RESEARCH

Improving the performance of relaxed clock model

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 computational evolutionary. However, due the huge amount of phylogenetic data, the efficiency becomes a great problem. Based on an uncorrelated relaxed model, this paper introduces a new operator to change the rates and maintain the genetic distances constant at the same time. For internal nodes, the proposed operator deals with three rates as the node time changes. For the root of the phylogenetic tree, three different strategies are discussed, including Simple distance, Small pulley and Big pulley. It should be noticed that Big pulley is able to change the tree topology, which enables the operator to sample all the possible trees under a given unrooted tree. As BEAST2 is a widely-used software platform for phylogenetic analysis, the proposed operator is implemented as a new package in BEAST2. To validate the effectiveness of the proposed, experiments of sampling from the prior is firstly conducted. Besides, a real data set is also utilised to illustrate the efficiency of the operator.

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 tree that shows the relationships among species and organisms, which is generally called a phylogenetic tree. The research based on phylogenetic analysis is aimed at inferring the phylogenies by constructing the phylogenetic trees. Until now, a lot of methods have been proposed and a majority of them have been implemented in the computer softwares, such as MRBAYES [6], PAML [7], APE [8] and so on [9].

In the beginning, the classical methods for tree construction are based on a distance matrix that is obtained by sequence alignment, after biologists extract DNA from organisms. Starting from a alignment approach, such as dynamic programming and backtracking routine [10], the distance-based algorithms tried to construct a tree with the distances on the branches (branch length) that fit the distance matrix. For example, UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is able to produce a tree for any distance matrix but fails to fit the tree with an additive matirx [11]. Another well-known clustering method is “neighbour joining”, in which neighbours are defined as the operational taxonomic units and is used to minimise the total branch length [12]. However, the trees constructed by a distance

matrix say nothing about ancestral states at internal nodes. In addition, it is probable to lost information when converting a multiple alignment to a distance matrix. Some scientists then suggested to reconstruct trees from biological characters that are considered as anatomical or physiological properties, where the species with similar character values occur to be near each other. But it is proved to violate the Dollo's law that evolution does not reinvent the same organ, when the phylogeny of insect wings is studied [13].

Over the last few years, more advanced techniques have been developed to reconstruct phylogenetic trees, due to the development of statistic 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 [14]. Specifically, Monte Carlo integration is used to integrate over the ancestral speciation times for particular trees and a MCMC method is used to generate the set of trees with the highest posterior probabilities. Based on this framework, much attention has been paid to estimating the divergence times [15, 16, 17]. In addition, many softwares have been developed, which plays an important role in Bayesian phylogenetic inference. Especially, a fast and flexible platform, called BEAST (Bayesian Evolutionary Analysis by Sampling Trees), which integrates a large number of models of sequence evolution and evolutionary trees, has been established with more and more packages added afterwards [18, 19] .

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 [20]. 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 [21]. So far, there have been already different kinds of operators proposed. According to the work in Ref. [22], 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, the operators are proposed in the case where genetic distances are exactly known. Namely, the proposed operators change the rates and times while maintaining the genetic distance constant, so that the likelihood of the phylogenetic tree is the same. It is believed that under these operations, the tree will be sampled more

efficiently in the MCMC runs. Besides, this paper provides different strategies for the root and internal nodes. It is worth noticing that the operator is able to sample all the tree topologies when working on the root, but meanwhile the unrooted tree topology is never changed.

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

Bayesian MCMC methods for inferring molecular evolution and phylogeny comes from a characterisation of the results of random or stochastic processes. Generally, it consists of a prior distribution, a phylogenetic likelihood and a posterior distribution 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. For instance, by taking into account species divergence and molecular evolution, Bayesian evolutionary analysis not only constructs a tree with genetic distance, but also contains divergence times and evolutionary rates, which makes phylogenetic inference more .

Eq.(1) shows a basic Bayesian framework for phylogenetic analysis. It is indicated that what a Bayesian phylogenetic analysis can give is a posterior probability of conditional on the data, which is inferred by a likelihood $\Pr(D|g,\Phi)$ and some priors, i.e. $\Pr(g)$ and $\Pr(\Phi)$. To be specific, the posterior probability is a statement of our beliefs about the parameter after evaluating the data at hand. The likelihood is the vehicle that conveys the information in the data. The prior probability specifies our beliefs about the parameter value before looking at the new data at hand. During an analysis, the prior beliefs are updated to become posterior beliefs by the information in the data via the likelihood function.

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

, where g denotes the phylogenetic tree, Φ is a set of parameters and D represents the data available. Besides, the marginal likelihood of the data $\Pr(D)$ simply serves to normalise the numerator. It is the sum of the equivalent term for all possible parameter values. For the phylogeny problem, this is a multi-dimensional integration of the likelihood function over the joint prior probability densities for all parameters, such as branch lengths and substitution model parameters for each tree, and a summation over all possible trees. The marginal likelihood is too nasty to be solved analytically. However, the rise of fast, cheap computing power and the development of some clever numerical methods made Bayesian phylogenetic inference possible, especially Markov Chain Monte Carlo (MCMC) algorithms which include

Gibbs samplers, Metropolis-coupled and reversible-jump MCMC and so on. Taking Metropolis-Hastings MCMC algorithm as an example, it allows us to sample from the posterior probability density without having to calculate the marginal likelihood.

Tree proposals

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

Basically, the random walk operator, which proposes a new state by adding a random number, and the scale operator, which multiplies the current state with a scale factor, can satisfy most sampling requirements. However, they are suitable for working on a single parameter or a set of parameters, such as population size or divergence times. 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.

Population model

A population model is a natural prior for the phylogenetic time tree, which specifies a distribution of divergence times in the tree. As a commonly-used model, a coalescent tree defines the probability that two individuals have a common ancestor in previous generation. Assuming that there is a population with a certain size N, then coalescent distribution of the tree g can be obtained by Eq.(2).

$$P(g|N) = \prod_{i=2}^{N} \left(\frac{1}{N} \times e^{-\frac{\binom{i}{2}}{N} \times \tau_i}\right) \tag{2}$$

, where τ_i is the time duration between two generations.

For the population size with uncertainty, a birth-death model builds a continuous-time Markov process with two types of state transitions by introducing birth rate and death rate. Specifically, birth rate is used to increases the state variable, and death rate is used to decreases the state variable. To simplify the process, the distribution of a tree based on Yule model uses birth rate only, as is given by Eq.3.

$$P(g|E) = (\lambda^{n-1} \times e^{-\lambda \times T}) \times \prod_{i=1}^{n-1} e^{-\lambda t_i} \tag{3}$$

, where n is the number of individuals, λ is the birth rate and E represents the nodes in the tree.

Substitution model

A substitution model defines the genetic distance by describing the changes between nucleotides. Based on continuous-time Markov chains, substitution models are used to estimate the number of substitutions, so that the expected number of nucleotide substitutions per site is defined as the genetic distance.

As an ideal case, JC69 model assumes that every nucleotide has the same rate of changing into any other nucleotide[23]. To be close to the reality, models like HKY and GTR are more frequently used by accomodating unequal base compositions and cancelling the reversibility in substitution process[24, 25]. What' more, since different sites in the sequence do not necessarily evolve at the same rate. So models that assumes the rate for any site to be a random variable are proposed, such as "HKY+Γ" and "GTR+Γ, which indicates the mutation rates follow a gamma distribution.

When the sequences are long enough, the genetic distances can be considered as exactly known. Since the genetic distances are constant, the calculation of the tree likelihood, which is a great burden for MCMC, only needs to be performed once. As a result, the efficiency of MCMC can be largely improved.

0.1 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. On the contrary to the unrooted model, 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.[26], 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. So the evolutionary rate on each branch has nothing to do with the rate on any other neighbouring branches. As a result, the rates can change faster than making a slight move over multiple adjoining branches.

The proposed operator

The flow chart of the proposed operators is illustrated in Fig.1. In a tree constructed by n taxa, there are $2n-1$ nodes. In the first place, the proposed operator generates a random integer x in the interval $[1, n]$, so that the corresponding node $\mathbf{X}$ with the node number x will selected. Then, the node time of the selected node is denoted by t_x. Afterwards, different strategies will be adopted in the case of a leaf node, internal node, or the root. Details are introduced step by step as follows.

The internal node operator

As can be seen in Fig.**??**, the original tree in the left represents the current state. The following 5 steps will propose a tree' (in the right) by changing the rates on branches.

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

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

Step 3 Propose a new node time for **X** by $t_X{}' = t_X + a$, where $a \sim Uniform[-w, +w]$. Make sure that $\max\{t_1, t_2\} < t_X{}' < t_P$.

Step 4 Propose new rates for **X**, **Ch1** and **Ch2** by using Eq.(4).

$$r_i{}' = \frac{d_i}{t_P - t_X{}'} \tag{4a}$$

$$r_j{}' = \frac{d_j}{t_X - t_1{}'} \tag{4b}$$

$$r_k{}' = \frac{d_k}{t_X - t_2{}'} \tag{4c}$$

Step 5 Return the $HastingsRatio = 1$, since the movement of node **X** is symmetric.

0.2 The root operator

For the root of the tree, there are three ways to deal with the rates and node times below. "Simple distance" is a way of proposing a new root time. To sample the tree in a larger range, "Small pulley" not only proposes a new root time, but also changes the genetic distances on the each side of the root. However, since the unrooted tree is the only evidence that we have and should be maintained after the operation, "Big pulley" potentially proposes a new tree topology on top of the root time and genetic distances in previous two ways.

0.2.1 Simple distance

In spired by how the operator works on the internal nodes, we will use the following steps to propose a new root time and keep the genetic distance on each side of the root constant.

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 $t_X{}' > \max\{t_j, t_k\}$.

Step 3 Propose new rates for branches on each side of the root, as is indicated in Eq.(5).

$$r_i{}' = \frac{d_i}{t_X{}' - t_j} \tag{5a}$$

$$r_x{}' = \frac{d_x}{t_X{}' - t_k} \tag{5b}$$

Step 4 Return the $HastingsRatio = 1$, since the proposal of $t_X{}'$ is symmetric.

0.2.2 Small pulley

Different from "Simple distance", a new genetic distance of branch on one side of the root is proposed in "Small pulley". Depending on whether to propose a new root time, there are two versions of "Small pulley". Version 1 only proposes a new distances, while Version 2 proposes a new distance and a new root time.

Version 1

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 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 $0 < d_i{}' < D$, where $D = d_i + d_x$.

Step 3 Propose new rates for branches on each side of the root by using Eq.(6).

$$r_i{}' = \frac{d_i{}'}{t_X - t_j} \tag{6a}$$

$$r_x{}' = \frac{D - d_i{}'}{t_X - t_k} \tag{6b}$$

Step 4 Return the $HastingsRatio = 1$, since the proposal of $d_i{}'$ is symmetric.

Version 2

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 $t_X{}' > \max\{t_j, t_k\}$.

Step 3 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 $0 < d_i{}' < D$, where $D = d_i + d_x$.

Step 4 Propose new rates for branches on each side of the root by using Eq.(7).

$$r_i{}' = \frac{d_i{}'}{t_X{}' - t_j} \tag{7a}$$

$$r_x{}' = \frac{D - d_i{}'}{t_X{}' - t_k} \tag{7b}$$

Step 5 Return the $HastingsRatio = 1$, since the proposal of both $t_X{}'$ and $d_i{}'$ are symmetric.

Big Pulley

First of all, the method of changing the tree topology is introduced. As is shown in Fig.**??**, once the method *Exchange (**B**, **C**)* is called, the following operations will be performed.

(1) Propose $d_i{}'$ by $d_i{}' = d_i + b$, where $b \sim Uniform[-v, +v]$. Make sure that $0 < d_i{}' < D$, where $D = d_i + d_x$.

(2) To maintain the genetic distances constant, d_j, d_k and d_x will be revised according to Eq.(8).

$$d_j{}' = d_j \tag{8a}$$

$$d_k{}' = d_k - d_i{}' \tag{8b}$$

$$d_x{}' = d_x + d_i \tag{8c}$$

(3) The new rates on corresponding branches are obtained by Eq.(9).

$$r_i{}' = \frac{d_i{}'}{t_E - t_D} \tag{9a}$$

$$r_j{}' = \frac{d_j{}'}{t_D - t_A} \tag{9b}$$

$$r_k{}' = \frac{d_k{}'}{t_E - t_B} \tag{9c}$$

$$r_x{}' = \frac{d_x{}'}{t_D - t_C} \tag{9d}$$

Hence, the process of Big pulley can be detailed as follows:

Step 1 Get the child nodes of **X**, denoted by **son** and **dau**.

Step 2 Get the node times of **X**, **son** and **dau**, denoted by t_X, t_j and t_k.

Step 3 Get the rates for **son** and **dau**, denoted by r_i and r_x.

Step 4 Propose new node times and rates and change the tree topology.

(1) Propose the node time for **X** by **X** by $t_X{}' = t_X + a$, where $a \sim Uniform[-w, +w]$.

(2) As is shown in Fig.**??**, there are two different tree shapes to be taken into consideration. For a symmetric tree, the proposed tree will be one in the four possible trees (illustrated in Fig.**??**). But for an asymmetric tree, there two directions to go (illustrated in Fig.**??**). Details are as follows.

Symmetric tree: both son and dau have child nodes

- 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 $t_k < t_j{}' < t_X{}'$.
  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 $t_j < t_k{}' < t_X{}'$.
  ③: With 0.5 probability to *Exchange (**Ch3** and **son**)*
  ④: With 0.5 probability to *Exchange (**Ch4** and **son**)*

Asymmetric tree: either son or dau has child nodes

- Pick the older child of the root, denoted by **O**. And its node time is t_O t_{O1} t_{O2}. The younger child of the root is denoted by **Y**.

- Propose new node time for **O** by $t_O{}' = t_O + a_3$, where $a_3 \sim Uniform[-w, +w]$.
  if($t_O{}' > \max\{t_{O1}, t_{O2}\}$ or $t_{O1} = t_{O2}$)
  ⑤: With 0.5 Probability to *Exchange (**O1** and **Y**)*
  ⑥: With 0.5 Probability to *Exchange (**O2** and **Y**)*
  if($\min\{t_{O1}, t_{O2}\} < t_O{}' < \max\{t_{O1}, t_{O2}\}$)
  ⑦: Exchange the older child of **O** and **Y**, here *Exchange (**O1** and **Y**)*

Step 5 Return the Hastings Ratio

Firstly, for all the cases that are not included in the above operations, we will reject the proposal by returning the HastingsRatio equalling to negative infinity. Secondly, for other cases in which the operations are not symmetric, the HastingsRatio will be calculated by the probability of prosing the new tree and going back exactly to the original tree. The details are provided in Algorithm 1.

Algorithm 1 Return HastingsRatio 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 $HastingsRatio = \frac{\alpha}{0.25}$
end for
for ③ ④ **do**
  Return $HastingsRatio = \frac{\beta}{0.25}$
end for
for ⑤ ⑥ **do**
  Return $HastingsRatio = \frac{\gamma}{0.5}$
end for
for ⑦ **do**
  Return $HastingsRatio = \frac{0.25}{1}$
end for

Experimental results and analysis

In this section, detailed experiments are conducted, together with some analysis and discussions, to validate the effectiveness of the proposed operator. At first, we performed MCMC in BEAST2 by sampling from prior, in which no alignments are included and only the proposed operator is used. In addition, the proposed operator is also applied to analyse the real data set by sampling more parameters and considering the phylogenetic likelihood.

Sample from the prior

As is shown in Fig.**??**, Tree1 is set as the initial state. Besides, a LogNormal distribution is used as the rate prior, as is shown below:

$$r = [r_i \quad r_j \quad r_k \quad r_x] \sim LogNormal(M = -3, S = 0.5) \tag{10}$$

The Coalescent model with constant population size (N) is used to describe the tree prior. To be specific, since Tree1 has 3 taxa, the probability of node times given the tree is calculated by Eq.(11).

$$P(t_E, t_D) = (\frac{1}{N} \times e^{-\frac{1}{N}(t_E - t_D)}) \times (\frac{1}{N} \times e^{-\frac{3}{N}t_D}), N = 0.3 \tag{11}$$

After the prior is specified, the distribution of the parameter is exactly known. In other words, the rates are the functions of its genetic distance and times, among which the distances are known in advance and the rates and times follow certain distributions, as is shown in Eq.(12).

$$\begin{aligned} P(r,t) &= P(t_E, t_D) \times P(r_i) \times P(r_j) \times P(r_k) \times P(r_x) \\ &= P(t_E, t_D) \times P(\frac{d_i}{\Delta t1}) \times P\frac{d_j}{\Delta t2}) \times P(\frac{d_k}{\Delta t3}) \times P(\frac{d_x}{\Delta t4}) \end{aligned} \tag{12}$$

, where $t = [t_E, t_D]$, Δt represents the time duration of the corresponding branch, and $P(r)$ is the probability density of the LogNormal distribution. Therefore, the exact distribution can be obtained by conducting numerical integration on Eq.(12).

Internal nodes

(1) Initial settings

The genetic distances, node times and rates 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. What's more, to make sure that the results is robust, two different MCMC chain lengths are performed in each scenario.

2. Results and comparisons

The mean, mean error and the standard deviation of the sampled distribution are summarised in Table 2. To evaluate the results, the correct distribution is obtained by using Eq.(13), according to Eq.(12). Besides, the histograms of MCMC samples, as well as the curves of the numerical integration, are shown in Fig.**??**. It is safe to conclude that the two distributions are consistent, which proves that the proposed operators works properly on the internal nodes.

$$P(r, t_D) = \int_{t_D=0}^{t_E} P(t_E, t_D) \times P(\frac{d_j}{t_D}) \times P(\frac{d_k}{t_D}) \times P(\frac{d_i}{t_E - t_D}) \times P(\frac{d_x}{t_E}) \tag{13}$$

Root

The initial settings for Tree1 with genetic distances, node times and rates are detailed in Table 3.

(1) Simple distance

To test the simple distance strategy, the root time t_E is sampled by the proposed operator, which ranges from 1 to positive infinity theoretically. Namely, all the genetic distances and the node time t_D are fixed during the experiment. Hence, similar to Eq.(13), the sampled distribution of t_E can be obtained by Eq.(14).

$$P(r, t_E) = \int_{t_E=1}^{+\infty} P(t_E, t_D) \times P(\frac{d_j}{t_D}) \times P(\frac{d_k}{t_D}) \times P(\frac{d_i}{t_E - t_D}) \times P(\frac{d_x}{t_E}) \tag{14}$$

The results are given in Table 4 and Fig.**??**. The mean and the standard deviation are close enough to confirm that the two distribution are the same, which also shows that simple distance is correct.

(2) Small pulley

Considering that there are two versions in small pulley, the sum of d_x and d_i are kept 0.67 in both versions, as is shown in Table 3. What differs is that, in Version 1, only the distances of d_x and d_i are changed during the sampling process. But both distances and root time t_E have a proposal in each state.

Based on Eq.(12), the exact distribution of t_E and d_i in Version1 and Version2 can be obtained by Eq.(15) and Eq.(16) respectively. The sampled distribution are compared in Table 4 and Fig.**??**, **??** ,**??**. Even though there exist some errors, the two parameters can be considered to follow the same distribution. So the small pulley is able to provide the correct samples.

$$P(r, d_i) = \int_{d_i=1}^{0.67} P(t_E, t_D) \times P(\frac{d_j}{t_D}) \times P(\frac{d_k}{t_D}) \times P(\frac{d_i}{t_E - t_D}) \times P(\frac{0.67 - d_i}{t_E}) \quad (15)$$

$$\begin{aligned} P(r, d_i) = \int_{t_E}^{+\infty} \int_{d_i=1}^{0.67} & P(t_E, t_D) \\ & \times P(\frac{d_j}{t_D}) \times P(\frac{d_k}{t_D}) \times P(\frac{d_i}{t_E - t_D}) \times P(\frac{0.67 - d_i}{t_E}) \end{aligned} \quad (16)$$

(3) Big pulley

When testing the big pulley strategy, a new tree is proposed by changing t_E, t_D and d_i. In this case, the initial state Tree1 will probably go to Tree2 and Tree3, as is shown in Fig.**??**. According to the initial settings in Table 3, the genetic distances $d_{AB} = 1$, $d_{AC} = 1$ and $d_{BC} = 1$ remains unchanged. Hence, the distribution we are about to achieve is calculated by Eq.(17).

According the statistical measurements, i.e. mean and standard deviation, in Table 4, the two distributions are consistent within the acceptable error range. Therefore, the big pulley strategy that samples all the tree topologies can also give the right combinations of rates and node times given constant the genetic distances among taxa.

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

A real data analysis

To verify the performance of the proposed operator when it is used to sample from the posterior distribution, analysis based on a real data set is provided in the section. As is shown in Fig.**??**, Alan Cooper et al. utilised mitochondrial genome sequences of two extinct moas and constructed the phylogenetic tree of ratites evolution [**?**]. The sequences have 10,767 bp, which is enough long and is thus suitable for the proposed operator to work on. To fully compare the results, there are three tests conducted.

Test1: A standard BEAST test

This test includes a birth-death model (as tree prior) where birth rate follows $LogNormal(M = 2, S = 1)$and death rate follows $Uniform(0, 1)$, and a substitution model where bModeltest is used. Besides, standard operators are adopted to sample the parameters and trees. Namely, Swap Operator, Uniform Operator Randomwalk Operator and Scale Operator are used for rates. Uniform Operator, Subtreeslide Operator, Narrow Operator, Wide Operator, Wilsonbalding Operator and ScaleOperator are used for the tree.

Test2: Adding the ConstantDistance Operator

The ConstantDistance Operator is added based on the settings in previous Test1.

By specifying the setting in two seperate XML files, the two tests are running in BEAST 2, after which the log files and tree files are obtained. Firstly, we summarised all the sampled trees in the tree files, by using TreeAnnotator which averages the rates and node times. As is presented in Fig.**??**, the tree topology obtained inTest1 and Test2 is exactly the same, which is also consistent with the tree in Fig.**??**. To be clear, the rates and node times on the summarised trees are detailed in Fig.**??**. Intuitively, the values are very similar. After performing t test, the p values (0.9627 for rates and 0.9865 for node times) further prove that there is no significant difference between rates and node times obtained in Test1 and Test2. In addition, the distributions of the mean of rates and the height of the tree are also compared, as is shown in Fig.**??**. The box plots again demonstrate that Test1 and Test2 provide the same distributions. Finally, in order to validate the efficiency, the effective sample size (ESS) in the above tests are compared in Fig.**??**, including 33 parameters sampled in MCMC chains. Overall, most ESS in Test2 are larger than those in Test1. Particularly, the red arrows indicate that the ConstantDistance Operator is able to sample the rates and tree height more efficiently with larger ESS, because the proposed operator works directly on the rates and node times. Therefore, it is concluded that the proposed operator is correct and more efficient.

Test3: Sampling a fixed unrooted tree

An extra test is conducted in order to elucidate the correctness of the proposed operator. As is shown in Fig.**??**, 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, and then randomly assign the divergence times for each node under the given topology and distances, so that the rates on each branch is calculated. Afterwards, the rooted time tree shown in Fig.**??** is used as the initial state in MCMC chain. Still based on the settings in Test1and Test2, the molecular models are the same. But the tree is sampled by only using the proposed operator in this test, which means the standard operators are excluded. In other words, this test is aimed at examining whether the proposed operator is able to provide the correct samples when given a exactly known tree.

Likewise, the summary tree is obtained after running the specified XML file in BEAST2, which is pictured in Fig.**??**. As can be seen, the result is consistent with Test1 and Test2, compared to the tree in Fig.**??**. 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.**??**. 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.

Conclusion

The efficiency is of great significance in phylogenetic analysis based on Bayesian MCMC. This paper discussed how to make a proposal to deal with rates and node times when the genetic distances is exactly known. In all, there are five operations proposed for internal nodes and the root in a phylogenetic tree. Then, a series of tests based on sampling from prior are performed to show the functionality of each operation. Finally, a novel operator, called ConstantDistance operator, is defined by combining the operation for internal nodes and big pulley strategy. The ConstantDistance operator is utilised to analyse a real data set so that the efficiency is verified by comparing to the standard BEAST test.

Acknowledgements

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. Huelsenbeck, J.P., Ronquist, F.: Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics **17**(8), 754–755 (2001)
7. Yang, Z.: Paml: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics **13**(5), 555–556 (1997)
8. Paradis, E., Claude, J., Strimmer, K.: Ape: analyses of phylogenetics and evolution in r language. Bioinformatics **20**(2), 289–290 (2004)
9. Kumar, S., Stecher, G., Tamura, K.: Mega7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular biology and evolution **33**(7), 1870–1874 (2016)
10. Wang, L., Jiang, T.: On the complexity of multiple sequence alignment. Journal of computational biology **1**(4), 337–348 (1994)
11. Sokal, R.R.: The principles and practice of numerical taxonomy. Taxon, 190–199 (1963)
12. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution **4**(4), 406–425 (1987)
13. Lynch, V.J., Wagner, G.P.: Did egg-laying boas break dollo's law? phylogenetic evidence for reversal to oviparity in sand boas (eryx: Boidae). Evolution: International Journal of Organic Evolution **64**(1), 207–216 (2010)
14. 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)
15. 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)
16. 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)
17. 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)
18. Drummond, A.J., Rambaut, A.: Beast: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology **7**(1), 214 (2007)
19. 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)
20. Kobert, K., Stamatakis, A., Flouri, T.: Efficient detection of repeating sites to accelerate phylogenetic likelihood calculations. Systematic biology **66**(2), 205–217 (2017)
21. 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)
22. Höhna, S., Drummond, A.J.: Guided tree topology proposals for bayesian phylogenetic inference. Systematic biology **61**(1), 1–11 (2011)
23. Jukes, T.H., Cantor, C.R., *et al.*: Evolution of protein molecules. Mammalian protein metabolism **3**(21), 132 (1969)
24. Hasegawa, M., Kishino, H., Yano, T.-a.: Dating of the human-ape splitting by a molecular clock of mitochondrial dna. Journal of molecular evolution **22**(2), 160–174 (1985)

25. Zharkikh, A.: Estimation of evolutionary distances between nucleotide sequences. Journal of molecular evolution **39**(3), 315–329 (1994)
26. Drummond, A.J., Ho, S.Y., Phillips, M.J., Rambaut, A.: Relaxed phylogenetics and dating with confidence. PLoS biology **4**(5), 88 (2006)

Figures

Figure 1 The flow chart of the constant distance operator.

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

	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.??
	20000000	3.271	6.1e-3	0.5616				Fig.??
Senario 2	10000000	0.4677	3.9e-04	0.0265	0.4667	3.5e-05	0.0262	Fig.??
	20000000	0.4672	2.8e-04	0.0262				Fig.??

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

Strategyh	Variable	Sample from MCMC		R curve		Plot
		Mean	StdEv	Mean	StdEv	
Simple distance	t_E	7.8081	1.2884	7.818691	1.299236	Fig.??
Version 1	t_E	0.3480	0.0492	0.3476	0.0494	Fig.??
Version 2	d_i	0.3924	0.0690	0.3918	0.0693	Fig.??
	t_E	4.9953	0.6665	5.0122	0.6832	Fig.??
Big pulley	d_i	0.1016	0.0766	0.0960	0.0760	Fig.??
	t_E	3.3017	0.6908	3.3095	0.6912	Fig.??

Table 4 Results for root