Dating Phylogenies with Sequentially Sampled Tips

Author(s):
Stadler, Tanja; Yang, Ziheng

Publication date:
2013

Permanent link:
https://doi.org/10.3929/ethz-b-000071086

Originally published in:
Systematic Biology 62(5), https://doi.org/10.1093/sysbio/syt030

Syst. Biol. 62(5):674–688, 2013

DOI:10.1093/sysbio/syt030
Advance Access publication April 28, 2013

Dating Phylogenies with Sequentially Sampled Tips

TANJA STADLER[1] AND ZIHENG YANG[2,*]

[1]*Institut für Integrative Biologie, Eidgenössiche Technische Hochschule Zürich, 8092 Zürich, Switzerland; and* [2]*Department of Biology, University College London, London WC1E 6BT, UK*

**Correspondence to be sent to Department of Biology, Darwin Building, University College London, London WC1E 6BT, UK; E-mail: z.yang@ucl.ac.uk.*

Received 21 November 2012; reviews returned 15 February 2013; accepted 23 April 2013
Associate Editor: Laura Kubatko

Abstract.—We develop a Bayesian Markov chain Monte Carlo (MCMC) algorithm for estimating divergence times using sequentially sampled molecular sequences. This type of data is commonly collected during viral epidemics and is sometimes available from different species in ancient DNA studies. We derive the distribution of ages of nodes in the tree under a birth–death-sequential-sampling (BDSS) model and use it as the prior for divergence times in the dating analysis. We implement the prior in the MCMCtree program in the PAML package for divergence dating. The BDSS prior is very flexible and, with different parameters, can generate trees of very different shapes, suitable for examining the sensitivity of posterior time estimates. We apply the method to a data set of SIV/HIV-2 genes in comparison with a likelihood-based dating method, and to a data set of influenza H1 genes from different hosts in comparison with the Bayesian program BEAST. We examined the impact of tree topology on time estimates and suggest that multifurcating consensus trees should be avoided in dating analysis. We found posterior time estimates for old nodes to be sensitive to the priors on times and rates and suggest that previous Bayesian dating studies may have produced overconfident estimates. [Bayesian inference; MCMC; molecular clock dating; sampled tips; viral evolution.]

The distance information in molecular sequences can be translated into absolute times and rates if information about the ages of some nodes in the phylogeny is available. This strategy has been used to date species divergences, with the fossil record used to inform the ages of certain nodes and thus to calibrate the molecular phylogeny. The Bayesian method (Thorne et al. 1998) provides a powerful general framework for integrating such different sources of information. Recent developments in Bayesian molecular clock dating include soft bounds and flexible statistical distributions to deal with uncertainties in fossil calibrations (Yang and Rannala 2006; Drummond and Rambaut 2007) and flexible prior models to describe the drift of the evolutionary rate across lineages (Rannala and Yang 2007; Guindon 2013).

Absolute dates can also be estimated from molecular sequence data without fossil calibrations, if the sequences are sampled at different time points and if the evolutionary rate is high enough so that the time gap covered by the sampled sequences is enough for substantial evolution to occur. This is the case with viral gene sequences and also in a few ancient DNA studies (Drummond et al. 2003). The different sample dates allow evolutionary changes to be calibrated to generate estimates of absolute rates and times. The Bayesian Markov chain Monte Carlo (MCMC) program multidivtime (Thorne et al. 1998) appears to be the first to implement such models for non-contemporary sequences although this option has rarely been used. Currently, BEAST (Drummond and Rambaut 2007) appears to be the only program being used for dating viral divergences. BEAST implements a number of priors for divergence times, based on the neutral coalescent model either with or without population growth as well as birth–death process models.

In this article, we extend the Bayesian dating method of Yang and Rannala (2006) and Rannala and Yang (2007), implemented in the MCMCtree program in the PAML package (Yang 2007), to analyze sequentially sampled viral sequences. Although the pruning algorithm for likelihood calculation (Felsenstein 1981) is implemented in BEAST, MCMCtree in addition implements an approximate method for the likelihood calculation (Thorne et al. 1998; dos Reis and Yang 2011), which is much faster and can be used in analysis of large data sets (Battistuzzi et al. 2011). We extend the birth–death-sequential-sampling (BDSS) model of Stadler (2010) to specify a prior distribution of divergence times, which is combined with the prior on the evolutionary rates and with the likelihood of the sequence data to give the posterior distribution of divergence times. Our new prior is a generalization of the prior for trees with one sampling time point (Yang and Rannala 1997, 2006).

In the following, we first derive the prior distribution of divergence times under the BDSS model, and then describe our implementation of the prior in the MCMCtree program. We apply the new dating method to two data sets. The first consists of 33 SIV/HIV-2 sequences, compiled and analyzed by Lemey et al. (2003). We use this data set to assess the sensitivity of the posterior time estimates to the prior on times and rates and to compare our new method with the maximum-likelihood (ML) TipDate method of Rambaut (2000). The second data set consists of 289 influenza H1 genes from avian, swine, and human hosts (dos Reis et al. 2011). We analyze it to estimate the divergence times under different rate-drift models in comparison

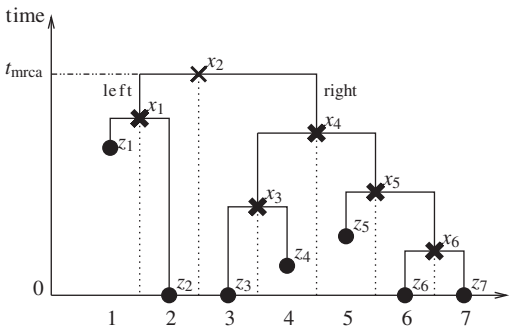

FIGURE 1. Example of a sampled tree for $N=7$ samples. The black circle at position i ($i=1,\ldots,N$) on the horizonal axis denotes the sampling time z_i (here we have $z_2=z_3=z_6=z_7=0$). The black cross at position $i+0.5$ ($i=1,\ldots,N-1$) on the horizonal axis denotes the branching times x_i. $\max(x)=x_2=t_{\mathrm{mrca}}$ is the time of the most recent common ancestor, that is, $k=2$. We have $z_1^*=z_1, z_2^*=0, z_3^*=z_4, z_4^*=z_5, z_5^*=z_5, z_6^*=0$. The probability density $f[x|z,k,t_{\mathrm{mrca}}]$ denotes the probability of branching times $x=(x_1,\ldots,x_{N-1})$, where z,k,t_{mrca} are fixed (i.e., only the bold crosses can be moved vertically). Note that different x may induce different tree topologies (e.g., $x_5<x_6$ induces a different tree than the tree in this figure).

with the Bayesian dating program BEAST (Drummond and Rambaut 2007).

METHODS

The BDSS Model

We assume a BDSS model of birth, death, and sampling, which gives rise to a prior distribution of binary trees with divergence times. The process starts with a first single lineage. A lineage gives birth to another lineage with rate λ and dies with rate μ. The term lineage here can mean a viral sequence in analysis of viral data with sample dates (as in this artcile) or a species in analysis of species phylogenies. A birth may correspond to a viral transmission or a speciation event, and a death may correspond to the host becoming noninfectious or to a species extinction. The two lineages originating from a birth event are distinguished and labeled "left" and "right." The process is stopped at the *present* time; see below for a specification of present time. Sampling takes place through time, described by two processes. First, with rate ψ per lineage, we sample lineages sequentially through time, yielding "extinct" samples (i.e., samples taken prior to the present). We assume that an extinct sample is not giving rise to further lineages (meaning it is essentially dead) and thus in particular an extinct sample is not ancestral to a descendant sample. Second, at the present time, we sample each extant lineage with probability ρ. These samples typically correspond to extant species. For analysis of viral sequences (as in this article), we set $\rho=0$. The parameters in the BDSS model are denoted $\eta=(\lambda,\mu,\psi,\rho)$. In this artcile, η is fixed, so we may suppress it in our notation.

The BDSS model produces a binary tree. Pruning all dead and nonsampled lineages, as well as the lineage ancestral to the first branching event yields the "sampled tree," for the sampled lineages only (Fig. 1). The sampled tree is drawn such that each branching event has the "left" descendant on the left and the "right" descendant on the right. The distinction between left and right is a convenient notation for the derivation of the probability density. Note that swapping the two sides of the tree does not change the density.

We label the sampled lineages from left to right by $1,\ldots,N$ and denote their sampling times by $z=(z_1,\ldots,z_N)$. Let the branching events be $1,\ldots,N-1$ and the branching times be $x=(x_1,\ldots,x_{N-1})$ from left to right. The last (most recent) sampling time is taken as time zero and is referred to as "the present." The age of the root (i.e., the time of the first branching event) is $t_{\mathrm{mrca}}=\max(x)$. With this definition, vectors x and z fully specify the sampled tree. Furthermore, each sampled tree has a unique representation through x and z. The birth–death process has no natural beginning or ending, so one has to condition on certain features of the process to apply the theory. One may condition on the time of the origination of the first lineage, as in (Stadler et al. 2013), on the number of sampled tips N, or on both.

For dating the branching times in a given sampled tree, we want to condition on the age of the root of the sampled tree (t_{mrca}). Second, we conditon on the number of the sampled tips (N). With such conditioning, the crosses representing the non-root nodes in Figure 1 may be moved up and down according to the prior distribution whereas the root and the sampling points are fixed. We require the user to specify a diffuse prior on the root age. While conditioning on the number of sampled tips N and the sampling times z is sufficient to generate a prior distribution on t_{mrca} from the BDSS model, the information provided by N and z may be too diffuse and misleading, running the risk that the posterior time estimates are unduely influenced by the prior without the user's knowledge. Similarly in the method of Yang and Rannala (2006) and Rannala and Yang (2007) for dating species divergences, the prior of times is conditioned on the root age and the user is required to specify a prior distribution for the root age. The joint prior of all node ages is generated by the so-called "conditional construction," which multiplies the user-specified prior on the root age and the prior for the ages of non-root nodes (conditioned on the root age) specified by the BDSS model (Yang and Rannala 2006). Third, we also condition on the topology of the sampled tree T (i.e., on the sampled tree ignoring branching times and sampling times, but preserving relatedness and orientation of "left" and "right"). The requirement for a given tree topology may be a drawback, especially for viral sequences that are highly similar and thus do not contain much phylogenetic information. For the present, it is unclear whether a model-averaging approach, which averages over different phylogenies to estimate clade ages, can produce more reliable time estimates than a fixed tree approach, which estimates divergence times on a fixed tree topology (such as the ML tree). However, fixing the tree topology allows us to use the approximate likelihood calculation (dos Reis and Yang 2011), which is

useful for analysis of large data sets. In practice, one may use alternative tree topologies to evaluate the impact of the tree on the posterior time estimates (see below).

Estimating Divergence Times

The main objective of this article is to derive the probability density of node ages (divergence times x) given the sampled tree topology T and sampling times z under the BDSS model. This density will be used as the prior on times, and together with the prior on rates r and substitution parameters θ and with the likelihood for the sequence data D, will generate the posterior distribution of divergence times x. By Bayes's theorem, the posterior distribution of x, r, θ is given as

$$f[x,r,\theta|D;T,z] = \frac{f[D|z,x,r,\theta]f[x|T,z,t_{\mathrm{mrca}}]f[t_{\mathrm{mrca}}|T,z]f[r|T,z]f[\theta|T,z]}{f[D|T,z]}, \quad (1)$$

where $f[D|z,x,r,\theta]$ is the likelihood of the sequence data given the tree (note that z and x specify the whole tree, in particular the topology T) and can be calculated using, for example, Felsenstein's pruning algorithm; $f[x|T,z,t_{\mathrm{mrca}}]$ is the prior on times specified under the BDSS model; $f[t_{\mathrm{mrca}}|T,z]$, $f[r|T,z]$, and $f[\theta|T,z]$ are prior distributions for the root age t_{mrca}, the rates r, and parameters θ, respectively. Finally, $f[D|T,z]$ is the normalizing constant and need not be calculated in MCMC algorithms. In the following, we derive the prior on times $f[x|T,z,t_{\mathrm{mrca}}]$.

Prior Distribution on Divergence Times x

Our objective in this section is to derive $f[x|T,z,t_{\mathrm{mrca}}]$ given the BDSS parameters $\eta=(\lambda,\mu,\psi,\rho)$. Let k be the number of samples in the left subtree descending from the root; in Figure 1, we have $k=2$. Note that k is known if T is given, but not vice versa. We can rewrite

$$f[x|T,z,t_{\mathrm{mrca}}] = f[x|T,z,k,t_{\mathrm{mrca}}] = \frac{f[x,T|z,k,t_{\mathrm{mrca}}]}{f[T|z,k,t_{\mathrm{mrca}}]} = \frac{f[x|z,k,t_{\mathrm{mrca}}]}{f[T|z,k,t_{\mathrm{mrca}}]}. \quad (2)$$

The last equality follows, because when z is given, x specifies the topology T. We now deal with the two terms $f[x|z,k,t_{\mathrm{mrca}}]$ and $f[T|z,k,t_{\mathrm{mrca}}]$ in Equation (2) separately.

The denominator and conditioning on the tree topology.—The denominator $f[T|z,k,t_{\mathrm{mrca}}]$ is the conditional probability of the tree topology (T) given the sample times (z), the number of samples in the left subtree (k), and the age of the root in the sampled tree (t_{mrca}). This probability is hard to calculate and is ignored in our implementation. As a result, our implementation of the BDSS prior is approximate in that the density for times is off by the factor $f[T|z,k,t_{\mathrm{mrca}}]$. Because $f[T|z,k,t_{\mathrm{mrca}}]$ is a function of t_{mrca} (note that T,z,k are fixed in the MCMC), ignoring it has the effect of changing the prior for t_{mrca}. The prior for t_{mrca} specified by the user, called the user-specified prior, is $f[t_{\mathrm{mrca}}|T,z]$, but the prior used by the computer program, called the effective or actual prior, is instead

$$\frac{f[t_{\mathrm{mrca}}|T,z]}{f[T|z,k,t_{\mathrm{mrca}}]} = \frac{f[t_{\mathrm{mrca}},T|z,k]/f[T|z,k]}{f[t_{\mathrm{mrca}},T|z,k]/f[t_{\mathrm{mrca}}|z,k]} = \frac{f[t_{\mathrm{mrca}}|z,k]}{f[T|z,k]} \propto f[t_{\mathrm{mrca}}|z,k] \quad (3)$$

as $f[T|z,k]$ is a constant.

Suppose the user specifies the prior on the root age $t_{\mathrm{mrca}}|T,z \sim U(100,500)$, flat between 100 and 500 years before present. This is the user-specified prior. If we run the MCMC algorithm without using the sequence data, but using the fixed tree topology (T) and the sampling times (z), and collect the MCMC sample for the root age to construct a histogram and estimate the empirical distribution, the distribution may be different from $U(100,500)$ and is instead given by Equation (3). The mismatch between the user-specified prior and the effective prior used by the MCMC algorithms is currently a common problem in Bayesian molecular clock dating methods (Inoue et al. 2010; Heled and Drummond 2012). We advise that the user should run the program without data to generate the effective prior and confirm that it is sensible for the data.

It turns out that if all samples were taken at the same time, we would not rely on an approximation at all. In that case, all rooted oriented tree topologies with interior nodes ordered by age [called labeled histories by Edwards (1970)] are equally likely (Aldous 2001; Ford et al. 2009). Then, $f[T|z,k,t_{\mathrm{mrca}}]$ would not depend on t_{mrca} and we would not alter the prior distribution of t_{mrca} by dropping $f[T|z,k,t_{\mathrm{mrca}}]$. This reasoning suggests that when the root is very old relative to the tips (so that the tips have almost the same sampling time relative to the root), the user-specified and effective priors will be very similar. This argument is supported by our tests (see Appendix B).

We note that it may be possible to estimate $f[T|z,k,t_{\mathrm{mrca}}]$ by an MCMC algorithm. For given z, k, and t_{mrca}, the MCMC should sample the times x and tree topology $\tilde{T}$ from $f[x,\tilde{T}|z,k,t_{\mathrm{mrca}}]$, which corresponds to moving the bold crosses in Figure 1 vertically under the only constraint that ancestral nodes are younger than $x_k = t_{\mathrm{mrca}}$. The proportion of samples in which $\tilde{T}=T$ will be an estimate of the probability $f[T|z,k,t_{\mathrm{mrca}}]$. As T,z, and k are fixed and only t_{mrca} varies, one can generate a look-up table for different t_{mrca} before running the MCMC using the sequence data. We suspect that small changes in t_{mrca} do not cause large changes to $f[T|z,k,t_{\mathrm{mrca}}]$, so that one does not need to tabulate many values for t_{mrca}. However, on a large data set with many sequences, there may be many trees that are comparable with the given z, k, and t_{mrca}. As a result, $f[T|z,k,t_{\mathrm{mrca}}]$

may be extremely small and difficult to estimate reliably by MCMC. This approach is not pursued further in this article.

The numerator and the distribution of the branching times.—The numerator in Equation (2), $f[x|z,k,t_{\text{mrca}}]$, is given by the following theorem, a proof of which is provided in Appendix A.

Theorem 1 *The probability density of the divergence times x, given z,k, and t_{mrca}, is*

$$f[x|z,k,t_{\text{mrca}}]=\prod_{i=1,i\neq k}^{N-1}\frac{c_1(1-c_2)e^{-c_1x_i}}{g(x_i)^2(1/g(t_{\text{mrca}})-1/g(z_i^*))}, \quad (4)$$

where $z_i^*=\max\{z_i,z_{i+1}\}$ and where

$$c_1=\sqrt{(\lambda-\mu-\psi)^2+4\lambda\psi},$$
$$c_2=-\frac{\lambda-\mu-2\lambda\rho-\psi}{c_1},$$
$$g(t)=e^{-c_1t}(1-c_2)+(1+c_2).$$

Thus, Theorem 1 provides a prior probability density for divergence times x, with the limitation that the effective prior for t_{mrca} is the user-specified prior distribution divided by $f[T|z,k,t_{\text{mrca}}]$. We refer to this prior as Approach 1.

Next, we show that the prior for the divergence times derived here is invariant to the choice of time scale.

Corollary 2 *When time is scaled by a factor $1/s$ (e.g., when one time unit is changed from 1 year to 100 years so that $s=100$), the rates λ,μ, and ψ are transformed accordingly to $s\lambda,s\mu$, and $s\psi$, and the sampling probabilty ρ remains constant, then the probability density of the speciation times x with the scaled parameters ($f_s[x|z,k,t_{\text{mrca}}]$) and the probability density of the speciation times x with the original parameters ($f[x|z,k,t_{\text{mrca}}]$) satisfy the following equality:*

$$f_s[x|z,k,t_{\text{mrca}}]=s^{N-2}f[x|z,k,t_{\text{mrca}}]. \quad (5)$$

Proof. Under the rate and time transformation, c_1 becomes sc_1 and c_2 remains constant. Also, $g(t)$ and c_1x_i remain constant, which establishes the corollary. ■

A consequence of this corollary is that for any time scale, the MCMC method estimates the same distribution for x, as s^{N-2} is a constant and cancels out in the prior ratio. The invariance to change of time scale does not hold anymore if the rate distribution is not invariant (e.g., when the log-normal distribution for rates is assumed), although we expect the effect to be minor.

The special case $\psi=0$.

Corollary 3 *For $\psi=0$, we have $z_i=0$ for all i, and Equation (4) simplifies to*

$$f[x|z,k,t_{\text{mrca}}]=\prod_{i=1,i\neq k}^{N-1}\frac{(\lambda-\mu)^2e^{-(\lambda-\mu)x_i}}{(\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)x_i})^2}\times\frac{\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t_{\text{mrca}}}}{1-e^{-(\lambda-\mu)t_{\text{mrca}}}}. \quad (6)$$

For $\mu=\lambda$ (in addition to $\psi=0$), the probability density $f[x|z,k,t_{\text{mrca}}]$ is obtained by taking the limit $\mu\to\lambda$ in Equation (6), using the property $e^{-\epsilon}\sim1-\epsilon$ for $\epsilon\to0$,

$$f[x|z,k,t_{\text{mrca}}]=\prod_{i=1,i\neq k}^{N-1}\frac{(1/t_{\text{mrca}})+\lambda\rho}{(1+\lambda\rho x_i)^2}.$$

Remark 1. Let t be the order statistic of x, that is $\{x_1,\ldots,x_{N-1}\}=\{t_1,\ldots,t_{N-1}\}$ and $t_1>\cdots>t_{N-1}$. For $\psi=0$ (implying that all $z_i=0$), the distribution for t (and implicitly also for x) is derived without conditioning on k in Yang and Rannala (1997, 2006). This distribution also follows from Theorem 1 and Corollary 3: For $z_1=\cdots=z_N=0$, it is easy to see that each permutation of x induces a unique tree and each permutation of x is equally likely (Gernhard 2008; Ford et al. 2009), thus we obtain,

$$f[t|z,k,t_{\text{mrca}}=t_1]=(N-2)!\prod_{i=2}^{N-1}\frac{(\lambda-\mu)^2e^{-(\lambda-\mu)t_i}}{(\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t_i})^2}\times\frac{\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t_1}}{1-e^{-(\lambda-\mu)t_1}}.$$

As each k is equally likely (Slowinski 1990),

$$\begin{aligned}f[t|z,t_{\text{mrca}}=t_1]&=\sum_{k=1}^{N-1}f[t|z,k,t_{\text{mrca}}=t_1]f[k|z,t_{\text{mrca}}=t_1]\\&=\sum_{k=1}^{N-1}f[t|z,k,t_{\text{mrca}}=t_1]\frac{1}{N-1}\\&=f[t|z,k,t_{\text{mrca}}=t_1].\end{aligned}$$

Thus,

$$f[t|z,t_{\text{mrca}}=t_1]=(N-2)!\prod_{i=2}^{N-1}\frac{(\lambda-\mu)^2e^{-(\lambda-\mu)t_i}}{(\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t_i})^2}\times\frac{\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t_1}}{1-e^{-(\lambda-\mu)t_1}}, \quad (7)$$

which corresponds to Equation (3) in Yang and Rannala (1997). For $\lambda=\mu$,

$$f[t|z,t_{\mathrm{mrca}}=t_1]=(N-2)!\prod_{i=2}^{N-1}\frac{(1/t_1)+\lambda\rho}{(1+\lambda\rho t_i)^2},$$

which corresponds to Equation (8) in Yang and Rannala (2006) (note that the same probability density in Yang and Rannala (1997), Equation (7), had a typo).

Note that when the prior of times is used to date the divergences of viral sequences with sample dates, as in this article, we should have $\rho=0$ and $\psi>0$. When the prior is used to date species divergences, for example, to analyze fossil occurrence data to derive calibration densities (Wilkinson et al. 2011), we should have $\rho>0$ and $\psi>0$. This will then be an extension to the density derived by Yang and Rannala (2006).

An Alternative Prior on Divergence Times

Besides the prior specified in Theorem 1 (Approach 1), we describe in the Online Supplementary Text (Doi:10.5061/dryad.9c568) a second prior for x and refer to it as Approach 2. This is based on rewritting the density as

$$f[x|T,z,t_{\mathrm{mrca}}]=\frac{f[x,z|n,t_{\mathrm{mrca}}]}{f[T,z|n,t_{\mathrm{mrca}}]}, \tag{8}$$

where n is the number of extant sampled lineages. The numerator, $f[x,z|n,t_{\mathrm{mrca}}]$, is given in Theorem 4 in the Online Supplementary Text. As we cannot calculate the denominator, $f[T,z|n,t_{\mathrm{mrca}}]$, it is ignored. As a result, instead of the user-specified prior $f(t_{\mathrm{mrca}}|T,z)$ for the root age, the effective prior used by the computer program is $(f[t_{\mathrm{mrca}}|T,z]/f[T,z|n,t_{\mathrm{mrca}}])$.

Approach 2 has a second slight difference from Approach 1 concerning the definition of t_{mrca}: in Approach 2, if $n>2$ we require both lineages at time t_{mrca} to have extant sampled descendants whereas in Approach 1, one lineage may only have extinct sampled descendants. If $\psi=0$, both priors simplify to the prior described by Yang and Rannala (1997, 2006).

All three priors are implemented in MCMCtree. In Appendix B, we present some results that validate our implementation. Furthermore, we show that in Approach 1, the effective prior for t_{mrca} is close to the user-specified prior if t_{mrca} is old, whereas Approach 2 does not have this property. This is the main reason for our preference for Approach 1.

Comparison of our New Prior with the Prior in BEAST

The prior of times developed in this article (Approach 1) and implemented in the MCMCtree program differs from the BDSS prior implemented in the software package BEAST (Drummond and Rambaut 2007). BEAST samples trees and parameters from the posterior distribution (Stadler et al. 2012),

$$f[x,z,t_0,r,\theta,\eta|D,\hat{z}]=\frac{f[D|z,x,r,\theta]f[x,z|t_0,\eta]f[t_0]f[\eta]f[r]f[\theta]}{f[D,\hat{z}]}, \tag{9}$$

where t_0 is the time of origin (the time of outbreak of an epidemic, or the stem age of a species clade) and $\hat{z}$ are the unordered sampling times. Note that BEAST typically samples tree topologies and birth–death parameters, although it is possible to fix the tree topology by applying constraints, and to essentially fix η by assuming a highly concentrated prior. Doing this will yield, for fixed η,

$$f[x,t_0,r,\theta|D,T,z]=\frac{f[D|z,x,r,\theta]f[x,z|t_0]f[t_0]f[r]f[\theta]}{f[D,T,z]}. \tag{10}$$

One could alter the formulae in BEAST to assume a prior for t_{mrca} instead of t_0, and thus sample only x without t_0. Our MCMCtree implementation samples from the posterior distribution $f[x,r,\theta|D;T,z]$ (Equation 1).

The major difference between the two implementations is that BEAST treats the unordered sampling times $\hat{z}$ as part of the data when it samples trees and the BDSS parameters, while MCMCtree conditions on the sampling times z. As a consequence, in BEAST, the prior on t_{mrca} or t_0 is not conditioned on the sampling times z and specification of the prior on t_{mrca} or on t_0 should not use any knowledge of the tree topology T or the sampling times z. For example, the sampling times for sequences in the influenza A H1 gene data set analyzed later in this article span 91 years, from 1918 to 2009, and the phylogeny consists of major clades such as the avian or avian-like clade, the classic swine clade, and the human clade (see Fig. 3 later in the article). Such information concerning the tree topology and sampling times should not be used when one specifies the prior for t_{mrca} or t_0 in the BEAST analysis, even if the tree topology is fixed and the Bayesian analysis is conditioned on the sampling times.

In comparison, the prior on times in MCMCtree is conditioned on the tree topology T and the sampling times z. For example, when one specifies the uniform bounds on the root age in the influenza tree, one should use the knowledge of the phylogeny and the sequence sampling times. Such knowledge may be the most important information that the biologist has to specify a prior on the root age.

Indeed, the main objective of this article has been to develop a prior that is properly conditioned on the information available to biologists when dating sequentially sampled phylogenies. Our BDSS prior is set up such that by using different parameters, it can generate flexible distributions of divergence times, which will be useful for examining the robustness of posterior time estimates to the prior on divergence times. Our preliminary tests on example data sets (see below) suggest that posterior time estimates may be very sensitive to the BDSS parameters. In contrast to our

implementation, BEAST allows estimation of important epidemiological parameters using molecular sequence data (Stadler 2013), by treating the sampling times as data and by assigning priors on the BDSS parameters. The robustness of Bayesian estimation of divergence times and of BDSS parameters to violation of the BDSS model is an interesting question that merits further study.

Implementation of the Divergence Time Prior in the MCMCtree Program

The new divergence time prior is implemented in the MCMCtree program in the PAML package (Yang 2007). Three models concerning the evolutionary rates are implemented, as described by Yang and Rannala (2006) and Rannala and Yang (2007): The strict molecular clock, the independent-rates model and the correlated-rates model. The likelihood, $f[D|z,x,r,\theta]$ in Equation (1), is calculated using either Felsenstein (1981)'s pruning algorithm or the large-sample approximation based on the Taylor expansion of the log likelihood (Thorne et al. 1998; dos Reis and Yang 2011). Details of those parts of the Bayesian analysis have been described before and are not repeated here.

The MCMC algorithms propose changes to the divergence times x, the evolutionary rates r (under both the clock and relaxed-clock models), and the parameters θ in the substitution model such as the transition/transversion rate ratio κ and the gamma shape parameter α for variable rates among sites. Those proposals are described in Yang and Rannala (2006). However, the so-called mixing proposal [Step 4 on page 225 of Yang and Rannala (2006)] works only for contemporary sequences sampled at the same time, and not for sequentially sampled sequences. In Appendix C, we describe a modification to the algorithm so that it works for our new dating analysis.

Results: Real-Data Analysis

In this section, we apply our new dating method based on the BDSS prior to two data sets. The first one is a small data set of 33 SIV/HIV-2 sequences. We use the data for comparison with the ML TipDate method of Rambaut (2000). The second is a large data set of 289 influenza A H1 gene sequences from the human, swine, and avian hosts. We analyze the data for comparison with the Bayesian dating program BEAST (Drummond et al. 2006; Drummond and Rambaut 2007).

Divergence Times of HIV-2

We applied our new dating method based on the BDSS prior to a data set of SIV/HIV-2 sequences with known isolation dates, aligned and analyzed by Lemey et al. (2003). Previous phylogenetic analysis indicates that HIV-2 originated through multiple interspecies transmissions from sooty mangabeys. In contrast to HIV-1, which has spread globally, HIV-2 is mainly restricted to West Africa, possibly due to its lower viral load and lower transmissibility. HIV-2 subtypes A and B are epidemic whereas subtypes C–G are nonepidemic. Lemey et al. (2003) analyzed the data set to date the introduction of HIV-2 into the human population and to estimate the epidemic history of HIV-2 subtype A in Guinea-Bissau, the putative geographic origin of HIV-2. The alignment has 33 sequences, consisting of partial *gag* and *env* genes, with 1107 nucleotide sites.

Lemey et al. (2003) used the TipDate ML method of Rambaut (2000), implemented in the BASEML program in PAML (Yang 2007), to estimate the divergence times under the molecular clock. The phylogenetic tree used is a bootstrap consensus tree inferred using PAUP* under the TN93+Γ model (Tamura and Nei 1993; Yang 1994b). We have used RAxML (Stamatakis et al. 2005) under GTR+Γ (Yang 1994a, 1994b) to infer the ML tree (shown in Fig. 2) and use it in our dating analysis. This is very similar to the ML tree inferred using PhyML (Guindon and Gascuel 2003) under HKY85+Γ (Hasegawa et al. 1985; Yang 1994b).

We note a few differences between our tree and the tree used by Lemey et al. First, the relationship among subtypes A, B, and C is (A, (B, C)) in our tree while it is ((A, B), C) in Lemey et al. Second, in our tree, the sequence SIVSTM89 is sister to the clade (A, (B, C)), whereas it is not in the Lemey et al. tree. The substitution models used in the three analyses mentioned above are very similar, so we suspect that the difference is due to the search algorithms implemented in PAUP* being less efficient than those in RAxML or PhyML.

More importantly, Lemey et al. used a bootstrap consensus tree with multifurcating nodes for divergence time dating while we use the ML tree from RAxML instead. We suggest that the multifurcating tree is a poor choice for two reasons. First, the polytomy is a biologist's intuitive way of representing phylogenetic uncertainty, but, when used in molecular clock dating, is treated as a precise mathematical model of the evolutionary relationships among the sequences. In contrast, the ML tree has some chance of being the true tree. Second, we suspect that an inferred binary tree (such as the ML tree), even if incorrect, should produce more reliable time estimates than a consensus tree with multifurcations. Consider the estimation of the age of the root in the tree of sequences 1, 2, and 3, which has the true relationship ((1, 2), 3). We suggest that use of the ML tree should provide more reliable estimate of the root age than the star tree (1, 2, 3), since consideration of pairwise distances d_{12}, d_{23}, d_{31} appears to suggest that using the star tree may lead to underestimates of the root age. However, complex patterns may result if there are several multifurcating nodes in different parts of the tree.

Table 1 shows the ML estimates (MLEs) and the 95% confidence intervals for the ages of a few key nodes in the tree when the analysis is conducted using the multifurcating tree of Lemey et al., and the ML trees and the bootstrap consensus trees obtained using RAxML

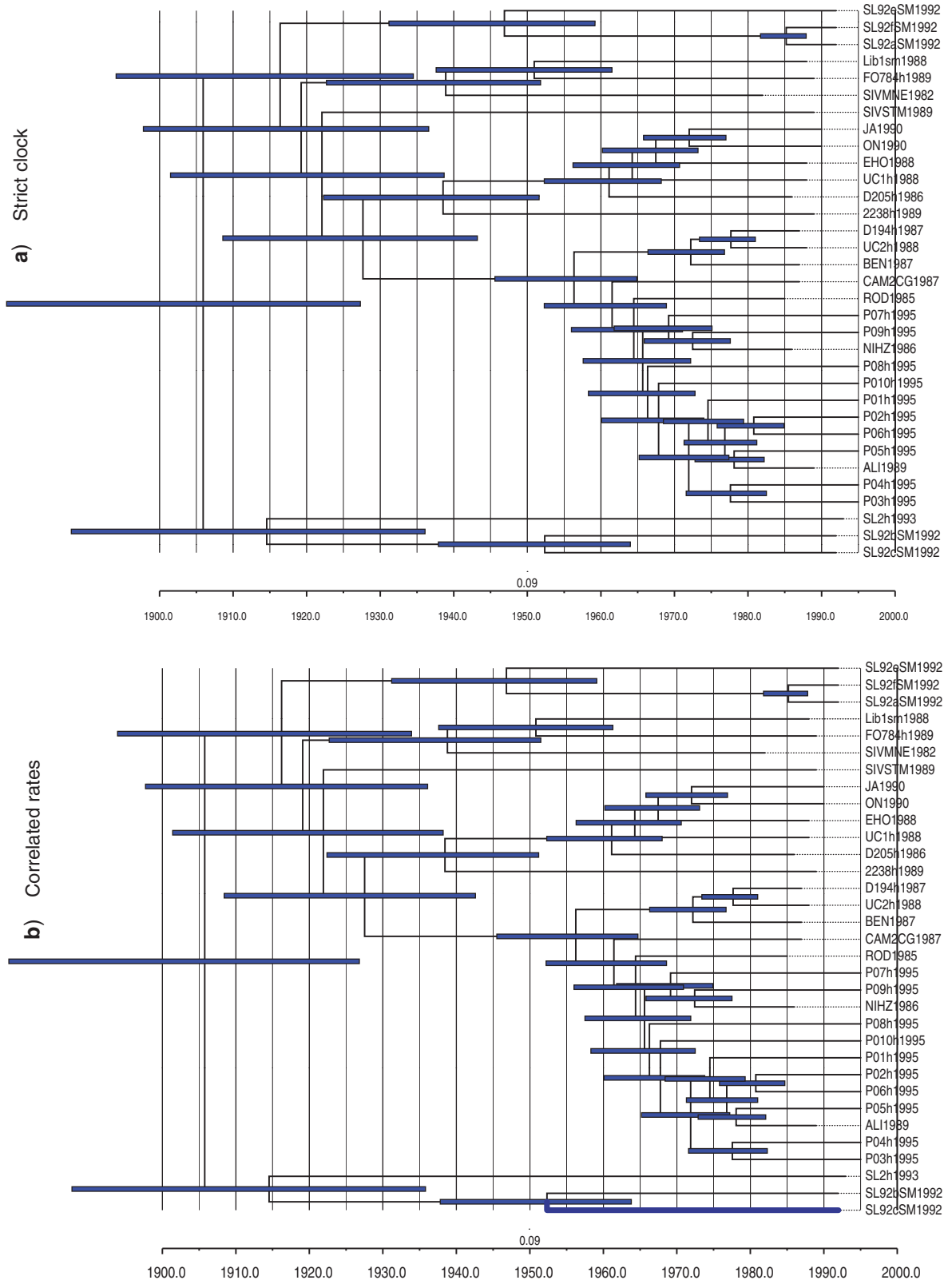

FIGURE 2. HIV-2/SIV ML tree obtained using RAxML with divergence time estimates obtained using the new BDSS prior under a) the strict clock or b) the correlated-rates models. The branch lengths represent the posterior means of time estimates, while the node bars represent the 95% posterior credibility intervals.

TABLE 1. MLEs and 95% confidence intervals of divergence dates under the clock using different tree topologies

Tree	Subtype A	Subtype B	Subtypes B–C	Root	μ	ℓ
Lemey et al. tree	1940 (1905, 1975)	1945 (1914, 1974)	NA	1842 (1740, 1943)	0.23 (0.13, 0.33)	−12 394.63
RAxML ML tree	1958 (1945, 1972)	1959 (1947, 1972)	1932 (1907, 1957)	1896 (1856, 1936)	0.21 (0.11, 0.31)	−12 352.11
RAxML consensus tree	1955 (1937, 1972)	1957 (1941, 1972)	1926 (1895, 1957)	1885 (1834, 1936)	0.24 (0.13, 0.34)	−12 410.38
PhyML ML tree	1959 (1945, 1973)	1959 (1947, 1972)	1931 (1905, 1957)	1898 (1857, 1938)	0.19 (0.08, 0.29)	−12 355.59
PhyML consensus tree	1950 (1928, 1972)	1954 (1935, 1973)	1916 (1876, 1956)	1873 (1809, 1937)	0.15 (0.05, 0.25)	−12 411.96

Notes: Rate μ is measured by the number of substitutions per site per 100 years. The analysis is conducted using ML under the HKY+Γ_5 model, with ℓ to be the log likelihood under the model.

TABLE 2. Bayesian estimates and 95% credibility intervals of HIV-2 divergence times under the clock using the RAxML tree and different priors on the root age

Model and root-age prior	Subtype A	Subtype B	Subtypes B–C	Root	μ
Prior					
$t_{mrca} \sim U(0.5, 2.0)$ (1795, 1945)	1942 (1916, 1961)	1961 (1936, 1980)	1950 (1922, 1973)	1872 (1797, 1943)	0.20 (0.02, 0.56)
$t_{mrca} \sim U(0.2, 1.5)$ (1845, 1975)	1952 (1922, 1978)	1967 (1941, 1983)	1959 (1928, 1981)	1918 (1848, 1975)	0.20 (0.02, 0.56)
$t_{mrca} \sim U(0.8, 2.5)$ (1745, 1915)	1938 (1913, 1957)	1960 (1934, 1979)	1948 (1920, 1971)	1831 (1748, 1912)	0.20 (0.02, 0.56)
Posterior—strict clock					
$t_{mrca} \sim U(0.5, 2.0)$ (1795, 1945)	1956 (1946, 1965)	1961 (1953, 1968)	1939 (1923, 1951)	1906 (1879, 1927)	0.23 (0.17, 0.29)
$t_{mrca} \sim U(0.2, 1.5)$ (1845, 1975)	1956 (1946, 1965)	1961 (1952, 1968)	1939 (1922, 1952)	1906 (1879, 1927)	0.23 (0.17, 0.29)
$t_{mrca} \sim U(0.8, 2.5)$ (1745, 1915)	1955 (1945, 1962)	1960 (1952, 1966)	1937 (1922, 1947)	1902 (1878, 1916)	0.22 (0.17, 0.26)
Posterior—correlated rates model					
$t_{mrca} \sim U(0.5, 2.0)$ (1795, 1945)	1956 (1944, 1967)	1964 (1952, 1974)	1942 (1922, 1958)	1901 (1861, 1933)	0.26 (0.15, 0.42)
$t_{mrca} \sim U(0.2, 1.5)$ (1845, 1975)	1957 (1944, 1967)	1964 (1952, 1974)	1942 (1923, 1958)	1901 (1864, 1933)	0.26 (0.15, 0.42)
$t_{mrca} \sim U(0.8, 2.5)$ (1745, 1915)	1955 (1943, 1964)	1963 (1951, 1972)	1940 (1922, 1954)	1895 (1861, 1915)	0.24 (0.15, 0.36)

Notes: Rate μ is measured by the number of substitutions per site per time unit (100 years). The RAxML ML tree (Table 1) is used in the Bayesian analysis. The results for the prior $t_{mrca} \sim U(0.5, 2.0)$ are shown in Figure 2.

and PhyML. The confidence intervals are calculated as MLE$\pm$2 SE, with the Hessian matrix (the observed information matrix), calculated numerically, used to approximate the variance–covariance matrix. The results for the Lemey et al. tree are identical to those published by Lemey et al. (2003). The RAxML and PhyML trees are very similar and the results for them are also very similar, with the consensus trees having larger confidence intervals. The age estimates obtained from the Lemey et al. tree are older. The most conspicuous difference is that the use of the multifurcating trees led to much wider confidence intervals. For example, the t_{mrca} of HIV-2 subtype A is estimated to be 1940 (1905–1975) by Lemey et al., with 70 years of uncertainty, whereas our estimate is 1958 (1945–1972), with only 27 years of uncertainty (and when using the consensus tree our uncertainty is 35 years). Similarly, for other node ages, our confidence intervals are much narrower than and nested within those of Lemey et al.

We then used the ML tree (the RAxML tree) for Bayesian divergence time estimation, using the method developed in this study. The sequence dates range from 1995 to 1982, so that the most recent date is set to 0, while the oldest sequence has time 0.13, with one time unit to be 100 years. We set $\lambda = 2, \mu = 1, \rho = 0$, and $\psi = 1.8$. We specify a soft uniform prior on the age of the root: (0.5, 2.0), which means that the root age is from 1795 to 1945. We refer to those prior settings as the standard prior. We introduce some variations to the standard prior to evaluate the impact of the prior on posterior time estimation. For example, we used two other uniform priors on the root age: (0.2, 1.5), meaning that the root age is from 1845 to 1975; and (0.8, 2.5), meaning that the root age is from 1745 to 1915. In each case, the bounds are soft with the tail probability set at 1% [see Yang and Rannala (2006); Fig. 2c]. We calculate the likelihood under the HKY85+Γ_5 model using the likelihood approximation of dos Reis and Yang (2011).

First, we assumed the molecular clock. We used the gamma prior $G(2, 10)$ for the substitution rate, with mean at $2/10 = 0.2$ changes per site per time unit or 0.002 changes per site per year. The results are summarized in Table 2. The first two priors on the root age, $U(0.5, 2.0)$ and $U(0.2, 1.5)$, produced nearly identical posterior time estimates. These are also close to the MLEs (Table 1, RAxML tree). For example, the MLE of the age of subtype A is 1958 (1945, 1972), whereas it is 1956 (1946, 1965) in the Bayesian analysis. The 95% Bayesian credibility intervals (CIs) are in general narrower than the confidence intervals, probably due to the use of the prior on the evolutionary rate in the Bayesian analyses. Note that the ML confidence intervals and the Bayesian credibility intervals are based on different philosophical interpretations, although they may be numerically similar in many applications. The third prior on root age, $U(0.8, 2.5)$, is somewhat unrealistic as its younger age bound (1915) is too old. The posterior CI for the root age slightly violates this bound. Nevertheless,

the estimates are very similar to those obtained from the other two priors, especially for the non-root node ages. The posterior estimates appear to be quite robust to the prior on the root age in this analysis.

Relaxing the clock assumption by using the correlated-rates model caused very small changes to the posterior time estimates for the major clades, with slightly wider CIs (Table 2). For example, the age of subtype A became 1956 (1944, 1967), compared with 1956 (1946, 1965) under the clock, and the age of subtype B became 1964 (1952, 1974), compared with 1961 (1953, 1968) under the clock. The rate-drift parameter σ^2 is estimated to be about 0.26 (0.15, 1.8).

We also examined the impact of the BDSS prior on the posterior time estimates by changing the parameters in the BDSS model. We multiplied the parameters in the BDSS model λ, μ, and ψ by 2 and 0.5 to generate two new priors. The parameters are thus (a) $\lambda = 2, \mu = 1, \psi = 1.8$; (b) $\lambda = 4, \mu = 2, \psi = 3.6$; and (c) $\lambda = 1, \mu = 0.5, \psi = 0.9$. The prior means of node ages and the 95% CIs are shown in Supplementary Figure S1, while the corresponding posterior results are shown in Supplementary Figure S2. Increasing the parameters caused the ages of the non-root nodes to become younger, so that the posterior age estimates for prior (b) are younger than those for prior (a) although the intervals overlap. Decreasing the parameters cause the variance in the prior to increase so that the prior intervals for prior (c) are wide. As a result, the posterior intervals are also much wider for prior (c) than those for the standard prior (a). The results suggest that the posterior time estimates are sensitive to the parameters in the BDSS model or to the prior of divergence times.

The results suggest (i) that the BDSS model is very flexible and, with different parameter values, can generate widely different prior distributions for times, and (ii) that our approximation strategy in conditioning on the tree topology has been successful, so that the effective marginal prior on the root age is often very close to the user-specified prior on the root age.

Compared with the ML analysis of Lemey et al. (2003), our Bayesian estimates in Table 2 are much more precise, with the credible intervals nested within the confidence intervals of Lemey et al. (2003). Our results are consistent with the hypothesis that the expansion of HIV-2 clades coincided with the independence war in Guinea-Bissau (1963–1974), suggesting that war-related changes in sociocultural patterns may have had a major impact on the HIV-2 epidemic.

Divergence Times of Influenza Viral Strains

The second data set we analyze consists of 289 influenza A H1 gene sequences from the human, swine, and avian hosts, compiled by dos Reis et al. (2011). The alignment has 1710 sites. The sampling spans 91 years from the earliest sequence of 1918 human pandemic virus to 2009. The human viruses apparently became extinct in 1957 and reappeared in 1977, with newly appearing strains being nearly identical to a Russian virus from 1954 (Smith et al. 2009). To account for this lack of evolution when the virus was frozen in the laboratory, we subtracted 23 years from all modern human viruses sampled after 1977. Nevertheless, two human viral sequences from the 2009 pandemic are part of the classic swine clade so their ages were not reduced. The data were analyzed using both MCMCtree and BEAST.

For MCMCtree analysis, we used the ML tree inferred using PhyML (Guindon and Gascuel 2003) by dos Reis et al. (2011). BEAST estimates the tree topology during the MCMC, and the generated posterior tree had the same major clades as the ML tree. As far as possible we used the same substitution model and priors in the two programs. We used the HKY+Γ_5 model of nuclear substitution (Hasegawa et al. 1985; Yang 1994b), although the likelihood is calculated using an approximate algorithm in MCMCtree (dos Reis et al. 2011) and the exact pruning algorithm in BEAST (Felsenstein 1981; Yang 1994b). The independent-rates model is used to relax the clock, with rates to be random variables from a log-normal distribution. The overall rate is assigned a gamma prior $r \sim G(2, 1000)$, with the mean mutation rate to be 0.002 substitutions/site/year. The rate-drift parameter σ^2 is parameterized differently in the two programs, with BEAST using σ and MCMCtree using σ^2. We assign a gamma prior $\sigma \sim G(4, 20)$, with mean 0.2 for BEAST and a gamma prior $\sigma^2 \sim G(1, 20)$ with mean 0.05 for MCMCtree. The time prior is generated by the BDSS process with $\lambda = 2$, $\mu = 1$, $\psi = 1.8$, $\rho = 0$ in MCMCtree, while for BEAST, it is generated using a constant-population coalescent process, which was one of the standard tree-generating models in BEAST [note that recently an alternative birth–death-based prior became available though (Stadler et al. 2013)]. The root age (root height) is assigned a uniform prior between 100 and 500 years before 2009. Those bounds are hard in BEAST and sharp (with tail probabilities 0.1%) in MCMCtree. We used preliminary runs to determine the length of the Markov chain to ensure that different runs of the same analysis produced consistent results. For MCMCtree, we ran the chain for 2×10^5 iterations, sampling every two iterations. For BEAST, we ran 10^8 iterations, sampling every 10^4 iterations. Those numbers are not comparable between programs as they depend on the details of the MCMC algorithms which may have very different mixing efficiencies. The same analysis is run at least twice, to confirm that the results are consistent between runs. Running time is several hours for MCMCtree and 1–2 weeks for BEAST.

The time tree with posterior means and 95% CIs of divergence times obtained using our new BDSS prior and the independent-rates model is shown in Figure 3. Table 3 lists the posterior means and the 95% CIs for several key nodes obtained under different rate-drift models implemented in MCMCtree, including the strict clock, the independent-rates model and the auto-correlated rates model. Estimates from BEAST under the independent-rates model are listed in the table as

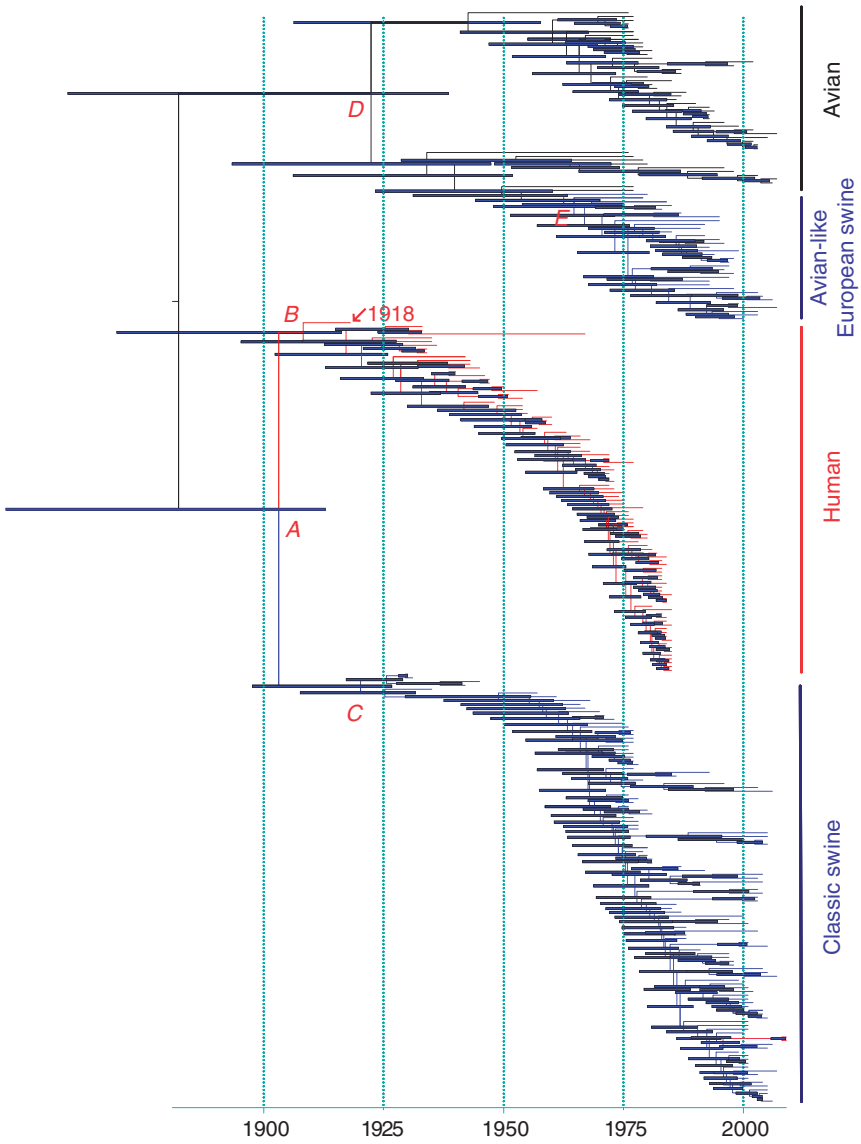

FIGURE 3. Time tree showing posterior estimates of divergence times obtained under the BDSS prior and the independent-rates model implemented in MCMCtree. A few major clades are labeled, such as the Human-Classical Swine clade (*A*), the Human clade (*B*), the Classical Swine clade (*C*), the Avian-European Swine clade (*D*), and European Swine clade (*E*). Estimates of the ages of those clades are shown in Table 3. Sampling times range from 2009 to 1918 (which is indicated by arrow).

TABLE 3. Bayesian estimates and 95% credibility intervals of influenza divergence times estimated using MCMCtree and BEAST

Node[a]	MCMCtree Prior	MCMCtree clock	MCMCtree independent rates	MCMCtree auto-correlated rates	BEAST independent rates
Root	1868 (1569, 1910)	1878 (1864, 1890)	1867 (1595, 1910)	1733 (1584, 1817)	1886 (1850, 1909)
Human-classical swine (node *A*)	1895 (1806, 1911)	1907 (1903, 1910)	1898 (1840, 1913)	1813 (1760, 1855)	1910 (1903, 1917)
Human clade (node *B*)	1901 (1841, 1914)	1909 (1906, 1913)	1905 (1866, 1916)	1832 (1787, 1867)	1914 (1911, 1918)
Classical swine (node *C*)	1903 (1847, 1917)	1926 (1924, 1928)	1918 (1895, 1927)	1889 (1857, 1911)	1925 (1923, 1929)
Avian-European swine (node *D*)	1903 (1840, 1922)	1941 (1936, 1946)	1917 (1852, 1938)	1863 (1815, 1899)	1953 (1945, 1961)
European swine (node *E*)	1940 (1921, 1956)	1973 (1971, 1975)	1961 (1939, 1970)	1942 (1923, 1959)	1977 (1975, 1978)
Rate (μ)[b]	0.200 (0.024, 0.556)	0.170 (0.161, 0.180)	0.137 (0.087, 0.170)	0.174 (0.050, 0.393)	
Rate-drift parameter (σ^2)			0.502 (0.341, 0.773)	1.597 (1.193, 2.073)	

[a]Node refers to the tree of Figure 3.
[b]Rate μ is measured by the number of substitutions per site per 100 years.

well. The prior means and 95% CIs of divergence times used in the MCMCtree and BEAST analyses are shown in Supplementary Figures S3 and S4. The MCMCtree prior is generated by using the control variable usedata = 0, while the BEAST prior is generated by replacing each sequence in the alignment by a question mark. As no topological constraints are applied in the BEAST analysis, most nodes in the tree of Supplementary Figure S4 have low prior probabilities. However, it is interesting to note that the sampling times in the sequences have caused BEAST to assign fairly strong preferences for rooting the tree around the earliest sampled sequences: the root splits the single earliest 1918 sequence from the rest, with a prior probability 0.90.

Node *A* represents the divergence of the human viral sequences from the classical swine clade (Fig. 3) and may indicate the jump of the virus from the swine host to the human before the 1918 pandemic. Under the molecular clock, MCMCtree estimated the date of node *A* to be around 1907, with the 95% CIs to be (1903, 1910) (Table 3 and Supplementary Fig. S5). Under the independent-rates model, MCMCtree dated node *A* to 1898 (1840, 1913), with much greater uncertainty. The correlated-rates model produced much older and more uncertain estimates: 1813 (1760, 1855) (Table 3 and Supplementary Fig. S6). As the posterior CI of σ^2 excludes 0, there appear to be considerable rate variation among lineages, so the estimates under the strict clock may be unreliable.

We also run the MCMCtree analyses using BDSS parameters $\lambda = 0.2, \mu = 0.1$, and $\psi = 10^{-6}$. The small sampling intensity may be considered more realistic than the value used above. Those BDSS parameters lead to posterior time estimates that are more recent, although the estimates, especially those under the relaxed-clock models, involve considerable uncertainties (results not shown). Overall, the posterior estimates are sensitive to the BDSS parameters used.

As mentioned earlier, the posterior tree from BEAST agrees with the ML tree we used and shares all major clades. Under the independent-rates model, BEAST produced date estimates that are similar to the estimates under the strict clock in MCMCtree, with very litle uncertainty (Supplementary Fig. S7). The age of the European swine clade is estimated to be 1977 (1975, 1978), which is extremely precise. The date for node *A* is estimated to be 1910 (1903, 1917) (Table 3). The estimates are similar to those obtained from a different data set by Smith et al. (2009).

All analyses summarized in Table 3 suggest that the influenza virus may have entered the human population at least several years prior to the 1918 pandemic, consistent with the conclusion of Smith et al. (2009). The MCMC time estimates under relaxed clock models, however, involve much greater uncertainties than the BEAST estimates. The large differences in date estimates under the different rate-drift models implemented in MCMCtree highlight the sensitivity of posterior date estimates to the prior model assumptions about rates. Estimates of divergence times by BEAST appear to be overconfident with too narrow credibility intervals. A few recent studies argue that rate shifts at deep time scales may mislead inferences of absolute rates and ages by BEAST, producing time estimates that are precise but not accurate (Dornburg et al. 2012; Wertheim et al. 2012). Our results appear to be consistent with those studies.

Discussion

The prior of times we developed in this article is conditioned on the root age t_{mrca}, and the user is required to specify a diffuse prior on t_{mrca}. In dating species phylogenies using uncertain fossil calibrations under relaxed clock models, we found that the number of sampled tips alone provides only very weak information about the root age, and specifying a diffuse prior on t_{mrca} is deemed beneficial (Rannala and Yang 2007). Here in dating viral divergences using sequentially sampled sequences, the sampling times (z) may be informative about the root age, if the root is not much older relative to the oldest sampled sequences, and if the sampling model or the assumption of a constant sampling intensity (ψ) is realistic. In situations like this, the alternative of using the BDSS model to specify the distribution of the root age, which will remove the burden for the user to specify such a prior, may be attractive and merits further investigation.

Sampling intensity may be stronger now than a few decades ago. Furthermore, samples appear to be taken in batches. The process may be better described by a compound Poisson process, with a Poisson process of sampling events and a distribution of the number of samples given that a sampling event occurs. The reliability of the constant-rate sampling model and its impact on the prior of times is also an interesting question for further research.

We used different uniform calibrations on the root age and different values of parameters λ, μ, and ψ in the BDSS model in the analysis of the HIV-2 data set to evaluate the impact of the prior for times on posterior time estimation. The posterior time estimates are found to be quite robust to the prior on the root age but are sensitive to the time prior, especially to the shape of the tree as influenced by parameters in the BDSS model. This sensitivity may be the nature of this kind of dating analysis, in which the sequence data provide information about distances, which are resolved into absolute times and rates only through the assistance of the prior. For dating species divergences using uncertain fossil calibrations, the estimation problem is only semi-identifiable; even if the amount of sequence data approaches infinity, the posterior time estimates will still have considerable uncertainties (Yang and Rannala 2006; Rannala and Yang 2007; dos Reis and Yang 2013). The situation with dating viral divergences using sequentially sampled sequences appears to be slightly better: if the molecular clock holds, the problem is clearly identifiable and unlimited increase of sequence data will reduce the errors in posterior time estimates to zero. The

case is less clear when the clock is violated and a relaxed clock model is used. At any rate, the sensitivity of the posterior time estimates to the different rate-drift models found in the analysis of the influenza data set can be easily explained by this confounding effect of rates and times. In the HIV-2 data set, the sampling times in the sequences cover 13 years, and the age of the root is in the order of 100 years old. In the influenza data set, the sequences span 91 years and the age of the root is in the order of 100–150 years old. The situation should be much worse when one tries to date more ancient events. We suggest that it is important to assess the impact of the prior on the posterior time estimates, for example, by varying parameters in the BDSS model.

In summary, in our implementation in MCMCtree, the BDSS model is used to generate a flexible prior of times. Parameters in the BDSS model and the sampling times z are fixed. Use of different values for those parameters provides a way to generate different time priors to assess the robustness of Bayesian divergence time estimation.

Supplementary Material

Supplementary material, including data files and online-only appendices, can be found in the Dryad data repository (DOI:10.5061/dryad.9c568).

Funding

A Biotechnological and Biological Sciences Research Council (UK) Grant (to Z.Y.); and a Royal Society/Wolfson Merit Award (to Z.Y.) and a Swiss National Science Foundation grant (to T.S.).

Acknowledgments

We thank Drs Phillippe Lemey and Mario dos Reis for providing the SIV/HIV-2 and influenza A data sets analyzed in this article. Both data sets are included in the PAML release, which is available for download at http://abacus.gene.ucl.ac.uk/software/. We thank three anonymous referees and the Associate Editors (Olivier Gascuel and Laura Kubatko) for many constructive comments.

Appendix A

Proof of Theorem 2.1

Consider the branching event x_i with the two sampling points at time z_i, z_{i+1} (see Fig. 1). Let $s_j = \psi$ if z_j is an extinct sample and $s_j = \rho$ if z_j is an extant sample. The probability density of a lineage at time t_{mrca} giving rise to two sampled descendants at time points z_i and z_{i+1} with branching time x_i is derived in Stadler (2010), Theorem 3.5 (except that here we ignore the term $p_0(t)$ as we assume that sampled lineages have no sampled descendants):

$$h[x_i, z_i, z_{i+1}|t_{\mathrm{mrca}}] = \lambda s_i s_{i+1} \frac{q(z_i)q(z_{i+1})}{q(t_{\mathrm{mrca}})q(x_i)}, \tag{A.1}$$

with

$$q(t) = \left(e^{-\frac{c_1 t}{2}}(1-c_2) + e^{\frac{c_1 t}{2}}(1+c_2)\right)^2. \tag{A.2}$$

Let $z_i^* = \max\{z_i, z_{i+1}\}$. The probability density of a lineage at time t_{mrca} giving rise to two sampled descendants at z_i and z_{i+1} with an arbitrary branching time is

$$\begin{aligned} H[z_i, z_{i+1}|t_{\mathrm{mrca}}] &= \int_{z_i^*}^{t_{\mathrm{mrca}}} h[x_i, z_i, z_{i+1}|t_{\mathrm{mrca}}]dx_i \\ &= \lambda s_i s_{i+1} \frac{q(z_i)q(z_{i+1})}{q(t_{\mathrm{mrca}})} \left(Q(t_{\mathrm{mrca}}) - Q(z_i^*)\right), \end{aligned} \tag{A.3}$$

where

$$Q(x) = \int \frac{1}{q(x)} \mathrm{d}x = \frac{1}{c_1(1-c_2)\left(e^{-c_1 x}(1-c_2) + (1+c_2)\right)}. \tag{A.4}$$

Thus, given the samples at time z_i and z_{i+1} diverged from a common ancestor more recent than t_{mrca}, the time of divergence has probability density function

$$\begin{aligned} f[x_i|z_i, z_{i+1}, t_{\mathrm{mrca}}] &= \frac{h[x_i, z_i, z_{i+1}|t_{\mathrm{mrca}}]}{H[z_i, z_{i+1}|t_{\mathrm{mrca}}]} \\ &= \frac{1}{q(x_i)(Q(t_{\mathrm{mrca}}) - Q(z_i^*))}. \end{aligned} \tag{A.5}$$

Thus, the probability density of x is

$$\begin{aligned} f[x|z, k, t_{\mathrm{mrca}}] &= \prod_{i=1, i\neq k}^{N-1} f[x_i|z_i, z_{i+1}, t_{\mathrm{mrca}}] \\ &= \prod_{i=1, i\neq k}^{N-1} \frac{1}{q(x_i)(Q(t_{\mathrm{mrca}}) - Q(z_i^*))} \\ &= \prod_{i=1, i\neq k}^{N-1} \frac{c_1(1-c_2)e^{-c_1 x_i}}{g(x_i)^2((1/g(t_{\mathrm{mrca}})) - (1/g(z_i^*)))}, \end{aligned}$$

which establishes the theorem.

Appendix B

Validation of Implementation

The BDSS prior is mathematically complex so that it is not a trivial task to validate the correctness of the theoretical derivations or of the program implementation. We have conducted numerous tests, which indicate the correctness of the theory and program implementation. Some of those are described here. Note, however, that they are not a proper evaluation of the

statistical performance of the dating method. The latter may require analysis of many empirical data sets or well-planned computer simulation, which is beyond the scope of this article.

Comparison of the three priors for $\psi=0$. When $\psi=0$, all sampling times (z) will be 0, and thus $f[T|z,k,t_{\text{mrca}}]$ and $f[T,z|n,t_{\text{mrca}}]$ will be uniform distributions, as each ranked tree is equally likely (Aldous 2001); that is, the probability of the tree will be independent of t_{mrca}. Thus, the user-specified prior is not changed, and we expect the same results as with using the method in Yang and Rannala (1997).

Indeed, the prior density for the three priors should be the same for different T,z (up to a constant as the densities are not normalized but the ratios of the densities should be the same among the approaches). MCMC runs with the different approaches yielded the same distribution of x. In particular, the user-specified prior for t_{mrca} was recovered.

Comparison of Approaches 1 and 2 for $\psi>0$. Approaches 1 and 2 must yield the same results when t_{mrca} is fixed. We calculated the prior with fixed t_{mrca}, and, as expected, Approaches 1 and 2 yielded the same density values (up to a constant as the densities are not normalized so that only the ratios of densities are the same between approaches). Note that when t_{mrca} is different, the density ratios are different between approaches, as the densities are calculated only up to a function of t_{mrca} ($f[T|z,k,t_{\text{mrca}}]$ and $f[T,z|n,t_{\text{mrca}}]$).

In order to investigate how different the effective prior for t_{mrca} is from the user-specified prior, we ran several MCMC chains using different priors. Our test tree has five sequences, with the topology ((((a@1990, b@2000), c@1950), d@1990), e@1940), so the sampling times of the sequences are $z=(0.1,0,0.5,0.1,0.6)$. The node ages are $t_1=t_{\text{mrca}}=1.0$, $t_2=0.8$, $t_3=0.6$, and $t_4=0.4$.

- Using a uniform prior on $[0.6,1.4]$ on the root age yielded a mean t_{mrca} of 1.034 with Approach 1 and 0.884 with Approach 2.
- Using a uniform prior on $[5.6,6.4]$ yielded a mean t_{mrca} of 6.006 with Approach 1 and 5.805 with Approach 2.

In general, the effective prior using Approach 1 was closer to the user-specified prior than using Approach 2. Furthermore, for older t_{mrca}, (i) the effective prior approaches the user-specified prior in Approach 1; (ii) the effective prior predicts younger t_{mrca} than the user-specified prior in Approach 2. The reason for (i) is that for older t_{mrca}, all branching events occur at more ancient times than any sampling dates, and thus the probabilities for the ranked trees (labeled histories) approach the uniform distribution (Aldous 2001). Thus, $f[T|z,k,t_{\text{mrca}}]$ approaches a constant, independent of t_{mrca}. The reason for (ii) is that for very old t_{mrca}, the quantity $f[T,z|n,t_{\text{mrca}}]$ becomes very small (compared with smaller t_{mrca}) as with old t_{mrca} we expect older sampling times z and more extinct samples $N-n$. Thus, the effective prior is pushed to be younger than the user-specified prior.

The pattern is similar on our empirical trees. In the analysis of the influenza gene sequences, we specified the prior on root age as $t_1\ U(1,5)$. The mean and 95% CI for the effective prior, generated by running MCMCtree without using sequence data, is 1.41 (0.99, 4.40) or 1868 (1569, 1910) (Table 3). Although the CI covers most of the range of the specified bounds, the effective prior is shifted to young ages, concentrated around the prior mean, and differs considerably from the specified prior. If we specify much older bounds, for example, $t_1\ U(2,5)$ or $t_1\ U(4,6)$, the effective prior is nearly flat between the bounds and matches the specified prior.

APPENDIX C

MCMC Proposal for Updating Node Ages

A mixing step used in MCMCtree for the divergence times x, which multiplies all node ages by the same random variable c that is close to 1 and which thus proportionally expands or shrinks all node ages (Thorne et al. (1998); Yang and Rannala (2006); Yang (2006), p. 170–171), does not work anymore when the sequences are sequentially sampled. This move has been noted to be effective in improving convergence of the chain, as it can bring all node ages to the right range if they are all too small or too large. It also improves mixing after the algorithm has converged, because the move accounts for the positive correlation among the node ages. When all sequences are contemporary, multiplying all node ages by $c>0$ and dividing all rates by c will keep the likelihood of the sequence data unchanged. For non-contemporary sequences considered in this article, such changes to times and rates do change the likelihood. Nonetheless, strong positive correlation among the node ages and strong negative correlation between times and rates are still expected.

We have modified the proposal so that it works with models for sequentially sampled sequences. Again, let t be the order statistic of the branching times x, that is, $t_1>\cdots>t_{N-1}$, with t_1 to be the age of the root. Suppose the minimum age bound for the branching event at time t_j is b_j: this is the maximum (oldest) sample date among the descendent sequences of that node (b_j is one of the sample dates in vector z defined above). For each non-root node j, define the "relative age" as

$$h_j=\frac{t_j-b_j}{t_{a(j)}-b_j},\quad j=2,3,\ldots,(N-1), \tag{C.1}$$

where $a(j)$ is the ancestral (mother) node of node j. Note that t_j lies in the interval $(b_j,t_{a(j)})$, and its relative position in the interval is represented by h_j, with $0<h_j<1$. Our proposal changes the root age t_1 but keeps the relative ages $h_j, j=2,3,\ldots,(N-1)$, unchanged. The proposal is thus one-dimensional even if it changes all node ages. We generate a new root age through a sliding window

at the logarithmic scale:

$$t_1^* = t_1 \cdot c = t_1 \cdot e^{\epsilon(u-\frac{1}{2})}, \quad \text{(C.2)}$$

where $u \sim U(0,1)$ is a random number, and ϵ is a fine-tuning step length. The new ages of the non-root nodes are generated by using h_j of Equation (C.1), with ancestral nodes visited before descendent nodes:

$$t_j^* = b_j + h_j(t_{a(j)}^* - b_j), \quad j = 2,3,\ldots,(N-1). \quad \text{(C.3)}$$

To derive the proposal ratio, consider $(t_1, h_2, h_3, \ldots, h_{N-1})$ as a variable transform or re-parameterization of the original parameters $(t_1, t_2, \ldots, t_{N-1})$, with the Jacobi of the transform to be

$$J(h) = \left| \frac{\partial(t_1, t_2, \ldots, t_{N-1})}{\partial(t_1, h_2, \ldots, h_{N-1})} \right| = \prod_{j=2}^{N-1} (t_{a(j)} - b_j). \quad \text{(C.4)}$$

For the transformed variables, the move from the current state $(t_1, h_2, h_3, \ldots, h_{N-1})$ to the new state $(t_1^*, h_2, h_3, \ldots, h_{N-1})$ incurs a proposal ratio $c = e^{\epsilon(u-\frac{1}{2})}$. Thus, the proposal ratio for the original variables $(t_1, t_2, \ldots, t_{N-1})$ is,

$$c \times \frac{J(h^*)}{J(h)} = c \times \prod_{j=2}^{N-1} \frac{t_{a(j)}^* - b_j}{t_{a(j)} - b_j}. \quad \text{(C.5)}$$

See, for example, Yang (2006): Equation (A.10) for the theory for deriving the proposal ratio using transformed variables.

If all sequences are contemporary, with $b_j = 0$ for all j, this proposal becomes the old proposal that multiplies all $(N-1)$ node ages by c, as $t_j^* = ct_j$ for all j, with the proposal ratio to be c^{N-1}. The new proposal is thus an extension to the old proposal.

A variation to this proposal is to divide all rates in the model by c, in addition to changing the node ages as above, in which case the proposal ratio of Equation (C.5) should be divided by c^r, where r is the number of rate parameters in the model (Yang 2006, p. 170–171). When all sequences are contemporary, changing the rates this way will ensure that the branch lengths and thus the likelihood function remain unchanged. When the sequences have sample dates, the branch lengths and likelihood function do change. Nevertheless, changing the rates at the same time appears to cause smaller changes to the log likelihood and to lead to a more efficient proposal. For the analyzed SIV/HIV-2 data set, it was noted that the optimal step length (that which achieves the acceptance proportion of about 0.3) is about $\epsilon = 0.08$ when the proposal changes the times but not the rates, and is about $\epsilon = 0.20$ when the proposal changes the rates as well as the times. Larger steps mean better traversal of the posterior density and thus a more efficient proposal. The difference is expected to be greater with larger data sets in which the likelihood surface is sharper.

REFERENCES

Aldous D.J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statist. Sci. 16:23–34.

Battistuzzi, F., Billing-Ross P., Paliwal A., Kumar S. 2011. Fast and slow implementations of relaxed-clock methods show similar patterns of accuracy in estimating divergence times. Mol. Biol. Evol. 28: 2439–2442.

Dornburg A., Brandley M.C., McGowen M.R., Near T.J. 2012. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (mammalia: Cetacea). Mol. Biol. Evol. 29:721–736.

dos Reis M., Tamuri A.U., Hay A.J., Goldstein R.A. 2011. Charting the host adaptation of influenza viruses. Mol. Biol. Evol. 28: 1755–1767.

dos Reis M., Yang Z. 2011. Approximate likelihood calculation for Bayesian estimation of divergence times. Mol. Biol. Evol. 28: 2161–2172.

dos Reis M., Yang Z. 2013. The unbearable uncertainty of Bayesian divergence time estimation. J. Syst. Evol. 51:30–43.

Drummond A., Pybus O., Rambaut A., Forsberg R., Rodrigo A. 2003. Measurably evolving populations. Trends Ecol. Evol. 18:481–488.

Drummond A., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Edwards A.W.F. 1970. Estimation of the branch points of a branching diffusion process (With discussion). J. R. Stat. Soc. B 32:155–174.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Ford D., Matsen E., Stadler T. 2009. A method for investigating relative timing information on phylogenetic trees. Syst. Biol. 58:167–183.

Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–778.

Guindon S. 2013. From trajectories to averages: an improved description of the heterogeneity of substitution rates along lineages. Syst. Biol. 62:22–34.

Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Heled J., Drummond A. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. Syst. Biol. 61: 138–149.

Inoue J., Donoghue P., Yang Z. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. Syst. Biol. 59:74–89.

Lemey P., Pybus O., Wang B., Saksena N., Salemi M., Vandamme A. 2003. Tracing the origin and history of the HIV-2 epidemic. Proc. Natl Acad. Sci. U. S. A. 100:6588–6592.

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

Rannala B., Yang Z. 2007. Inferring speciation times under an episodic molecular clock. Syst. Biol. 56:453–466.

Slowinski J. 1990. Probabilities on n-trees under two models: a demonstration that asymmetrical interior nodes are not improbable. Syst. Zool. 39:89–94.

Smith G.J., Bahl J., Vijaykrishna D., Zhang J., Poon L.L., Chen H., Webster R.G., Peiris J.S., Guan Y. 2009. Dating the emergence of pandemic influenza viruses. Proc. Natl Acad. Sci. U.S.A. 106: 11709–11712.

Stadler T. 2010. Sampling-through-time in birth-death trees. J. Theor. Biol. 267:396–404.

Stadler T. 2013. How can we improve accuracy of macroevolutionary rate estimates? Syst. Biol. 62:321–329.

Stadler T., Kouyos R., von Wyl V., Yerly S., Böoni J., Bürgisser P., Klimkait T., Joos B., Rieder P., Xie D., Günthard H.F., Drummond A.J., Bonhoeffer S.; Swiss HIV Cohort Study. 2012. Estimating the basic reproductive number from viral sequence data. Mol. Biol. Evol. 29:347–357.

Stadler T., Kühnert D., Bonhoeffer S., Drummond A.J. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis c virus (HCV). Proc. Natl Acad. Sci. U. S. A. 110: 228–233.

Stamatakis A., Ludwig T., Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.

Tamura K., Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

Thorne J., Kishino H., Painter I. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.

Wertheim J.O., Fourment M., Kosakovsky Pond S.L. 2012. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. Mol. Biol. Evol. 29:451–456.

Wilkinson R., Steiper M., Soligo C., Martin R., Yang Z., Tavaré S. 2011. Dating primate divergences through an integrated analysis of palaeontological and molecular data. Syst. Biol. 60: 16–31.

Yang Z. 1994a. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

Yang Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z. 2006. Computational molecular evolution. Oxford, UK: Oxford University Press.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z., Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 17:717–724.

Yang Z., Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. 23:212–226.