RESEARCH

Improving the performance of relaxed clock model in phylogenetic analysis

Alexei Drummond* and Rong Zhang

*Correspondence:
alexei@cs.auckland.ac.nz
Department of Computer Science, University of Auckland, Princes Street, 1010 Auckland, New Zealand
Full list of author information is available at the end of the article

Abstract

Bayesian MCMC plays an important role in phylogenetic analysis. However, due the huge amount of phylogenetic data, the efficiency becomes a great problem. Based on uncorrelated relaxed clock model, this paper introduces a new operator to propose evolutionary rates and divergence times while maintaining the genetic distance constant. Specifically, the proposed operator deals with three rates when changing the node time for an internal node. For the root of a phylogenetic tree, there are three strategies discussed, including Simple Distance, Small Pulley and Big Pulley. It is noticed that Big Pulley is able to change the tree topology, which enables the operator to sample all the possible trees under an unrooted tree. To validate the effectiveness, experiments have been performed by implementing the operator in BEAST2 software. The results prove that the proposed operator is able to improve the performance by giving better estimation and less running time.

Keywords: Bayesian MCMC; Operator; Phylogenetic trees; Genetic distances; Divergence times; Evolutionary rates

Introduction

Phylogenetics has attracted much research interest over the years. More and more scientists are becoming keen to discover the evolutionary history of life, such as birds [1], primates [2], grasses [3] and so on [4, 5]. One fundamental concept in phylogenetics is a graph model that shows the relationships among species and organisms, which is called a phylogenetic tree. The main task for phylogenetic analysis is aimed at inferring the phylogenies by constructing the phylogenetic trees. Until now, a lot of methods have already been proposed and a majority of them have been implemented in the computer softwares, such as BEAST2 [6, 7] , MRBAYES [8] and APE [9].

Traditional methods for constructing phylogenetic trees are based on a distance matrix that is obtained by sequence alignment, after biologists extract DNA from organisms. In the last few decades, more advanced techniques have been developed to reconstruct phylogenetic trees, due to the development of statistics and computer science. One popular research area called Bayesian phylogenetics puts an emphasis on a sampling process to construct the phylogenetic tree, in which the methods based on Markov Chain Monte Carlo (MCMC) provide a computational tool for sampling problems. For example, Yang and Rannala presented a Bayesian framework that includes the specified prior for phylogenies and ancestral speciation times, and the posterior probabilities of phylogenies to be inferred as the maximum posterior probability (MAP) tree [10]. Specifically, Monte Carlo integration is used

to integrate over the ancestral speciation times for particular trees and a MCMC method is used to sample a set of trees with probabilities. Based on this framework, much attention has been paid to estimating the divergence times [11, 12, 13].

For phylogenetic analysis based on Bayesian MCMC, the efficiency of a sampling method is always the main issue that should be carefully dealt with. In particular, the calculation of the phylogenetic likelihood is a large burden of the efficiency. Hence, some researchers have tried to improve the likelihood calculations, such as detection of repeating sites [14]. On the other hand, in MCMC methods, the operators that propose a new state based on the current state also have a leading influence on the overall performance. As is discussed by Lakner et al., the major limitation in Bayesian MCMC analysis of phylogeny lies in the efficiency with which operators sample the tree space [15]. So far, there have been already many different kinds of operators proposed. According to the work in Ref. [16], the two common operators, i.e. prune-and-regraft and subtree-swap, both contribute to a tree with low likelihood since they propose a new tree by random movement of the current tree. So the authors introduced two new operators by proposing a state from a discrete set of possible proposals and narrowing the proposal distribution to the more likely proposals. Their experimental results proved that the two operator have faster average run time and more accurate predictability.

Nevertheless, it is acknowledged that a faster and more reliable performance is also dependent on a good mixture of operators and different operators may be applicable for different objects. In this paper, a novel operator is proposed in the case where genetic distances are constant. Namely, the proposed operator is used to change the divergence times of nodes and evolutionary rates on the branches at the same time without changing the genetic distances, so that the likelihood of the phylogenetic tree remains unchanged. According to a series of simulations and experiments, it is concluded that by using the proposed operator, the tree will be sampled more efficiently in the MCMC runs.

The following of this paper is constructed as follows: Section 2 gives some preliminary theories of the work related to this paper. Section 3 introduces the proposed operator, which includes the mechanism for internal nodes and 3 different strategies for the root. The experiments and discussions are detailed in the 4th section to validate the efficiency of the proposed operator. Section 5 ends the paper with a short conclusion.

Prelimiaries

Bayesian MCMC

Eq.(1) shows a basic Bayesian framework for phylogenetic analysis. It consists of prior distributions for the tree and parameters of interest $\Pr(g)$ and $\Pr(\Phi)$, a phylogenetic likelihood $\Pr(D|g,\Phi)$ and the posterior distribution $f(g,\Phi|D)$ that is to be inferred. So from a Bayesian perspective, the biological data, phylogenetic trees and all the parameters in the model exist in the form of a distribution and are related by the Bayes' formula. As a result, it is able to combine different models that describe various biological and evolutionary process at the same time.

$$f(g,\Phi|D) = \frac{\Pr(D|g,\Phi) \times \Pr(g) \times \Pr(\Phi)}{\Pr(D)} \tag{1}$$

, where g denotes the phylogenetic tree, Φ is a set of parameters and D represents the data available.

Moreover, due to the huge amount of data and the marginal likelihood being difficulty to calculate, Markov Chain Monte Carlo (MCMC) algorithms are adopted to get samples from the posterior distribution.

Tree proposals

The term "operator" used in this paper is defined as the tree proposal in Bayesian MCMC to propose a new state that is only dependent on the current state.

Generally, some classical operators can satisfy most sampling requirements, such as a random walk operator which proposes a new state by adding a random number, and a scale operator which multiplies the current state with a scale factor. But they are suitable for working on a single parameter or a set of parameters, such as population size. To make a proposal about the tree topology, it is also necessary to include other operators, such as subtree slide operator, swap operator narrow operator and so on. On top of this, there are also some extended operators available to help give the better performance. For instance, the prune-and-regraft operator selects a random subtree and reattaches the subtree at a new random branch. And the subtree-swap operator exchanges two random subtended subtrees.

What matters for developing an operator is that the proposal should be reversible. As is discussed in Green's paper [17], the constructed Markov transition kernel $P(x, dx')$ should satisfy the detailed balance, as is shown by Eq.(2). Consequently, the probability that the operator propose a new state from the current state is required to be equal to the probability that the proposed state goes back to current state. Therefore, in Eq.(3), to ensure the detailed balance in the MCMC chain, a ratio $q(x', dx)/q(x, dx')$ should always be included in the probability $\alpha(x, x')$ that the proposal made by the operator is accepted, which is defined as *Hasgtings ratio*.

$$\int_A \int_B \pi(dx) P(x, dx') = \int_B \int_A \pi(dx') P(x', dx) \tag{2}$$

, where $\pi(dx)$ is the target distribution, A and B are Borel sets in parameter space φ.

$$\alpha(x, x') = \min \left\{ 1, \frac{\pi(dx') q(x', dx)}{\pi(dx) q(x, dx')} \right\} \tag{3}$$

Uncorrelated relaxed clock model

A clock model is used to model how rates evolve along branches in the phylogenetic tree, so that a time tree can reconcile with the genetic distances between sequences described by a substitution model. For example, a strict clock model assumes evolution rates to be the same at every branch. However, a relaxed clock model allows rates to vary across lineages so that better estimates of divergence times can be obtained. As a incomplete way of relaxing, a random local clock model only allows rates to be different within a subregion of the tree.

In this paper, the proposed operator is based on an uncorrelated relaxed model, where the rates vary and follow a certain distribution. As is detailed in Ref.[18],

uncorrelated rates indicate that the rate on each branch is identically distributed and will be independently drawn from a certain distribution such as a log-normal distribution. As a result, the rates can change faster than making a slight move over multiple adjoining branches.

The proposed operator

We define the proposed operator as ConstantDistance Operator (cons for abbreviation).The flow chart of the proposed operators is shown in Fig.1. In a phylogenetic tree, the node operated on is denoted by **X**. And the proposed operator works differently on the internal nodes and the root of the tree. The details of the operations are introduced step by step as follows.

The internal node operator

Fig.2 represents the tree (or subtree of a phylogenetic tree) with the node **X**, which is randomly selected among the internal nodes, as well as its parental node and two child nodes. The original tree on the left is denoted by Tree. And the following 5 steps will propose a new tree, denoted by Tree', on the right.

Step 1 Get the parent node and two child nodes of **X**, denoted by **P**, **Ch1** and **Ch2** respectively.

Step 2 Get the nodes times of **X**, **P**, **Ch1** and **Ch2** , denoted by t_X, t_P, t_1, t_2, as well as the rates on the branches above the nodes, denoted by r_i, r_j, r_k.

Step 3 Propose a new node time for **X** by ${t_X}' = t_X + a$, where a follows a Uniform distribution with a symmetric window size, i.e. $a \sim Uniform[-w, +w]$. Make sure that the proposed time is valid, i.e. $\max\{t_1, t_2\} < {t_X}' < t_P$ holds.

Step 4 Propose new rates by using Eq.(4).

$$ {r_i}' = \frac{r_i \times (t_P - t_X)}{t_P - {t_X}'} \quad {r_j}' = \frac{r_j \times (t_X - t_1)}{{t_X}' - t_1} \quad {r_k}' = \frac{r_k \times (t_X - t_2)}{{t_X}' - t_2} \tag{4} $$

Step 5 Return the $HastingsRatio$.

The root operator

For the root of the tree, there are three strategies to propose the new rates and node times. To be more specific, Simple Distance is a way of proposing a new root time only. Considering the genetic distance, Small Pulley makes changes for branches on each side of the root. Moreover, under the constraint that the unrooted tree is fixed, Big Pulley proposes a new tree topology by rearranging the root. Details are discussed below

Simple Distance

Fig.3 shows the trees that are rooted at the selected node X. The initial tree in Fig.3(a) represents the current state, which is denoted by Tree. Inspired by how the operations on internal nodes, we will use the following steps to propose a new tree, denoted by Tree', and keep the genetic distances d_i and d_x constant at the same time, as is shown in Fig.3(b).

Step 1 Get the child nodes of the root **X**, denoted by **son** and **dau**. Their corresponding node times and rates are t_X, t_j, t_k and r_i, r_x.

Step 2 Propose a new node time for the root **X** by **X** by $t_X{}' = t_X + a$, where $a \sim Uniform[-w, +w]$. Make sure that $t_X{}' > \max\{t_j, t_k\}$ holds.

Step 3 Propose new rates for branches on each side of the root by using Eq.(5).

$$r_i{}' = \frac{r_i \times (t_X - t_j)}{t_X{}' - t_j} \qquad r_x{}' = \frac{r_x \times (t_X - t_k)}{t_X{}' - t_k} \tag{5}$$

Step 4 Return the $HastingsRatio$.

Small Pulley

Different from Simple Distance, a new genetic distance of branch on one side of the root is proposed in Small Pulley. As is indicated in Fig.3, Small Pulley proposed a new tree in Fig.3(c), based on the current state in Fig.3(a). In order to remain the total genetic distance of d_i and d_x unchanged, once $d_i{}'$ is proposed, d_x will be adjusted by $d_x{}'$ simultaneously. The detailed process includes the following four steps.

Step 1 Get the child nodes of the root **X**, denoted by **son** and **dau**. Their corresponding node times and rates are t_X, t_j, t_k and r_i, r_x. So the genetic distances can be calculated according to Eq.(6).

$$d_i{}' = r_i \times (t_X - t_j) \qquad D = d_i + r_x \times (t_X - t_k) \tag{6}$$

Step 2 Propose a new genetic distance for d_i by adding a random number that follows a Uniform distribution, i.e. $d_i{}' = d_i + b$, where $b \sim Uniform[-v, +v]$. Make sure that $0 < d_i{}' < D$ holds, where $D = d_i + d_x$.

Step 3 Propose new rates for branches on each side of the root by using Eq.(7).

$$r_i{}' = \frac{d_i{}'}{t_X - t_j} \qquad r_x{}' = \frac{D - d_i{}'}{t_X - t_k} \tag{7}$$

Step 4 Return the $HastingsRatio$.

Big Pulley

Big Pulley is used to sample the rates and times from a fixed unrooted tree so that the genetic distances among taxon are constant, which means the location of the root will be rearranged.

Firstly, a method called *Exchange (Node1,Node2)* is designed to propose a new tree topology in this circumstances. For the tree in Fig.4, once the method *Exchange (**B**,**C**)* is called, the following operations will be performed to proposed a new tree, as is shown on the right in the figure.

(1) Propose $d_i{}'$ by $d_i{}' = d_i + b$, where $b \sim Uniform[-v, +v]$. Make sure that $0 < d_i{}' < D$ holds, where $D = d_i + d_x$.
(2) Exchange the node **B** and **C** by pruning and grafting, i.e. cutting **B** (**C**) at its original position and attaching it to the original position of **C** (**B**).
(3) To maintain the genetic distances d_{AB}, d_{AC} and d_{BC} constant, the distances on the other three branches d_j, d_k and d_x will be adjusted by using Eq.(8).

$$d_j{}' = d_j \qquad d_k{}' = d_k - d_i{}' \qquad d_x{}' = d_x + d_i \tag{8}$$

As can be seen from the above descriptions, the method *Exchange (Node1,Node2)* is actually aimed at swapping two nodes and reassigning distances on the four branches.

Secondly, before applying this method in Big Pulley, there are two different tree shapes to take into consideration. In Fig.5, a symmetric tree is shown on the left, in which both the child nodes of the root have child nodes. But in an asymmetric tree, only one of the child nodes of the root has child nodes below it, which means the other child node of the root is a leaf node. Hence, Big Pulley also differs when it works on a symmetric and asymmetric tree. The corresponding operations are detailed in the following two parts.

Symmetric tree shape For the symmetric tree in Fig.5, the operations are illustrated in Fig.6, after which one of the four possible trees will be proposed.

Step 1 Get the child nodes of the root **X**, denoted by **son** and **dau**. And the child nodes below them are denoted by $Ch1$, $Ch2$, $Ch3$ and $Ch4$. Correspondingly, the node times are denoted by t_X, t_j, t_k.

Step 2 Propose a new node time for the root **X** by **X** by ${t_X}' = t_X + a$, where $a \sim Uniform[-w, +w]$.

Step 3 Propose a new node time either for **son** or **dau**. And apply the method using **dau** and either child node of **son**.

- With 0.5 probability to pick **son** and propose a new node time by ${t_j}' = t_j + a_1$, where $a_1 \sim Uniform[-w, +w]$. Make sure that $t_k < {t_j}' < {t_X}'$ holds. Then, there are two options to apply the method, i.e.
  ①: With 0.5 probability to *Exchange (**Ch1** and **dau**)*
  ②: With 0.5 probability to *Exchange (**Ch2** and **dau**)*
- With 0.5 probability to pick **dau** and propose a new node time by ${t_k}' = t_k + a_2$, where $a_2 \sim Uniform[-w, +w]$. Make sure that $t_j < {t_k}' < {t_X}'$ holds. Similarly, there are two options to apply the method, i.e.
  ③: With 0.5 probability to *Exchange (**Ch3** and **son**)*
  ④: With 0.5 probability to *Exchange (**Ch4** and **son**)*

Step 4 Update the rates using the adjusted genetic distances divided by the proposed node times.

Step 5 Return the $HastingsRatio$.

Asymmetric tree shape How to operate on the asymmetric tree in Fig.5 is illustrated Fig.7, in which there are three possible trees.

Step 1 Get the older child of the root, denoted by **O**, and the younger child of the root is denoted by **Y**. The node times of the root and its child nodes are denoted by t_O, t_{O1} and t_{O2}.

Step 2 Propose a new node time for **O** by ${t_O}' = t_O + a_3$, where $a_3 \sim Uniform[-w, +w]$.

Step 3 Apply the method using **Y** and either child nodes of **O**, which is dependent on ${t_O}'$.

- if ${t_O}'$ satisfies ${t_O}' > \max\{t_{O1}, t_{O2}\}$ or $t_{O1} = t_{O2}$, then there are two options, i.e.
  ⑤: With 0.5 Probability to *Exchange (**O1** and **Y**)*
  ⑥: With 0.5 Probability to *Exchange (**O2** and **Y**)*

- if $t_O{}'$ satisfies $\min\{t_{O1}, t_{O2}\} < t_O{}' < \max\{t_{O1}, t_{O2}\}$, then there is only one option, i.e.

  ⑦: Exchange the older child of **O** and **Y**. In the example here, we apply *Exchange (**O1** and **Y**)*.

Step 4 Update the rates using the adjusted genetic distances divided by the proposed node times.

Step 5 Return the *HastingsRatio*.

Calculate the HastingsRatio As is mentioned in the previous section, the operator in phylogenetic analysis based on Bayesian MCMC should make a reversible proposal to satisfy the detailed balance (Eq.(2)). Therefore, the last step in the ConstantDistance Operator is to return the *HastingsRatio* to the acceptance probability (Eq.(3)).

In the operations for internal nodes, the rates on the three branches that are linked the internal node are proposed by one random number a. Apparently, the dimension in the two parameter subspaces is different. To solve this problem, a Jacobian matrix $\mathbf{J_1}$ is constructed which takes the vector $\mathbf{X} = [t_x, r_i, r_j, r_k]$ as input and outputs the vector $\mathbf{Y} = [t_x{}', r_i{}', r_j{}', r_k{}']$, as is shown in Eq.(9).

$$\mathbf{J_1} = \left[\begin{array}{cccc} \frac{\partial \mathbf{f}}{\partial t_x} & \frac{\partial \mathbf{f}}{\partial r_i} & \frac{\partial \mathbf{f}}{\partial r_j} & \frac{\partial \mathbf{f}}{\partial r_k} \end{array} \right] = \left[\begin{array}{cccc} \frac{\partial f_1}{\partial t_x} & \frac{\partial f_1}{\partial r_i} & \frac{\partial f_1}{\partial r_j} & \frac{\partial f_1}{\partial r_k} \\ \frac{\partial f_2}{\partial t_x} & \frac{\partial f_2}{\partial r_i} & \frac{\partial f_2}{\partial r_j} & \frac{\partial f_2}{\partial r_k} \\ \frac{\partial f_3}{\partial t_x} & \frac{\partial f_3}{\partial r_i} & \frac{\partial f_3}{\partial r_j} & \frac{\partial f_3}{\partial r_k} \\ \frac{\partial f_4}{\partial t_x} & \frac{\partial f_4}{\partial r_i} & \frac{\partial f_4}{\partial r_j} & \frac{\partial f_4}{\partial r_k} \end{array} \right] \tag{9}$$

, where the functions f_1, f_2, f_3 and f_4 represent how the operator makes an proposal. After substituting Eq.(4) in Eq.(9), the *HastingsRatio* for the internal nodes can be derived by Eq.(10).

$$r = \frac{q(x', dx)}{q(x, dx')} = \frac{\Pr(-a)}{\Pr(a)} \left|\mathbf{J_1}\right| \tag{10}$$

Since the random number a is drawn from a Uniform distribution, the probability $\Pr(-a)$ is equal to $\Pr(a)$, which makes Eq.(10) simply become $r = |\mathbf{J_1}|$.

For the root of the phylogenetic tree, the operations are concerned with tree topology. To make the proposed topology reversible, a factor μ is defined and will be included in the *HastingsRatio*, which is calculated by Algorithm 1.

Experimental results and analysis

In this section, a series of experiments are conducted, together with some analysis and discussions, to validate the effectiveness and efficiency of the proposed operator. Specifically, by sampling from the prior distributions, in which no alignments are involved, the correctness of the operator is proved. After the well-calibrated simulation study, it turns out that the operator is able to work properly with other operators and sample the simulated data correctly. Besides, comparing ESS and running time, it is demonstrated that the performance is improved when the proposed operator is included.

Algorithm 1 Calculation of μ for Big pulley

if the node that has been exchanged with **dau** or **dau** has child nodes **then**
  $\alpha = \beta = 0.25$
else if $t_R > t_L$ **then**
  $\alpha = 1, \beta = 0.5$
else if $t_R < t_L$ **then**
  $\alpha = 0.5, \beta = 1$
else if $t_R = t_L$ **then**
  $\alpha = \beta = 1$
end if
if the node that has been exchanged with **O** has child nodes **then**
  $\gamma = 0.25$
else
  $\gamma = 0.5$
end if
for ① ② **do**
  Return $\mu = \frac{\alpha}{0.25}$
end for
for ③ ④ **do**
  Return $\mu = \frac{\beta}{0.25}$
end for
for ⑤ ⑥ **do**
  Return $\mu = \frac{\gamma}{0.5}$
end for
for ⑦ **do**
  Return $\mu = \frac{0.25}{1}$
end for

Sample from the prior

In Fig.8, a tree with three taxa A, B and C is used as a small example in this experiment, where Tree1 is set as the initial state. First of all, a LogNormal distribution is set as the rate prior, given by Eq.(11).

$$r = [r_i \quad r_j \quad r_k \quad r_x] \sim LogNormal(M = -3, S = 0.5) \tag{11}$$

Additionally, a Coalescent model with constant population size ($N = 0.3$) is used to describe the tree prior. Hence, for a tree with 3 taxa, the probability of node times given the tree is calculated by Eq.(12).

$$P(t_E, t_D) = (\frac{1}{N} \times e^{-\frac{1}{N}(t_E - t_D)}) \times (\frac{1}{N} \times e^{-\frac{3}{N}t_D}) \tag{12}$$

After the prior is specified, the distribution to sample can be exactly known, since the samples are from the prior distributions. In other words, as the rates are the functions of its genetic distance and times, the joint distribution to sample can be represented by Eq.(13).

$$\begin{aligned} P(r,t) &= P(t_E, t_D) \times P(r_i) \times P(r_j) \times P(r_k) \times P(r_x) \\ &= P(t_E, t_D) \times P(\frac{d_i}{\Delta t1}) \times P(\frac{d_j}{\Delta t2}) \times P(\frac{d_k}{\Delta t3}) \times P(\frac{d_x}{\Delta t4}) \end{aligned} \tag{13}$$

, where Δt represents the time duration along the corresponding branch, and $P(.)$ is the probability of the LogNormal distribution. Therefore, the whole probability can be obtained by conducting numerical integration on Eq.(13), which shows the average over all the possible values of parameters.

Test the operator on internal nodes
The genetic distances, node times and rates for the Tree1 in Fig.8 are given in Table 1. To test roundly, two scenarios are considered. In each scenario, the genetic distances are fixed, the node time t_D starts from the initial value and is changed by the operator during the sampling process, so that node D moves between node A and E. Besides, to make sure that the result is robust, two different MCMC chain lengths are performed in each scenario.

The mean, mean error and the standard deviation of the sampled distribution are summarised in Table 2. By using Eq.(14), the actual joint distribution is obtained according to Eq.(13), and is used to evaluate the results, which is also included in Table 2. Besides, the histograms of MCMC samples, as well as the curves of the numerical integration of Eq.(14), are shown in Fig.9. It can be seen that the red curves well fit the histograms, and the mean values and standard deviations are consistent, which makes it safe to conclude that the proposed operators works properly on the internal nodes.

$$P(r,t_D)=\int_{t_D=0}^{t_E} P(t_E,t_D)\times P(\frac{d_j}{t_D})\times P(\frac{d_k}{t_D})\times P(\frac{d_i}{t_E-t_D})\times P(\frac{d_x}{t_E})d_{t_D} \quad (14)$$

Test the operator on root
Still starting from Tree1 in Fig.8, the initial settings for testing the root are given in Table 3. And the three strategies are tested separately in the following parts.

Using Simple Distance The root time t_E is sampled by Simple Distance, which ranges from 1 to positive infinity theoretically. Namely, all the genetic distances and the node time t_D are fixed. Hence, similar to Eq.(14), the sampled distribution of t_E can be obtained by Eq.(15).

$$P(r,t_E)=\int_{t_E=1}^{+\infty} P(t_E,t_D)\times P(\frac{d_j}{t_D})\times P(\frac{d_k}{t_D})\times P(\frac{d_i}{t_E-t_D})\times P(\frac{d_x}{t_E})d_{t_E} \quad (15)$$

The results are given in Table 4 and Fig.10(a). Obviously, the mean and the standard deviation are close enough, which confirms that the two distribution are the same. Thus, Simple Distance is proved to be correct.

Using Small Pulley Although both d_x and d_i are changed during the sampling process when using Small Pulley, the sum of d_x and d_i are kept 0.67 in this test, as is shown in Table 3. To make it simple to represent, only d_i is compared in this subsection.

Then, based on Eq.(13), the exact distribution of d_i can be obtained by Eq.(16), which is compared with the sampled distribution in Table 4 and Fig.10(b). Even though there exist some errors, the sampled parameters can be considered to follow the same distribution. So the Small Pulley also is able to provide the correct samples.

$$P(r,d_i)=\int_{d_i=1}^{0.67} P(t_E,t_D)\times P(\frac{d_j}{t_D})\times P(\frac{d_k}{t_D})\times P(\frac{d_i}{t_E-t_D})\times P(\frac{0.67-d_i}{t_E})d_{d_i} \quad (16)$$

Using Big Pulley For Tree1 in Fig.8 with tree taxa, a new tree, together with the root time t_E and it older child node time t_D, as well as a genetic distance d_i, is proposed by Big Pulley. In this case, the initial state Tree1 will either go to Tree2 or Tree3, as is shown in Fig.8. According to the initial settings in Table 3, the genetic distances remain unchanged during the process, i.e. $d_{AB} = 1$, $d_{AC} = 1$ and $d_{BC} = 1$ hold. Hence, the distribution we are about to achieve can be calculated by Eq.(17).

The statistical measurements, i.e. mean and standard deviation, are summarised in Table 4. And the histogram of samples and theoretical distribution are pictured in Fig. It is shown that the two distributions are consistent within the acceptable error range. Therefore, Big Pulley can also give the right combinations of rates and node times, under the condition that the genetic distances among taxa are constant.

$$\begin{aligned} P(t_E, t_D, d_i) = \int_{t_E=0}^{+\infty} \int_{t_D=0}^{t_E} \int_{d_i=0}^{0.5} P(t_E, t_D) \times P(\frac{0.5}{t_D}) \\ \times P(\frac{0.5}{t_D}) \times P(\frac{d_i}{t_E - t_D}) \times P(\frac{0.5 - d_i}{t_E}) d_{d_i} d_{t_D} d_{t_E} \end{aligned} \tag{17}$$

Well-calibrated simulation study

Fig.11 shows the models and prior distributions used in this study. In the first place, independent samples from the prior distributions are set to be the values of random parameters, such as κ in HKY model, UcldStdev $S1$ for rate prior, and so on. In this test, there are 100 independent samples obtained which are utilised to simulate 100 sets of sequence alignment. To make this study more robust, two groups of sequence data are simulated, one group with 20 taxa and the other with 120 taxa.

In the second place, the two groups of simulated data are sampled by using ConstantDistance Operator in BEAST2. Acoording to models in Fig.11 , there are 7 parameters sampled. The sampled values are compared to the real values that refer to the independent samples from the prior distributions that are used to simulate each sequence data. The results of the two groups are shown in Fig.12 and Fig. 13. Besides, Table 5 shows the number of real values that are within the 95% HPD of the sampled distribution.

Performance comparison

To evaluate the performance, only one set of the 100 simulated data sets in the two groups with 20 and 120 taxa is sampled. And each data set from the two groups is sampled 100 times.

To make a comparison, the same data set will be run 100 times with and without using the ConstantDistance Operator. In addition, as rates can be also modeled by discrete categories in BEAST2, the same data set is also run 100 times by using rate categories.

Effective Sample Size (ESS) of a parameter sampled from an MCMC chain in BEAST2 is the number of effectively independent draws from the posterior distribution that the Markov chain is equivalent to. The larger ESS indicates the better estimation of the parameter. In this study, the ESS of UcldStdev in all the simulations are compared in order to measure the predictability of ConstantDistance Operator.

Meanwhile, all the sampling process are run in NeSI, using 1 CPU and one single thread. And the user time is recorded as the running time of each simulation.

The results are summarised in Fig.14 and Table 6.

A real data analysis

In this section, a real data set with 83 primates is used to further evaluate the performance of ConstantDistance Operator. The ESS and running time are compared in Fig.15.

Sampling a fixed unrooted tree

The test in this subsection is to elucidate the correctness of the proposed operator. As is shown in Fig.16(a), this test starts from an unrooted tree constructed by Neighbour-Joining(NJ) algorithm and will be fixed during the test. Specifically, after the NJ algorithm provides the distances on the branch, we adopted the midpoint method to root the tree. Then randomly assign the divergence times for each node under the given topology and distances, so that the rates on each branch can be calculated in accordance with the node times and the genetic distances. Afterwards, the rooted time tree shown in Fig.16(b) is used as the initial state in MCMC chain, and is sampled by only using the proposed operator.

The summary tree is obtained after running the simulation in BEAST2, which is pictured in Fig.17(a). As can be seen, the result is consistent with, compared to the tree in Fig.17(b). In other words, the proposed operator is able to provide the correct samples when given a exactly known tree.

However, it is noticed that there exits some uncertainties in the node times, according to the non-1 posterior (0.9433) labeled on the node in Fig.17(a). That is to say, even though the summary tree gives the most possible rooted tree, it is probable for the root to locate on any other branches.

Correlation analysis of rates and times

To understand how the operator works on rates and node times, this section visualises the correlations of rates and times. Take an internal node as an example. Assuming the operator increases its current node time, to maintain the distances on the three branches linked to this node, the rates are supposed to decrease or increase according to its original state. If the rate increases (decrease) along with the node time, then they have a positive (negative) correlation.

By analysing the seven ratites data set, we filter the tree in the sampled state by the shared common ancestor for each taxa, as well as the rates and node times. And we conducted s pairwise comparison between each rate and node time to see how they are correlated, as is shown in Fig.18. Remember that this is only an average pairwise comparison. For higher dimensions, the results will be different.

Conclusion

The efficiency is of great significance in phylogenetic analysis based on Bayesian MCMC. This paper discussed how to make a tree proposal to deal with rates and node times on condition that the genetic distances are constant. In all, the proposed operator integrates four strategies to work on internal nodes and the root in a

phylogenetic tree. By the tests that sample from known prior distribution, the ConstantDistance Operator shows the correct functionality. Then, after comparing the results of a series of simulations using both simulated data and real data, it is verified that the ConstantDistance Operator is valid and more efficient. It is believed that the work in this paper will make some contributions to the research in phylogenetic analysis by providing more efficient methods.

Acknowledgements
The work is partially supported by scholarship from China Scholarship Council (File No.201706990021).

References

1. Hackett, S.J., Kimball, R.T., Reddy, S., Bowie, R.C., Braun, E.L., Braun, M.J., Chojnowski, J.L., Cox, W.A., Han, K.-L., Harshman, J., *et al.*: A phylogenomic study of birds reveals their evolutionary history. science **320**(5884), 1763–1768 (2008)
2. Szalay, F.S., Delson, E.: Evolutionary History of the Primates. Academic Press, ??? (2013)
3. Kellogg, E.A.: Evolutionary history of the grasses. Plant physiology **125**(3), 1198–1205 (2001)
4. Sawabe, T., Kita-Tsukamoto, K., Thompson, F.L.: Inferring the evolutionary history of vibrios by means of multilocus sequence analysis. Journal of bacteriology **189**(21), 7932–7936 (2007)
5. Garnery, L., Cornuet, J.-M., Solignac, M.: Evolutionary history of the honey bee apis mellifera inferred from mitochondrial dna analysis. Molecular ecology **1**(3), 145–154 (1992)
6. Drummond, A.J., Rambaut, A.: Beast: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology **7**(1), 214 (2007)
7. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J.: Beast 2: a software platform for bayesian evolutionary analysis. PLoS computational biology **10**(4), 1003537 (2014)
8. Huelsenbeck, J.P., Ronquist, F.: Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics **17**(8), 754–755 (2001)
9. Paradis, E., Claude, J., Strimmer, K.: Ape: analyses of phylogenetics and evolution in r language. Bioinformatics **20**(2), 289–290 (2004)
10. Yang, Z., Rannala, B.: Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. Molecular biology and evolution **14**(7), 717–724 (1997)
11. Rannala, B., Yang, Z.: Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. Genetics **164**(4), 1645–1656 (2003)
12. Yang, Z., Yoder, A.D.: Comparison of likelihood and bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Systematic biology **52**(5), 705–716 (2003)
13. Reis, M.d., Yang, Z.: Approximate likelihood calculation on a phylogeny for bayesian estimation of divergence times. Molecular Biology and Evolution **28**(7), 2161–2172 (2011)
14. Kobert, K., Stamatakis, A., Flouri, T.: Efficient detection of repeating sites to accelerate phylogenetic likelihood calculations. Systematic biology **66**(2), 205–217 (2017)
15. Lakner, C., Van Der Mark, P., Huelsenbeck, J.P., Larget, B., Ronquist, F.: Efficiency of markov chain monte carlo tree proposals in bayesian phylogenetics. Systematic biology **57**(1), 86–103 (2008)
16. Höhna, S., Drummond, A.J.: Guided tree topology proposals for bayesian phylogenetic inference. Systematic biology **61**(1), 1–11 (2011)
17. Green, P.J.: Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika **82**(4), 711–732 (1995)
18. Drummond, A.J., Ho, S.Y., Phillips, M.J., Rambaut, A.: Relaxed phylogenetics and dating with confidence. PLoS biology **4**(5), 88 (2006)
19. Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., Ward, R.: Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. Nature **409**(6821), 704 (2001)

Figures

Tables

	genetic distances (fixed)				t_D	t_E	initial rates			
	d_j	d_k	d_x	d_i	initial	(fixed)	r_j	r_k	r_x	r_i
Scenario 1	0.1	0.2	0.4	0.27	1	10	0.1	0.2	0.04	0.03
Scenario 2	0.4	0.8	2.4	1.6	0.4	0.8	1	2	3	4

Table 1 Initial settings for internal nodes

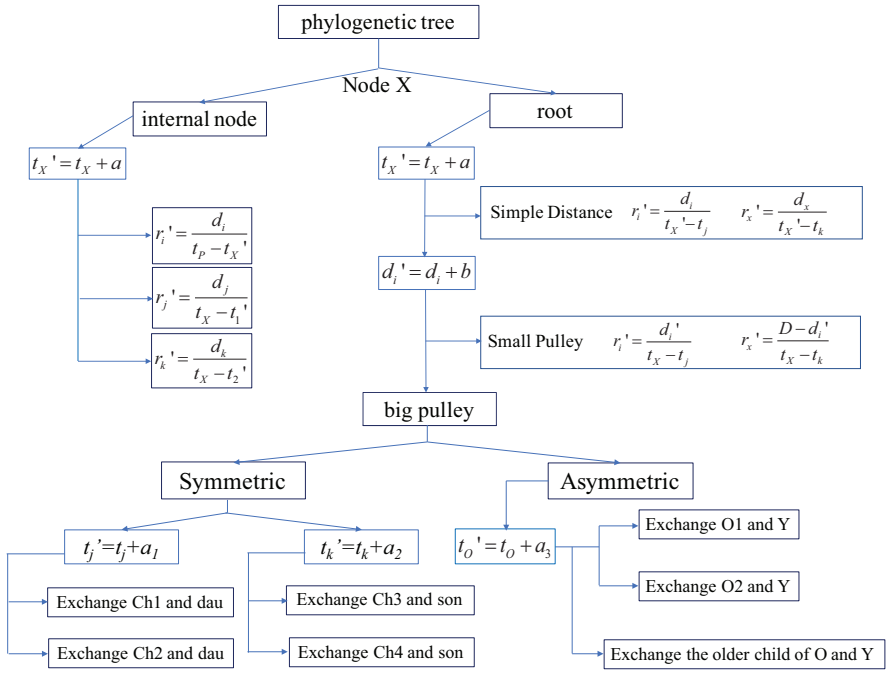


Figure 1 The flow chart of the constant distance operator.

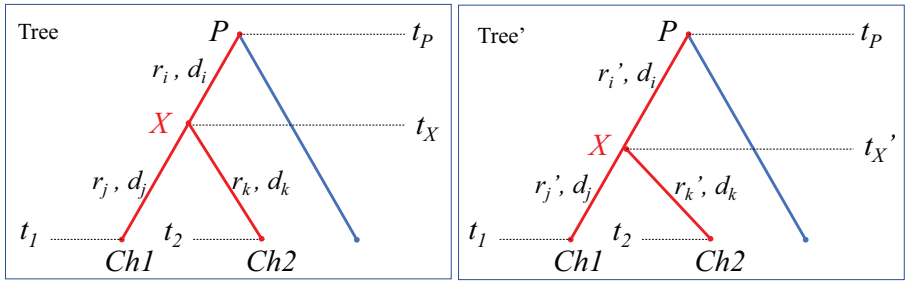
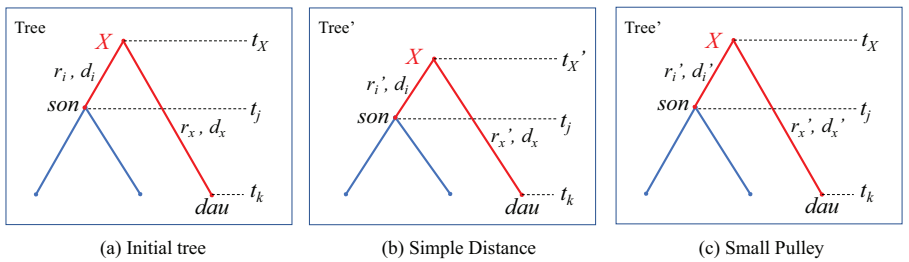


Figure 2 The illustration of internal nodes.



(a) Initial tree (b) Simple Distance (c) Small Pulley

Figure 3 The illustration of root using simple distance.

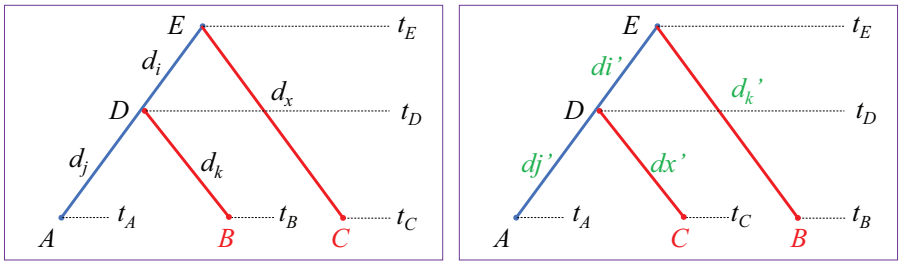


Figure 4 The illustration of Exchange (B, C) method.

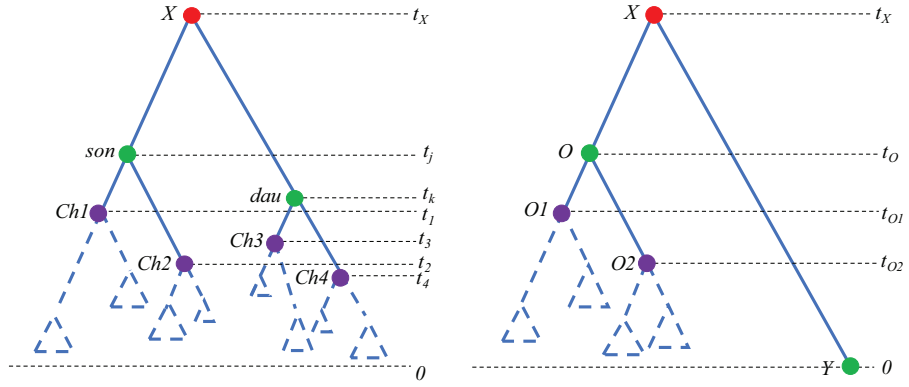


Figure 5 The illustration of tree shapes. The symmetric tree is on the left and the asymmetric tree is on the right. The dashed triangles represent the potential subtrees rooted at the nodes.

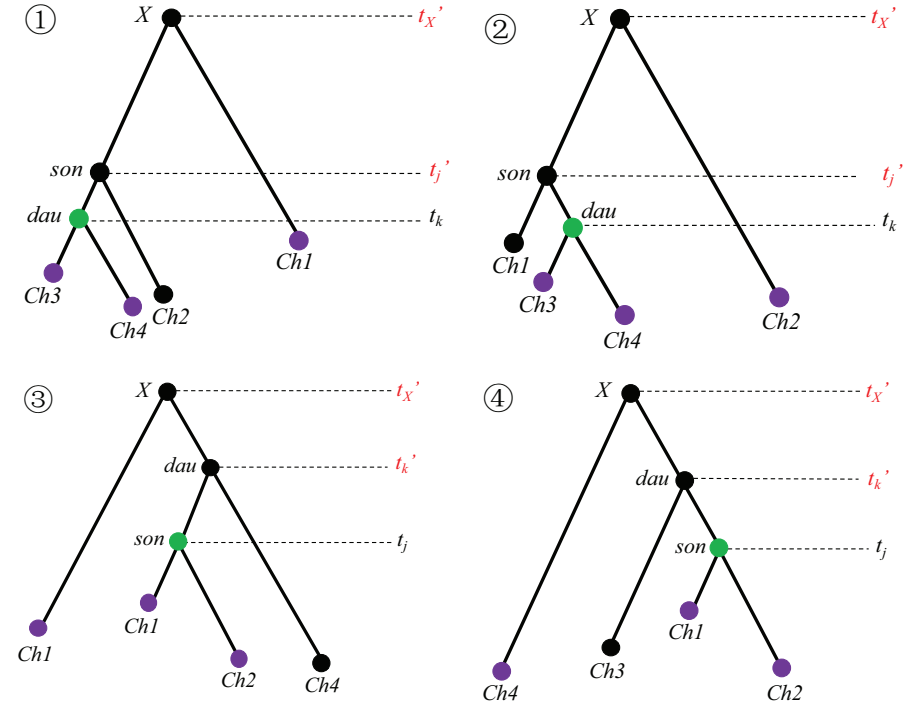


Figure 6 The illustration of symmetric trees.

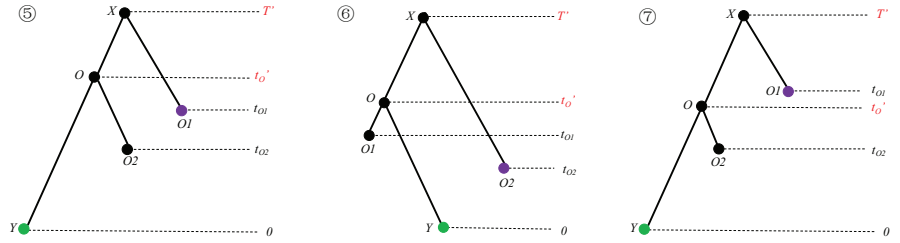


Figure 7 The illustration of asymmetric trees.

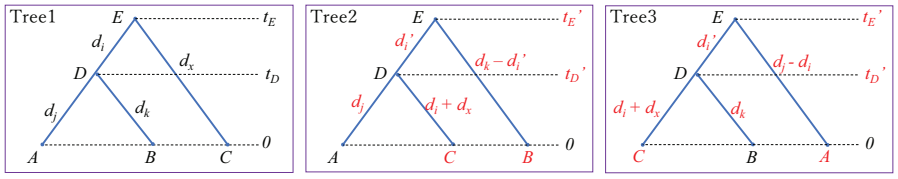
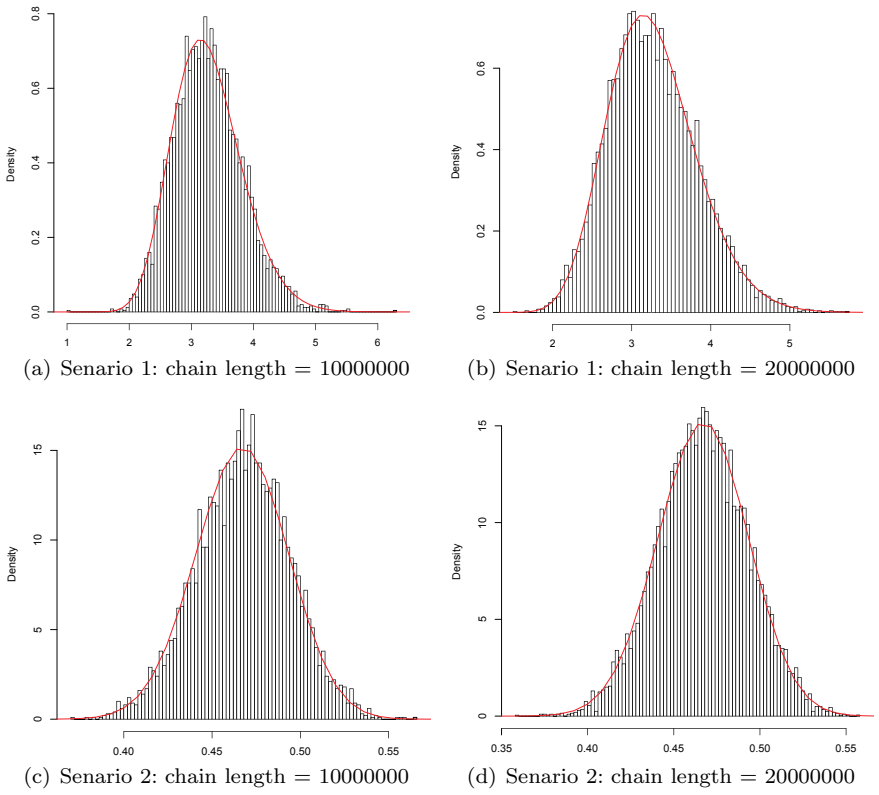


Figure 8 The illustration of sampling from prior.



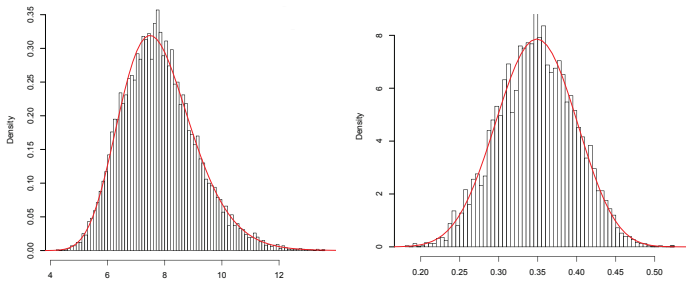
(a) Senario 1: chain length = 10000000

(b) Senario 1: chain length = 20000000

(c) Senario 2: chain length = 10000000

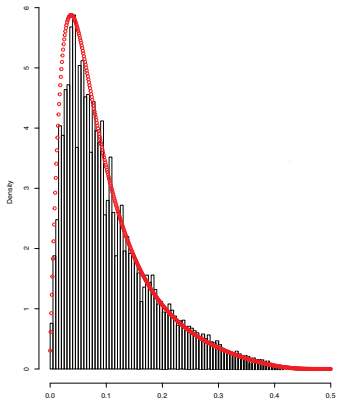
(d) Senario 2: chain length = 20000000

Figure 9 Distributions of internal nodes

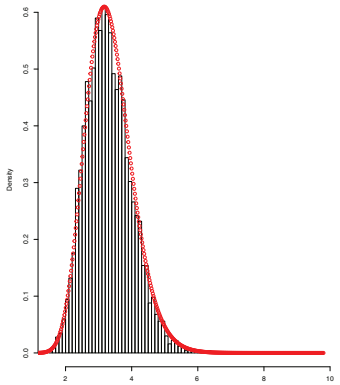


(a) t_E in Simple Distance

(b) d_i in Small Pulley



(c) d_i in Big pulley



(d) t_E in Big pulley

Figure 10 Distributions of root

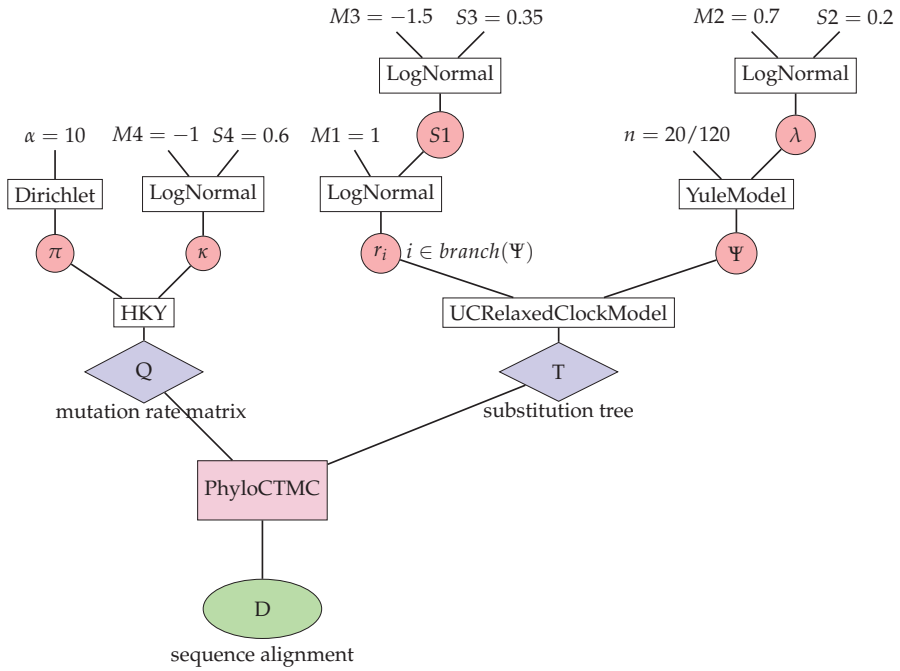


Figure 11 The models and prior distributions to simulate the sequence data.

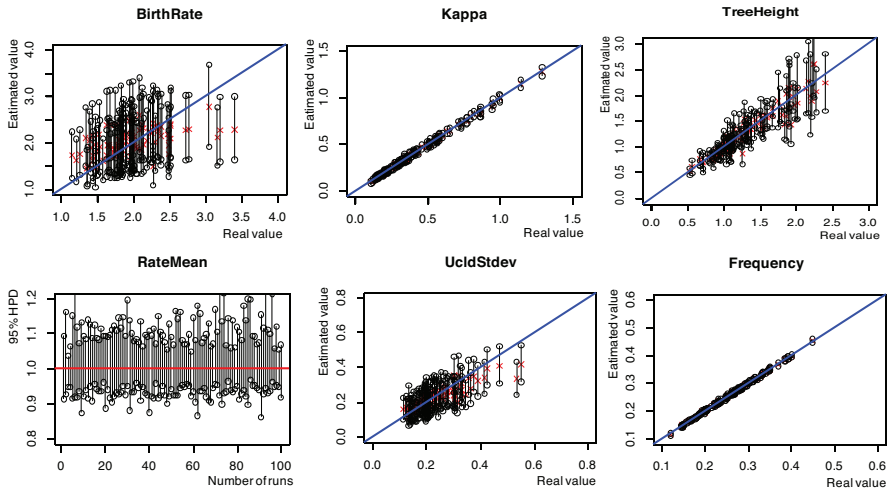


Figure 12 Comparison of simulation study with 20 taxa.

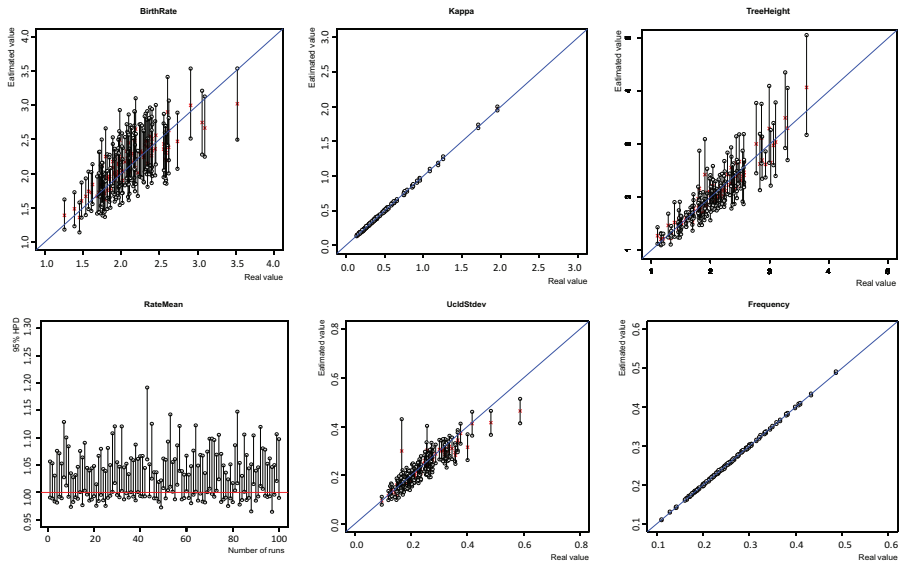


Figure 13 Comparison of simulation study with 120 taxa.

	Chain Length	Sample from MCMC			R curve			Plot
		Mean	Err	StdEv	Mean	Err	StdEv	
Senario 1	10000000	3.2727	8.3e-3	0.5467	3.2669	1.3e-06	0.5553	Fig.9(a)
	20000000	3.271	6.1e-3	0.5616				Fig.9(b)
Senario 2	10000000	0.4677	3.9e-04	0.0265	0.4667	3.5e-05	0.0262	Fig.9(c)
	20000000	0.4672	2.8e-04	0.0262				Fig.9(d)

Table 2 Results of internal nodes

strategy	genetic distances				t_D	t_E	initial rates			
	d_j	d_k	d_x	d_i			r_j	r_k	r_x	r_i
Simple distance	0.1	0.2	0.4	0.27	1	10	0.1	0.2	0.04	0.03
Small pulley	0.1	0.2	0.67		1	10	0.1	0.2	0.04	0.03
Big pulley	0.5	0.5	0.5		5	10	0.1	0.1	0.03	0.04

Table 3 Initial settings for simple distance

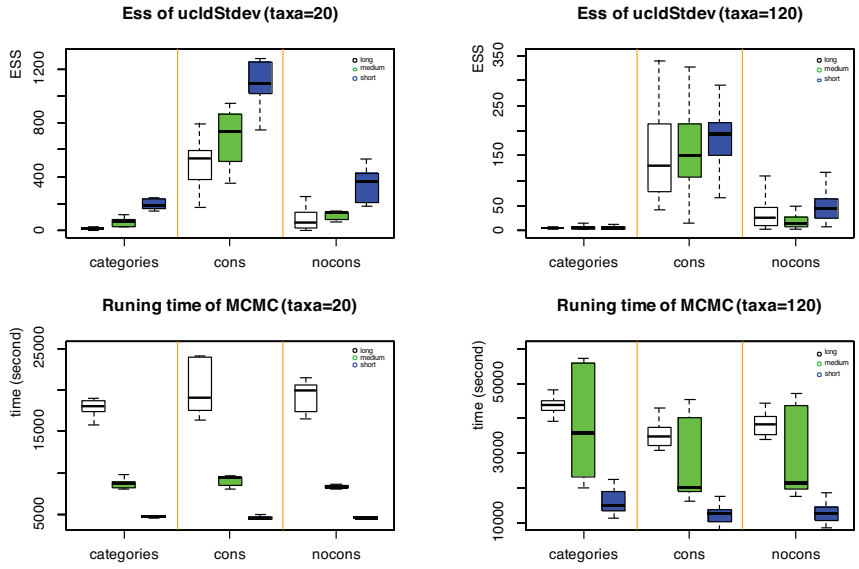


Figure 14 Comparison of ESS and running time using simulated data.

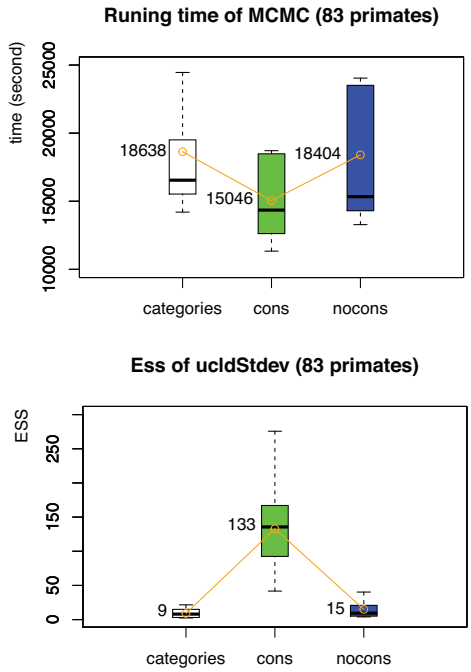
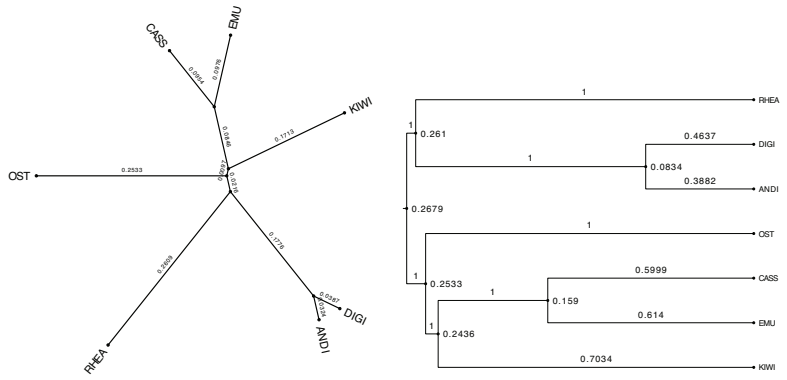


Figure 15 Comparison the ESS and ESS using primates data

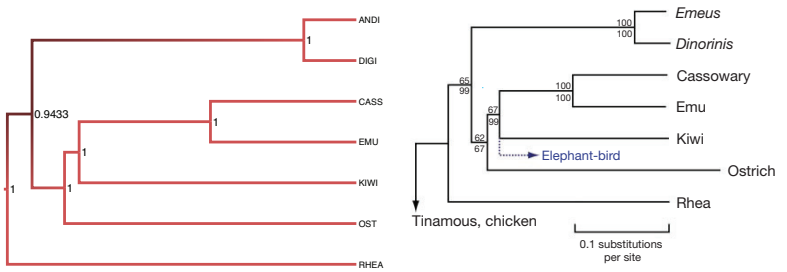
Strategy	Variable	Sample from MCMC		R curve		Plot
		Mean	StdEv	Mean	StdEv	
Simple distance	t_E	7.8081	1.2884	7.818691	1.299236	Fig.10(a)
Small Pulley	d_i	0.3480	0.0492	0.3476	0.0494	Fig.10(b)
Big Pulley	d_i	0.1016	0.0766	0.0960	0.0760	Fig.10(c)
	t_E	3.3017	0.6908	3.3095	0.6912	Fig.10(d)

Table 4 Results for root



(a) Unrooted tree (b) Rooted time tree

Figure 16 Initialising the rooted time tree based on an unrooted tree



(a) Uncertainties in summary tree (b) Ratite phylogeography [19]

Figure 17 Comparison of ratites phylogenetic trees

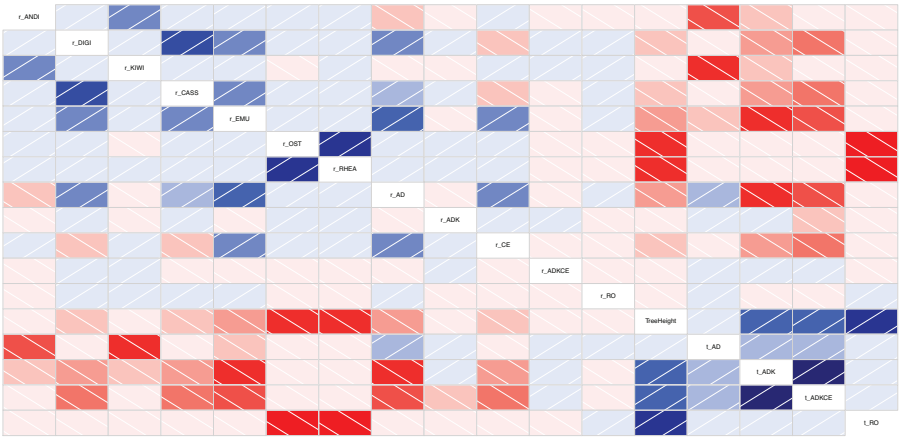


Figure 18 Correlation between rates and times in the ratites tree.

	BirthRate	TreeHeight	RateMean	UcldStdev	Kappa	Frequency
20 taxa	93	98	100	95	100	100
120 taxa	100	98	85	94	100	100

Table 5 Number of real values lying in the 95% HPD

		ESS		Running time	
	Length	20 taxa	120 taxa	20 taxa	120 taxa
categories	20000	12.80167	4.405382	18635	44155
	10000	57.99529667	6.316446	8660	36406
	5000	170.7803167	8.395184	4690	15956
Average		80.52576111	6.372337333	10662	32172
cons	20000	615.8584	146.700042	20207	35406
	10000	645.72143	161.220762	8967	25589
	5000	993.1825567	186.265154	4581	12487
Average		751.5874622	164.7286527	11252	24494
nocons	20000	86.46476333	62.77579	19344	38245
	10000	152.89205	19.598514	8361	30521
	5000	296.23683	47.514258	4499	12940
Average		178.5312144	43.29618733	10735	27236

Table 6 Results of ESS and running time