sets). Equal amounts of computational work (MCMC iterations) were run for all estimators, as well as equal numbers of posterior samples being used to estimate the marginal likelihood. Only PS and SS are able to yield a consistent model classification across both fittings, thereby generating the same overall ranking as in the original publication (Gray et al. 2011). The HME and AICM are only able to generate a consistent and correct classification in one out of three data sets.

			Fitting 1				Fitting 2			
Data	Clock	Coalescent	HME	AICM	PS	SS	HME	AICM	PS	SS
Full	SC	Constant	-28420.5 (4)	56865.6 (4)	-28738.2 (4)	-28735.9 (4)	-28418.2 (3)	56865.2 (4)	-28735.5 (4)	-28734.2 (4)
Full	SC	BSP	-28419.4 (3)	56860.7 (3)	-28724.9 (3)	-28723.2 (3)	-28420.8 (4)	56860.2 (3)	-28723.8 (3)	-28722.3 (3)
Full	UCLD	Constant	-28304.2 (2)	56681.8 (2)	-28641.1 (2)	-28638.3 (2)	-28308.2 (2)	56682.4 (2)	-28647.5 (2)	-28644.2 (2)
Full	UCLD	BSP	-28304.1 (1)	56679.6 (1)	-28635.6 (1)	-28631.9 (1)	-28304.4 (1)	56680.1 (1)	-28631.8 (1)	-28628.2 (1)
Intergenic	SC	Constant	-6493.7 (4)	13016.8 (2)	-6749.5 (4)	-6749.3 (4)	-6495.9 (4)	13016.7 (2)	-6750.0 (4)	-6749.6 (4)
Intergenic	SC	BSP	-6489.4 (3)	13001.4 (1)	-6740.0 (3)	-6739.7 (3)	-6488.9 (3)	13001.4 (1)	-6742.3 (3)	-6742.0 (3)
Intergenic	UCLD	Constant	-6479.9 (1)	13037.7 (3)	-6730.1 (2)	-6729.4 (2)	-6481.9 (1)	13038.3 (3)	-6725.2 (2)	-6724.8 (2)
Intergenic	UCLD	BSP	-6480.6 (2)	13048.2 (4)	-6716.7 (1)	-6716.1 (1)	-6482.0 (2)	13043.7 (4)	-6717.1 (1)	-6716.5 (1)
Synonymous	SC	Constant	-6563.9 (4)	13149.7 (4)	-6816.3 (4)	-6815.8 (4)	-6561.9 (4)	13149.2 (4)	-6816.8 (4)	-6816.3 (4)
Synonymous	SC	BSP	-6556.1 (3)	13133.1 (2)	-6806.4 (3)	-6806.0 (3)	-6558.5 (3)	13133.8 (2)	-6806.6 (3)	-6806.1 (3)
Synonymous	UCLD	Constant	-6541.7 (2)	13138.7 (3)	-6787.4 (2)	-6786.7 (2)	-6538.6 (2)	13138.5 (3)	-6786.8 (2)	-6786.1 (2)
Synonymous	UCLD	BSP	-6533.6 (1)	13122.8 (1)	-6780.8 (1)	-6780.3 (1)	-6536.1 (1)	13123.9 (1)	-6780.9 (1)	-6780.0 (1)