RESEARCH

Improving the performance of relaxed clock model in phylogenetic analysis

Rong Zhang and Alexei Drummond*

*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 to 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 at the same time, under the condition that the genetic distances are 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, a series of experiments have been performed by implementing the proposed operator in BEAST2 software. The results prove that the proposed operator is able to improve the performance by giving better estimations 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. Early in 1997, 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]. By virtue of a sampling method, a set of trees with probabilities is obtained, so that Monte Carlo integration can be used to integrate over the ancestral speciation times for particular trees. 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 handled. 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. In addition, the proposed operator has been implemented as a new package in BEAST2 software. After a series of simulations and experiments, the results prove that the proposed operator can provide correct samples and work properly in Bayesian phylogenetic analysis. According to the comparisons, it is concluded that by using the proposed operator, the efficiency of sampling parameters can be increased by.

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 g and a set of parameters of interest Φ, a phylogenetic likelihood that conveys information from data D, and the posterior distribution to be inferred, which is denoted in the form of probability density by $p(g)$, $p(\Phi)$, $p(D|g, \Phi)$, $f(g, \Phi|D)$ correspondingly. From a Bayesian perspective, the biological data, phylogenetic trees and the parameters in the model exist with some probabilities and are related by the Bayes' formula. As a result, it is able to jointly infer evolutionary history of species with various models that describe the evolutionary process.

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

On the other hand, 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 phylogenetics, and is used to propose a new state that is only dependent on the current state.

For now, there are some classical and frequently-used operators, 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 represented by Eq.(2). In other words, 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, as is shown 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.

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

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

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

, where x x' dx dx'.The ratio $q(x', dx)/q(x, dx')$ is called Hasgtings ratio.

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.

This paper is aimed at improving the performance of relaxed clock model in Bayesian phylogenetic analysis. 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

In this section, we define the proposed operator as ConstantDistance Operator. Fig.1 illustrates the flow chart of the proposed operator. In a phylogenetic tree, the node to operate 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 in the following subsections.

The internal node operator

Fig.2 represents the tree (or subtree of a phylogenetic tree) with the node **X** that is randomly selected among the internal nodes. The original tree in the current state is denoted by g_{in}. And the following 5 steps will propose a new tree ${g_{in}}'$.

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. $a \sim U[-w, +w]$. 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).

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

Step 5 Return the Hastings ratio.

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 adjusts the distances of 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 node X. The original tree in the current state is shown in Fig.3(a), which is denoted by g_r. Inspired by the operations on internal nodes, we will use the following steps to propose a new tree ${g_{r1}}'$, and keep the genetic distances d_i and d_x constant at the same time, as is illustrated 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 $a \sim U[-w, +w]$. 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).

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

Step 4 Return the Hastings ratio.

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 illustrated in Fig.3, Small Pulley proposes a new tree ${g_{r1}}'$ in Fig.3(c), based on the original tree g_r in Fig.3(a). In order to maintain 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 4 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).

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

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 $b \sim U[-v, +v]$. 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).

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

Step 4 Return the Hastings ratio.

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. to be specific, for the original tree g in Fig.4, once the method *Exchange (**B**,**C**)* is called, the following operations will be performed to proposed a new tree g'.

- 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**).
- Propose ${d_i}'$ by ${d_i}' = d_i + b$, where $b \sim U[-v, +v]$. Make sure that $0 < {d_i}' < D$ holds, where $D = d_i + d_x$.
- 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. That is to say, after using *Exchange (**B**,**C**)*, the distances d_i, d_j, d_k and d_k will be adjusted to maintain the distances between node **A**, **B** and **C**, as the tree topology changes.

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 the asymmetric tree on the right, 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.6, the operations are illustrated in Fig.6, after which one of the four possible trees (① ② ③ ④) 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 ${t_X}' = t_X + a$, where $a \sim U[-w, +w]$.

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 $a_1 \sim U[-w, +w]$. 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. For example, suppose we propose tree ①, *Step 5* Return the Hastings ratio.

Asymmetric tree shape How to operate on the asymmetric tree in Fig.5 is illustrated Fig.7, in which there are three possible trees (⑤ ⑥ ⑦).

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

Step 2 Propose a new node time for the root $\mathbf{X}$ by $t_X{}' = t_X + a$, where $a \sim U[-w, +w]$. Moreover, propose a new node time for $\mathbf{O}$ by $t_O{}' = t_O + a_3$, where $a_3 \sim U[-w, +w]$. To make it valid, make sure that $t_O{}' < t_X{}'$ holds.

Step 3 Apply the method using $\mathbf{Y}$ and either child nodes of $\mathbf{O}$, 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 $\mathbf{O}$ and $\mathbf{Y}$. 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. To give an example, assume we propose tree ⑤, *Step 5* Return the Hastings ratio.

Calculate the Hastings ratio

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 calculate the Hastings ratio for the acceptance probability (Eq.(3)).

According to the third and forth steps in the operations for internal nodes, three rates on the branches linked to the selected internal node are proposed by one random number a that is used to change the node time. There are four parameters involved in this proposal, including 3-dimensional rate space and 1-dimensional time space. The proposed operator utilises one random number in time space and changes both time space and rate space, which makes the parametric spaces inconsistent. To solve this dimension-matching problem, as is mentioned in Green's paper [17], it is necessary to construct a Jacobian matrix. In Eq.(9), $\mathbf{J_1}$ deals with the parametric spaces before the proposal in vector $\mathbf{X} = [t_x, r_i, r_j, r_k]$ and after the proposal in vector $\mathbf{Y} = [t_x{}', r_i{}', r_j{}', r_k{}']$.

$$\mathbf{J_1} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial t_x} & \frac{\partial \mathbf{f}}{\partial r_i} & \frac{\partial \mathbf{f}}{\partial r_j} & \frac{\partial \mathbf{f}}{\partial r_k} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial t_x} & \frac{\partial f_1}{\partial r_i} & \frac{\partial f_1}{\partial r_j} & \frac{\partial f_1}{\partial r_k} \\ \frac{\partial f_2}{\partial t_x} & \frac{\partial f_2}{\partial r_i} & \frac{\partial f_2}{\partial r_j} & \frac{\partial f_2}{\partial r_k} \\ \frac{\partial f_3}{\partial t_x} & \frac{\partial f_3}{\partial r_i} & \frac{\partial f_3}{\partial r_j} & \frac{\partial f_3}{\partial r_k} \\ \frac{\partial f_4}{\partial t_x} & \frac{\partial f_4}{\partial r_i} & \frac{\partial f_4}{\partial r_j} & \frac{\partial f_4}{\partial r_k} \end{bmatrix} \tag{9}$$

, where the functions f_1, f_2, f_3 and f_4 represent how the operator makes an proposal, i.e. the way of proposing new values. After substituting Eq.(4) in Eq.(9), the

Hastings ratio for the internal nodes can be derived by Eq.(10).

$$r_1 = \frac{q(x', dx)}{q(x, dx')} = \frac{\Pr(-a)}{\Pr(a)} \left|\mathbf{J_1}\right| \tag{10}$$

, where the probability $\Pr(-a)$ is equal to $\Pr(a)$ since the random number a is drawn from a Uniform distribution.

Likewise, the Jocobian matrix for Simple Distance, Small Pulley and Big Pulley can be obtained by . For the strategy of Big Pulley, the operations are concerned with tree topology. To make the proposed topology reversible, a factor μ is defined and will be included in the Hastings ratio, which is calculated by Algorithm 1. Therefore, the Hastings ratio when the proposed operator is working on the root of a phylogenetic tree is derived by Eq.

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
```

Experimental results and analysis

In this section, a series of experiments are conducted by implementing the proposed operator in BEAST2. Some analysis and discussions are also included to validate the effectiveness and efficiency of the proposed operator. Firstly, 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, by comparing ESS and running time, it is demonstrated that the performance is improved when using the proposed operator.

Sample from the prior

Fig.8 shows a tree with three taxa A, B and C used as a small example in this experiment. In the figure, g_1 is set as the initial tree. Firstly, a LogNormal distribution is used as the rate prior in the uncorrelated relaxed clock model, given by

Eq.(11).

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

In addition, 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 in Fig.8 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 drawn from the prior distributions. In other words, according to the rates being 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 $\Delta t1$, $\Delta t2$, $\Delta t3$, $\Delta t4$ represents the time duration along the corresponding branch, and $P(.)$ is the probability of certain rate values in the LogNormal distribution. Therefore, the whole probability can be obtained by conducting numerical integration on Eq.(13), which shows the probability distribution over all the possible values of parameters.

Test the operator on internal nodes

The genetic distances, node times and rates for the g_1 in Fig.8 are given in Table 1. To test roundly, two scenarios are designed. 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 MCMC samples are summarised in Table 2. Besides, according to Eq.(13), the actual joint distribution is obtained by using Eq.(14), and is used to evaluate the results, which is also included in Table 2. Moreover, the histograms of MCMC samples that indicate the sampled distributions, as well as the curves of the numerical integration of Eq.(14), are shown in Fig.9. From Table 2 and Fig.9, it can be seen that the red curves well fit the black 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.

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

Test the operator on root

Still starting from g_1 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 joint distribution of t_E and rates to sample 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, 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 is also 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 g_1 in Fig.8 with tree taxa, a new tree, together with the root time t_E and node time of its older child t_D, as well as a genetic distance d_i, is proposed by Big Pulley. In this case, the initial tree g_1 will either go to g_2 or g_3, as is shown in Fig.8. So the samples are repeatedly drawn from the 3 trees. Besides, 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).

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

The statistical measurements, i.e. mean and standard deviation, are compared in Table 4. The histograms of samples and theoretical distributions of d_i and t_E are pictured in Fig.10(c) and Fig.10(d). 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.

Well-calibrated simulation study

Fig.11 shows the framework used in this study, including the evolutionary models and parametric distributions. In the first place, 100 independent samples of Ucld-Stdev $S1$, frequency π, BirthRate λ in Yule model, κ in HKY model are obtained from their prior distributions. Then, the samples are utilised to simulate nucleotide sequences. To make this study more robust, two groups of sequence data are simulated, one group with 20 taxa and the other with 120 taxa ($n = 2/120$). So in each group, there are 100 sets of sequence alignment, each with length of 10 thousand. In the second place, the two groups of 200 simulated data sets are specified in 200 seperate XML files that include the same models in Fig.11 and ConstantDistance Operator. Afterwards, the XML files are run by BEAST2.

Finally, the parameters estimated by BEAST2 that integrates the proposed operator are compared to the real values that are used to simulate each sequence data set. The results of the two groups of simulated data sets are shown in Fig.12 and Fig.Acoording to models in Fig.11 , there are 7 parameters sampled. 13. As is shown in the figures, there are four random variables sampled, i.e. BithRate, UcldStdev, Kappa and Frequency, among which BithRate determines TreeHeight. And RateMean is fixed to 1, because of which the mean values of rates for the 100 data sets are all compared with 1. Besides, Table 5 also lists the number of real values that are within the 95% HPD of the sampled distribution. Indicated in the figures and table, 95 percent of estimated values are close to their corresponding real values, which proves the correctness of the estimation in both small and large data sets.

Performance comparison

The performance of a phylogenetic analysis in BEAST2 is concerned with how well the parameters are estimated and how much time it requires to run an MCMC chain. Effective Sample Size (ESS) of a parameter 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. Besides, whether the ConstantDistance Operator is time-consuming or not is quantified by the time for BEAST2 to finish running the input XML file. Therefore, ESS and running time are adopted to valuate the predictability of the ConstantDistance Operator.

On the other hand, to make comparisons, the exact same data set will be specified in XML with three different situations, 1) with using the ConstantDistance Operator (cons), 2) without using the ConstantDistance Operator, and 3) using operators that sample discrete rate categories. In addition, because of the randomness of an MCMC chain, the XML file for each situation will be run 100 times. Last but not Least, all the running jobs are submitted to Mahuika, a platform of NeSI that provides high performance computing [19]. By using one CPU and one single thread, the user time for performing each job is recorded as the running time of the corresponding simulation. And ESS of the parameters are obtained by input the log file into Tracer, a programme for analysing log files output by softwares such as BEAST2.

A simulated data set analysis

We use the same models in Fig.11 to simulate 3 data sets with different length for each group with 20 and 120 taxa. So there are 6 data sets in total, i.e. 2 long sequences with 20 thousand sites, 2 medium with 10 thousand sites and 2 short with 5 thousand sites. And each data set will be run 100 times in each situation.

The ESS and running time are summarised in Fig.14 and Table 6. In Fig.14, by comparing the ESS of using the proposed operator with other two situations, no matter how many taxa are involved, the ESS tends to be larger when the proposed operator is included. When comparing the running time, it is more obvious for long sequence data sets that the running time is deceased in the situation where the proposed operator is used. According to Table 6,

A real data set analysis

We also use a real data set with 83 primates to further evaluate the performance of ConstantDistance Operator. The data set is assembled from. Similarly, the data set will be specified in 3 different situations and will be run 100 times in each situation. The ESS and running time are shown in Fig.15. As can be seen from the figure, the analysis with using the ConstantDistance Operator costs less time and achieves much larger ESS. After calculating the mean value of the 100 runs in each situation, the ConstantDistance Operator improves the performance by

To draw a conclusion from the experiments above, the ConstantDistance Operator is effective and able to more efficient.

Discussions and future work

This subsection further discusses how the ConstantDistance Operator makes a proposal and elucidate the correctness of the proposal in the aspect of a certain tree. [20]

Correlation analysis of rates and times

As is described in the methodology section, the key mechanism of the proposed operator is to increase or decrease a node time randomly and adjust the rates according to the constant product of rates and time, i.e. the genetic distances. In this way, there exist some correlations between rates and times. Take an internal node as an example. Assuming the proposed operator increases its current node time, to maintain the distances on the three branches linked to this node, the rate on the branch above this node is supposed to increase as well. But the two rates on the branches that are below this node are supposed to decrease. If the rate increases (decreases) along with the node time, then they have a positive (negative) correlation.

After analysing the seven ratites data set in BEAST2 with the ConstantDistance Operator, we filter the trees in the output tree file by the shared common ancestor for each taxa, as well as the rates and node times in the log file. Then, we conducted a pairwise comparison between each rate and node time in the filtered trees to see how they are correlated. The results are shown in Fig.18. As can be seen from upper right of the figure, to a large degree, the rates have a negative correlation with the node times. But in the upper left and bottom right of the figure, most rates (all the node times) have a positive correlation with each other. Remember that this is only an average pairwise comparison. For higher dimensions, the results will be different.

Sampling a fixed unrooted tree

The tests that sample from the prior distributions in the previous section start from a tree with arbitrary genetic distances. Here, still by sampling from the prior distributions, the ratites data set is used to specify the genetic distances on an unrooted tree, so that the data is actually involved and likelihood is not included.

First of all, we use the ratites data set to construct an unrooted tree with an online program [21], the method of which is detailed in Ref.[22]. As is shown in Fig.16(a), the unrooted tree with the genetic distances on the branches will be fixed when sampled by the proposed operator in BEAST2.

Then, we adopt the midpoint method to root the tree randomly assign the divergence times for each node under the given topology and distances. As a result, the rates on each branch are obtained. In Fig.16(b), the rooted time tree is specified as the initial state of MCMC chain in XML file, and will be sampled by only using the proposed operator.

Finally, after running the XML file, we use TreeAnnotator to get the summary tree of all the samples in the output tree file [23], which is shown in Fig.17(a). As can be seen, the tree topology is consistent with 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.

In the future, we will continue to improve the performance of the ConstantDistance Operator.

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, this paper proposed an operator, called ConstantDistance Operator, which integrates four different strategies to work on internal nodes and the root in a phylogenetic tree. And the proposed operator has been developed as a new package for BEAST2. 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. NeSI's Platform Mahuika. https://www.nesi.org.nz/services/high-performance-computing/platforms
20. 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)
21. PhyML3.0: New Algorithms, Methods and Utilities. http://www.atgc-montpellier.fr/phyml/
22. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O.: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML3.0. Systematic biology **59**(3), 307–321 (2010)
23. TREEANNOTATOR. https://beast2.blogs.auckland.ac.nz/treeannotator/

Figures

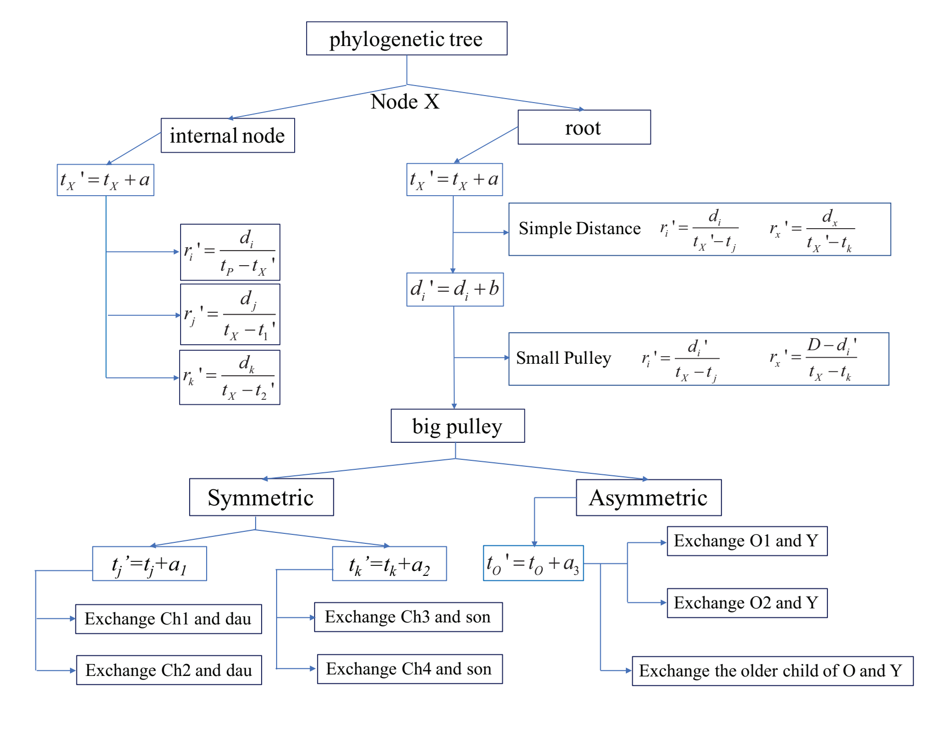

phylogenetic tree
Node X
internal node
root
$t_X' = t_X + a$
$r_i' = \frac{d_i}{t_P - t_X'}$
$r_j' = \frac{d_j}{t_X - t_1'}$
$r_k' = \frac{d_k}{t_X - t_2'}$
$t_X' = t_X + a$
Simple Distance $r_i' = \frac{d_i}{t_X' - t_j}$ $r_x' = \frac{d_x}{t_X' - t_k}$
$d_i' = d_i + b$
Small Pulley $r_i' = \frac{d_i'}{t_X - t_j}$ $r_x' = \frac{D - d_i'}{t_X - t_k}$
big pulley
Symmetric
Asymmetric
$t_j' = t_j + a_1$
$t_k' = t_k + a_2$
$t_O' = t_O + a_3$
Exchange O1 and Y
Exchange O2 and Y
Exchange Ch1 and dau
Exchange Ch2 and dau
Exchange Ch3 and son
Exchange Ch4 and son
Exchange the older child of O and Y

Figure 1 The flow chart of the constant distance operator.

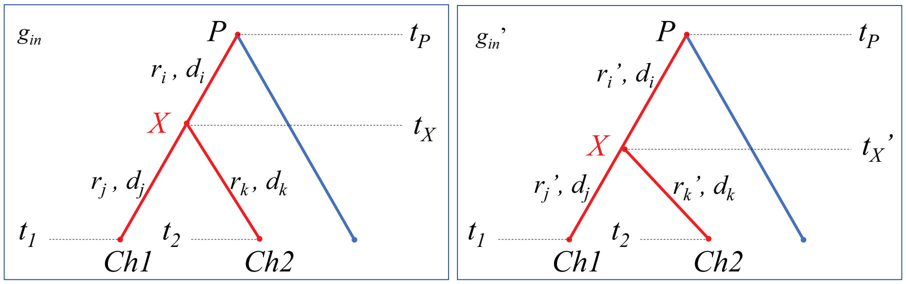
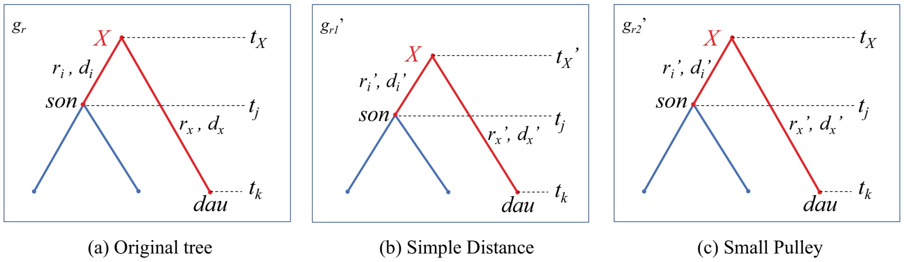


Figure 2 The illustration of operations on internal nodes. The operator proposes a tree g_{in}' based on tree g_{in}, during which d_i, d_j, d_k are constant.



(a) Original tree

(b) Simple Distance

(c) Small Pulley

Figure 3 The illustration of Simple Distance and Small Pulley. Simple Distance proposes g_{r1}' and keeps d_i, d_x constant. Small Pulley proposes g_{r2}' and $d_i + d_x$ remains constant.

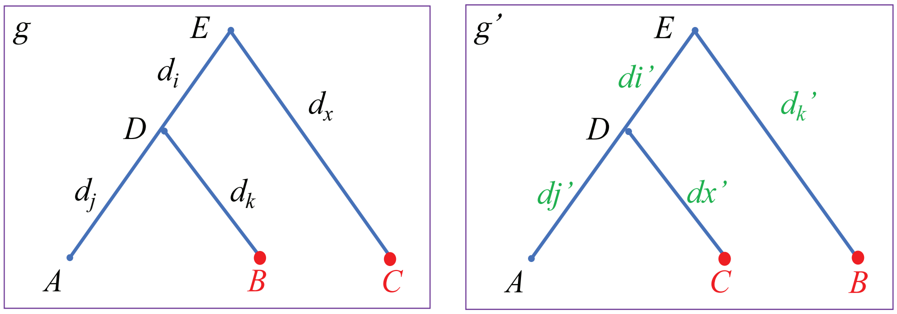


Figure 4 The illustration of Exchange (B, C) method. This method is applied to tree g and proposes g' by swapping node B and C, so that the four distances are adjusted to maintain d_{AB}, d_{AC} and d_{BC}.

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

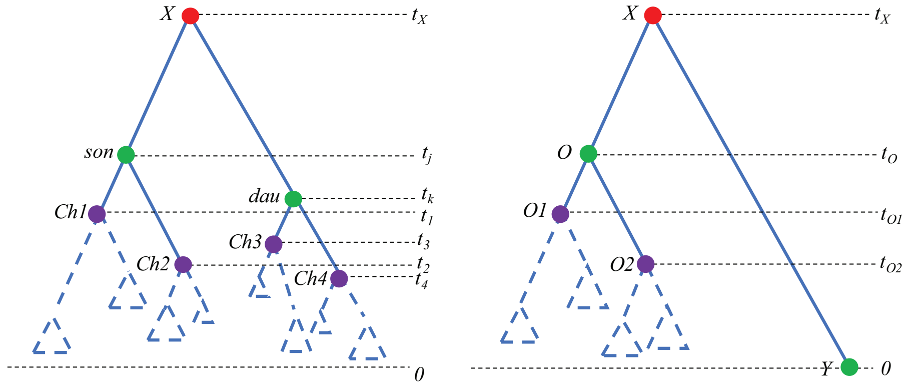


Figure 5 Two different tree shapes. The symmetric tree is on the left and the asymmetric tree is on the right. The dashed triangles represent the potential subtrees rooted at the nodes.

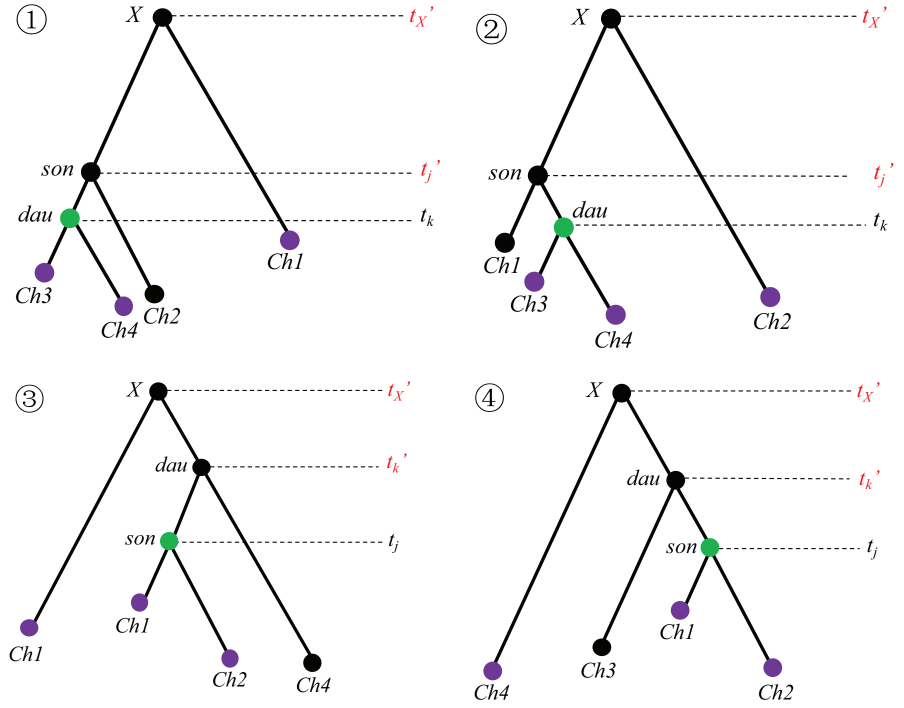


Figure 6 The illustration of operations on symmetric tree in Fig.5. The proposed operator will propose one of the four possible trees, each with 0.25 probability.

	Chain Length	Sample from MCMC			R curve			Plot
		Mean	Err	StdEv	Mean	Err	StdEv	

Senario 1	10000000
	20000000
		3.2727
		3.271
			8.3e-3
			6.1e-3
				0.5467
				0.5616
					3.2669	1.3e-06	0.5553	Fig.9(a)
								Fig.9(b)

Senario 2	10000000
	20000000
		0.4677
		0.4672
			3.9e-04
			2.8e-04
				0.0265
				0.0262
					0.4667	3.5e-05	0.0262	Fig.9(c)
								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

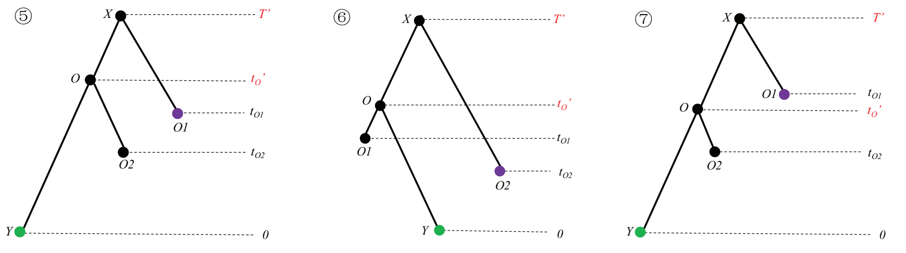


Figure 7 The illustration of operations on asymmetric tree in Fig.5. The proposed operator will propose one of the three possible trees. Depending on $t_o{}'$, ⑤ has 1 probability, ⑥ and ⑦ have 0.5 probability each.

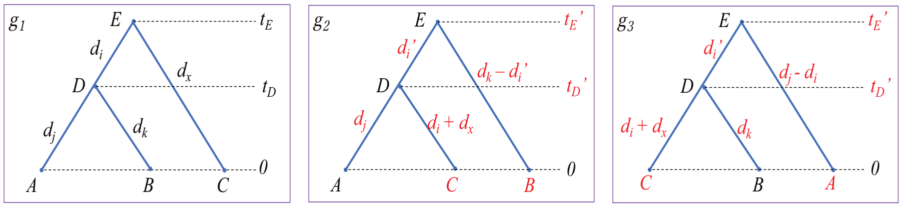
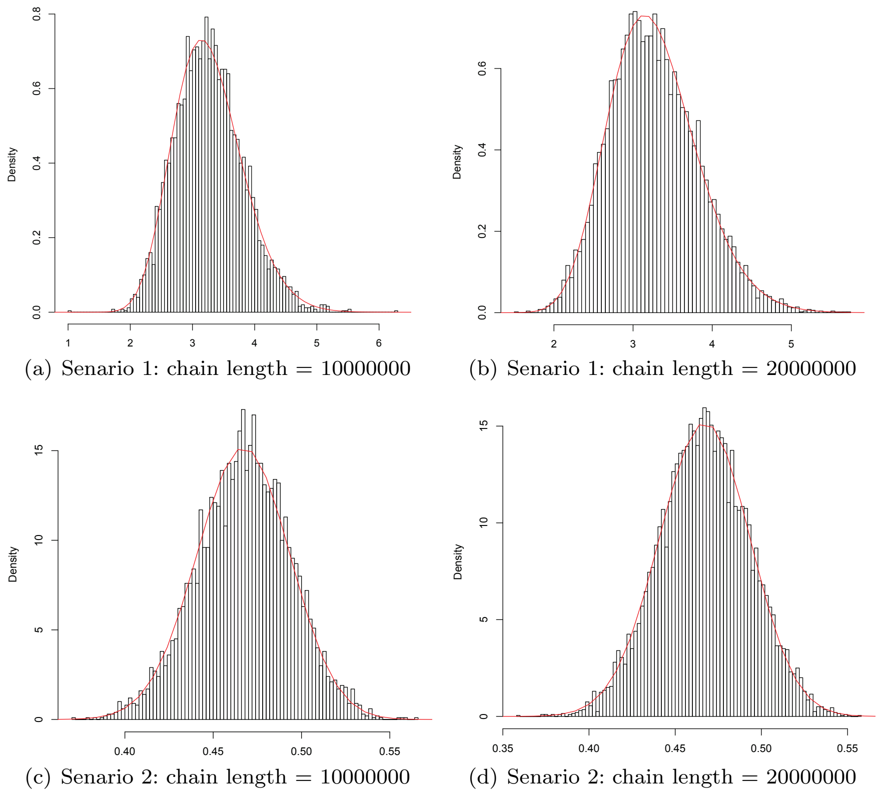


Figure 8 The illustration of sampling from prior. g_1 is set to be the original tree where an MCMC chain starts. When testing Big Pulley, the proposed operator samples the trees among g_1, g_2 and g_3.



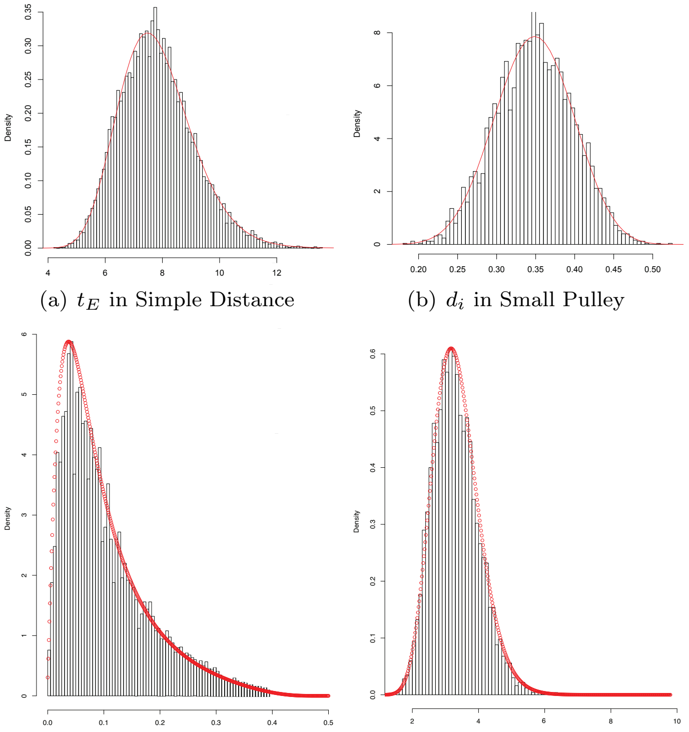
(a) Senario 1: chain length = 10000000

(b) Senario 1: chain length = 20000000

(c) Senario 2: chain length = 10000000

(d) Senario 2: chain length = 20000000

Figure 9 Sampled parameters in tests of internal nodes. The horizontal axis represents the node time of D in Fig.8. The two scenarios sample two trees with different distances specified in Table 1 . And each scenario has two different lengths of MCMC chains.



(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 Sampled parameters in test of the root. For the trees in Fig.8, Simple Distance samples the root time t_E only, Small Pulley samples the distance d_i only, and Big Pulley samples t_E, t_D, d_i. To make it simple, t_E and d_i are compared.

Strategy	Variable	Sample from MCMC		R curve		Plot
		Mean	StdEv	Mean	StdEv	
Simple Distance	t_E	7.8081	1.2884	7.8187	1.2992	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

	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 in Fig.12 and Fig.13

		ESS		Running time	
	Length	20 taxa	120 taxa	20 taxa	120 taxa
categories	20000	13	4	18635	44155
	10000	58	6	8660	36406
	5000	171	8	4690	15956
Average		81	6	10662	32172
cons	20000	616	147	20207	35406
	10000	646	161	8967	25589
	5000	993	186	4581	12487
Average		752	165	11252	24494
nocons	20000	86	63	19344	38245
	10000	153	20	8361	30521
	5000	296	48	4499	12940
Average		179	43	10735	27236

Table 6 Results of ESS and running time

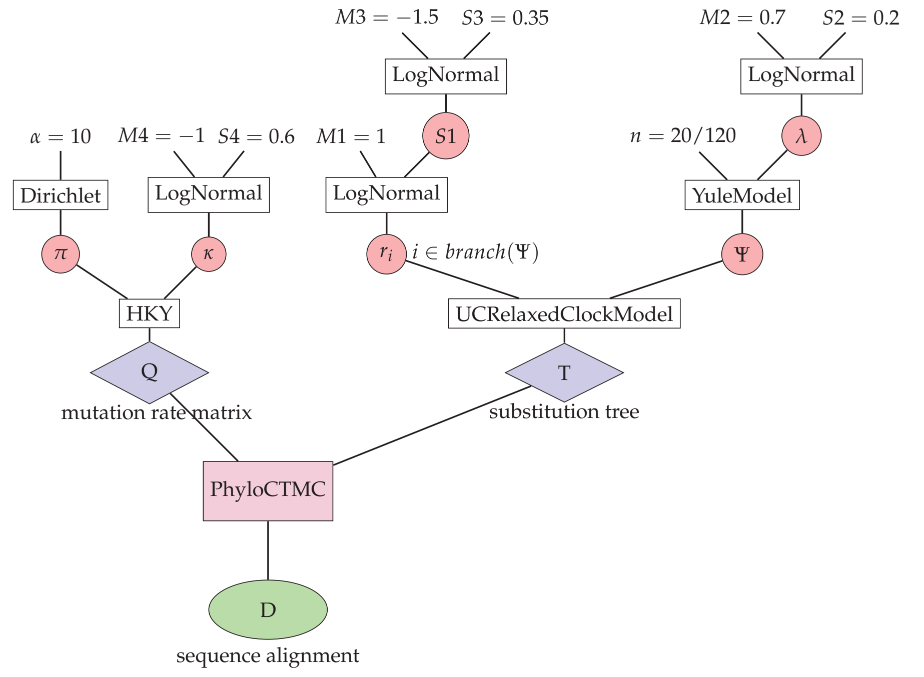


Figure 11 The models and prior distributions to simulate the sequence data. The sequence alignment is simulated through a phylogenetic continuous-time Markov Chain that consists of a substitution model (HKY) and a uncorrelated relaxed clock model. The random variables in HKY model, including frequency and ratio, construct the mutation rate matrix. The rates and phylogenetic time trees specified by a Yule model construct the substitution tree. The standard deviation of rates and the birth rate in Yule model are both random variables following LogNormal distributions.

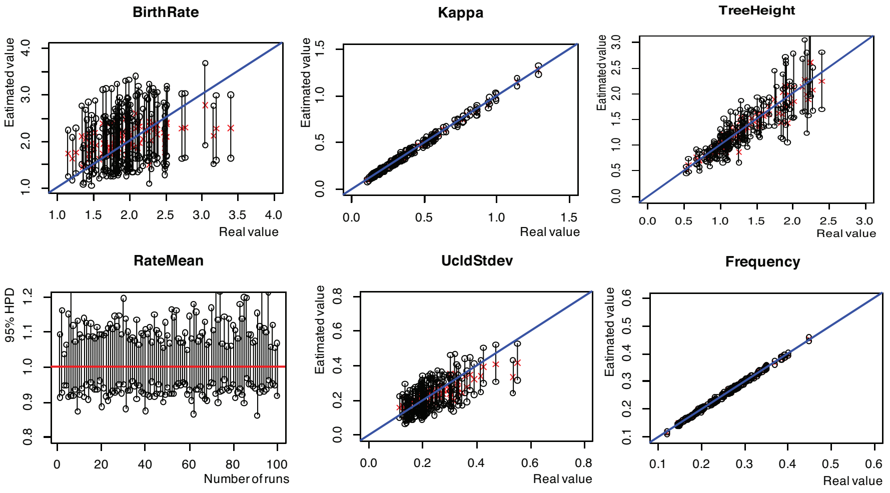


Figure 12 Comparing the sampled parameters in simulation study with 20 taxa. The red crossings represent the coordinates of real and estimated value. The vertical lines with circles show the 95% HPD of the sampled parameters. The blue lines indicate that the estimated value is equal to the real value. The red line represents the fixed value of rate mean.

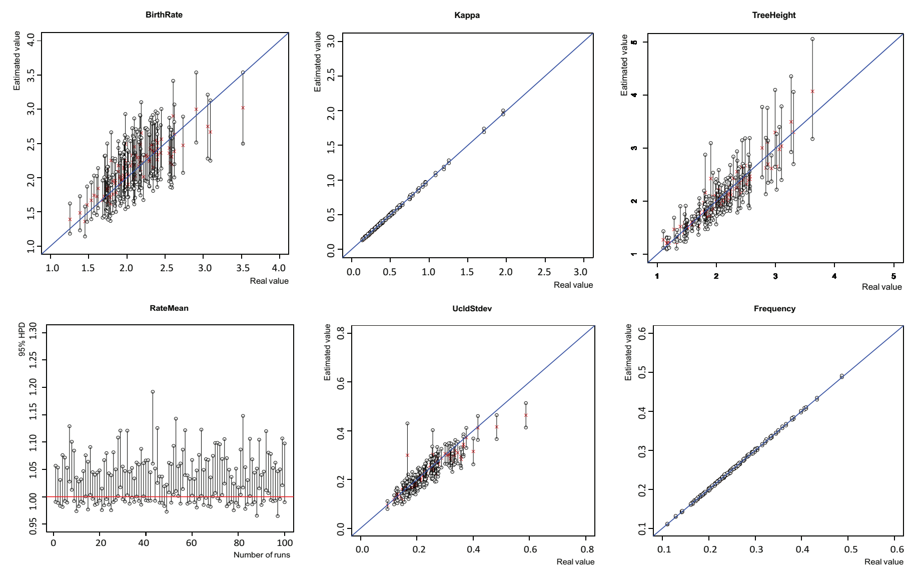


Figure 13 Comparing the sampled parameters in simulation study with 120 taxa.

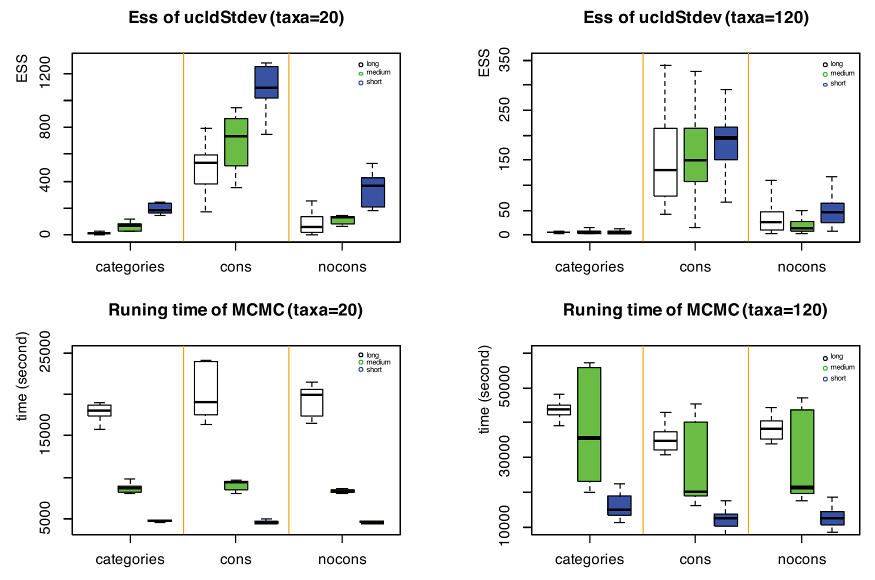


Figure 14 Comparison of ESS and running time using simulated data. The term long, medium and short represent the length of sequence with 20 thousand, 10 thousand and 5 thousand respectively.

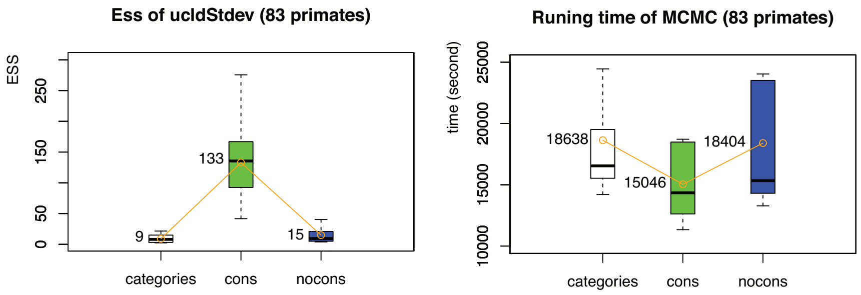
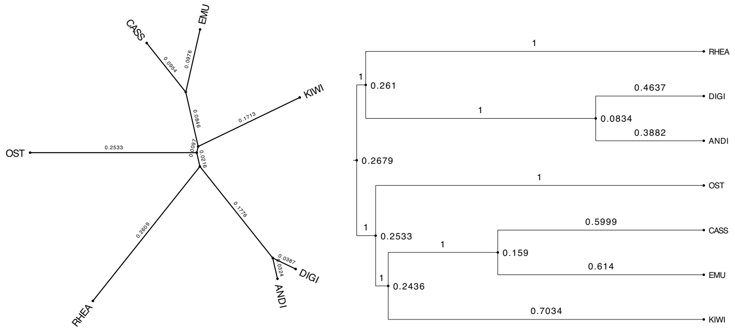
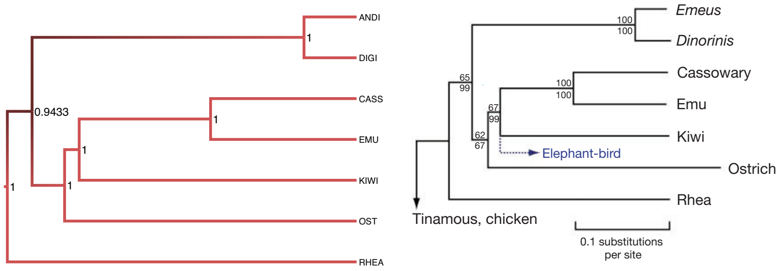


Figure 15 Comparison of ESS and running time using primates data.



(a) Unrooted tree (b) Rooted time tree

Figure 16 Illustration of sampling a fixed unrooted tree. The unrooted tree is obtained from the ratites data set. The rooted time tree is the initial state in MCMC, with time on the nodes and rates on the branches.



(a) Summary tree of sampled trees (b) Ratite phylogeography in Ref. [20]

Figure 17 Comparison of ratites phylogenetic trees.

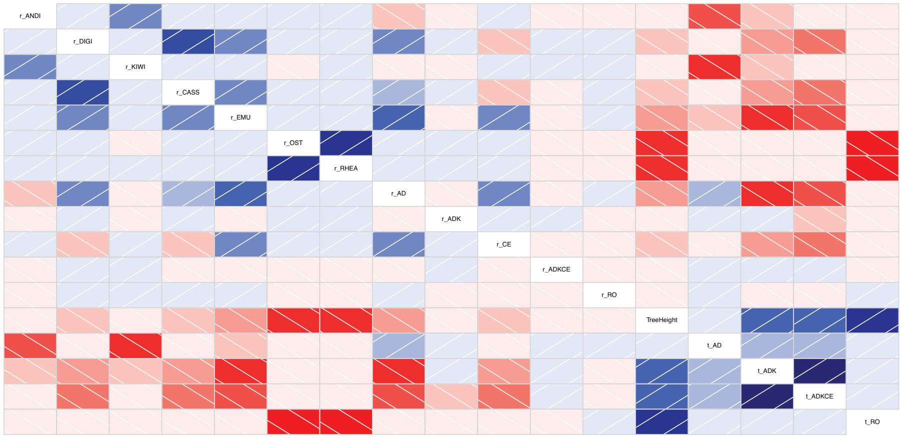


Figure 18 Correlation between rates and times in the ratites tree. Blue indicates the positive relation and red indicates the negative relation. The darker the colour is, the stronger the relation tends to be.