

## A Simulation Approach for Change-Points on Phylogenetic Trees

ADAM PERSING,<sup>1</sup> AJAY JASRA,<sup>2</sup> ALEXANDROS BESKOS,<sup>1</sup>  
DAVID BALDING,<sup>3</sup> and MARIA DE IORIO<sup>1</sup>

### ABSTRACT

We observe  $n$  sequences at each of  $m$  sites and assume that they have evolved from an ancestral sequence that forms the root of a binary tree of known topology and branch lengths, but the sequence states at internal nodes are unknown. The topology of the tree and branch lengths are the same for all sites, but the parameters of the evolutionary model can vary over sites. We assume a piecewise constant model for these parameters, with an unknown number of change-points and hence a transdimensional parameter space over which we seek to perform Bayesian inference. We propose two novel ideas to deal with the computational challenges of such inference. Firstly, we approximate the model based on the time machine principle: the top nodes of the binary tree (near the root) are replaced by an approximation of the true distribution; as more nodes are removed from the top of the tree, the cost of computing the likelihood is reduced linearly in  $n$ . The approach introduces a bias, which we investigate empirically. Secondly, we develop a particle marginal Metropolis-Hastings (PMMH) algorithm, that employs a sequential Monte Carlo (SMC) sampler and can use the first idea. Our time-machine PMMH algorithm copes well with one of the bottlenecks of standard computational algorithms: the transdimensional nature of the posterior distribution. The algorithm is implemented on simulated and real data examples, and we empirically demonstrate its potential to outperform competing methods based on approximate Bayesian computation (ABC) techniques.

**Key words:** approximate Bayesian computation, binary trees, change-point models, particle marginal Metropolis-Hastings, sequential Monte Carlo samplers, time machine.

**A** PHYLOGENY (OR EVOLUTIONARY TREE) can explain the ancestral relationships among species based on similarities in their genetic sequences (e.g., DNA). A phylogenetic model will typically be parameterized by rates that can correspond to genetic mutations that occur within populations as they evolve over time. In many applications, it is reasonable to assume that genetic sequences share a common phylogenetic structure across all of their sites. However, to allow for greater modeling flexibility, it is often times desirable to allow for the evolutionary rate parameters to change across the length of the sequences. The main focus of

---

<sup>1</sup>Department of Statistical Science, University College London, London, United Kingdom.

<sup>2</sup>Department of Statistics & Applied Probability, National University of Singapore, Singapore, Singapore.

<sup>3</sup>Genetics Institute, University College London, London, United Kingdom.

this article is estimating that rate variation across the sites. We consider a model that allows for neighboring blocks of sites to be parametrized by different evolutionary rates (i.e., a change-point model), and we propose a novel computational scheme that enables a practitioner to fit this model to genetic sequences when the true number of change-points is unknown. Thus, our contribution is a methodology that infers both the number of distinct blocks along the length of the sequence and the values of the rates themselves. We empirically demonstrate the potential of our algorithm to outperform competing computational methods.

## 2. INTRODUCTION

A phylogeny (or evolutionary tree) is the most common structure employed to explain the evolutionary relationships among species (taxa) based on similarities in their physical or (more usually) genetic characteristics. The branching pattern of the tree is usually referred to as its topology and describes shared and independent periods of evolution of different taxa. The leaves of the tree correspond to observations on the taxa. In a rooted phylogenetic tree (Fig. 1 of Appendix A), each internal node corresponds to a speciation event and represents the most recent common ancestor of all the taxa descended from that node. The length of the edges connecting the nodes (called branches) can be interpreted as the time between speciation events.

The evolutionary analysis of molecular sequence variation is statistically challenging. Parsimony methods were among the first approaches for inferring phylogenies, but in recent years, great research effort has been devoted to likelihood-based methods, both in the frequentist (Felsenstein, 1981) and Bayesian framework.

DNA sequences occupy one of four states (A, C, G, T) at each site, and so specifying the likelihood function requires a model for how these change over time at each site. The simplest such model is the Jukes-Cantor, in which each state is substituted by any other state at the points of a homogeneous Poisson process. The Kimura model has a rate for transitions ( $A \leftrightarrow G$  or  $C \leftrightarrow T$ ) that can differ from the rate for transversions (all other substitutions); see chapter 13 of Felsenstein (2004). Objects of inference can include the topology of the phylogenetic tree (here regarded as known), the relative branch lengths on the tree, and the substitution rates.

Likelihood-based approaches usually assume that substitution rates are the same at all sites, so that the likelihood is obtained as a product across sites. However, variation in substitution rates along DNA sequences is well established (Huelsenbeck and Suchard, 2007). This variation can be explained by variation in functional constraint across the genes encoded in the sequences. If the DNA sequence is from a coding region, natural selection may constrain variability at some sites more than others and therefore sites might exhibit different rates of evolution. Therefore, it is important to accommodate rate variation across sites in phylogenetic inference (Huelsenbeck and Hillis, 1993; Wakeley, 1994). One possibility is to estimate a different rate for each site (Swofford et al., 1996), but this is computationally demanding because of the large number of parameters, and the limited information per parameter leads to poor inferences. A better alternative is to assume that the rates at different sites are independent draws from a distribution, typically either a Gamma (Uzzell and Corbin, 1971; Nei et al., 1976) or a Log-Normal distribution (Olsen, 1987). A more realistic model would assume that the rates are auto-correlated along the sequence. One possible solution is offered by phylogenetic hidden Markov (phylo-HMM) models, which allow for correlated rates between nearby sites (Yang, 1995; Felsenstein and Churchill, 1996): the rate of evolution is modeled as a Markov process operating along the sequence, and site specific rates are drawn from a finite set of values. The discrete number of “rate categories” represents one limitation of the phylo-HMM approach (Yang et al., 1994; Siepel and Haussler, 2005), while another is the small number of taxa that can be accommodated with reasonable computational resources (Yang, 1993). Alternatively, Suchard et al. (2003) have developed a Bayesian multiple change-point model of rate variation along the DNA sequence, which assumes that sites are grouped into an unknown number of contiguous segments, each with possibly a different tree topology, as well different substitution rates and branch lengths. Several recent proposals involve finite mixtures of distributions to model heterogeneity across sites. In this case, the distribution of each site on the sequence is a mixture of multiple processes, each of which may have its own tree topology, branch lengths, and substitution rates (e.g., Pagel and Meade, 2004; Huelsenbeck and Suchard, 2007; Loza-Reyes et al., 2014). Wu et al. (2013) extend these ideas to infinite mixtures assuming a Dirichlet process prior.

The main focus of this article is estimating evolutionary rate variation across sites assuming that the tree topology and branch lengths are known and the same at every site under analysis. The latter assumption is not very restrictive in most applications, which involve taxa that are separated by enough time that within-taxon coalescent variation is unimportant. Although substitution rates can vary along the sequence, they are assumed to be the same across all taxa at each site. Our proposed time-machine PMMH model is able to account for quantitative differences in rates of substitutions (e.g., sites with high rates versus sites with low rates), and can also allow different rates for different types of substitution (such as transitions and transversions).

Recently there has been a revival of interest in models that allow for variation in evolutionary rates due to an explosion in the availability of comparative sequence data, and consequent interest in comparative methods for the detection of functional elements (e.g., Boffelli et al., 2003; Gibbs et al., 2004; Chinwalla et al., 2002). The model proposed in this article is similar in spirit to early work on spatial variation of evolutionary rates (e.g., Yang, 1995; Felsenstein and Churchill, 1996), which maintains a single consistent topology along the sequence but allows changes in evolutionary rates. In this framework, given the rate at each site, each site is then assumed to evolve independently along the true phylogeny with that rate and the correlation between sites arises from the clustering of high and low rates at adjacent sites. However, most of these models allow for a small discrete number of “rate categories” into which sections of the sequences are sorted (Yang et al., 1994; Siepel et al., 2005), and many methods are limited to two-species comparisons as they become increasingly computationally expensive when more species are included. Our proposed model overcomes both these difficulties, as the model for evolutionary rates, based on a multiple change-point model, is structurally simple and flexible so that the rates are not restricted to a finite set but estimated online. Moreover, the use of the “time machine” significantly speeds up computations.

### 2.1. Specific contributions

Several negative mathematical results exist in the literature (e.g., Mossel and Vigoda, 2006) for Markov chain Monte Carlo (MCMC) inference when the tree topology (and branch lengths) is unknown, and these have spurred the development of highly sophisticated Monte Carlo-based algorithms (Bouchard-Côté et al., 2012). Here, the tree topology and branch lengths are assumed to be known, but the position and number of change-points for the rates are unknown. In addition, as we will explain later, the cost of evaluating the likelihood will be an  $\mathcal{O}(mp^2n^2)$  operation ( $p$  is the number of states at each site,  $m$  the number of sites, and  $n$  the number of sequences). It follows that parameter inference requires expectations with regard to a probability on a transdimensional state-space. Constructing efficient MCMC algorithms on transdimensional spaces is a notoriously challenging problem, and the standard approach is to use reversible jump MCMC (RJMCMC) (Green, 1995). Typically, and especially for our model, it is difficult to develop moves on the transdimensional state-space that are likely to be accepted, which is important here because likelihood computations are expensive.

To deal with some of these inferential and computational issues, we propose the following:

- To reduce the cost of computing the likelihood and assist the mixing of MCMC, through a likelihood approximation based on the time-machine principle (Jasra et al., 2011).
- To improve mixing compared to standard RJMCMC, by adapting an idea in Karigiannis and Andrieu (2013), developing a particle marginal Metropolis-Hastings (PMMH) algorithm (Andrieu et al., 2010), based on the sequential Monte Carlo (SMC) samplers method in Del Moral et al. (2006). This approach can benefit from the time-machine approach.

In the time-machine approach, the unobserved sequence at the root, and possibly also other top-most nodes of the tree are replaced with the stationary distribution of the substitution process. This can reduce the cost of computing the likelihood by a linear factor in  $n$ ; this can allow larger datasets than would otherwise be manageable. The resulting estimates are biased, but in the examples below we find the bias to be smaller than for competitive methods. Indeed, we conjecture (and this is supported by empirical results) that our approach is competitive with other approximate methods, in particular approximate Bayesian computation (ABC); this latter method is often not appropriate for model selection problems as we describe in section 3.3. An important point here is that the time-machine performs a “principled” approximation of the mathematical model. This is based on the general understanding that most of the information in the data is at the lower part of the tree, thus contrasting with an often ad-hoc selection of summary statistics in ABC approaches.

Our PMMH algorithm extends the idea in Karigiannis and Andrieu (2013), both with regard to the methodology and the context of phylogenetic trees with change-points. The MCMC method will often generate (as we will explain in section 3.2) transdimensional proposals that are more likely to be accepted than standard RJMCMC algorithms. This is further aided by using the time-machine, which results in a less complex posterior with faster likelihood evaluations. The combination of the above factors can lead to reliable, but biased, inference from moderate sized data sets. As mentioned above, we expect the bias to be minimal relative to ABC methods.

This article is structured as follows. In section 3, the model and methods are described; this includes our mathematical result on the bias. In section 4, our empirical results are given. In section 5, we conclude the article with a discussion. The appendixes provide further details of the methods.

### 3. MODEL AND METHODS

We first describe our change-point model and the associated Bayesian inference problem, then the time machine approximation and the PMMH algorithm. The end of this section then briefly discusses some competing ABC methods that can also be used to perform Bayesian inference, but we make a case against using such algorithms in this context. Throughout the article, given a vector  $(x_1, \dots, x_n)$ , we define  $x_{k:l} := (x_k, x_{k+1}, \dots, x_l)$ ,  $k \leq l \leq n$ ; also, we use the notation  $[k] = \{1, 2, \dots, k\}$ .

#### 3.1. Phylogenetic model

We observe  $n$  sequences of length  $m$ , such that each observation is  $x_{ij} \in \{1, \dots, p\}$  with  $1 \leq i \leq n$ ,  $1 \leq j \leq m$ . Similar to as in Ma (2008), it is assumed that the data originate from a rooted binary tree (ch. 1 of Felsenstein, 2004) of known topology and branch lengths with the  $n$  leaves being the observed sequences. The sequences at the other  $n-1$  nodes are unobserved. Nodes are numbered backward in time, starting from the observed leaves (numbered 1 to  $n$ ) to the root  $2n-1$ . Let  $\nu : [2n-2] \rightarrow \{n+1, \dots, 2n-1\}$  map nodes other than the root onto their parent node. It is assumed that we are given a Markov model on the tree describing the evolution of states over time on each branch of the tree and at each site; each site evolves independently given the branch lengths. Treating the sequence states at internal nodes as missing data, we can write the full-data likelihood as:

$$p(x_{1:2n-1,1:m}|\theta) = \prod_{j=1}^m \mu_\theta(x_{(2n-1)j}) \prod_{i=1}^{2n-2} f_\theta(x_{ij}|x_{\nu(i)j}) \quad (1)$$

where  $\theta \in \Theta \subseteq \mathbb{R}^d$  is an unknown parameter,  $f_\theta$  a Markov transition probability, and  $\mu_\theta$  a probability distribution on the state space at a site. Note that  $x_{n+1:2n-1,1:m}$  are unobserved, whereas  $x_{1:n,1:m}$  are observed. Note also that  $f_\theta(x_{ij}|x_{\nu(i)j})$  can depend on the (known) length of the branch connecting the node  $i$  with node  $\nu(i)$ . For convenience in subsequent formulas we will write  $f_\theta(x_{(2n-1)j}|x_{\nu(2n-1)j})$  in place of  $\mu_\theta(x_{(2n-1)j})$ , even though  $\nu(2n-1)j$  is undefined.

The observed-data likelihood can be written as a sum over the missing data:

$$p(x_{1:n,1:m}|\theta) = \prod_{j=1}^m \left[ \sum_{x_{n+1:2n-1,j} \in [p]^{n-1}} \prod_{i=1}^{2n-1} f_\theta(x_{ij}|x_{\nu(i)j}) \right]. \quad (2)$$

Using belief propagation (Pearl, 1982) (also called the sum and products algorithm), the cost of computing (2) is  $\mathcal{O}(mp^2n^2)$ .

Our model generalizes (1) to allow  $\theta$  to vary along the sequence at a set of change-points  $1 = s_0 < s_1 < \dots < s_{k+1} = m$ . Then the full-data likelihood for the change-point model is:

$$p(x_{1:2n-1,1:m}|k, s_{1:k}, \theta_{1:k+1}) = \prod_{j=1}^{k+1} \prod_{l=s_{j-1}}^{s_j-1} \prod_{i=1}^{2n-1} f_{\theta_j}(x_{il}|x_{\nu(i)l}),$$

and, as in (2), one can sum over  $x_{n+1:2n-1,1:m}$  to obtain an observed-data likelihood:

$$p(x_{1:n,1:m}|k, s_{1:k}, \theta_{1:k+1}) = \prod_{j=1}^{k+1} \prod_{l=s_{j-1}}^{s_j-1} p(x_{1:n,l}|\theta_j)$$

where

$$p(x_{1:n,l}|\theta_j) = \sum_{x_{n+1:2n-1,l} \in [p]^{n-1}} \left\{ \prod_{i=1}^{2n-1} f_{\theta_j}(x_{ii}|x_{\nu(i)l}) \right\}.$$

*3.1.1. Bayesian inference.* For  $0 \leq k < m$ , let

$$\mathcal{S}_k = \{s_{1:k} \in [m] : 1 < s_1 < \dots < s_k \leq m\}.$$

Then we will define a posterior probability on the space

$$\mathbb{E} := \bigcup_{k=0}^{m-1} (\{k\} \times \mathcal{S}_k \times \Theta^{k+1}).$$

Let  $p(k, s_{1:k}, \theta_{1:k+1})$  be any proper prior probability on  $\mathbb{E}$ . Our objective is then to consider the posterior

$$\pi(k, s_{1:k}, \theta_{1:k+1}) \propto p(x_{1:n,1:m}|k, s_{1:k}, \theta_{1:k+1}) p(k, s_{1:k}, \theta_{1:k+1}) \quad (3)$$

which can be computed pointwise up to a normalizing constant in  $\mathcal{O}(mp^2n^2)$  steps. We assume that we know how to calculate the priors  $p(s_{1:k}, \theta_{1:k}|k)$  and  $p(k)$ .

*3.1.2. Time machine.* One way to cut the cost of the  $\mathcal{O}(mp^2n^2)$  calculation of the likelihood is to remove the top of the tree (a related idea is used in Jasra et al., 2011, for the standard coalescent). Suppose we only consider the tree backward in time until the parent of node  $2n-g$ ,  $g \in \{2, \dots, n\}$ . We propose the model:

$$p^B(x_{1:2n-g,1:m}, x_{2n-g+1:2n-1,1:m}|k, s_{1:k}, \theta_{1:k+1}) = \prod_{j=1}^{k+1} \prod_{l=s_{j-1}}^{s_j-1} \left( \left\{ \prod_{i=1}^{2n-g} f_{\theta_j}(x_{ii}|x_{\nu(i)l}) \right\} \eta_{\theta_j}(x_{\mathcal{B}}(2n-g+1:2n-1), l) \right),$$

where  $\mathcal{B}(2n-g+1:2n-1)$  denotes the nodes in the cut-off part of the tree  $2n-g+1:2n-1$  that are parents to at least one of the nodes in  $1:2n-g$ , and  $\eta_{\theta_j}(\cdot)$  is a joint probability distribution over sequences on these ‘‘boundary’’ nodes. Thus, the joint distribution of a number of the upper-most  $g-1$  nodes, for the  $j$  site, is replaced by the approximation  $\eta_{\theta_j}$ . Then one can perform inference from the relevant posterior

$$\pi^B(k, s_{1:k}, \theta_{1:k+1}) \propto p^B(x_{1:n,1:m}|k, s_{1:k}, \theta_{1:k+1}) p(k, s_{1:k}, \theta_{1:k+1})$$

using the PMMH method described below. The cost of computing the new likelihood is now  $\mathcal{O}(mp^2n(n-g))$ .

### 3.2. Particle marginal Metropolis-Hastings (PMMH)

To sample from the transdimensional state-space of (3), we first consider an SMC sampler that only samples on  $\mathcal{S}_k \times \Theta^{k+1}$  for  $k$  fixed. We then show how the SMC sampler can be embedded within a PMMH algorithm to target (3). The SMC sampler will be necessary to ensure a good acceptance probability for transdimensional moves. Our approach has the advantage over alternative simulation techniques for model selection (see Zhou et al., 2013) that the model selection and parameter estimates are simultaneous, which helps to focus computational resources on the important model(s).

For  $1 \leq k < m$ , and a user-specified  $T \geq 1$ , let  $\{\xi_{t,k}\}_{0 \leq t \leq T}$  be a sequence of probabilities on  $\mathcal{S}_k \times \Theta^{k+1}$ , such that  $\xi_{0,k}(s_{1:k}, \theta_{1:k+1}) = p(s_{1:k}, \theta_{1:k+1}|k)$  and

$$\xi_{T,k}(s_{1:k}, \theta_{1:k+1}) \propto p(x_{1:n,1:m}|k, s_{1:k}, \theta_{1:k+1}) p(s_{1:k}, \theta_{1:k+1}|k).$$

The remaining sequence of targets  $\{\xi_{t,k}\}_{1 \leq t \leq T-1}$  interpolate between the (conditional) posterior and the prior, for example, via the tempering procedure:

$$\xi_{t,k}(s_{1:k}, \theta_{1:k+1}) \propto p(x_{1:n,1:m}|k, s_{1:k}, \theta_{1:k+1})^{\kappa_t} p(s_{1:k}, \theta_{1:k+1}|k)$$

with  $0 < \kappa_1 < \dots < \kappa_{T-1} < 1$ . The SMC sampler will propagate a collection of  $N$  particles from the prior  $\xi_{0,k}$  all the way to the posterior  $\xi_{T,k}$  via the bridging densities  $\xi_{t,k}$  by means of importance sampling, resampling, and MCMC move steps. The tempering procedure aims at controlling the variability of the incremental importance weights, for instance, providing robust estimates of the normalizing constants  $p(x_{1:n,1:m}|k)$ , which is an important attribute for the overall algorithm. The sampler propagates the particles by using a sequence of MCMC kernels of invariant densities  $\xi_{t,k}$  (which operate on a fixed dimensional space). All the details of the specific steps of the SMC sampler are given in the Supplementary Material (available online of [www.liebertpub.com/cmb](http://www.liebertpub.com/cmb)). We write the probability of all the variables associated with the SMC sampler (which resamples  $N > 1$  ‘‘particles’’ at every time except at time  $T$ ) as

$$\Psi_{k,N}(a_{0:T-1}^{1:N}, \phi_{0:T}^{1:N}(k)),$$

where  $a_{0:T-1}^{1:N} = (a_0^1, \dots, a_0^N, \dots, a_{T-1}^1, \dots, a_{T-1}^N)$  are the resampled indices and  $\phi_{0:T}^{1:N}(k) = (\phi_0^1(k), \dots, \phi_0^N(k), \dots, \phi_T^1(k), \dots, \phi_T^N(k))$  with  $\phi_t^i(k) = (s_{t,1:k}^i, \theta_{t,1:k+1}^i)$ ,  $i \in [N]$ ,  $t \in \{0\} \cup [T]$  is the collections of the  $N$  particles as propagated through the sequence  $\xi_{t,k}$ .

One can use this SMC sampler within a broader PMMH algorithm to sample from the true target of interest (3). The specific steps of PMMH are given in the Supplementary material, but briefly, a single iteration of the algorithm is as follows. Given the current state of the Markov chain, one proposes to change  $k$  with some proposal kernel  $q(k'|k)$ . Conditional on this  $k'$ , we run an SMC sampler  $\Psi_{k',N}(\cdot)$  and choose a particle  $\phi_T^l(k')$ , for some  $1 \leq l \leq N$ , with probability proportional to a weight. Acceptance of both the model index  $k'$  and of the proposed change-point times and rates  $\phi_T^l(k')$  happens with probability

$$1 \wedge \frac{p^N(x_{1:n,1:m}|k')p(k')}{p^N(x_{1:n,1:m}|k)p(k)} \times \frac{q(k|k')}{q(k'|k)},$$

where  $p^N(x_{1:n,1:m}|k)$  is the SMC (unbiased) estimate of  $p(x_{1:n,1:m}|k)$ , the normalizing constant of  $\xi_{T,k}$  (see Del Moral, 2004). The Supplementary Material presents the formula used to calculate  $p^N(x_{1:n,1:m}|k)$ . Note that whilst there are a lot of user set parameters (namely, the temperatures and tuning parameters for the MCMC kernels), their choice can be done adaptively to reduce user involvement (see Jasra et al., 2014). In this article we tune the parameters by trial and error.

The advantages of our procedure is that it mitigates having to construct transdimensional proposals that need to mix well (see Karigiannis and Andrieu, 2013, for another recent work that attempts to deal with this issue). We note, however, that the cost of each proposal will be  $\mathcal{O}(NTmp^2n^2)$ , as  $\xi_{t,k}$  must be obtained at each time step of the SMC sampler (see Supplementary Material). In addition, note that tailored methods for change-point models (e.g., Fearnhead and Liu, 2007) do not apply here as one does not have a convenient way to integrate the likelihood.

### 3.3. Approximate Bayesian computation (ABC)

ABC is another methodology that avoids exact computation of the likelihood, at the cost of a biased approximation of the posterior; see, for instance, Marin et al. (2012) for a review. The method is based on accepting simulated data sets that are similar to the observed data set, where ‘‘similar’’ is usually assessed using summary statistics sensitive to the parameter(s) of interest.

ABC can be unreliable as a tool for model selection. According to Marin et al. (2013), the best summary statistics to be used in ABC approximation to a Bayes factor are ancillary statistics with different mean values under two competing models. Otherwise, the summary statistic must have enough components to prohibit a parameter under a wrong model from generating summary statistics that are plausible under the true model. However, summary statistics satisfying the conditions of Marin et al. (2013) for model choice in ABC is not easy (or even possible) to verify in our context.

In the numerical examples of section 4.1, we consider two ABC algorithms that approximate the same ABC posterior. The first algorithm is a PMMH that replaces the SMC sampler of Del Moral et al. (2006) with the SMC sampler of section 3.3 of Del Moral et al. (2012); see Supplementary Material for details. The second ABC algorithm is the ABC-SMC algorithm for model selection appearing on page 190 of Toni et al. (2009).

## 4. RESULTS

### 4.1. Comparison of computational methods on simulated data

We compared three algorithms on their performance in Bayesian model selection for four simulated DNA data sets. Within each data set, the DNA sequences shared a common ancestral binary tree with known topology, unknown sequence states at ancestral nodes, and unknown substitution rates and branch lengths. The first algorithm was our proposed PMMH algorithm outlined in section 3.2, and we employed three versions of the time machine. Using the notation of section 3.1.2, these used  $g = 1$  (so in effect the time machine was not implemented at all),  $g = 4$ , and  $g = 8$ . We also used two ABC algorithms described in section 3.3. The PMMH algorithms were not run until they converged fully, but were compared on the basis of results achieved after 6 hours of computation. Other implementation details of the algorithms may be found in the Supplementary Material.

*4.1.1. Base data set.* The base data set consists of  $n = 8$  simulated DNA sequences ( $p = 4$  types of nucleotide), each of  $m = 50$  sites. The sequences evolved according to a binary tree under a Jukes-Cantor model of DNA evolution with one substitution rate up to site  $s_1 = 25$ , and a second rate beyond this single change-point (so  $k = 1$ , but for inference we assumed only  $k \in \{0, 1\}$ ). In the standard Newick notation, the structure of the tree was:

```
(( (Taxon0:1.0, Taxon1:1.0) :1.0, (Taxon2:1.0, Taxon3:1.0) :1.0) :1.0,
 ( (Taxon4:1.0, Taxon5:1.0) :1.0, (Taxon6:1.0, Taxon7:1.0) :1.0) :1.0) :1.0
```

We ran the three algorithms to infer  $k$ , location of the change-point,  $s_1$  given  $k = 1$ , and the substitution rate(s)  $\theta_{1:k+1}$ . The prior on  $k$  was uniform on  $\{0,1\}$ ; the prior on  $s_1$  was  $1/m - 1$  (change-points occur immediately before a site so cannot occur at site 1); finally, all substitution rates had a gamma prior, with shape = 2 and scale = 0.4, and so expected value of 0.8 mutations per generation per site.

The results in the top quadrants of Tables 1 and 2 in Appendix B show that our time-machine PMMH algorithm with  $g = 4$  outperformed all other algorithms. It sampled from the true model (i.e.,  $k = 1$ ) much more frequently than the incorrect  $k = 0$  model (Appendix Table 1). In comparison  $g = 8$  performed poorly, as expected since  $n = 8$  for this data set so  $g = 8$  implies removing all internal nodes and assuming independent evolution of each sequence. The ABC algorithms did not perform well. The PMMH-ABC algorithm sampled from the two models almost evenly, while the ABC-SMC algorithm had low effective sample sizes (Kong et al., 1994; Liu, 1996) and actually preferred the wrong model.

In Appendix Table 2, we give 95% confidence intervals of estimates of  $s_1$  and of the rates given  $k = 1$ . The time-machine PMMH algorithms again provide the best inferences and were able to find the change-point. The PMMH-ABC was more accurate for the substitution rates but less precise. The ABC-SMC algorithm gave unusable output.

We do not present the output for the  $g = 1$  version of the time machine because it performed very poorly. Without removing any nodes from the top of the tree, the variability of  $p^N(x_{1:n,1:m}|k')$ / $p^N(x_{1:n,1:m}|k)$  in the acceptance probability of the PMMH was very high when  $k \neq k'$  (Fig. 2 in Appendix A). Thus, the algorithm accepted jumps between models only rarely and the output was very “sticky.” This phenomenon illustrates that the time machine is a cost-saving technique by two measures. First, it reduces the computational complexity of the algorithm. Second, it aids in mixing and facilitates jumping between models.

*4.1.2. Further tests.* We repeated the above experiment for three more data sets that differed only slightly from the base data set. We found the results to be similar across the data sets (see Tables 1 and 2 in Appendix B). Collectively, these results suggest that when doing Bayesian model selection under these scenarios, ABC approximations should be avoided and instead our PMMH method used instead, with the time machine but removing as few nodes as computational considerations permit.

### 4.2. Application to a real dataset

Using the publicly available database of Weiss et al. (2013), we assembled a data set consisting of  $n = 6$  ACT1 gene DNA sequences ( $m = 540$  sites). We assumed the tree structure given in Figure 1 of Appendix A, and a Jukes-Cantor model of DNA evolution. We implemented our time-machine PMMH algorithm to infer  $k$ ,  $s_{1:k}$ , and  $\theta_{1:k+1}$  for cut-off parameter  $g = 4$  (see Supplementary Material for further details). The

prior on  $k$  was a discrete uniform distribution on  $\{0, 1, 2\}$ , the prior on  $s_{1:k}$  was uniform on  $k$ -subsets of  $[m-1]$ , and each substitution rate had a gamma prior with shape = 1 and scale = 0.3.

We ran the algorithm for 10,589 iterations (23 days) on a Linux workstation that used 12 Intel Xeon E5-1650 3.20GHz CPUs. We monitored convergence via autocorrelation and trace plots (Appendix Fig. 3). We also monitored convergence of each model individually using the diagnostic in Geweke (1992); that is, we obtained a Z-score for each model parameter per each value of  $k$  to get a sense of the algorithm's ability to fully explore the state space of each model (only some values are reported below). Appendix Figure 3 suggests good exploration of the state space of  $k$ , resulting in an estimated distribution: 0.17 ( $k = 0$ ), 0.47 ( $k = 1$ ), and 0.36 ( $k = 2$ ).

From the 4,946 samples with  $k = 1$ , we estimated a 95% highest posterior density interval of (194,199) (see also the histogram in Appendix Fig. 4). The Z-score for  $k$  was  $-0.60$ , suggesting that we were still some way off convergence (values close to 0 imply convergence). For the rates  $\theta_1$  (before the change-point) and  $\theta_2$  (after the change-point), the Z-scores of 0.11 and 0.23, respectively, give stronger evidence for convergence (estimated densities of these parameters are shown in Appendix Fig. 5).

## 5. DISCUSSION

We considered sequence data that originates from a rooted binary tree (ch. 1 of Felsenstein, 2004) of known topology and branch lengths but unknown sequence states at internal nodes, and the substitution rates in the DNA evolution model allowed to have change-points. We detailed Bayesian parameter inference from such a model with an unknown number of change-points, implying a transdimensional posterior density. Computational inference from this model is challenging, and we introduced two novel contributions to facilitate sampling.

Firstly, based on the time machine principle of Jasra et al. (2011), we showed how the top-most nodes of the binary tree can be replaced with a probability distribution of the sequence evolution model to reduce the cost of computing the likelihood linearly in  $n$  (the number of sequences). This approach introduces a bias, but this was found in practice to have a small effect on inferences.

Secondly, we developed a particle marginal Metropolis-Hastings (PMMH) algorithm (section 3.2) that mitigates having to construct transdimensional proposals that need to mix well. We first developed a sequential Monte Carlo (SMC) sampler that only samples on a fixed-dimensional subspace of the full transdimensional state-space. We then showed how that SMC sampler can be embedded within the PMMH algorithm to target the full posterior. By employing the time machine within this PMMH, we attained an algorithm that could run with a reduced computational cost and easily jump between models with different numbers of change-points.

We successfully implemented our PMMH to perform inference from the model in a reliable fashion for small to moderately sized data sets. We empirically demonstrated that our PMMH can outperform approximate Bayesian computation (ABC) techniques (Tavare et al., 1997) in terms of precision and accuracy, and we showed that our algorithm can successfully be used to carry out reliable inference on real data. The success of our PMMH algorithm is largely due to the time machine, which, as we witnessed in section 4.1, reduces the variance of the acceptance probability and enables the algorithm to jump easily between models. However, based on the output of section 4.1, it seems that bias introduced by the time machine reduces the accuracy of the inferred substitution rates.

In future work, one might want to extend the methodology to allow for unknown tree topologies, similar to Suchard et al. (2003). Also, a future work could attempt to use a more appropriate distribution to approximate the distribution at the top of the tree. We attempted to find approximations in the point processes and coalescent literature, but we were unable to find a better approximation than that, which we employed here. From the computational point of view, it will certainly be important to further speed up the algorithm, and great savings could be made by parallelizing calculations within the SMC particle method and carefully investigating adaptive procedures for fine-tuning the temperatures and the MCMC kernels. All such efforts could have a big effect on reducing the variance of the estimate of  $p(x_{1:n,1:m}|k)$ , thus further improving the mixing of PMMH, even with fewer removed nodes. Also, there could then be great scope to apply the method for larger numbers of potential change-points compared to the relatively small one we tried here.



## ACKNOWLEDGMENTS

This research was funded by the EPSRC grant “Advanced Stochastic Computation for Inference from Tree, Graph and Network Models” (Ref: EP/K01501X/1). A.J. was additionally supported by Singapore MOE grant R-155-000-119-133 and is also affiliated with the risk management institute and the centre for quantitative finance at the National University of Singapore.

## AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

## REFERENCES

- Andrieu, C., Doucet, A., and Holenstein, R. 2010. Particle Markov chain Monte Carlo methods. *J. R. Statist. Soc. Ser. B* 72, 269–342.
- Boffelli, D., McAuliffe, J., Ovcharenko, D., et al. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299, 1391–1394.
- Bouchard-Côté, A., Sankararaman, S., Jordan, M.I. 2012. Phylogenetic inference via sequential Monte Carlo. *System. Biol.* 61, 579–593.
- Chinwalla, A.T., Cook, L.L., Delehaunty, K.D., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562.
- Del Moral, P. 2004. *Feynman-Kac Formuale*. Springer, New York.
- Del Moral, P., Doucet, A., and Jasra, A. 2006. Sequential Monte Carlo samplers. *J. R. Statist. Soc. Ser. B* 68, 411–436.
- Del Moral, P., Doucet, A., and Jasra, A. 2012. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statist. Comp.* 22, 1009–1020.
- Fearnhead, P and Liu, Z. 2007. Online Inference for Multiple Changepoint Problems. *J. R. Statist. Soc. Ser. B* 69, 589–605.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.
- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer, Sunderland, Massachusetts.
- Felsenstein, J., and Churchill, G.A. 1996. A Hidden Markov Model approach to variation of evolutionary rates among sites. *Mol. Biol. Evol.* 13, 93–104.
- Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M. *Bayesian Statistics 4*. Clarendon Press, Oxford, United Kingdom.
- Gibbs, R.A., Weinstock, G.M., Metzker, M.L., et al. 2004. Genome sequence of the brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493–521.
- Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–732.
- Huelsenbeck, J.P., and Hillis, D.M. 1993. Success of phylogenetic methods in the four taxon case. *System. Biol.* 42, 247–264.
- Huelsenbeck, J.P., and Suchard, M.A. 2007. A nonparametric method for accommodating and testing across-site rate variation. *System. Biol.* 56, 975–987.
- Jasra, A., De Iorio, M., and Chadeau-Hyam, M. 2011. The time machine: a simulation approach for stochastic trees. *Proc. R. Soc. A* 467, 2350–2368.
- Jasra, A., Kantas, N., and Persing, A. 2014. Bayesian inference for partially observed stopped processes. *Stat. Comp.* 24, 1–20.
- Karigiannis, G., and Andrieu, C. 2013. Annealed importance sampling reversible jump MCMC algorithms. *J. Comp. Graph. Statist.* 22, 623–648.
- Kong, A., Liu, J.S., and Wong, W.H. 1994. Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* 89, 278–288.
- Liu, J.S. 1996. Metropolisized independent sampling with comparison to rejection sampling and importance sampling. *Statist. Comp.* 6, 113–119.
- Loza-Reyes, E., Hurn, M.A., and Robinson A. 2014. Classification of molecular sequence data using Bayesian phylogenetic mixture models. *Comp. Statist. Data Anal.* 75, 81–95.
- Ma, Y. 2008. Bayesian and MCMC methods for phylogenetic footprinting [PhD Thesis]. Imperial College London.

- Marin, J.M., Pillai, N.S., Robert, C.P., and Rousseau, J. 2013. Relevant statistics for Bayesian model choice. *J. R. Statist. Soc. Ser. B* [Epub ahead of print]; doi: 10.1111/rssb.12056.
- Marin, J.-M., Pudlo, P., Robert, C.P., and Ryder, R. 2012. Approximate Bayesian computational methods. *Statist. Comp.* 22, 1167–1180.
- Mossel, E., and Vigoda, E. 2006. Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Ann. Appl. Probab* 16, 2215–2234.
- Nei, M., Chakraborty, R., and Fuerst, P.A. 1976. Infinite allele model with varying mutation rate. *Proc. Natl. Acad. Sci.* 73, 4164–4168.
- Olsen, G.J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symp Quant. Biol.* 52, 825–837.
- Pagel, M., and Meade, A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* 53, 571–581.
- Pearl, J. 1982. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second National Conference on Artificial Intelligence AAAI-82* AAAI Press, Menlo Park, California.
- Siepel, A., Bejerano, G., Pedersen, J.S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050.
- Siepel, A. and Haussler, D. 2005. Phylogenetic hidden Markov models, 325–351. In Nielsen, R., ed. *Statistical Methods in Molecular Evolution*. Springer, New York.
- Suchard, M.A., Weiss, R.E., Dorman, K.S., and Sinsheimer, J.S. 2003. Inferring spatial phylogenetic variation along nucleotide sequences: A multiple change-point model. *J. Amer. Statist. Ass.* 98, 427–437.
- Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. 1996. Phylogenetic Inference, 407–514. In *Molecular Systematics*, 2nd edition. Sinauer and Associates, Sunderland, Massachusetts.
- Tavare, S., Balding, D.J., Griffiths, R.C., and Donnelly, P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* 145, 505–518.
- Toni, T., Welch, D., Strelkowa, N., et al. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Statist. Soc. Interface* 6, 187–202.
- Uzzell, T., and Corbin, K.W. 1971. Fitting discrete probability distributions to evolutionary events. *Science* 172, 1089–1096.
- Wakeley, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* 11, 436–442.
- Weiss, S., Samson, F., Navarro, D., and Casaregola, S. 2013. YeastIP: a database for identification and phylogeny of ascomycetous yeasts. *FEMS Yeast Res.* 13, 117–125.
- Wu, C.H., Suchard, M.A., and Drummond, A.J. 2013. Bayesian selection of nucleotide substitution models and their site assignments. *Mol. Biol. Evol.* 3, 669–688.
- Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10, 1396–1401.
- Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139, 993–1005.
- Yang, Z., Goldman, N., and Friday, A. 1994. Comparison for models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11, 316–324.
- Zhou, Y., Johansen, A.M., and Aston, J.A.D. 2013. Towards Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach. arXiv preprint.

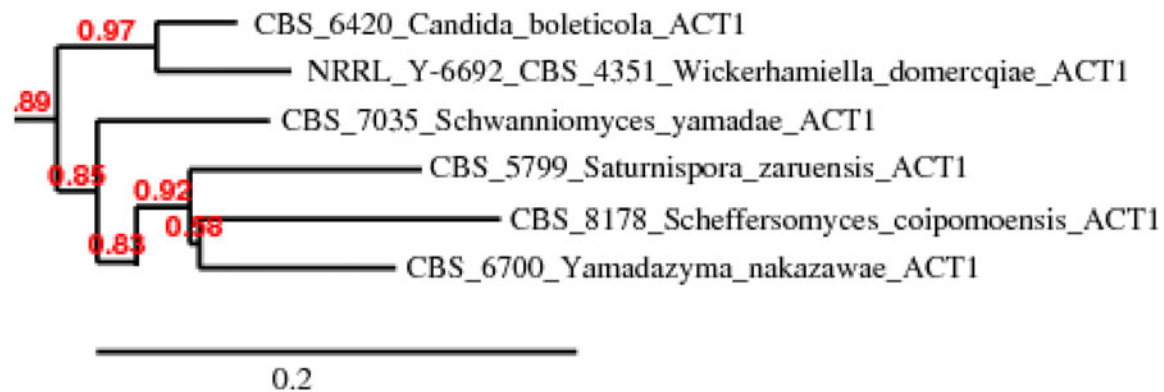
Address correspondence to:  
 Dr. Adam Ford Persing  
 Department of Statistical Science  
 University College London  
 1-19 Torrington Place London WC1E 7HB  
 United Kingdom

E-mail: a.persing@ucl.ac.uk

(Appendix follows →)

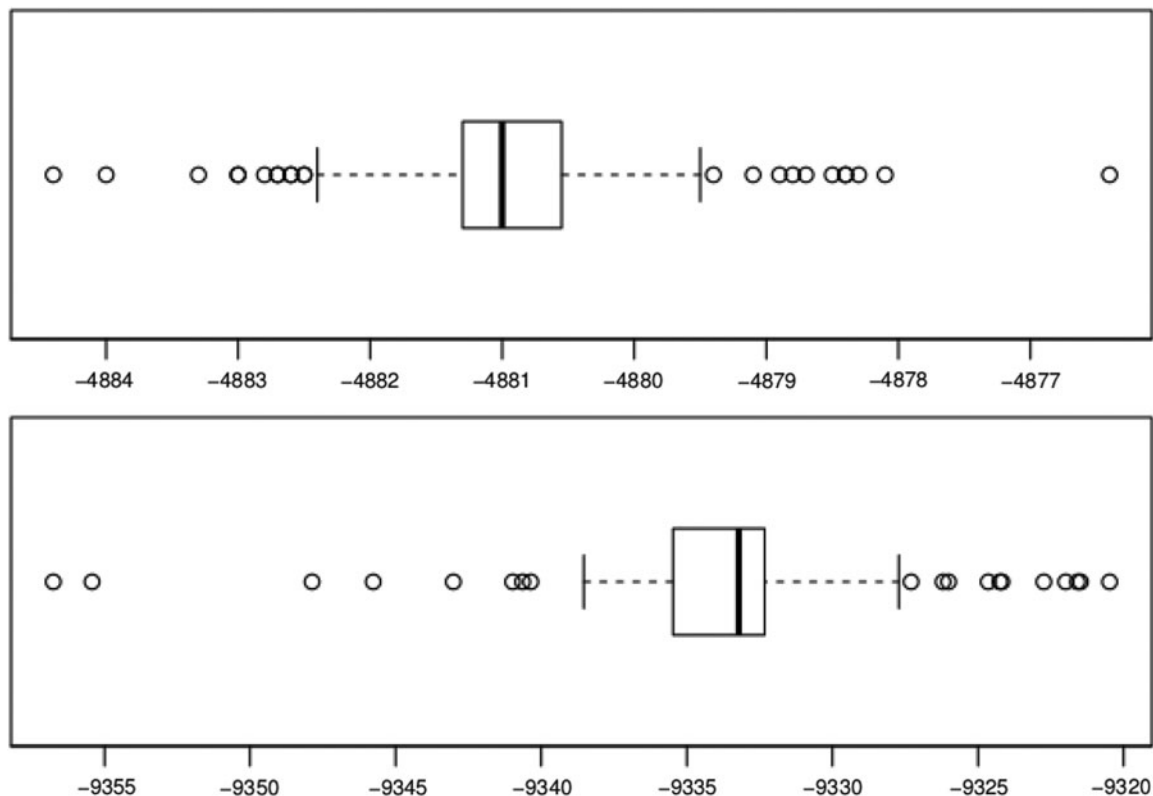
## APPENDIX A. FIGURES

Real data case: Phylogeny of a subset of the *Saccharomycotina* subphylum



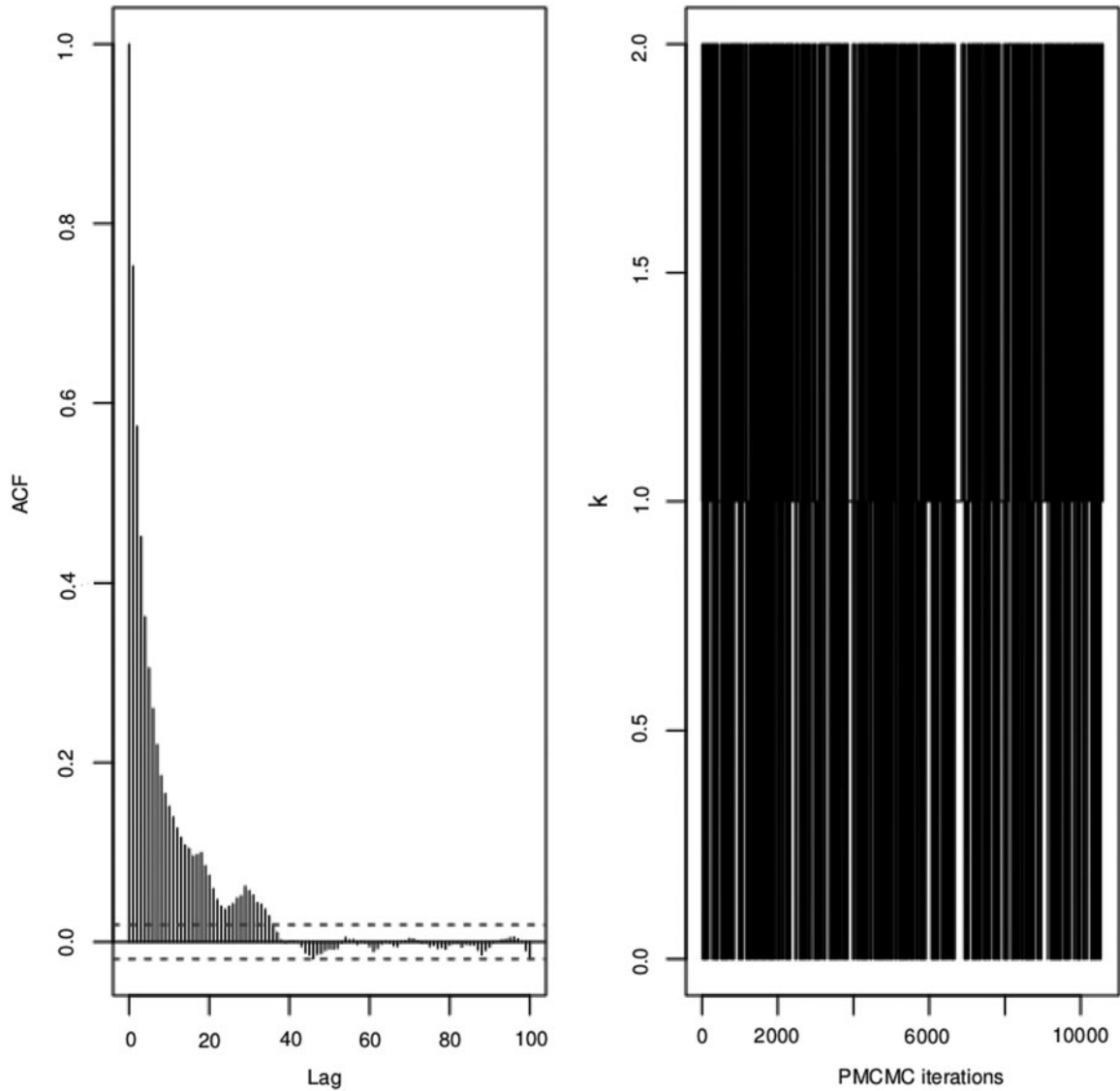
**FIG. 1.** This tree is provided by Weiss et al. (2013). The labels on the leaves at the far right are in the format STRAIN\_taxa\_GENE; for example, CBS\_6420\_Candida\_boleticola\_ACT1 is the label for the ACT1 gene of *Candida boleticola*, strain CBS 6420. The edge at the bottom of the figure with the number 0.2 shows the branch length that represents an amount of genetic change equal to 0.2 nucleotide substitutions per site. The red numbers represent a measure of evidence for the node, with values closer to 1.0, meaning that there is stronger evidence for the genes to the right of the node clustering together; in this case, those values are computed using a bootstrapping method.

Sim. database data set: Variability of  $\log [p^N(x_{1:n,1:m}|k)]$  for  $g = 1$  versus  $g = 4$



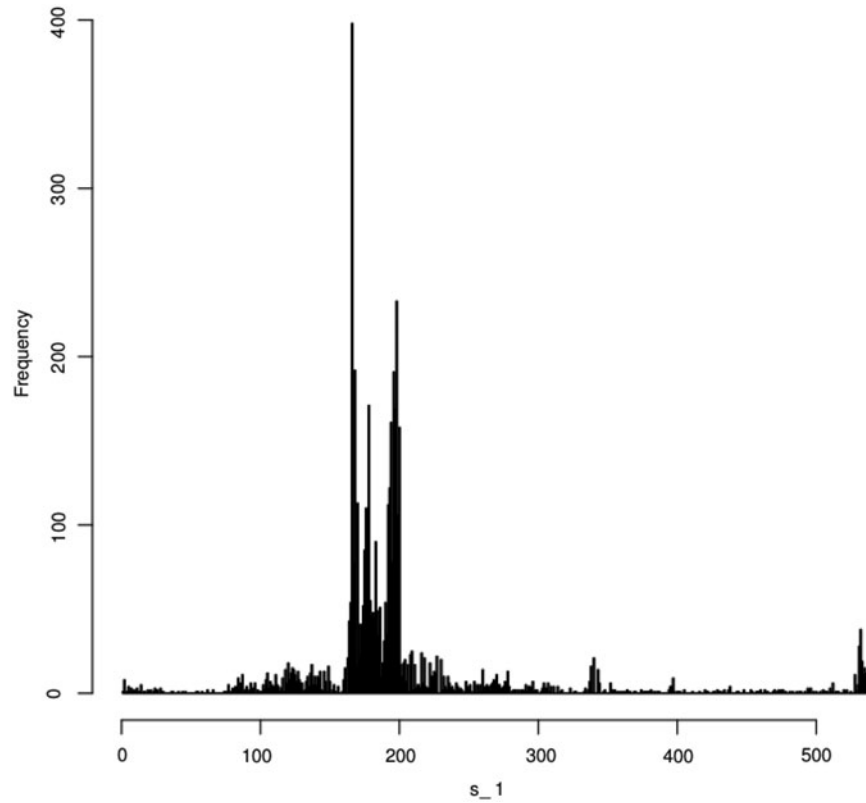
**FIG. 2.** The top box plot illustrates the variability of  $\log [p^N(x_{1:n,1:m}|k)]$  for the  $g = 1$  version of our algorithm, and the bottom plot gives the same for the  $g = 4$  version of our algorithm;  $\log [p^N(x_{1:n,1:m}|k)]$  runs along the horizontal axis of each plot. One can see that the variability in the bottom plot is much less than that of the top plot.

*Real data case: Autocorrelation and trace plot of sampled  $k$*



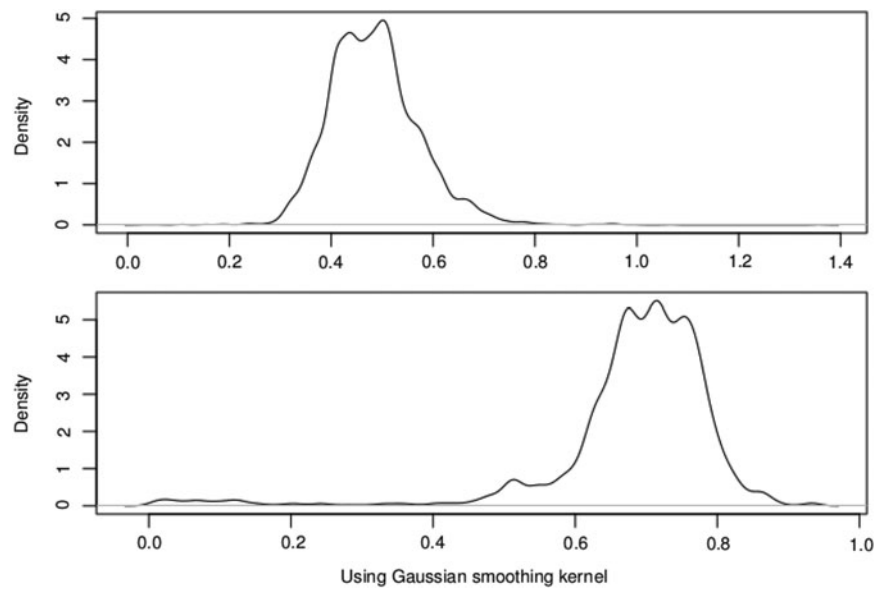
**FIG. 3.** We monitored the convergence of the PMMH implemented in section 4.2 via autocorrelation and trace plots. The plots illustrate a nonsticky algorithm and low autocorrelation. PMMH, particle marginal Metropolis-Hastings.

*Real data case: Histogram of sampled  $s_1$  given  $k = 1$*



**FIG. 4.** The histogram strongly indicates the position of the single change-point in the numerical example of section 4.2.

*Real data case: Kernel density plots of sampled substitution rates given  $k = 1$*



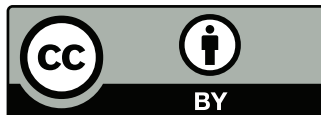
**FIG. 5.** At top, we have the kernel density plot of  $\theta_1$ , which is obtained using all samples from the PMMH implemented in section 4.2, where  $k = 1$ . Values for  $\theta_1$  run along the horizontal axis. We have the same plot for  $\theta_2$  on the bottom.

## APPENDIX B. TABLES

TABLE 1. ALGORITHM DIAGNOSTICS AND MODEL CHOICE

Example	Algorithm	Samples	Accp. ratio	Autocorrelation of $k$			ESS per model			$\mathbb{P}(k   \text{data})$		
				@25	@100	$k=0$	$k=1$	$k=2$	$k=0$	$k=1$	$k=2$	
Base dataset ( $k=1$ , $n=8$ , $redm=50$ )	Time machine, $g=4$	4255	0.214	0.029	0.035	—	—	—	0.261	0.739	—	
	Time machine, $g=8$	7237	0.325	-0.001	-0.006	—	—	—	0.477	0.523	—	
	PMMH with ABC ABC-SMC	1070 5000	0.305 —	-0.012 —	0.025 —	6	7	—	0.482 0.538	0.518 0.462	—	
Two change-points ( $k=2$ , $n=8$ , $m=50$ )	Time machine, $g=4$	4213	0.121	0.184	0.036	—	—	—	—	0.159	0.841	
	Time machine, $g=8$	7198	0.287	-0.004	-0.001	—	—	—	—	0.472	0.528	
	PMMH with ABC ABC-SMC	1021 5000	0.394 —	0.010 —	0.011 —	—	1	1	—	0.562 0.477	0.438 0.523	
Subtle change-point ( $k=1$ , $n=8$ , $m=50$ )	Time machine, $g=4$	4264	0.108	0.017	0.027	—	—	—	0.079	0.921	—	
	Time machine, $g=8$	7229	0.306	0.005	0.014	—	—	—	0.493	0.507	—	
	PMMH with ABC ABC-SMC	1058 5000	0.329 —	0.008 —	0.046 —	2	1	—	0.486 0.624	0.514 0.376	—	
More sites ( $k=1$ , $n=8$ , $m=80$ )	Time machine, $g=4$	530	0.106	-0.015	-0.037	—	—	—	0.075	0.925	—	
	Time machine, $g=8$	904	0.293	-0.023	-0.080	—	—	—	0.461	0.539	—	
	PMMH with ABC ABC-SMC	132 5000	0.530 —	0.003 —	0.013 —	22	7	—	0.469 0.508	0.531 0.492	—	

All algorithms ran for 6 hours on a Linux workstation that used five Intel Core i5-2600 CPUs, each at 3.40GHz (see important note regarding ABC-SMC in the Supplementary Material). The column labeled “Accp. ratio” gives the frequency at which a new sample is accepted in the particle marginal Metropolis-Hastings (PMMH) algorithms (Andrieu et al., 2010). The columns labeled “Autocorrelation of  $k$ ” give the autocorrelation at time delays of 25 and 100 (when available). ESS is the abbreviation for the effective sample size (Kong et al., 1994; Liu, 1996) of the ABC-SMC for model selection of Toni et al. (2009). Note that for the two change-point case, the prior on  $k$  was a discrete uniform distribution on  $\{1,2\}$ . ABC, approximate Bayesian computation; SMC, sequential Monte Carlo.



This work is licensed under a Creative Commons Attribution 3.0 United States License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as "Journal of Computational Biology." Copyright 2014 Mary Ann Liebert, Inc. <http://liebertpub.com/cmb>, used under a Creative Commons Attribution License: <http://creativecommons.org/licenses/by/3.0/us/>

TABLE 2. INFERENCE FOR TRUE MODEL: 95% CONFIDENCE INTERVALS

Example	Algorithm	Samples	$s_1$	$s_2$	CI of $\mu_1$	CI of $\mu_2$	CI of $\mu_3$
Base dataset ( $s_1 = 25, \mu_1 = 0.75, \mu_2 = 0.85$ )	Time machine, $g = 4$	3144	(22,23)	—	(0.484,0.494)	(0.663,0.671)	—
	Time machine, $g = 8$	3784	(25,26)	—	(0.322,0.327)	(0.432,0.438)	—
	PMMH with ABC	554	(24,26)	—	(0.709,0.811)	(0.702,0.794)	—
	ABC-SMC	2310	(1,1)	—	(1.069,1.093)	(0.454,0.488)	—
Two change-points ( $s_1 = 15, s_2 = 35, \mu_1 = 0.75, \mu_2 = 0.85, \mu_3 = 0.75$ )	Time machine, $g = 4$	3543	(10,11)	(43,45)	(0.340,0.343)	(0.590,0.599)	(0.239,0.243)
	Time machine, $g = 8$	3801	(17,18)	(32,33)	(0.290,0.294)	(0.414,0.428)	(0.285,0.290)
	PMMH with ABC	447	(15,22)	(34,42)	(0.712,0.807)	(0.698,0.795)	(0.708,0.799)
	ABC-SMC	2615	(1,1)	(1,2)	(1e-7,1e-7)	(1e-7,1e-7)	(1e-7,1e-7)
Subtle change-point ( $s_1 = 25, \mu_1 = 0.75, \mu_2 = 0.8$ )	Time machine, $g = 4$	3927	(21,22)	—	(0.435,0.441)	(0.343,0.353)	—
	Time machine, $g = 8$	3665	(22,24)	—	(0.297,0.304)	(0.319,0.326)	—
	PMMH with ABC	543	(22,25)	—	(0.678,0.774)	(0.690,0.788)	—
	ABC-SMC	1880	(2,2)	—	(1.278,1.287)	(0.449,0.463)	—
More sites ( $s_1 = 40, \mu_1 = 0.75, \mu_2 = 0.85$ )	Time machine, $g = 4$	490	(41,46)	—	(0.432,0.449)	(0.393,0.418)	—
	Time machine, $g = 8$	487	(39,43)	—	(0.426,0.439)	(0.406,0.425)	—
	PMMH with ABC	70	(34,45)	—	(0.682,0.991)	(0.669,0.906)	—
	ABC-SMC	2460	(1,1)	—	(1.289,1.345)	(0.505,0.531)	—

The leftmost column contains the true parameter values for each example. We also record the number of samples on which each inference is based (i.e., we record how many samples from the true model each algorithm obtained).