Auto-validating von Neumann rejection sampling from small phylogenetic tree spaces

Raazesh Sainudiin[*,1,2] , Thomas York[3,4]

[1]Department of Statistics, University of Oxford, Oxford, OX1 3TG, United Kingdom
[2]Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand
[3]Department of Biological Statistics and Computational Biology and [4] Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, New York 14853, United States of America

Email: RS: r.sainudiin@math.canterbury.ac.nz; TY: tly2@cornell.edu;

[*]Corresponding author

Abstract

Background: In phylogenetic inference one is interested in obtaining samples from the posterior distribution over the tree space on the basis of some observed DNA sequence data. One of the simplest sampling methods is the rejection sampler due to von Neumann. Here we introduce an auto-validating version of the rejection sampler, via interval analysis, to rigorously draw samples from posterior distributions over small phylogenetic tree spaces.

Results: The posterior samples from the auto-validating sampler are used to rigorously (i) estimate posterior probabilities for different rooted topologies based on mitochondrial DNA from human, chimpanzee and gorilla, (ii) conduct a non-parametric test of rate variation between protein-coding and tRNA-coding sites from three primates and (iii) obtain a posterior estimate of the human-neanderthal divergence time.

Conclusions: This solves the open problem of rigorously drawing independent and identically distributed samples from the posterior distribution over rooted and unrooted small tree spaces (3 or 4 taxa) based on any multiply-aligned sequence data.

Background

Obtaining samples from a real-valued target density $f^{\cdot}(t)$ is a basic problem in statistical estimation. The target $f^{\cdot}(t) : \mathbb{T} \mapsto \mathbb{R}$ maps n-dimensional real points in $\mathbb{R}^n$ to real numbers in $\mathbb{R}$, i.e. $t \in \mathbb{T} \subset \mathbb{R}^n$. In

Bayesian phylogenetic estimation, we want to draw independent and identically distributed samples from a target posterior density on the space of phylogenetic trees. The standard point-valued or punctual Monte Carlo methods via conventional floating-point arithmetic are typically non-rigorous as they do not account for all sources of numerical errors and are limited to evaluating the target at finitely many points. The standard approaches to sampling from the posterior density, especially over phylogenetic trees, rely on Markov chain Monte Carlo (MCMC) methods. Despite their asymptotic validity, it is nontrivial to guarantee that an MCMC algorithm has converged to stationarity [1], and thus MCMC convergence diagnostics on phylogenetic tree spaces are heuristic [2].

A more direct sampler that is capable of producing independent and identically distributed samples from the target density $f^{\cdot}(t) := f(t)/(N_f)$, by only evaluating the target shape $f(t)$ without knowing the normalizing constant $N_f := \int_{\mathbb{T}} f(t)dt$, is the von Neumann rejection sampler [3]. However, the limiting step in the rejection sampler is the construction of an envelope function $\widehat{g}(t)$ that is not only greater than the target shape $f(t) := N_f f^{\cdot}(t)$ at every $t \in \mathbb{T}$, but also easy to normalize and draw samples from. Moreover, a practical and efficient envelope function has to be as close to the target shape as possible from above. When an envelope function is constructed using point-valued methods, except for simple classes of targets, one cannot guarantee that the envelope function dominates the target shape globally.

None of the available samplers can rigorously produce independent and identically distributed samples from the posterior distribution over phylogenetic tree spaces, even for 3 or 4 taxa. We describe a new approach for rigorously drawing samples from a target posterior distribution over small phylogenetic tree spaces using the theory of *interval analysis*. This method can circumvent the problems associated with (i) heuristic convergence diagnostics in MCMC samplers and (ii) pseudo-envelopes constructed via non-rigorous point-valued methods in rejection samplers.

Informally, our method partitions the domain into boxes and uses interval analysis to rigorously bound the target shape in each box; then we use as envelope the simple function which takes on in each box the upper bound obtained for that box. It is easy to draw samples from the density corresponding to this step function envelope. More formally, the method employs an interval extension of the target posterior shape $f(t) : \mathbb{T} \mapsto \mathbb{R}$ to produce rigorous enclosures of the range of f over each interval vector or box in an adaptive partition $\mathfrak{T} := \{\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, \ldots, \boldsymbol{t}^{(|\mathfrak{T}|)}\}$ of the tree space $\mathbb{T} = \cup_i \boldsymbol{t}^{(i)}$. This partition is adaptively constructed by a priority queue. The interval extended target shape maps boxes in $\mathbb{T}$ to intervals in $\mathbb{R}$. This image interval provides an upper bound for the global maximum and a lower bound for the global minimum of f over each element of the partition of $\mathbb{T}$. We use this information to construct an envelope as

a simple function over the partition $\mathfrak{T}$. Using the Alias method [4] we efficiently propose samples from this normalized step-function envelope for von Neumann rejection sampling.

We call our method auto-validating because we employ interval methods to rigorously construct the envelope for a large class of target densities. The method was described in a more rudimentary form in [5]. Unlike many conventional samplers, each sample produced by our method is equivalent to a computer-assisted proof that it is drawn from the desired target, up to the pseudo-randomness of the underlying, deterministic, pseudo-random number generator. `MRS 0.1.2`, a `C++` class library for statistical set processing is available from www.math.canterbury.ac.nz/~r.sainudiin/codes/mrs under the terms of the GNU General Public License.

The rest of the paper is organized as follows. In the Methods Section, we introduce (i) von Neumann rejection sampler (RS), (ii) phylogenetic estimation problem, (iii) interval analysis and (iv) an interval extension of the rejection sampler called the Moore rejection sampler (MRS) in honor of Ramon E. Moore. Moore was one of the influential founders of interval analysis [6]. In Results Section, we employ MRS to rigorously draw samples from the posterior density over small tree spaces. Using one of the earliest primate mitochondrial DNA data sets we use the posterior samples to estimate the posterior probability of each rooted tree topology and conduct a non-parametric test of rate variation between protein-coding and tRNA-coding sites. Using one of the latest data sets we obtain a rigorous posterior estimate of the human-neanderthal divergence time. We can also draw samples from the space of unrooted triplet and quartet trees. We conclude after a discussion of the method.

Methods

In the following sections, we first introduce the rejection sampler (RS) due to von Neumann [3]. Secondly, we describe the basic phylogenetic inference problem (e.g. [7–9]). Then, we introduce the basic principles of interval methods (e.g. [6, 10–13]). Finally, we construct interval extensions of RS to rigorously draw independent and identically distributed samples from small phylogenetic tree spaces. We leave the formal proofs to the Appendix for completeness.

Rejection sampler (RS)

Rejection sampling [3] is a Monte Carlo method to draw independent samples from a target random variable or random vector T with density $f^{\cdot}(t) := f(t)/N_f$, where $t \in \mathbb{T} \subset \mathbb{R}^n$, i.e. $T \sim f^{\cdot}$. The challenge is to draw the samples without any knowledge of the normalizing constant $N_f := \int_{\mathbb{T}} f(t)\, dt$. Typically the

target $f^{\cdot}(t)$ is any density that is absolutely continuous with respect to the Lebesgue measure. The von Neumann rejection sampler (RS) can produce samples from $T \sim f^{\cdot}$ according to Algorithm 1 when provided with (i) a fundamental sampler that can produce independent samples from the Uniform$[0,1]$ random variable M with density given by the indicator function $\mathbf{1}_{[0,1]}(m) : \mathbb{R} \mapsto \mathbb{R}$, (ii) a target shape $f(t) : \mathbb{T} \mapsto \mathbb{R}$, (iii) an envelope function $\widehat{g}(t) : \mathbb{T} \mapsto \mathbb{R}$, such that,

$$\widehat{g}(t) \geq f(t) \quad \text{for all} \quad t \in \mathbb{T} \ , \tag{1}$$

(iv) a normalizing constant $N_{\widehat{g}} := \int_{\mathbb{T}} \widehat{g}(t)dt$, (v) a proposal density $g(t) := (N_{\widehat{g}})^{-1}\widehat{g}(t)$ over $\mathbb{T}$ from which independent samples can be drawn and finally (vi) $f(t)$ and $\widehat{g}(t)$ must be computable for any $t \in \mathbb{T}$.

Algorithm 1: von Neumann RS

input : (i) f; (ii) samplers for $V \sim g$ and $M \sim \mathbf{1}_{[0,1]}$; (iii) $\widehat{g}$; (iv) integer MaxTrials;
output : (i) possibly one sample t from $T \sim f^{\cdot}$ and (ii) Trials

initialize: Trials $\leftarrow 0$; Success $\leftarrow$ false; $t \leftarrow \emptyset$;

repeat // propose at most MaxTrials times until acceptance
  $v \leftarrow$ sample(g) ; // draw a sample v from RV V with density g
  $u \leftarrow \widehat{g}(v)$ sample($\mathbf{1}_{[0,1]}$); // draw a sample u from RV U with density $\mathbf{1}_{[0,\widehat{g}(v)]}$
  if $u \leq f(v)$ **then** // accept the proposed v and flag Success
    $t \leftarrow v$; Success $\leftarrow$ true
  end
  Trials $\leftarrow$ Trials $+1$; // track the number of proposal trials so far
until Trials $\geq$ MaxTrials *or* Success = true;
return t *and* Trials

We use the Mersenne Twister pseudo-random number generator [14] to imitate independent samples from $M \sim \mathbf{1}_{[0,1]}$. The random variable T, if generated by Algorithm 1, is distributed according to $f^{\cdot}$ (e.g. [15]). Let $\boldsymbol{A}(\widehat{g})$ be the probability that a point proposed according to g gets accepted as an independent sample from $f^{\cdot}$ through the envelope function $\widehat{g}$. Observe that the envelope-specific acceptance probability $\boldsymbol{A}(\widehat{g})$ is the ratio of the integrals

$$\boldsymbol{A}(\widehat{g}) = \frac{N_f}{N_{\widehat{g}}} := \frac{\int_{\mathbb{T}} f(t)\, dt}{\int_{\mathbb{T}} \widehat{g}(t)\, dt} \ ,$$

and the probability distribution over the number of samples from g to obtain one sample from $f^{\cdot}$ is geometrically distributed with mean $1/\boldsymbol{A}(\widehat{g})$ (e.g. [15]).

Phylogenetic estimation

In this section we briefly review phylogenetic estimation. A more detailed account can be found in [7–9]. Inferring the ancestral relationship among a set of extant species based on their DNA sequences is a basic

problem in phylogenetic estimation. One can obtain the likelihood of a particular phylogenetic tree that relates the extant species of interest at its leaves by superimposing a continuous time Markov chain model of DNA substitution upon that tree. The length of an edge (branch length) connecting two nodes (species) in the tree represents the amount of evolutionary time (divergence) between the two species. The internal nodes represent ancestral species. During the likelihood computation, one needs to integrate over all possible states at the unobserved ancestral nodes.

Next we give a brief introduction to some phylogenetic nomenclature. A phylogenetic tree is said to be rooted if one of the internal nodes, say node r, is identified as the root of the tree, otherwise it is said to be unrooted. The rooted tree is conventionally depicted with the root node r at the top. The four topology-labeled, three-leaved, rooted trees, namely, ${}^0t$, ${}^1t$, ${}^2t$ and ${}^3t$, with leaf label set $\{1, 2, 3\}$, are depicted in Figure 1(i)–(iv). The unrooted, three-leaved tree with topology label 4 or the unrooted triplet ${}^4t$ is shown in Figure 1(v). For each tree, the terminal branch lengths, i.e. the branch lengths leading to the leaf nodes, have to be strictly positive and the internal branch lengths have to be non-negative. Our rooted triplets (Figure 1(i)–(iv)) are said to satisfy the molecular clock, since the branch lengths of each ${}^kt$, where $k \in \{0, 1, 2, 3\}$, satisfy the constraint that the distance from the root node r to each of the leaf nodes is equal to ${}^kt_0 + {}^kt_1$ with ${}^kt_1 > 0$ and ${}^kt_0 \geq 0$.

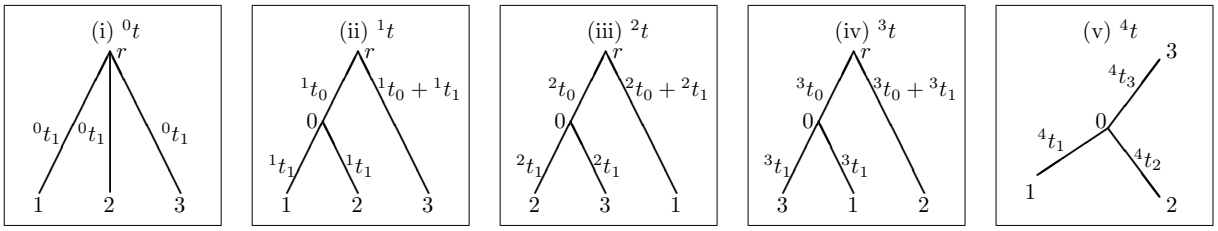

Figure 1: Tree space with three labeled leaves. Space of phylogenetic trees with three labeled leaves $\{1, 2, 3\}$. See text for description.

Likelihood of a tree

Let d denote a homologous set of sequences of length v with character set $\mathfrak{A} = \{a_1, a_2, \ldots, a_{|\mathfrak{A}|}\}$ from n taxa. We think of d as an $n \times v$ matrix with entries from $\mathfrak{A}$. We are interested in estimating the branch lengths and topologies of the tree underlying our observed d. Let b_k denote the number of branches and s_k denote the number of nodes of a tree with a specific topology or branching order labeled by k. Thus, for a given topology label k, n labeled leaves and b_k many branches, the labeled tree ${}^kt$ is the topology-labeled

vector of branch lengths $({}^k t_1, \ldots, {}^k t_{b_k})$ contained in the topology-labeled tree space ${}^k\mathbb{T}$, i.e.,

$$ {}^k\mathbb{T} := \{ {}^k t := ({}^k t_1, \ldots, {}^k t_{b_k}) \in \mathbb{R}_+^{b_k} : {}^k t_i > 0 \text{ for terminal branches} \} \ . $$

Any subset of the tree space with $|\mathfrak{K}|$ many topologies in the topology label set $\mathfrak{K}$ can be defined as follows:

$$ {}^{\mathfrak{K}}\mathbb{T} := \bigcup_{k \in \mathfrak{K}} {}^k\mathbb{T} \ . $$

An explicit model of sequence evolution is prescribed in order to obtain the likelihood of observing data d at the leaf nodes as a function of the parameter ${}^k t \in {}^{\mathfrak{K}}\mathbb{T}$ for each topology label $k \in \mathfrak{K}$. Such a model prescribes $P_{a_i,a_j}(t)$, the probability of mutation from a character $a_i \in \mathfrak{A}$ to another character $a_j \in \mathfrak{A}$ in time t. Using such a transition probability we may compute $\ell_q({}^k t)$, the log-likelihood of the data d at site $q \in \{1, \ldots, v\}$ or the q-th column of d, via the post-order traversal over the labeled tree with branch lengths ${}^k t := ({}^k t_1, {}^k t_2, \ldots, {}^k t_{b_k})$. This amounts to the sum-product Algorithm 2 [16] that associates with each node $h \in \{1, \ldots, s_k\}$ of ${}^k t$ subtending $\hbar$ many descendants, a partial likelihood vector, $l_h := (l_h^{(a_1)}, l_h^{(a_2)}, \ldots, l_h^{(a_{|\mathfrak{A}|})}) \in \mathbb{R}^{|\mathfrak{A}|}$, and specifies the length of the branch leading to its ancestor as ${}^k t_h$.

Algorithm 2: Likelihood by post-order traversal

input : (i) a labeled tree with branch lengths ${}^k t := ({}^k t_1, {}^k t_2, \ldots, {}^k t_{b_k})$, (ii) transition probability $P_{a_i,a_j}(t)$ for any $a_i, a_j \in \mathfrak{A}$, (iii) stationary distribution $\pi(a_i)$ over each character $a_i \in \mathfrak{A}$, (iv) site pattern or data $d_{\bullet,q}$ at site q

output : $l_{d_{\bullet,q}}({}^k t)$, the likelihood at site q with pattern $d_{\bullet,q}$

initialize: For a leaf node h with observed character $a_i = d_{h,q}$ at site q, set $l_h^{(a_i)} = 1$ and $l_h^{(a_j)} = 0$ for all $j \neq i$. For any internal node h, set $l_h := (1, 1, \ldots, 1)$.

recurse : compute l_h for each sub-terminal node h, then those of their ancestors recursively to finally compute l_r for the root node r to obtain the likelihood for site q,

$$ l_{d_{\bullet,q}}({}^k t) = l_r = \sum_{a_i \in \mathfrak{A}} \left(\pi(a_i) \cdot l_r^{(a_i)} \right) \ . $$

For an internal node h with descendants $s_1, s_2, \ldots, s_{\hbar}$,

$$ l_h^{(a_i)} = \sum_{j_1, \ldots, j_{\hbar} \in \mathfrak{A}} \{ \, l_{s_1}^{(j_1)} \cdot P_{a_i,j_1}({}^k t_{s_1}) \cdot l_{s_2}^{(j_2)} \cdot P_{a_i,j_2}({}^k t_{s_2}) \ldots l_{s_{\hbar}}^{(j_{\hbar})} \cdot P_{a_i,j_{\hbar}}({}^k t_{s_{\hbar}}) \, \}. $$

Assuming independence across all v sites we obtain the likelihood function for the given data d, by multiplying the site-specific likelihoods

$$ l_d({}^k t) = \prod_{q=1}^{v} l_{d_{\bullet,q}}({}^k t) \ . \tag{2} $$

The maximum likelihood estimate is a point estimate (single best guess) of the unknown phylogenetic tree on the basis of the observed data d and it is

$$\underset{{}^k t \in {}^{\mathfrak{K}}\mathbb{T}}{\operatorname{argmax}}\, l_d({}^k t) \ .$$

The simplest probability models for character mutation are continuous time Markov chains with finite state space $\mathfrak{A}$. We introduce three such models employed in this study next. We only derive the likelihood functions for the simplest model with just two characters as it is thought to well-represent the core problems in phylogenetic estimation (see for e.g. [17]).

Posterior density of a tree

The posterior density $f^{\cdot}({}^k t)$ conditional on data d at tree ${}^k t$ is the normalized product of the likelihood $l_d({}^k t)$ and the prior density $p({}^k t)$ over a given tree space ${}^{\mathfrak{K}}\mathbb{T}$:

$$f^{\cdot}({}^k t) = \frac{l_d({}^k t) p({}^k t)}{\int_{{}^{\mathfrak{K}}\mathbb{T}} l_d({}^k t) p({}^k t)\ \partial({}^k t)} \ . \tag{3}$$

We assume a uniform prior density over a large box or a union of large boxes in a given tree space ${}^{\mathfrak{K}}\mathbb{T}$. Typically, the sides of the box giving the range of branch lengths, are extremely long, say, $[0, 10]$ or $[10^{-10}, 10]$. The branch lengths are measured in units of expected number of DNA substitutions per site and therefore the support of our uniform prior density over ${}^{\mathfrak{K}}\mathbb{T}$ contains the biologically relevant branch lengths. If ${}^{\mathfrak{K}}\mathbb{T}$ is a union of distinct topologies then we let our prior be an equally weighted finite mixture of uniform densities over large boxes in each topology. Naturally, other prior densities are possible especially in the presence of additional information. We choose flat priors for the convenient interpretation of the target posterior shape $f({}^k t) = f^{\cdot}({}^k t) \int_{{}^{\mathfrak{K}}\mathbb{T}} l_d({}^k t) p({}^k t)\ \partial({}^k t)$ to be the likelihood function in the absence of prior information beyond a compact support specification.

Likelihood of a triplet under Cavender-Farris-Neyman (CFN) model

We now describe the simplest model for the evolution of binary sequences under a symmetric transition matrix over all branches of a tree. This model has been used by authors in various fields including molecular biology, information theory, operations research and statistical physics; for references see [7, 18]. This model is referred to as the Cavender-Farris-Neyman (CFN) model in molecular biology, although in other fields it has been referred to as 'the on-off machine', 'symmetric binary channel' and the 'symmetric two-state Poisson model'. Although the relatively tractable CFN model itself is not popular in applied

molecular evolution, the lessons learned under the CFN model often extend to more realistic models of DNA mutation (e.g. [17]). Thus, our first stop is the CFN model.

Model 1 (Cavender-Farris-Neyman (CFN) model) *Under the CFN mutation model, only pyrimidines and purines, denoted respectively by* $\mathtt{Y} := \{\mathtt{C}, \mathtt{T}\}$ *and* $\mathtt{R} := \{\mathtt{A}, \mathtt{G}\}$*, are distinguished as evolutionary states among the four nucleotides* $\{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$*, i.e.* $\mathfrak{A} = \{\mathtt{Y}, \mathtt{R}\}$*. Time* t *is measured by the expected number of substitutions in this homogeneous continuous time Markov chain with rate matrix:*

$$Q = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} ,$$

and transition probability matrix $P(t) = e^{Qt}$ *:*

$$P(t) = \begin{pmatrix} 1-(1-e^{-2t})/2 & (1-e^{-2t})/2 \\ (1-e^{-2t})/2 & 1-(1-e^{-2t})/2 \end{pmatrix} .$$

Thus, the probability that $\mathtt{Y}$ *mutates to* $\mathtt{R}$*, or vice versa, in time* t *is* $a(t) := (1-e^{-2t})/2$*. The stationary distribution is uniform on* $\mathfrak{A}$*, i.e.* $\pi(\mathtt{R}) = \pi(\mathtt{Y}) = 1/2$*.*

When there are only three taxa, there are five tree topologies of interest as depicted in Figure 1. There are $2^3 = 8$ possible site patterns, i.e. for each site $q \in \{1, 2, \ldots, v\}$, the q-th column of the data d, denoted by $d_{\bullet,q}$, is one of eight possibilities, numbered $0, 1, \ldots, 7$ for convenience:

$$d_{\bullet,q} \in \left\{ \begin{array}{ccccccccccccccc} 0 & , & 1 & , & 2 & , & 3 & , & 4 & , & 5 & , & 6 & , & 7 \\ \\ \mathtt{R} & & \mathtt{Y} & & \mathtt{R} & & \mathtt{Y} & & \mathtt{R} & & \mathtt{Y} & & \mathtt{R} & & \mathtt{Y} \\ \mathtt{R} & , & \mathtt{Y} & , & \mathtt{R} & , & \mathtt{Y} & , & \mathtt{Y} & , & \mathtt{R} & , & \mathtt{Y} & , & \mathtt{R} \\ \mathtt{R} & & \mathtt{Y} & & \mathtt{Y} & & \mathtt{R} & & \mathtt{Y} & & \mathtt{R} & & \mathtt{R} & & \mathtt{Y} \end{array} \right\} . \tag{4}$$

Given a multiple sequence alignment data d from 3 taxa at v homologous sites, i.e. $d \in \{\mathtt{Y}, \mathtt{R}\}^{3\times v}$, the likelihood function over the tree space ${}^k\mathbb{T}$ is simplified from (2) as follows:

$$l_d({}^k t) = \prod_{q=1}^{v} l_{d_{\bullet,q}}({}^k t) = \prod_{i=0}^{7} \left(l_i({}^k t)\right)^{c_i} , \tag{5}$$

where $l_i({}^k t)$ is the likelihood of the the i-th site pattern as in (4) and c_i is the count of sites with pattern i. In fact, $l_i({}^k t) = P(i|{}^k t)$ is the probability of observing site pattern i given topology label k and branch lengths t and similarly $l_d({}^k t) = P(d|{}^k t)$.

Consider the unrooted tree-space with a single topology labeled 4 and three non-negative terminal branch lengths ${}^4 t = ({}^4 t_1, {}^4 t_2, {}^4 t_3) \in \mathbb{R}^3_+$ as shown in Figure 1(v). An application of Algorithm 2 to compute the likelihoods $l_0({}^4 t), l_1({}^4 t), \ldots, l_7({}^4 t)$, as derived in (19)-(25), reveals symmetry. There are in fact four

minimally sufficient site pattern classes, namely, xxx, xxy, yxx and xyx, where x and y simply denote distinct characters in the alphabet set $\mathfrak{A} = \{\mathtt{R}, \mathtt{Y}\}$. The corresponding likelihoods are:

$$\begin{aligned}
l_{\mathtt{xxx}}({}^4t) := l_0({}^4t) = l_1({}^4t) &= \frac{1}{8}\left(1 + e^{-2({}^4t_1 + {}^4t_2)} + e^{-2({}^4t_2 + {}^4t_3)} + e^{-2({}^4t_1 + {}^4t_3)}\right) \\
l_{\mathtt{xxy}}({}^4t) := l_2({}^4t) = l_3({}^4t) &= \frac{1}{8}\left(1 + e^{-2({}^4t_1 + {}^4t_2)} - e^{-2({}^4t_2 + {}^4t_3)} - e^{-2({}^4t_1 + {}^4t_3)}\right) \\
l_{\mathtt{yxx}}({}^4t) := l_4({}^4t) = l_5({}^4t) &= \frac{1}{8}\left(1 - e^{-2({}^4t_1 + {}^4t_2)} + e^{-2({}^4t_2 + {}^4t_3)} - e^{-2({}^4t_1 + {}^4t_3)}\right) \\
l_{\mathtt{xyx}}({}^4t) := l_6({}^4t) = l_7({}^4t) &= \frac{1}{8}\left(1 - e^{-2({}^4t_1 + {}^4t_2)} - e^{-2({}^4t_2 + {}^4t_3)} + e^{-2({}^4t_1 + {}^4t_3)}\right) \quad . \qquad (6)
\end{aligned}$$

Therefore, the multiple sequence alignment data d from three taxa evolving under Model 1 can be summarized by the minimal sufficient site pattern counts

$$(c_{\mathtt{xxx}}, c_{\mathtt{xxy}}, c_{\mathtt{yxx}}, c_{\mathtt{xyx}}) := (c_0 + c_1, c_2 + c_3, c_4 + c_5, c_6 + c_7) \ ,$$

which simplifies (5) to:

$$l_d({}^kt) = \prod_{q=1}^{v} l_{d_{\cdot,q}}({}^kt) = \prod_{i=0}^{7} \left(l_i({}^kt)\right)^{c_i} = \prod_{\mathtt{s}=\mathtt{xxx},\mathtt{xxy},\mathtt{yxx},\mathtt{xyx}} \left(l_{\mathtt{s}}({}^kt)\right)^{c_{\mathtt{s}}} \quad . \qquad (7)$$

Note that the probability of our sample space with eight patterns given in (4) is $\sum_{i=0}^{7} l_i({}^4t) = 1$. Our likelihoods are half of those in [17] that are prescribed over a sample space of only four classes of patterns: $\{0,1\}$, $\{2,3\}$, $\{4,5\}$ and $\{6,7\}$. This is because we distinguish between the sample space of data from that of the minimal sufficient statistics. We compute the rooted topology-specific likelihood functions, i.e. $l({}^kt)$ for $k \in \{0,1,2,3\}$ (Figure 1) by substituting the appropriate constraints on branch lengths in ${}^4\mathbb{T} = \mathbb{R}^3_+$, the space of unrooted triplets.

Likelihood of a triplet under Jukes-Cantor (JC) model

The r-state symmetric model introduced in [19] is specified by the $r \times r$ rate matrix with equal off-diagonal entries over an alphabet set $\mathfrak{A}$ of size r. The stationary distribution under this model is the uniform distribution on $\mathfrak{A}$. Thus, CFN model is the 2-state symmetric model over $\mathfrak{A} = \{\mathtt{Y}, \mathtt{R}\}$. The Jukes-Cantor (JC) model [20] is the 4-state symmetric model over $\mathfrak{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. This is perhaps the simplest model on four characters.

Model 2 (Jukes-Cantor (JC) model) *All four nucleotides form the state space for this mutation model, i.e.* $\mathfrak{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. *Once again, evolutionary time t is measured by the expected number of*

substitutions in the homogeneous continuous time Markov chain with rate matrix:

$$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix} \quad .$$

The transition probability matrix $P(t) = e^{Qt}$ is also symmetric. The probability that any given nucleotide mutates to any other nucleotide in time t is $P_{\mathtt{x},\mathtt{y}}(t)$ and that it is found in the same state is $P_{\mathtt{x},\mathtt{x}}(t)$. These transition probabilities are:

$$a(t) := P_{\mathtt{x},\mathtt{y}}(t) = \frac{1}{4} - \frac{1}{4}\exp\left(-\frac{4}{3}t\right), \qquad b(t) := P_{\mathtt{x},\mathtt{x}}(t) = \frac{1}{4} + \frac{3}{4}\exp\left(-\frac{4}{3}t\right) \quad .$$

The stationary distribution is uniform, i.e. $\pi(\mathtt{A}) = \pi(\mathtt{C}) = \pi(\mathtt{G}) = \pi(\mathtt{T}) = \mathtt{1/4}$.

Consider the three non-negative terminal branch lengths ${}^4t = ({}^4t_1, {}^4t_2, {}^4t_3) \in \mathbb{R}^3_+$ of an unrooted tree ${}^4t$ of Figure 1(v). An application of Algorithm 2 to compute the likelihoods of the 64 possible site patterns (see for e.g. [21–24]), reveals five minimally sufficient site pattern classes. Let `x`, `y` and `z` simply denote distinct characters from the alphabet set $\mathfrak{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ at taxon 1, 2 and 3, respectively. The minimally sufficient site pattern classes `xxx`, `xyz`, `xxy`, `yxx` and `xyx` encode 4, 24, 12, 12 and 12 nucleotide site patterns, respectively. By a computation similar to that in (19)-(25), the likelihoods are:

$$\begin{aligned}
l_{\mathtt{xxx}}({}^4t) &= \frac{1}{4}\left(\prod_{i=1}^{3} b({}^4t_i) + 3\prod_{i=1}^{3} a({}^4t_i)\right) \\
l_{\mathtt{xyz}}({}^4t) &= \frac{1}{4}\left(b({}^4t_1)a({}^4t_2)a({}^4t_3) + a({}^4t_1)\left(b({}^4t_2)a({}^4t_3) + a({}^4t_2)\left(b({}^4t_3) + a({}^4t_3)\right)\right)\right) \\
l_{\mathtt{xxy}}({}^4t) &= \frac{1}{4}\left(b({}^4t_1)b({}^4t_2)a({}^4t_3) + a({}^4t_1)a({}^4t_2)\left(b({}^4t_3) + 2a({}^4t_3)\right)\right) \\
l_{\mathtt{yxx}}({}^4t) &= \frac{1}{4}\left(b({}^4t_1)a({}^4t_2)b({}^4t_3) + a({}^4t_1)a({}^4t_3)\left(b({}^4t_2) + 2a({}^4t_2)\right)\right) \\
l_{\mathtt{yxx}}({}^4t) &= \frac{1}{4}\left(a({}^4t_1)b({}^4t_2)b({}^4t_3) + a({}^4t_2)a({}^4t_3)\left(b({}^4t_1) + 2a({}^4t_1)\right)\right) \quad .
\end{aligned}$$

Notice that the probability of observing one of the 64 possible site patterns is 1 for any ${}^4t \in (0,\infty)^3$:

$$4\, l_{\mathtt{xxx}}({}^4t) + 24\, l_{\mathtt{xyz}}({}^4t) + 12\, l_{\mathtt{xxy}}({}^4t) + 12\, l_{\mathtt{yxx}}({}^4t) + 12\, l_{\mathtt{yxx}}({}^4t) = 1 \quad .$$

Let $c_{\mathtt{ijk}}$ denote the number of sites with the site pattern $\mathtt{ijk} \in \{\mathtt{xxx}, \mathtt{xyz}, \mathtt{xxy}, \mathtt{yxx}, \mathtt{xyx}\}$. Then, under the assumption of independence across sites, we obtain the likelihood of a given data d by multiplying the site-specific likelihoods:

$$l_d({}^4t) = \left(l_{\mathtt{xyz}}({}^4t)\right)^{c_{\mathtt{xyz}}} \left(l_{\mathtt{xxy}}({}^4t)\right)^{c_{\mathtt{xxy}}} \left(l_{\mathtt{yxx}}({}^4t)\right)^{c_{\mathtt{yxx}}} \left(l_{\mathtt{yxx}}({}^4t)\right)^{c_{\mathtt{yxx}}} \left(l_{\mathtt{xxx}}({}^4t)\right)^{c_{\mathtt{xxx}}} \quad .$$

Once again, the likelihood of a rooted tree or the star tree can be obtained from that of the unrooted tree by substituting the appropriate constraints on branch lengths in the above equations or by directly applying Algorithm 2 with the appropriate input tree with its topology and branch lengths.

Model 3 (Hasegawa-Kishino-Yano (HKY) model) *The Hasegawa-Kishino-Yano or HKY model [25] has all four nucleotides in the state space, i.e.* $\mathfrak{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. *There are five parameters in this more flexible model. Transitions are changes within the purine* $\{\mathtt{A}, \mathtt{G}\}$ *or pyrimidine* $\{\mathtt{C}, \mathtt{T}\}$ *state subsets, while transversions are changes from purine to pyrimidine or from pyrimidine to purine. In this model, we have a mutational parameter* κ *that allows for transition:transversion bias and four additional parameters* $\pi_{\mathtt{A}}$, $\pi_{\mathtt{C}}$, $\pi_{\mathtt{G}}$ *and* $\pi_{\mathtt{T}}$ *that explicitly control the stationary distribution. The entries of the rate matrix are:*

$$q_{x,y} = \begin{cases} \kappa\pi_y & \text{for transitions} \\ \pi_y & \text{for transversions} \\ -\sum_{z\in\mathfrak{A}, z\neq x} q_{x,z} & \text{if } x = y \ . \end{cases}$$

The transition probabilities are known analytically for this model (see for e.g. [8, p. 203]). We can use these expressions when evaluating the likelihood of a rooted or unrooted tree along with the five mutational parameters via Algorithm 2. For simplicity we set the stationary distribution parameters to the empirical nucleotide frequencies and κ *to be* 2.0 *in this study.*

Interval analysis

Let $\mathbb{IR}$ denote the set of closed and bounded real intervals. Let any element of $\mathbb{IR}$ be denoted by $\boldsymbol{x} := [\underline{x}, \overline{x}]$, where, $\underline{x} \leq \overline{x}$ and $\underline{x}, \overline{x} \in \mathbb{R}$. Next we define arithmetic over $\mathbb{IR}$.

Definition 1 (Interval Operation) *If the binary operator* $\star$ *is one of* $+, -, \times, /$, *then we define an arithmetic on operands in* $\mathbb{IR}$ *by*

$$\boldsymbol{x} \star \boldsymbol{y} := \{x \star y : x \in \boldsymbol{x}, y \in \boldsymbol{y}\} \ ,$$

with the exception that $\boldsymbol{x}/\boldsymbol{y}$ *is undefined if* $0 \in \boldsymbol{y}$.

Theorem 1 (Interval arithmetic) *Arithmetic on the pair* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{IR}$ *is given by:*

$$\begin{aligned} \boldsymbol{x} + \boldsymbol{y} &= [\underline{x} + \underline{y}, \overline{x} + \overline{y}] \\ \boldsymbol{x} - \boldsymbol{y} &= [\underline{x} - \overline{y}, \overline{x} - \underline{y}] \\ \boldsymbol{x} \times \boldsymbol{y} &= [\min\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}, \max\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}] \\ \boldsymbol{x}/\boldsymbol{y} &= \boldsymbol{x} \times [1/\overline{y}, 1/\underline{y}], \text{ provided, } 0 \notin \boldsymbol{y} \ . \end{aligned}$$

When computing with finite precision, say in floating-point arithmetic, directed rounding must be taken into account (see e.g., [6, 10]) to contain the solution. Interval multiplication is branched into nine cases, on

the basis of the signs of the boundaries of the operands, such that only one case entails more than two real multiplications. Therefore, a rigorous computer implementation of an interval operation mostly requires two directed rounding floating-point operations. Interval addition and multiplication are both commutative and associate but not distributive. For example,

$$[-1,2]\times([1,2]+[-2,1])=[-1,2]\times[-1,3]=[-3,6]\ ,$$

$$\text{but,}\quad [-1,2]\times[1,2]+[-1,2]\times[-2,1]=[-2,4]+[-4,2]=[-6,6]\ .$$

Interval arithmetic satisfies a weaker rule than distributivity called sub-distributivity:

$$\boldsymbol{x}(\boldsymbol{y}+\boldsymbol{z})\subseteq \boldsymbol{x}\boldsymbol{y}+\boldsymbol{x}\boldsymbol{z}\ .$$

An extremely useful property of interval arithmetic that is a direct consequence of Definition 1 is summarized by the following theorem.

Theorem 2 (Fundamental property of interval arithmetic) *If* $\boldsymbol{x}\subseteq\boldsymbol{x}'$ *and* $\boldsymbol{y}\subseteq\boldsymbol{y}'$ *and* $\star\in\{+,-,\times,/\}$*, then*

$$\boldsymbol{x}\star\boldsymbol{y}\subseteq\boldsymbol{x}'\star\boldsymbol{y}'\ ,$$

where we require that $0\notin\boldsymbol{y}'$ *when* $\star=/$*.*

Note that an immediate implication of Theorem 2 is that when $\boldsymbol{x}=[x,x]$ and $\boldsymbol{y}=[y,y]$ are thin intervals, i.e. $\underline{x}=\overline{x}=x$ and $\underline{y}=\overline{y}=y$ are real numbers, then $\boldsymbol{x}'\star\boldsymbol{y}'$ will contain the result of the real arithmetic operation $x\star y$.

Let $\underline{x},\overline{x}\in\mathbb{R}^n$ be real vectors such that $\underline{x}_i\leq\overline{x}_i$, for all $i=1,2,\ldots,n$, then $\boldsymbol{x}:=[\underline{x},\overline{x}]$ is an *interval vector* or a *box*. The set of all such boxes is $\mathbb{IR}^n$. The i-th component of the box $\boldsymbol{x}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ is the interval $\boldsymbol{x}_i=[\underline{x}_i,\overline{x}_i]$ and the interval extension of a set $\mathbb{D}\subseteq\mathbb{R}^n$ is $\mathbb{ID}:=\{\boldsymbol{x}\in\mathbb{IR}^n:\underline{x},\overline{x}\in\mathbb{D}\}$. We write $\inf\boldsymbol{x}:=\underline{x}$ for the *lower bound*, $\sup\boldsymbol{x}:=\overline{x}$ for the *upper bound*. Let the maximum norm of a vector $x\in\mathbb{R}^n$ be $\|x\|_\infty:=\max_k|x_k|$. Let the vector valued hyper-metric between boxes $\boldsymbol{x}$ and $\boldsymbol{y}$ be

$$\operatorname{dist}(\boldsymbol{x},\boldsymbol{y})=\sup\{|\underline{x}-\underline{y}|,|\overline{x}-\overline{y}|\}\ ,$$

and the Hausdorff distance between the boxes $\boldsymbol{x}$ and $\boldsymbol{y}$ in the metric given by the maximum norm is then

$$\operatorname{dist}_\infty(\boldsymbol{x},\boldsymbol{y})=\|\operatorname{dist}(\boldsymbol{x},\boldsymbol{y})\|_\infty\ .$$

We can make $\mathbb{IR}^n$ a metric space by equipping it with the Hausdorff distance.

Our main motivation for the extension to intervals is to enclose the *range*:

$$\text{range}(f; S) := \{f(x) : x \in S\} \ ,$$

of a real-valued function $f : \mathbb{R}^n \mapsto \mathbb{R}$ over a set $S \subseteq \mathbb{R}^n$. Except for trivial cases, few tools are available to obtain the range.

Definition 2 (Directed acyclic graph (DAG) expression of a function) *One can think of the process by which a function $f : \mathbb{R}^m \mapsto \mathbb{R}$ is computed as the result of a sequence of recursive operations with the sub-expressions f_i of its expression f where, $i = 1, \ldots, n < \infty$. This involves the evaluation of the sub-expression f_i at node i with operands s_{i_1}, s_{i_2} from the sub-terminal nodes of i given by the directed acyclic graph (DAG) for f*

$$s_i = \odot \mathsf{f}_i := \begin{cases} \mathsf{f}_i(s_{i_1}, s_{i_2}) & : \textit{if node } i \textit{ has 2 sub-terminal nodes } s_{i_1}, s_{i_2} \\ \mathsf{f}_i(s_{i_1}) & : \textit{if node } i \textit{ has 1 sub-terminal node } s_{i_1} \\ I(s_i) & : \textit{if node } i \textit{ is a leaf or terminal node, } I(x) = x. \end{cases} \tag{8}$$

The leaf or terminal node of the DAG is a constant or a variable and thus the f_i for a leaf i is set equal to the respective constant or variable. The recursion starts at the leaves and terminates at the root of the DAG. The DAG for an elementary f is simply its expression f with n sub-expressions $\mathsf{f}_1, \mathsf{f}_2, \ldots, \mathsf{f}_n$:

$$\{\odot \mathsf{f}_i\}_{i=1}^n \quad \rightarrowtail \quad \odot \mathsf{f}_n = \mathsf{f}(x) \ , \tag{9}$$

where each $\odot \mathsf{f}_i$ is computed according to (8).

We look at some DAGs for 0 functions to concretely illustrate these ideas.

Example 1 *Consider the constant zero function $f(x) = 0$ expressed as (i) $\mathsf{f}(x) = 0$, (ii) $\mathsf{f}'(x) = x \times 0$ and (iii) $\mathsf{f}''(x) = x - x$. The corresponding DAG expressions are shown in Figure 2.*

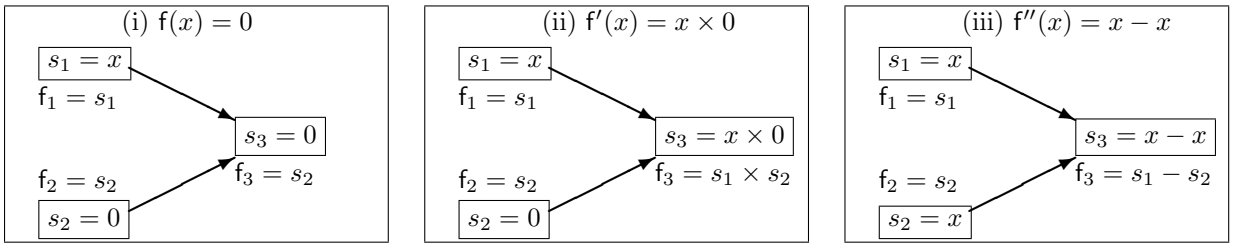

Figure 2: DAG expression for zero functions. The directed acyclic graph (DAG) expression for the three zero functions: (i) $\mathsf{f}(x) = 0$, (ii) $\mathsf{f}'(x) = x \times 0$ and (iii) $\mathsf{f}''(x) = x - x$.

Definition 3 (The natural interval extension) *Consider a real-valued function $f(x) : \mathbb{R}^n \mapsto \mathbb{R}^m$ given by a formula or a DAG expression $\mathsf{f}(x)$. If real constants, variables, and operations in $\mathsf{f}(x)$ are replaced by their interval counterparts, then one obtains*

$$\mathbf{f}(\boldsymbol{x}) : \mathbb{IR}^n \mapsto \mathbb{IR}^m \ .$$

$\mathbf{f}(x)$ is known as the natural interval extension of the expression $\mathsf{f}(x)$ for $f(x)$. This extension is well-defined if we do not run into division by zero.

Although the three distinct expressions $\mathsf{f}(x)$, $\mathsf{f}'(x)$ and $\mathsf{f}''(x)$ of the real function $f : \mathbb{R} \mapsto \mathbb{R}$ of Example 1 are equivalent upon evaluation in the reals, their respective interval extensions $\mathbf{f}(\boldsymbol{x}) = [0, 0]$, $\mathbf{f}'(\boldsymbol{x}) = \boldsymbol{x} \times [0, 0]$, and $\mathbf{f}''(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{x}$ are not. For instance, if $\boldsymbol{x} = [1, 2]$,

$$\begin{array}{lcl} \mathbf{f}([1,2]) & = & [0,0], \\ \mathbf{f}'([1,2]) & = & [1,2] \times [0,0] = [\min\{1\times 0, 1\times 0, 2\times 0, 2\times 0\}, \max\{1\times 0, 1\times 0, 2\times 0, 2\times 0\}] = [0,0] \\ \mathbf{f}''([1,2]) & = & [1,2] - [1,2] = [1-2, 2-1] = [-1,1] \ , \end{array}$$

and in general for any $\boldsymbol{x} := [\underline{x}, \overline{x}] \in \mathbb{IR}$,

$$\begin{array}{lcl} \mathbf{f}([\underline{x},\overline{x}]) & = & [0,0], \\ \mathbf{f}'([\underline{x},\overline{x}]) & = & [\underline{x},\overline{x}] \times [0,0] = [\min\{\underline{x}\times 0, \underline{x}\times 0, \overline{x}\times 0, \overline{x}\times 0\}, \max\{\underline{x}\times 0, \underline{x}\times 0, \overline{x}\times 0, \overline{x}\times 0\}] = [0,0] \\ \mathbf{f}''([\underline{x},\overline{x}]) & = & [\underline{x},\overline{x}] - [\underline{x},\overline{x}] = [\underline{x}-\overline{x}, \overline{x}-\underline{x}] \neq [0,0], \quad \text{unless } \underline{x} = \overline{x} \ . \end{array}$$

Thus, $\mathbf{f}(\boldsymbol{x}) = \mathbf{f}'(\boldsymbol{x}) \neq \mathbf{f}''(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{IR}$, albeit $\mathsf{f}(x) = \mathsf{f}'(x) = \mathsf{f}''(x)$ for any $x \in \mathbb{R}$.

Theorem 3 (Interval rational functions) *Consider the rational function $f(x) = p(x)/q(x)$, where p and q are polynomials. Let $\mathbf{f}$ be the natural interval extension of its DAG expression f such that $\mathbf{f}(\boldsymbol{y})$ is well-defined for some $\boldsymbol{y} \in \mathbb{IR}$ and let $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{IR}$. Then we have*

$$\begin{array}{ll} (i) & \textit{Inclusion isotony:} \quad \forall\, \boldsymbol{x} \subseteq \boldsymbol{x}' \subseteq \boldsymbol{y} \implies \mathbf{f}(\boldsymbol{x}) \subseteq \mathbf{f}(\boldsymbol{x}') \ , \textit{ and} \\ (ii) & \textit{Range enclosure:} \quad \forall\, \boldsymbol{x} \subseteq \boldsymbol{y} \implies \mathrm{range}(f; \boldsymbol{x}) \subseteq \mathbf{f}(\boldsymbol{x}) \ . \end{array}$$

Definition 4 (Standard functions) *Piece-wise monotone functions, including exponential, logarithm, rational power, absolute value, and trigonometric functions, constitute the set of standard functions*

$$\mathfrak{S} = \{\, a^x, log_b(x), x^{p/q}, |x|, \sin(x), \cos(x), \tan(x), \sinh(x), \ldots, \arcsin(x), \ldots \} \ .$$

Such functions have well-defined interval extensions that satisfy inclusion isotony and *exact range enclosure*, i.e. $\mathrm{range}(f; \boldsymbol{x}) = \mathbf{f}(\boldsymbol{x})$. Consider the following definitions for the interval extensions for some monotone functions in $\mathfrak{S}$ with $\boldsymbol{x} \in \mathbb{IR}$,

$$\begin{array}{lcll} \exp(\boldsymbol{x}) & = & [\exp(\underline{x}), \exp(\overline{x})] & \\ \arctan(\boldsymbol{x}) & = & [\arctan(\underline{x}), \arctan(\overline{x})] & \\ \sqrt{(\boldsymbol{x})} & = & [\sqrt{(\underline{x})}, \sqrt{(\overline{x})}] & \text{if } 0 \leq \underline{x} \\ \log(\boldsymbol{x}) & = & [\log(\underline{x}), \log(\overline{x})] & \text{if } 0 < \underline{x} \ , \end{array}$$

and a piece-wise monotone function in $\mathfrak{S}$ with $\mathbb{Z}_+$ and $\mathbb{Z}_-$ representing the set of positive and negative integers, respectively. Let the *mignitude* of an interval $\boldsymbol{x}$ be the number $\langle \boldsymbol{x} \rangle = \min\{|x| : x \in \boldsymbol{x}\}$ and the *absolute value* of $\boldsymbol{x}$ be the number $|\boldsymbol{x}| = \max\{|x| : x \in \boldsymbol{x}\} = \sup\{-\underline{x}, \overline{x}\}$. Then, the interval-extended power function that plays a basic role in product likelihood functions is:

$$\boldsymbol{x}^n = \begin{cases} [\underline{x}^n, \overline{x}^n] & : \text{if } n \in \mathbb{Z}_+ \text{ is odd,} \\ [\langle \boldsymbol{x} \rangle^n, |\boldsymbol{x}|^n] & : \text{if } n \in \mathbb{Z}_+ \text{ is even,} \\ [1, 1] & : \text{if } n = 0, \\ [1/\overline{x}, 1/\underline{x}]^{-n} & : \text{if } n \in \mathbb{Z}_-; 0 \notin \boldsymbol{x} \ . \end{cases}$$

Definition 5 (Elementary functions) *A real-valued function that can be expressed as a finite combination of constants, variables, arithmetic operations, standard functions and compositions is called an elementary function. The set of all such elementary functions is referred to as* $\mathfrak{E}$.

Example 2 (Probability of the pattern xxx under CFN star tree ^0t) *The trifurcating star-tree* $^0t := (^0t_1)$ *has topology label* 0 *and common branch length parameter* $^0t_1$ *as shown in Figure 1(i). Either a direct application of Algorithm 2 with input as* $^0t := (^0t_1)$ *or a substitution of* $^0t_1$ *for* $^4t_1$, $^4t_2$ *and* $^4t_3$ *in* (6)*, yields the likelihood for pattern* xxx *as:*

$$l_{\texttt{xxx}}(^0t) \quad = \quad (1 + 3e^{-4(^0t_1)})/8 \ .$$

The probability of the pattern xxx *under CFN star tree* ^0t *given by* $l_{\texttt{xxx}}(^0t)$ *with the corresponding DAG expression shown in Figure 3 is an elementary function.*

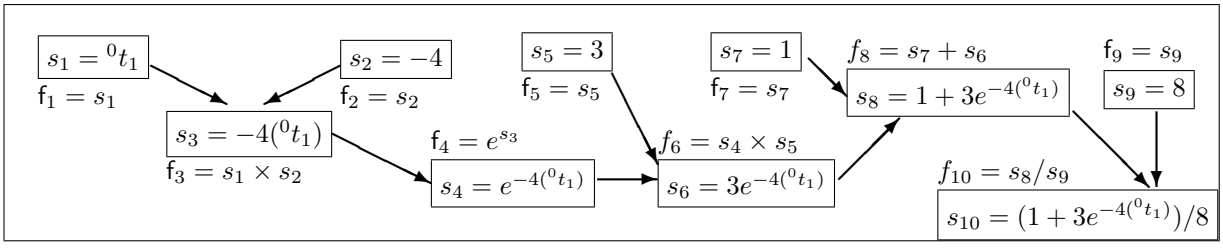

Figure 3: DAG expression for probability of the pattern xxx under a CFN star tree. The elementary function $l_0(^0t) = (1 + 3e^{-4(^0t_1)})/8$ can be obtained from the terminus $\odot \mathsf{f}_{10}$ of the recursion $\{\odot \mathsf{f}_i\}_{i=1}^{10}$ over the sub-expressions $\mathsf{f}_1, \ldots, \mathsf{f}_{10}$ in the above directed acyclic graph (DAG) expression of $l_0(^0t)$. Note that the leaf nodes are constants (s_2, s_5, s_7 and s_9) or variables (s_1).

It would be convenient if guaranteed enclosures of the range of an elementary f can be obtained by the natural interval extension $\mathbf{f}$ of one of its expressions f. The following Theorem 4 is the work-horse of interval Monte Carlo algorithms.

Theorem 4 (The fundamental theorem of interval analysis) *Consider any elementary function* $f \in \mathfrak{E}$ *with expression* f*. Let* $\mathsf{f} : \boldsymbol{y} \mapsto \mathbb{IR}$ *be its natural interval extension such that* $\mathsf{f}(\boldsymbol{y})$ *is well-defined for some* $\boldsymbol{y} \in \mathbb{IR}$ *and let* $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{IR}$*. Then we have*

$$\begin{array}{ll} (i) & \textit{Inclusion isotony:} \quad \forall\, \boldsymbol{x} \subseteq \boldsymbol{x}' \subseteq \boldsymbol{y} \implies \mathsf{f}(\boldsymbol{x}) \subseteq \mathsf{f}(\boldsymbol{x}')\,, \textit{ and} \\ (ii) & \textit{Range enclosure:} \quad \forall\, \boldsymbol{x} \subseteq \boldsymbol{y} \implies \mathrm{range}(f; \boldsymbol{x}) \subseteq \mathsf{f}(\boldsymbol{x})\ . \end{array}$$

The fundamental implication of the above theorem is that it allows us to enclose the range of any elementary function and thereby produces an upper bound for the global maximum and a lower bound for the global minimum over any compact subset of the domain upon which the function is well-defined. This is the work-horse for rigorously constructing an envelope for rejection sampling.

Unlike the natural interval extension of an $f \in \mathfrak{S}$ that produces exact range enclosures, the natural interval extension $\mathsf{f}(\boldsymbol{x})$ of an $f \in \mathfrak{E}$ often overestimates $\mathrm{range}(f; \boldsymbol{x})$, but can be shown under mild conditions to linearly approach the range as the maximal width of the box $\boldsymbol{x}$ goes to zero. This implies that a partition of $\boldsymbol{x}$ into smaller boxes $\{\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(m)}\}$ gives better enclosures of $\mathrm{range}(f; \boldsymbol{x})$ through the union $\bigcup_{i=1}^{m} \mathsf{f}(\boldsymbol{x}^{(i)})$ as illustrated in Figure 4. Next we make the above statements precise in terms of the *width* and *radius* of a box $\boldsymbol{x}$ defined by $\mathrm{wid}\, \boldsymbol{x} := \overline{x} - \underline{x}$ and $\mathrm{rad}\, \boldsymbol{x} := (\overline{x} - \underline{x})/2$, respectively.

Definition 6 *A function* $f : \mathbb{D} \mapsto \mathbb{R}$ *is Lipschitz if there exists a Lipschitz constant* K *such that, for all* $x, y \in \mathbb{D}$*, we have* $|f(x) - f(y)| \leq K|x - y|$*. We define* $\mathfrak{E}_{\mathfrak{L}}$ *to be the set of elementary functions whose sub-expressions* f_i*,* $i = 1, \ldots, n$ *at the nodes of its corresponding DAG* f *are all Lipschitz:*

$$\mathfrak{E}_{\mathfrak{L}} := \{f \in \mathfrak{E} : \textit{each sub-expression } \mathsf{f}_i \textit{ in the DAG expression } \mathsf{f} \textit{ for } f \textit{ is Lipschitz}\}\ .$$

Theorem 5 (Range enclosure tightens linearly with mesh) *Consider a function* $f : \mathbb{D} \mapsto \mathbb{R}$ *with* $f \in \mathfrak{E}_{\mathfrak{L}}$*. Let* f *be an inclusion isotonic interval extension of the DAG expression* f *of* f *such that* $\mathsf{f}(\boldsymbol{x})$ *is well-defined for some* $\boldsymbol{x} \in \mathbb{IR}$*. Then there exists a positive real number* K*, depending on* f *and* $\boldsymbol{x}$*, such that if* $\boldsymbol{x} = \cup_{i=1}^{k} \boldsymbol{x}^{(i)}$*, then*

$$\mathrm{range}(f; \boldsymbol{x}) \subseteq \bigcup_{i=1}^{k} \mathsf{f}(\boldsymbol{x}^{(i)}) \subseteq \mathsf{f}(\boldsymbol{x})\ ,$$

and

$$\mathrm{rad}\left(\bigcup_{i=1}^{k} \mathsf{f}(\boldsymbol{x}^{(i)})\right) \leq \mathrm{rad}\left(\mathrm{range}(f; \boldsymbol{x})\right) + K \max_{i=1,\ldots,k} \mathrm{rad}\left(\boldsymbol{x}^{(i)}\right)\ .$$

Likelihood of a box of trees

The likelihood function (2) over trees with a DAG expression that is directly or indirectly obtained via Algorithm 2 has a natural interval extension over boxes of trees [5, 26]. This interval extension of the

likelihood function allows us to produce rigorous enclosures of the likelihood over a box in the tree space. Next we give a concrete example of the natural interval extension of the likelihood function over an interval of trees ${}^0\boldsymbol{t}$ in the star-tree space ${}^0\mathbb{T}$. The same ideas extend to any labeled box of trees ${}^k\boldsymbol{t}$ when the number of branch lengths is greater than one and more generally to a finite union of labeled boxes with possibly distinct labels.

Example 3 (Posterior density over the CFN star-tree space ${}^0\mathbb{T}$) *The trifurcating star-tree* ${}^0t := ({}^0t_1)$ *has topology label* 0 *and common branch length* ${}^0t_1 > 0$. *Either a direct application of Algorithm 2 with input triplet* ${}^0t$ *or a substitution of* ${}^4t_1$, ${}^4t_2$ *and* ${}^4t_3$ *in* (6) *by* ${}^0t_1$ *yields the following* ${}^0\mathbb{T}$*-specific likelihoods:*

$$\begin{aligned} l_0({}^0t) = l_1({}^0t) &= (1 + 3e^{-4({}^0t_1)})/8 \ , \\ l_2({}^0t) = l_3({}^0t) = l_4({}^0t) = l_5({}^0t) = l_6({}^0t) = l_7({}^0t) &= (1 - e^{-4({}^0t_1)})/8 \ . \end{aligned} \tag{10}$$

Therefore, on the basis of (4), (5), (6) *and* (7), *the likelihood of the data at the star-tree* ${}^0t \in {}^0\mathbb{T}$ *is*

$$\begin{aligned} l_d({}^0t) &= \prod_{i=0}^{7} (l_i({}^0t))^{c_i} = \Big((1 + 3e^{-4({}^0t_1)})/8\Big)^{c_0+c_1} \Big((1 - e^{-4({}^0t_1)})/8\Big)^{\sum_{i=2}^{7} c_i} \\ &= \Big((1 + 3e^{-4({}^0t_1)})/8\Big)^{c_0+c_1} \Big((1 - e^{-4({}^0t_1)})/8\Big)^{v-(c_0+c_1)} \ , \end{aligned} \tag{11}$$

the posterior density (3) *based on a uniform prior* $p({}^0t_1) = 1/10$ *over* ${}^0\mathbb{T} = (0, 10]$ *is*

$$f^{\cdot}({}^0t) = \frac{l_d({}^0t)}{\int_0^{10} l_d({}^0t) \ \partial({}^0t)} \ . \tag{12}$$

Thus, under our conveniently chosen uniform prior, the target posterior shape (without the normalizing constant) is simply the likelihood function, i.e.

$$f({}^0t) = f^{\cdot}({}^0t) \int_0^{10} l_d({}^0t) \ \partial({}^0t) = l_d({}^0t) \ .$$

Observe that the minimal sufficient statistics over ${}^0\mathbb{T}$ *are the number of sites with the same character* $c_{\mathtt{xxx}} := c_0 + c_1$ *and the total number of sites* v. *Let the natural interval extension of the DAG expression for the posterior shape* $f({}^0t) : {}^0\mathbb{T} \mapsto \mathbb{R}$ *be:*

$$\mathbf{f}({}^0\boldsymbol{t}) : {}^0\mathbb{IT} \mapsto \mathbb{IR} \ .$$

Thus, $\mathbf{f}$ *maps an interval* ${}^0\boldsymbol{t}$ *in the tree space* ${}^0\mathbb{T}$ *to an interval in* $\mathbb{IR}$ *that encloses the target shape or likelihood of* ${}^0\boldsymbol{t}$.

For the human, chimpanzee and gorilla mitochondrial sequence data [27] analyzed in [17], $c_{\texttt{xxx}} = 762$ *and* $v = 895$*. Figure 4 shows* $\log(f({}^0t))$ *or the log-likelihood function for this data set as the white line. Evaluations of its interval extension over partitions by 3, 7 and 19 intervals are depicted by colored rectangles in Figure 4. Notice how the range enclosure by the interval extension of the log-likelihood function, our target shape, tightens with domain refinement as per Theorem 5. The maximum likelihood estimate derived in [17] (the red dot in Figure 4) is*

$$\widehat{{}^0t} = \underset{{}^0t \in {}^0\mathbb{T}}{\operatorname{argmax}} f({}^0t) = (0.055205) \ .$$

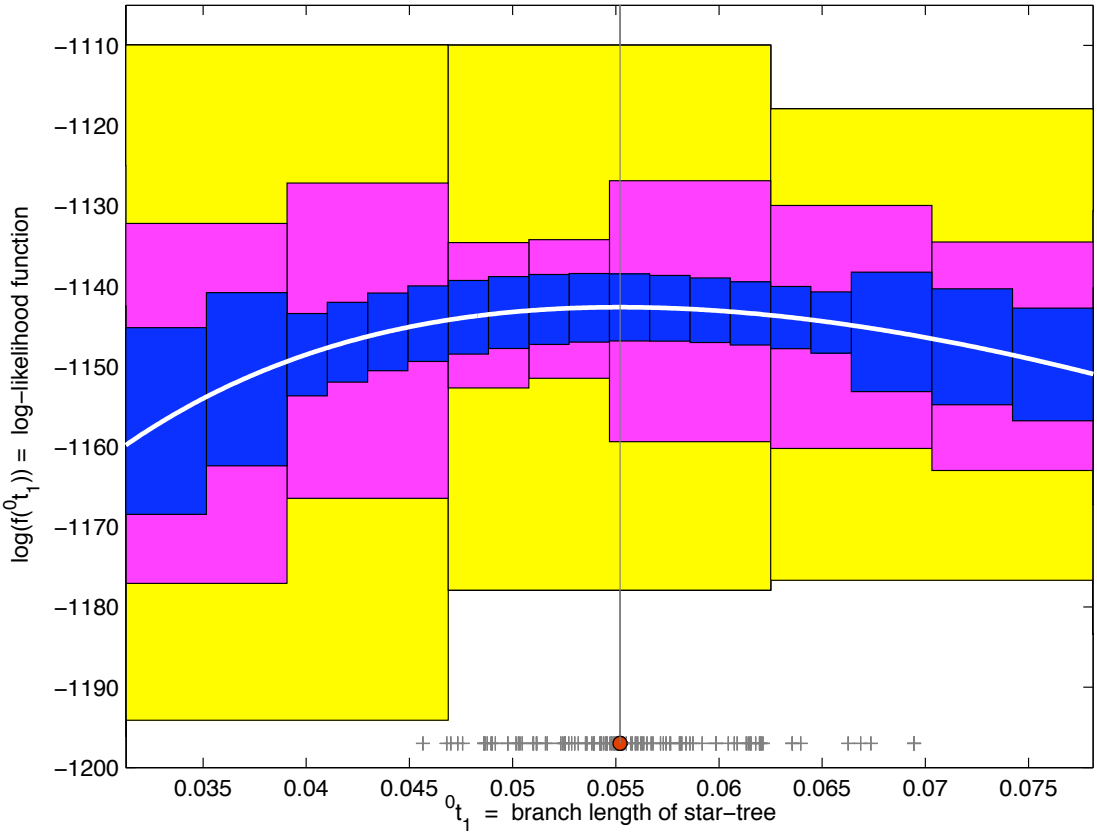

Figure 4: Adaptive range enclosure of the posterior density over the star-tree space. Range enclosure of the log-likelihood (white line) for the human, chimpanzee and gorilla mitochondrial sequence data [27] analyzed in [17], under the CFN model with $c_{\texttt{xxx}} = 762$ and $v = 895$ over star-trees, via its interval extension linearly tightens with the mesh. One hundred samples (+) from the MRS and the maximum likelihood estimate (red dot) are shown.

Moore rejection sampler (MRS)

Moore rejection sampler (MRS) is an auto-validating rejection sampler (RS). MRS is said to be auto-validating because it automatically obtains a proposal g that is easy to simulate from, and an envelope $\widehat{g}$ that is guaranteed to satisfy the envelope condition (1). MRS can produce independent samples from any target shape f whose DAG expression f has a well-defined natural interval extension $\mathbf{f}$ over a compact domain $\mathbb{T}$. In summary, the defining characteristics and notations of MRS are:

Compact domain	$\mathbb{T} = [\underline{t}, \overline{t}]$
Target shape	$f(t) : \mathbb{T} \mapsto \mathbb{R}$
Target integral	$N_f := \int_{\mathbb{T}} f(t)\, dt$
Target density	$f^{\cdot}(t) := (N_f)^{-1} f(t) : \mathbb{T} \mapsto \mathbb{R}$
DAG expression of f	$\mathsf{f}(t) : \mathbb{T} \mapsto \mathbb{R}$
Interval extension of f	$\mathbf{f}(\boldsymbol{t}) : \mathbb{IT} \mapsto \mathbb{IR}$
Envelope function	$\widehat{g}(t) : \mathbb{T} \mapsto \mathbb{R}$
Envelope integral	$N_{\widehat{g}} := \int_{\mathbb{T}} \widehat{g}(t)\, dt$
Proposal density	$g(t) := (N_{\widehat{g}})^{-1} \widehat{g}(t) : \mathbb{T} \mapsto \mathbb{R}$
Acceptance probability	$\boldsymbol{A}(\widehat{g}) = N_f / N_{\widehat{g}}$
Partition of $\mathbb{T}$	$\mathfrak{T} := \{\, \boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, ..., \boldsymbol{t}^{(\lvert\mathfrak{T}\rvert)} \,\}$.

Suppose f is an elementary function and its DAG expression f has a well-defined interval extension $\mathbf{f}$ on $\mathbb{T}$. If $\mathfrak{T} := \{\, \boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, ..., \boldsymbol{t}^{(\lvert\mathfrak{T}\rvert)} \,\}$ is a finite partition of $\mathbb{T}$, then by Theorem 4 we can enclose $\text{range}(f; \boldsymbol{t}^{(i)})$, i.e. the range of f over the i-th element of $\mathfrak{T}$, with the interval extension $\mathbf{f}$ of f:

$$\text{range}(f; \boldsymbol{t}^{(i)}) \subseteq \mathbf{f}(\boldsymbol{t}^{(i)}) := [\underline{\mathbf{f}}(\boldsymbol{t}^{(i)}), \overline{\mathbf{f}}(\boldsymbol{t}^{(i)})], \, \forall\, i \in \{1, 2, ..., \lvert\mathfrak{T}\rvert\} \ . \tag{13}$$

For a given partition $\mathfrak{T}$, we can construct a partition-specific envelope function:

$$\widehat{g}^{\mathfrak{T}}(t) = \sum_{i=1}^{\lvert\mathfrak{T}\rvert} \overline{\mathbf{f}}(\boldsymbol{t}^{(i)})\, \mathbf{1}_{\{t \,\in\, \boldsymbol{t}^{(i)}\}}, \qquad \mathbf{1}_{\{t \,\in\, \boldsymbol{t}^{(i)}\}} = \begin{cases} 1 & \text{if } t \ \in \ \boldsymbol{t}^{(i)} \\ 0 & \text{otherwise} \ . \end{cases} \tag{14}$$

The necessary envelope condition (1) is satisfied by $\widehat{g}^{\mathfrak{T}}(t)$ because of (13). We can obtain the corresponding proposal $g^{\mathfrak{T}}(t)$ as a normalized simple function over $\mathbb{T}$:

$$g^{\mathfrak{T}}(t) = \left(N_{\widehat{g}^{\mathfrak{T}}}\right)^{-1} \widehat{g}^{\mathfrak{T}}(t) = \left(N_{\widehat{g}^{\mathfrak{T}}}\right)^{-1} \sum_{i=1}^{\lvert\mathfrak{T}\rvert} \overline{\mathbf{f}}(\boldsymbol{t}^{(i)})\, \mathbf{1}_{\{t \,\in\, \boldsymbol{t}^{(i)}\}} \ , \tag{15}$$

where the normalizing constant $N_{\widehat{g}^{\mathfrak{T}}} := \sum_{i=1}^{\lvert\mathfrak{T}\rvert} \left(\text{vol}\, \boldsymbol{t}^{(i)} \cdot \overline{\mathbf{f}}(\boldsymbol{t}^{(i)})\right)$ and $\text{vol}\, \boldsymbol{t} := \prod_{i=1}^{n} \text{wid}\, \boldsymbol{t}_i$ is the *volume* of the box $\boldsymbol{t}$. The volume of an interval $\boldsymbol{x}$ is simply its width, i.e. $\text{vol}\, \boldsymbol{x} = \text{wid}\, \boldsymbol{x}$, if $\boldsymbol{x} \in \mathbb{IR}$. Now, we have all the ingredients to perform a more efficient, partition-specific, auto-validating von Neumann rejection sampling or simply Moore rejection sampling.

Before making formal statements about our sampler let us gain geometric insight into the sampler from Example 3 and Figure 4. The upper boundaries of rectangles of a given color, depicting a simple function

in Figure 4, is a partition-specific envelope function (14) for the logarithm of the posterior shape or the log-likelihood function of Example 3 over the prior-specified support $[10^{-10}, 10] \subset {}^0\mathbb{T}$. In Figure 4 only a small interval about the maximum likelihood estimate (red dot) that contains the posterior samples (gray '+' markers) is depicted since the likelihood falls sharply outside this range. Normalization of the envelope gives the corresponding proposal function (15). As the refinement of the domain proceeds through adaptive bisections (described later), the partition size increases. We show partitions of size 3,7 and 19 over an interval containing the posterior samples. These samples were obtained from the partition with 19 intervals. Each of the corresponding envelope functions (upper boundaries of rectangles of a given color) can be used to draw independent and identically distributed samples from the target posterior density. Note how the acceptance probability (ratio of the area below the target shape to that below the envelope) increases with refinement.

Theorem 6 shows that Moore rejection sampler (MRS) indeed produces independent samples from the desired target and Theorem 7 describes the asymptotics of the acceptance probability as the partition of the domain is refined. Proofs for both Theorems are included in the Appendix for completeness.

Theorem 6 *Suppose that the DAG expression* f *of the target shape* f *has a well-defined natural interval extension* $\mathbf{f}$ *over* $\mathbb{T} \in \mathbb{IR}^n$*. If* T *is generated according to Algorithm 1, and if the the envelope function* $\widehat{g}^{\mathfrak{T}}(t)$ *and the proposal density* $g^{\mathfrak{T}}(t)$ *are given by* (14) *and* (15)*, respectively, then* T *is distributed according to the target density* $f^{\cdot} : \mathbb{T} \mapsto \mathbb{R}$*.*

Next we bound the partition-specific acceptance probability $\boldsymbol{A}(\mathfrak{T}) := \boldsymbol{A}(\widehat{g}^{\mathfrak{T}})$ for this sampler. For simplicity, let the domain $\mathbb{T}$ of the target shape f be an interval. Due to the linearity of the integral operator and (13),

$$\begin{array}{rcl} N_f & := & \int_{\mathbb{T}} f(t)\,dt \\ & = & \sum_{i=1}^{|\mathfrak{T}|} \int_{\boldsymbol{t}^{(i)}} f(t)\,dt \\ & \in & \sum_{i=1}^{|\mathfrak{T}|} \left(\mathrm{wid}\left(\boldsymbol{t}^{(i)}\right) \cdot \mathbf{f}(\boldsymbol{t}^{(i)})\right) \\ & = & \left[\ \sum_{i=1}^{|\mathfrak{T}|} \left(\mathrm{wid}\left(\boldsymbol{t}^{(i)}\right) \cdot \underline{\mathbf{f}}(\boldsymbol{t}^{(i)})\right),\ \sum_{i=1}^{|\mathfrak{T}|} \left(\mathrm{wid}\left(\boldsymbol{t}^{(i)}\right) \cdot \overline{\mathbf{f}}(\boldsymbol{t}^{(i)})\right)\ \right]\ . \end{array}$$

Therefore,

$$\boldsymbol{A}(\mathfrak{T}) := \boldsymbol{A}(\widehat{g}^{\mathfrak{T}}) = \frac{N_f}{N_{\widehat{g}^{\mathfrak{T}}}} = \frac{N_f}{\sum_{i=1}^{|\mathfrak{T}|} \left(\mathrm{wid}\left(\boldsymbol{t}^{(i)}\right) \cdot \overline{\mathbf{f}}(\boldsymbol{t}^{(i)})\right)} \geq \frac{\sum_{i=1}^{|\mathfrak{T}|} \left(\mathrm{wid}\left(\boldsymbol{t}^{(i)}\right) \cdot \underline{\mathbf{f}}(\boldsymbol{t}^{(i)})\right)}{\sum_{i=1}^{|\mathfrak{T}|} \left(\mathrm{wid}\left(\boldsymbol{t}^{(i)}\right) \cdot \overline{\mathbf{f}}(\boldsymbol{t}^{(i)})\right)}\ . \tag{16}$$

If $f \in \mathfrak{E}_{\mathfrak{L}}$, the Lipschitz class of elementary functions (Definition 6), then we might expect the enclosure of N_f to be proportional to the *mesh* $w := \max_{i \in \{1,\ldots,\mathfrak{T}\}} \mathrm{wid}\left(\boldsymbol{t}^{(i)}\right)$ of the partition $\mathfrak{T}$.

Theorem 7 *Let $\mathfrak{U}_W$ be the uniform partition of $\mathbb{T} = [\underline{t}, \bar{t}]$ into W intervals each of width w*

$$\begin{array}{lcl} w & = & \frac{(\bar{t}-\underline{t})}{W} \\ \boldsymbol{t}_W^{(i)} & = & [\ \underline{t} + (i-1)w,\ \underline{t} + iw\ ], i = 1, \ldots, W \\ \mathfrak{U}_W & = & \{\boldsymbol{t}_W^{(i)}, i = 1, \ldots, W\}. \end{array}$$

and let $f \in \mathfrak{E}_{\mathfrak{L}}$, then

$$\boldsymbol{A}(\mathfrak{U}_W) = 1 - \mathcal{O}(1/W)$$

Theorem 7 shows that if $f \in \mathfrak{E}_{\mathfrak{L}}$ and $\mathfrak{U}_W$ is a uniform partition of $\mathbb{T}$ into W intervals, then the acceptance probability $\boldsymbol{A}(\mathfrak{U}_W) = 1 - \mathcal{O}(1/W)$. Thus, the acceptance probability approaches 1 at a rate that is no slower than linearly with the mesh.

Prioritized partitions and pre-processed proposals

We studied the efficiency of uniform partitions for their mathematical tractability. In practice, we may further increase the acceptance probability for a given partition size by adaptively partitioning $\mathbb{T}$. In our context, adaptive means the possible exploitation of any current information about the target. We can refine the current partition $\mathfrak{T}_\alpha$ and obtain a finer partition $\mathfrak{T}_{\alpha'}$ with an additional box by bisecting a box $\boldsymbol{t}^{(*)} \in \mathfrak{T}_\alpha$ along the midpoint of its side with the maximal width into a left box $\boldsymbol{t}_L^{(*)}$ and a right box $\boldsymbol{t}_R^{(*)}$. There are several ways to choose a box $\boldsymbol{t}^{(*)} \in \mathfrak{T}_\alpha$ for bisection. For instance, a relatively optimal choice is

$$\boldsymbol{t}^{(*)} = \underset{\boldsymbol{t}^{(i)} \in \mathfrak{T}_\alpha}{\operatorname{argmax}} \left(\operatorname{vol}(\boldsymbol{t}^{(i)}) \cdot \operatorname{wid}(\mathbf{f}(\boldsymbol{t}^{(i)}) \right) \ . \tag{17}$$

We employ a priority queue to conduct sequential refinements of $\mathbb{T}$ under this partitioning scheme. This approach avoids the exhaustive argmax computations to obtain the $\boldsymbol{t}^{(*)}$ for bisection at each refinement step. Thus, the current partition is represented by a queue of boxes that are prioritized in descending order by the the priority function $\operatorname{vol}(\boldsymbol{t}^{(i)}) \cdot \operatorname{wid}(\mathbf{f}(\boldsymbol{t}^{(i)})$ in (17). Therefore, the box with the largest uncertainty in the enclosure of the integral over it gets bisected first. There are several ways to decide when to stop refining the partition. A simple strategy is to stop when the number of boxes reaches a number that is well within the memory constraints of the computer, say 10^6, or when the lower bound of the acceptance probability given by (16) is above a desired threshold, say 0.1.

Once we have a partition $\mathfrak{T}$ of $\mathbb{T}$, we can sample t from the proposal density $g^{\mathfrak{T}}$ given by (15) in two steps:

1. Sample a box $\boldsymbol{t}^{(i)} \in \mathfrak{T}$ according to the discrete distribution:

$$\ddot{g}^{\mathfrak{T}}(\boldsymbol{t}^{(i)}) = \frac{\operatorname{vol} \boldsymbol{t}^{(i)} \bar{\mathbf{f}}(\boldsymbol{t}^{(i)})}{\sum_{i=1}^{|\mathfrak{T}|} \left(\operatorname{vol} \boldsymbol{t}^{(i)} \bar{\mathbf{f}}(\boldsymbol{t}^{(i)}) \right)}, \ \boldsymbol{t}^{(i)} \in \mathfrak{T} \ , \tag{18}$$

2. Sample a point t uniformly at random from the box $\boldsymbol{t}^{(i)}$.

Sampling from large discrete distributions (with million states or more) can be made faster by pre-processing the probabilities and saving the result in some convenient look-up table. This basic idea [28] allows samples to be drawn rapidly. We employ an efficient pre-processing strategy known as the Alias Method [4] that allows samples to be drawn in constant time even for very large discrete distributions as implemented in the GNU Scientific Library [29]. We also minimize the number of evaluations of the target shape f by saving the box-specific computations of $\underline{\mathbf{f}}(\boldsymbol{t}^{(i)})$ and $\bar{\mathbf{f}}(\boldsymbol{t}^{(i)})$ and exploiting the so-called "squeeze principle", i.e. immediately accepting those points proposed in the box $\boldsymbol{t}^{(i)}$ that fall below $\underline{\mathbf{f}}(\boldsymbol{t}^{(i)})$ when uniformly stretched toward $\bar{\mathbf{f}}(\boldsymbol{t}^{(i)})$.

Thus, by means of priority queues and look-up tables we can efficiently manage our adaptive partitioning of the domain for envelope construction, and rapidly draw samples from the proposal distribution. Our sampler class `MRSampler` implemented in `MRS 0.1.2`, a `C++` class library for statistical set processing, builds on `C-XSC 2.0`, a `C++` class library for extended scientific computing using interval methods [30]. All computations were done on a 2.8 GHz Pentium IV machine with 1GB RAM. Having given theoretical and practical considerations to our Moore rejection sampler, we are ready to draw samples from various targets over small tree spaces.

Results

The natural interval extension of the likelihood function over labeled boxes in the tree space allows us to employ the Moore rejection sampler to rigorously draw independent and identically distributed samples from the posterior distribution over a compact box in the tree space given by our prior distribution. We draw samples from the posterior distribution based on two mitochondrial DNA data sets and use these samples (i) to estimate the posterior probabilities of each of the three rooted topologies, (ii) to conduct a nonparametric test of rate homogeneity between protein-coding and tRNA-coding sites and (iii) to estimate the human-neanderthal divergence time.

Human, chimpanzee and gorilla

We revisit the data from a segment of the mitochondrial DNA of human, chimpanzee and gorilla [27] that was analyzed under the CFN model of DNA mutation (Model 1) within a point estimation setting [17]. The sufficient statistics of pattern counts for this data with total number of sites $v = 895$ under the CFN

model over the space of all three-leaved phylogenetic trees are:

$$(c_{\mathtt{xxx}}, c_{\mathtt{xxy}}, c_{\mathtt{yxx}}, c_{\mathtt{xyx}}) = (762, 54, 41, 38)$$

Let human, chimpanzee and gorilla be denoted by leaf labels 1, 2 and 3, or H, C and G, respectively. Let the set of rooted tree labels corresponding to (ii),(iii) and (iv) of Figure 1 be $\mathfrak{K} = \{1, 2, 3\}$. The maximum likelihood estimate over ${}^{\mathfrak{K}}\mathbb{T} := {}^1\mathbb{T} \cup {}^2\mathbb{T} \cup {}^3\mathbb{T}$, the rooted and clocked three-leaved phylogenetic tree space, is derived in [17] as

$$\widehat{{}^1t} := (\widehat{{}^1t_0}, \widehat{{}^1t_1}) = \underset{({}^it_0, {}^it_1) \in {}^{\mathfrak{K}}\mathbb{T}}{\operatorname{argmax}} f({}^it_0, {}^it_1) = (0.010036, 0.048559) \ .$$

Recall that due to our flat priors, our posterior shape $f({}^it) := f({}^it_0, {}^it_1)$ with $i \in \mathfrak{K} = \{1, 2, 3\}$ is our likelihood function over ${}^{\mathfrak{K}}\mathbb{T}$. Now, suppose ${}^{b_1}t^{(1)}, {}^{b_2}t^{(2)}, \ldots, {}^{b_n}t^{(n)}$ are n independent and identically distributed samples from the posterior density $f^{\cdot}$ over ${}^{\mathfrak{K}}\mathbb{T}$. We can obtain asymptotically consistent estimates of the posterior probabilities of ${}^1\mathbb{T}$, ${}^2\mathbb{T}$ and ${}^3\mathbb{T}$ from Monte Carlo integration of the indicator function of each of the three topology labels using

$$\widehat{{}^jP_n} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{b_i=j\}}({}^{b_i}t) \xrightarrow{P} {}^jP := \int_{{}^{\mathfrak{K}}\mathbb{T}} \mathbf{1}_{\{i=j\}}({}^it) f^{\cdot}({}^it) \ \partial({}^it), \qquad \mathbf{1}_{\{b_i=j\}}({}^{b_i}t^{(i)}) = \begin{cases} 1 & \text{if } \ b_i = j \\ 0 & \text{otherwise} \end{cases} .$$

The 95% confidence interval for ${}^jP$, based on asymptotic normality of the Monte Carlo estimator, is

$$\widehat{{}^jP_n} \pm 1.96\sqrt{\widehat{{}^jP_n}(1 - \widehat{{}^jP_n})/n} \ .$$

Point estimate and a symmetric 95% confidence interval for the posterior probability of each of the three topologies from $n = 10^6$ posterior samples are

$$\widehat{{}^1P}_{10^6} = 0.8875 \pm 0.0006 \ ,$$

$$\widehat{{}^2P}_{10^6} = 0.0646 \pm 0.0005 \ ,$$

$$\widehat{{}^3P}_{10^6} = 0.0479 \pm 0.0004 \ .$$

These point estimates are in agreement with estimates obtained in [31,32] through quadrature routines in `Mathematica`. The first 10,000 of these samples are shown in Figure 5 upon transforming the rooted and clocked trees, ${}^it := ({}^it_0, {}^it_1), i \in \{1, 2, 3\}$, into constrained unrooted trees, ${}^4t := ({}^4t_1, {}^4t_2, {}^4t_3)$, according to Table 1.

Obtaining confidence intervals from dependent MCMC samples requires nontrivial computations for the burn-in period and the thinning rate [1]. These are not readily available for phylogenetic MCMC samplers.

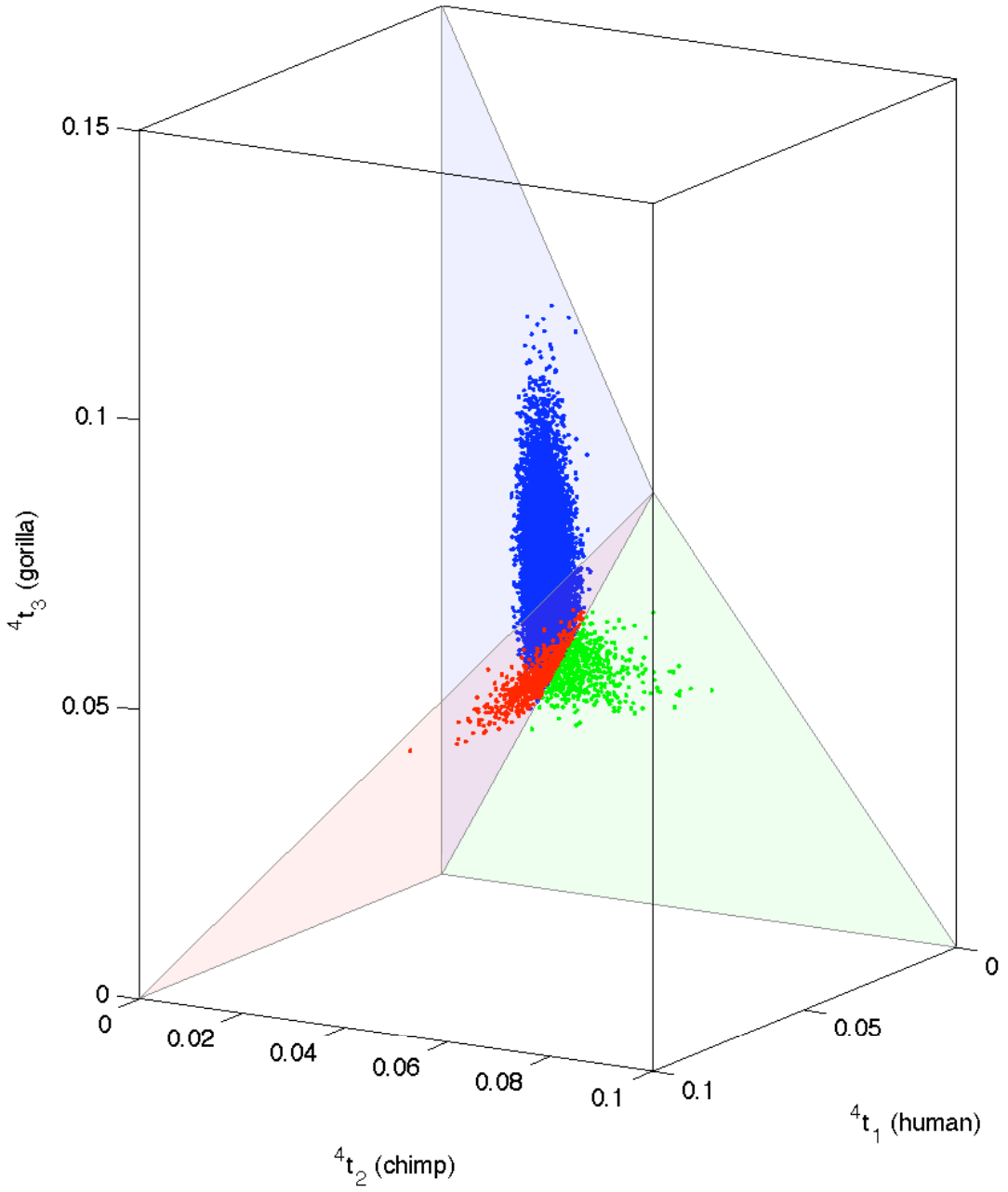

Figure 5: Posterior samples from the rooted tree space of human, chimpanzee and gorilla. Ten thousand independent and identically distributed posterior samples from the rooted and clocked binary tree space of human, chimpanzee and gorilla with topology label set $\{1 \cup 2 \cup 3\}$ (see Figure 1(ii),(iii),(iv)) on the basis of mitochondrial data [27] summarized by $(c_{\mathtt{xxx}}, c_{\mathtt{xxy}}, c_{\mathtt{yxx}}, c_{\mathtt{xyx}}) = (762, 54, 41, 38)$ under the Cavender-Farris-Neyman model (blue $\cup$ red $\cup$ green dots, respectively) are depicted.

Rooted and Clocked Trees, ${}^it := ({}^it_0, {}^it_1), i \in \{1,2,3\}$		Unrooted Trees ${}^4t := ({}^4t_1, {}^4t_2, {}^4t_3)$		
Labeled Tree	Newick Representation of ${}^it$	${}^4t_1$	${}^4t_2$	${}^4t_3$
${}^1t := ({}^1t_0, {}^1t_1)$	((H:${}^1t_1$,C:${}^1t_1$):${}^1t_0$,G:${}^1t_0 + {}^1t_1$))	${}^1t_1$	${}^1t_1$	${}^1t_1 + {}^1t_0 + {}^1t_0$
${}^2t := ({}^2t_0, {}^2t_1)$	((C:${}^2t_1$,G:${}^2t_1$):${}^2t_0$,H:${}^2t_0 + {}^2t_1$))	${}^2t_1 + {}^2t_0 + {}^2t_0$	${}^2t_1$	${}^2t_1$
${}^3t := ({}^3t_0, {}^3t_1)$	((H:${}^3t_1$,G:${}^3t_1$):${}^3t_0$,C:${}^3t_0 + {}^3t_1$))	${}^3t_1$	${}^3t_1 + {}^3t_0 + {}^3t_0$	${}^3t_1$

Table 1: Rooted triplets as constrained unrooted triplets. Any labeled, rooted and clocked tree with three leaves can be represented as a constrained unrooted tree according to the tabulated transformation.

Thus, the independent and identically distributed samples from our rejection sampler has the advantage of producing valid confidence intervals for our integrals of interest. The point estimate of the posterior mean $E({}^1T) := \int_{{}^1\mathbb{T}} {}^1t \; f({}^1t) \; \partial({}^1t)$ for topology label 1 is $(0.010863, 0.048994)$. This posterior mean is close to $(0.010036, 0.048559)$, the mode of our target shape or the maximum likelihood estimate derived in [17].

Chimpanzee, gorilla and orangutan

We focus here on the 895 bp long homologous segment of mitochondrial DNA from chimpanzee, gorilla and orangutan [27]. This gives us a greater phylogenetic depth than the human, chimpanzee and gorilla sequences that were just analyzed. These sequences encode the genes for three transfer RNAs and parts of two proteins. Under the assumption of independence across sites, the sufficient statistics, under the JC model of DNA mutation (Model 2) over triplets, are given in Table 2 for all of the data as well as a partition of the data into tRNA-coding and protein-coding sites.

Site type	v	$c_{\mathtt{xxx}}$	$c_{\mathtt{xxy}}$	$c_{\mathtt{yxx}}$	$c_{\mathtt{xyx}}$	$c_{\mathtt{xyz}}$
All	895	700	100	46	42	7
tRNA-coding	198	173	13	7	3	2
protein-coding	697	527	87	39	39	5

Table 2: Minimal sufficient statistics for the chimpanzee, gorilla and orangutan data. The minimal sufficient statistics under the JC model for all 895 sites, 198 tRNA-coding sites and 697 protein-coding sites based on the homologous segment of mitochondrial DNA from chimpanzee, gorilla and orangutan [27].

Ten thousand independent and identically distributed samples were drawn in 942 CPU seconds from the posterior distribution over JC triplets, i.e. unrooted trees with three edges corresponding to the three primates. Figure 6 shows these samples (blue dots) scattered about the verified global maximum likelihood estimate (MLE) of the triplet obtained in [5, 26] and subsequently confirmed algebraically in [23]. We also drew ten thousand independent and identically distributed samples from the posterior based on the 198 tRNA-coding DNA sites (green dots in Figure 6) as well as from that based on the remaining 697

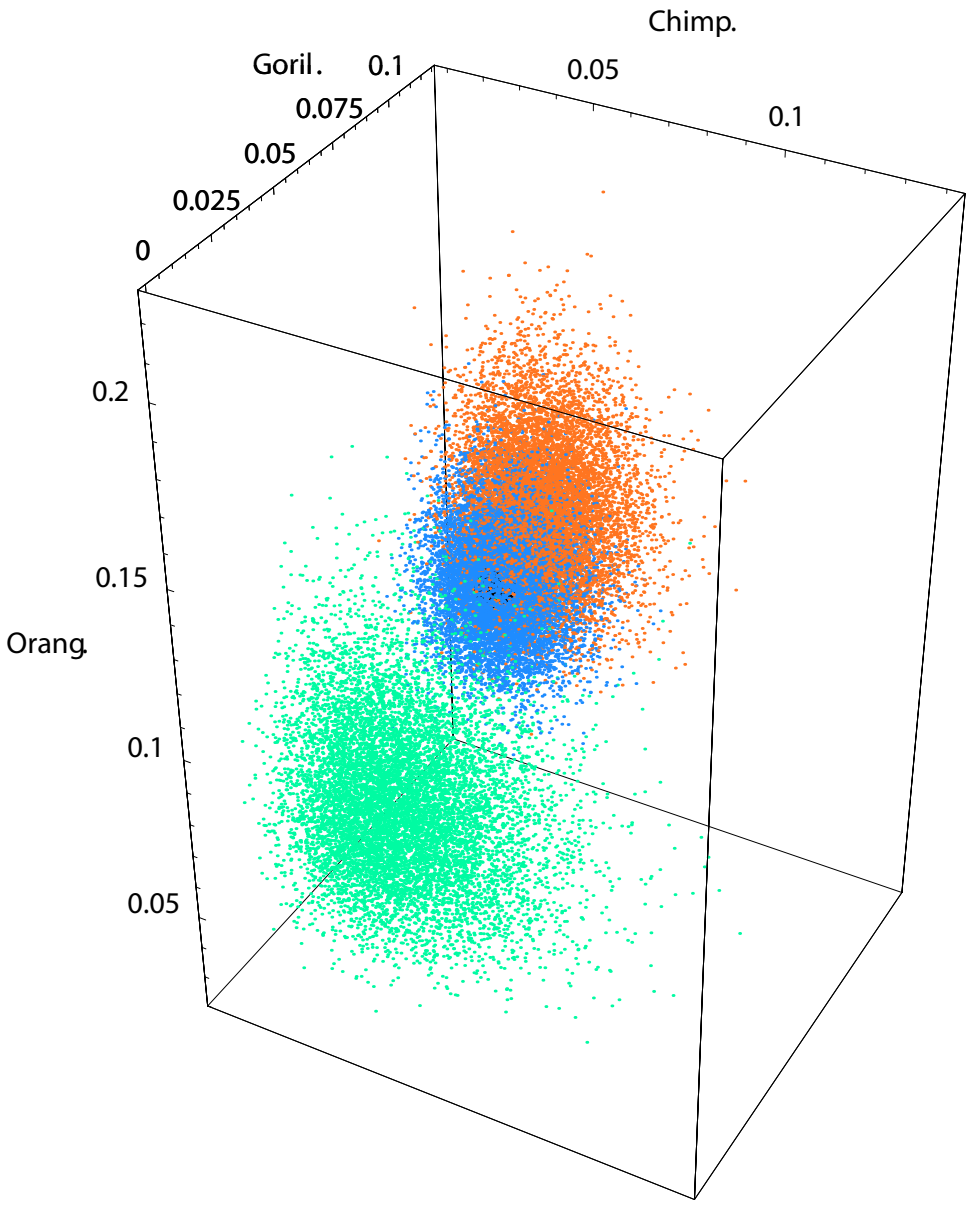

Figure 6: Posterior samples from the unrooted tree space of chimpanzee, gorilla and orangutan. Ten thousand Moore rejection samples from the posterior distribution over the three branch lengths of the unrooted tree space of chimpanzee, gorilla and orangutan based on their homologous mitochondrial DNA sequence of length 895 base pairs (blue dots), the tRNA-coding sequence with 198 base pairs (green dots) and the protein-coding sequence with 697 base pairs (red dots). The verified maximum likelihood estimate is the large black dot within the blue dots.

protein-coding sites (red dots in Figure 6). The former posterior samples, corresponding to the tRNA-coding sites, are more dispersed than the posterior samples based on the entire sequence. This is due to the smaller number of tRNA-coding sites making the posterior less concentrated. Moreover, the cluster of samples from the posterior based on tRNA-coding sites seem to be farther away from that based on protein-coding sites. Such a clustering of two sets of posterior samples is a signal of mutational rate heterogeneity between the two types of sites. Hotelling's trace statistics, being a natural measure of distance between two clusters of points, can be used as a test statistic to determine the significance of the observed test statistic. On the basis of 100 random permutations of the sites, we obtain the null distribution of Hotelling's trace statistics. We were able to reject the null hypothesis of rate homogeneity between the posterior samples based on the tRNA-coding sites and that based on the protein-coding sites at the 10% significance level using this permutation test (P-value = 0.06). Any biological interpretation of this test must be done cautiously since the JC model employed here forbids any transition:transversion bias that is reportedly relevant for this data [27].

Neanderthal, human and chimpanzee

We used the 15 site patterns and their counts in Table 3 to infer the human-neanderthal divergence time. These counts are obtained from a multiple sequence alignment of the data made available in [33]. Our alignment procedure is more robust at the ends of each locus than that of [33]. We do an ordered concatenation of all the loci for each species prior to a multiple sequence alignment. The alignment was further edited by hand to obtain the locus-specific alignments. Under the assumption of independence across sites, the sufficient statistics, under any Markov model of DNA mutation, is the set of distinct site patterns and their respective counts. They are given in Table 3 for this data set.

We drew $10,000$ samples that were independently and identically distributed from each of three posterior densities; (i) over the space of unrooted triplets under the JC model in 312 CPU seconds, (ii) over the clocked and rooted triplets under the JC model in 375 CPU seconds and (iii) over the clocked and rooted triplets under the HKY model in 1.2 CPU hours. In the HKY model we used the empirical nucleotide frequencies from the data ($\pi(\mathtt{T}) = 0.2588$, $\pi(\mathtt{C}) = 0.2571$, $\pi(\mathtt{A}) = 0.2916$, $\pi(\mathtt{G}) = 0.1925$) and a hominid-specific transition:transversion rate of 2.0. Unlike the JC model with five sufficient statistics $(c_{\mathtt{xxx}}, c_{\mathtt{xxy}}, c_{\mathtt{yxx}}, c_{\mathtt{xyx}}, c_{\mathtt{xyz}}) = (2343, 56, 2, 4, 0)$, all 15 distinct site patterns are required for the likelihood computations under the HKY model and this is reflected in its longer CPU time. Both models gave similar posterior samples over rooted triplets, as shown in Figure 7.

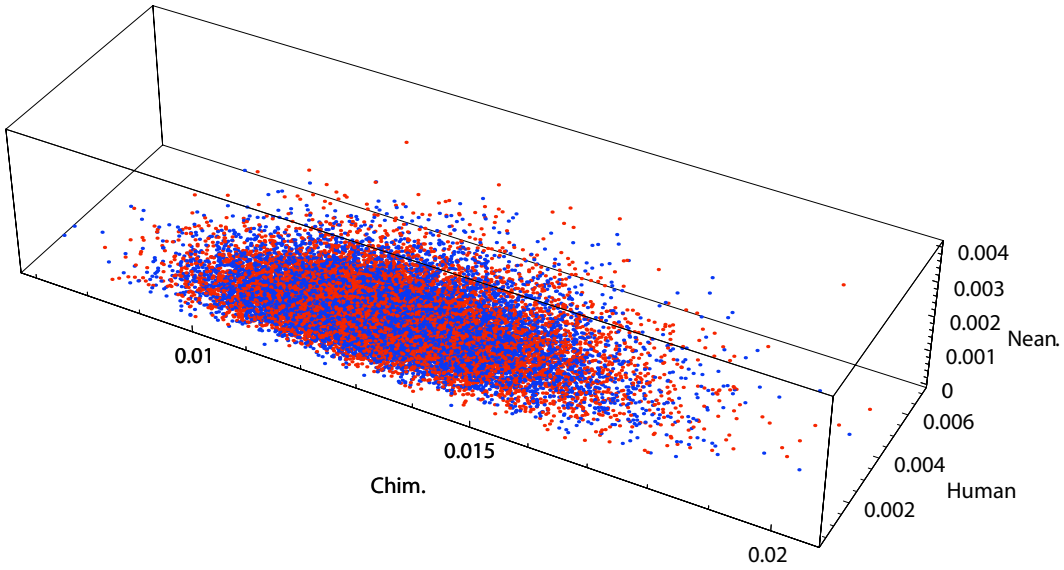

Figure 7: Posterior samples from the unrooted tree space of neanderthal, human and chimpanzee. Ten thousand Moore rejection samples each from the posterior distribution over the three branch lengths of the unrooted tree space of neanderthal, human and chimpanzee under the JC model (blue dots) and the HKY model (red dots)

We transformed the three posterior distributions over the triplet spaces; (i) unrooted JC triplets that were rooted using the mid-point rooting method, (ii) rooted JC triplets and (iii) rooted HKY triplets, respectively, into three posterior distributions over the human-neanderthal divergence time relative to the human-chimp divergence time. The corresponding posterior quantiles ($\{5\%, 50\% , 95\%\}$) for the human-neanderthal divergence time in units of human-chimp divergence time are $\{0.0643 , 0.125 , 0.214\}$, $\{0.0694 , 0.142 , 0.263\}$ and $\{0.0682 , 0.143 , 0.268\}$, respectively. We constrained the neanderthal lineage to be a fraction of the human lineage in branch length in order to estimate the age of the neanderthal fossil from the rooted HKY triplets. The posterior quantiles of the fossil date in units of human-chimp divergence is $\{0.00685 , 0.0666 , 0.195\}$. The estimate of $38,310$ years based on carbon-14 accelerator mass spectrometry [33] is within our $[5\%, 95\%]$ posterior quantile interval for the fossil date, provided the human-chimp divergence estimate ranges in $[196103, 5.6 \times 10^6]$. Thus, reasonable bounds for the human-chimp divergence are 4×10^6 and 5.6×10^6 years, under the assumption that 4×10^6 is an acceptable lower-bound. Based on these two calendar year estimates, we transformed the posterior quantiles of the human-neanderthal divergence times from the rooted HKY triplets into $\{272680 , 571124 , 1073375\}$ and $\{381752 , 799574 , 1502724\}$ years, respectively. Our $[5\%, 95\%]$ posterior intervals contain the interval estimate of $[461000, 825000]$ years reported in [33]. However, our confidence intervals are from

```
site       : 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1
pattern    : 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
 . . . . . . . . . . . . . . . . . . . . .
neanderthal: a t c g a t c g t t g a c a a
human      : a t c g a t c g t c a g t a g
chimpanzee : a t c g g c t a a t a a c t g
 . . . . . . . . . . . . . . . . . . . . .
site       : 6 6 6 4 1 1 1 1 2 2 1 1 1 1 1
pattern    : 8 0 0 5 5 4 4 0
counts     : 5 5 3 0
```

Table 3: Minimal sufficient statistics for the neanderthal, human and chimpanzee data. Site patterns and their counts from a multiple sequence alignment of the whole mitochondrial genome shotgun sequence (gi|115069275) of a neanderthal fossil Vi-80, from Vindija cave, Croatia [33], and its homologous sequence in a human (gi|13273200) and a chimpanzee (gi|1262390). The first column, $(\texttt{01.aaa.685})^T$, expresses that there are `685` sites with nucleotide `a` in all three species, ..., and the fifteenth column, $(\texttt{15.agg.1})^T$, expresses that there is `1` site with nucleotide `g` in human and chimpanzee and nucleotide `a` in neanderthal.

perfectly independent samples from the posterior and account for the finite number of neanderthal sites that were successfully sequenced, unlike those obtained on the basis of a bootstrap of site patterns [34] or heuristic MCMC [1]. Unfortunately, our human-neanderthal divergence estimates are overestimates as they ignore the non-negligible time to coalescence of the human and neanderthal homologs within the human-neanderthal ancestral population. Improvements to our estimates based on the other 310 human and 4 chimpanzee homologs reported in [33] may be possible with more sophisticated models of populations within a phylogeny and needs further investigation.

Chimpanzee, gorilla, orangutan and gibbon

We were able to draw samples from JC quartets on the basis of the mitochondrial DNA of chimpanzee, gorilla, orangutan and gibbon [27]. The data for all four primates can be summarized by 61 distinct site patterns [5]. Now, the problem is more challenging because there are three distinct tree topologies in the unrooted, bifurcating, quartet tree space, and each of these topologies has five edges. Thus, the domain of quartets is a piecewise Euclidean space that arises from a fusion of 3 distinct five dimensional orthants. Since the post-order traversals (Algorithm 2) specifying the likelihood function are topology-specific, we extended the likelihood over a compact box of quartets in a topology-specific manner. The computational time was about a day and a half to draw 10000 samples from the quartet target due to low acceptance probability of the naive likelihood function based on the 61 distinct site patterns. All the samples had the topology which grouped Chimp and Gorilla together, i.e. ((chimpanzee, gorilla), (orangutan, gibbon)). The

samples (results not shown) were again scattered about the verified global MLE of the quartet [5]. This quartet likelihood function has an elaborate DAG with numerous operations. When the data got compressed into sufficient statistics, the efficiency increased tremendously (e.g. for triplets the efficiency increases by a factor of 3.7). This is due to the number of leaf nodes in the target DAG, which encode the distinct site patterns of the observed data into the likelihood function, getting reduced from 29 to 5 for the triplet target and from 61 to 15 for the quartet target [24].

Discussion

Interval methods provide for a rigorous sampling from posterior target densities over small phylogenetic tree spaces. When one substitutes conventional floating-point arithmetic for real arithmetic in a computer and uses discrete lattices to construct the envelope and/or proposal, it is generally not possible to guarantee the envelope property, and thereby ensure that samples are drawn from the desired target density, except in special cases [35]. Thus, the construction of the Moore rejection sampler through interval methods, that enclose the target shape over the entire real continuum in any box of the domain with machine-representable bounds, in a manner that rigorously accounts for all sources of numerical errors (see [36] for a discussion on error control), naturally guarantees that the Moore rejection samples are independent draws from the desired target. Moreover, the target is allowed to be multivariate and/or non-log-concave with possibly 'pathological' behavior, as long as it has a well-defined interval extension.
The efficiency of MRS is not immune to the curse of dimensionality and target DAG complexity. When the DAG expression for the likelihood gets large, its natural interval extension can have terrible over-enclosures of the true range, which in turn forces the adaptive refinement of the domain to be extremely fine for efficient envelope construction. Thus, a naive application of interval methods to targets with large DAGs can be terribly inefficient. In such cases, sampler efficiency rather than rigor is the issue. Thus, one may fail to obtain samples in a reasonable time, rather than (as may happen with non-rigorous methods) produce samples from some unknown and undesired target.
There are several ways in which efficiency can be improved for such cases. First, the particular structure of the target DAG should be exploited to avoid any redundant computations. For example, sufficient statistics must be used to dissolve symmetries in the DAG. Second, we can further improve efficiency by limiting ourselves to differentiable targets in C^n. Tighter enclosures of the range of $f(\boldsymbol{t}^{(i)})$ with $\mathbf{f}(\boldsymbol{t}^{(i)})$ can come from the enclosures of Taylor expansions of f around the midpoint $\mathrm{mid}\,(\boldsymbol{t}^{(i)})$ through interval-extended automatic differentiation (e.g. [36]) that can then yield tighter estimates of the integral

enclosures [37]. Third, we can employ pre-processing to improve efficiency. For example, we can pre-enclose the range of a possibly rescaled f over a partition of the domain and then obtain the enclosure of $\mathbf{f}$ over some arbitrary $\boldsymbol{t}$ through a combination of hash access and hull operations on the pre-enclosures. Such a pre-enclosing technique reduces not only the overestimation of target shapes with large DAGs but also the computational cost incurred while performing interval operations with processors that are optimized for floating-point arithmetic. In the next version of the `MRS` library we plan to extend interval arithmetic beyond $\mathbb{IR}^n$ to a class of multi-dimensional data-structures related to regular sub-pavings (e.g. [38]) to improve the efficiency of our sampler. Fourth, various contractors can be used to improve the range enclosure in polynomial time (e.g. [38]). The most promising contractors employ interval constraint propagation. Finally, efficiency at the possible cost of rigor can also be gained (up to 30%) by foregoing directed rounding during envelope construction.

Poor sampler efficiency makes it currently impractical to sample from trees with five leaves and 15 topologies. However, one could use such triplets and quartets drawn from the posterior distribution to stochastically amalgamate and produce estimates of larger trees via fast amalgamating algorithms (e.g. [39, 40]), which may then be used to combat the slow mixing in MCMC methods [2] by providing a good set of initial trees. A collection of large trees obtained through such stochastic amalgamations would account for the effect of finite sample sizes (sequence length) as well as the sensitivity of the amalgamating algorithm itself to variation in the input vector of small tree estimates. It would be interesting to investigate if such stochastic amalgamations can help improve mixing of MCMC algorithms on large tree spaces, albeit auto-validating rejection sampling via the natural interval extension of the likelihood function may not be practical for trees with more than four leaves.

Conclusions

None of the currently available punctual samplers can rigorously produce independent and identically distributed samples from the posterior distribution over phylogenetic tree spaces, even for 3 or 4 taxa. We describe a new approach for rigorously drawing samples from a target posterior distribution over small phylogenetic tree spaces using the theory of *interval analysis*. Our Moore rejection sampler (MRS), being an auto-validating von Neumann rejection sampler (RS), can produce independent samples from any target shape f whose DAG expression f has a well-defined natural interval extension $\mathbf{f}$ over a compact domain $\mathbb{T}$. MRS is said to be auto-validating because it automatically obtains a proposal g that is easy to simulate from, and an envelope $\widehat{g}$ that is guaranteed to satisfy the envelope condition (1). MRS can circumvent the

problems associated with (i) heuristic convergence diagnostics in MCMC samplers and (ii) pseudo-envelopes constructed via non-rigorous punctual methods in rejection samplers. When the target DAG is large, MRS becomes inefficient and may fail to produce the desired samples in a reasonable time, rather than (as may happen with non-rigorous methods) produce samples from some unknown and undesired target. MRS solves the open problem of rigorously drawing independent and identically distributed samples from the posterior distribution over small rooted and unrooted phylogenetic tree spaces (3 or 4 taxa) based on any multiply-aligned sequence data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RS developed the basic algorithm, analyzed the data and wrote the first draft. TY improved the object-oriented interface and refined the final implementation of the algorithm. Both authors edited the manuscript.

Appendix

Likelihoods for the CFN model on unrooted triplets

Recall that the probability that `Y` mutates to `R`, or vice versa, in time t is $a(t) := (1 - e^{-2t})/2$ and the stationary distribution $\pi(\mathtt{R}) = \pi(\mathtt{Y}) = 1/2$. Next we apply Algorithm 2 to compute the likelihood $l_{d_{\bullet,q}}({}^4t)$ at a given site q which could be one of $l_0({}^4t), l_1({}^4t), \ldots, l_7({}^4t)$.

$$
\begin{aligned}
l_0({}^4t) &= \pi(\mathtt{R})P_{\mathtt{R},\mathtt{R}}({}^4t_1)P_{\mathtt{R},\mathtt{R}}({}^4t_2)P_{\mathtt{R},\mathtt{R}}({}^4t_3) + \pi(\mathtt{Y})P_{\mathtt{Y},\mathtt{R}}({}^4t_1)P_{\mathtt{Y},\mathtt{R}}({}^4t_2)P_{\mathtt{Y},\mathtt{R}}({}^4t_3) \\
&= \frac{1}{2}\left((1 - a({}^4t_1))(1 - a({}^4t_2))(1 - a({}^4t_3)) + a({}^4t_1)a({}^4t_2)a({}^4t_3)\right) \\
&= \frac{1}{16}\left((1 + e^{-2({}^4t_1)})(1 + e^{-2({}^4t_2)})(1 + e^{-2({}^4t_3)}) + (1 - e^{-2({}^4t_1)})(1 - e^{-2({}^4t_2)})(1 - e^{-2({}^4t_3)})\right) \\
&= \frac{1}{8}\left(1 + e^{-2({}^4t_1 + {}^4t_2)} + e^{-2({}^4t_2 + {}^4t_3)} + e^{-2({}^4t_1 + {}^4t_3)}\right) \qquad (19) \\
l_1({}^4t) &= \pi(\mathtt{R})P_{\mathtt{R},\mathtt{Y}}({}^4t_1)P_{\mathtt{R},\mathtt{Y}}({}^4t_2)P_{\mathtt{R},\mathtt{Y}}({}^4t_3) + \pi(\mathtt{Y})P_{\mathtt{Y},\mathtt{Y}}({}^4t_1)P_{\mathtt{Y},\mathtt{Y}}({}^4t_2)P_{\mathtt{Y},\mathtt{Y}}({}^4t_3) \\
&= \frac{1}{2}\left(a({}^4t_1)a({}^4t_2)a({}^4t_3) + (1 - a({}^4t_1))(1 - a({}^4t_2))(1 - a({}^4t_3))\right) \\
&= l_0({}^4t) \qquad (20)
\end{aligned}
$$

$$
\begin{aligned}
l_2({}^4t) &= \pi(\mathtt{R})P_{\mathtt{R},\mathtt{R}}({}^4t_1)P_{\mathtt{R},\mathtt{R}}({}^4t_2)P_{\mathtt{R},\mathtt{Y}}({}^4t_3) + \pi(\mathtt{Y})P_{\mathtt{Y},\mathtt{R}}({}^4t_1)P_{\mathtt{Y},\mathtt{R}}({}^4t_2)P_{\mathtt{Y},\mathtt{Y}}({}^4t_3) \\
&= \frac{1}{2}\left((1-a({}^4t_1))(1-a({}^4t_2))a({}^4t_3) + a({}^4t_1)a({}^4t_2)(1-a({}^4t_3))\right) \\
&= \frac{1}{16}\left((1+e^{-2({}^4t_1)})(1+e^{-2({}^4t_2)})(1-e^{-2({}^4t_3)}) + (1-e^{-2({}^4t_1)})(1-e^{-2({}^4t_2)})(1+e^{-2({}^4t_3)})\right) \\
&= \frac{1}{8}\left(1+e^{-2({}^4t_1+{}^4t_2)} - e^{-2({}^4t_2+{}^4t_3)} - e^{-2({}^4t_1+{}^4t_3)}\right) \qquad (21) \\
l_3({}^4t) &= \pi(\mathtt{R})P_{\mathtt{R},\mathtt{Y}}({}^4t_1)P_{\mathtt{R},\mathtt{Y}}({}^4t_2)P_{\mathtt{R},\mathtt{R}}({}^4t_3) + \pi(\mathtt{Y})P_{\mathtt{Y},\mathtt{Y}}({}^4t_1)P_{\mathtt{Y},\mathtt{Y}}({}^4t_2)P_{\mathtt{Y},\mathtt{R}}({}^4t_3) \\
&= \frac{1}{2}\left(a({}^4t_1)a({}^4t_2)(1-a({}^4t_3)) + (1-a({}^4t_1))(1-a({}^4t_2))a({}^4t_3)\right) \\
&= l_2({}^4t) \qquad (22)
\end{aligned}
$$

$$
\begin{aligned}
l_4({}^4t) &= \pi(\mathtt{R})P_{\mathtt{R},\mathtt{R}}({}^4t_1)P_{\mathtt{R},\mathtt{Y}}({}^4t_2)P_{\mathtt{R},\mathtt{Y}}({}^4t_3) + \pi(\mathtt{Y})P_{\mathtt{Y},\mathtt{R}}({}^4t_1)P_{\mathtt{Y},\mathtt{Y}}({}^4t_2)P_{\mathtt{Y},\mathtt{Y}}({}^4t_3) \\
&= \frac{1}{2}\left((1-a({}^4t_1))a({}^4t_2)a({}^4t_3) + a({}^4t_1)(1-a({}^4t_2))(1-a({}^4t_3))\right) \\
&= \frac{1}{16}\left((1+e^{-2({}^4t_1)})(1-e^{-2({}^4t_2)})(1-e^{-2({}^4t_3)}) + (1-e^{-2({}^4t_1)})(1+e^{-2({}^4t_2)})(1+e^{-2({}^4t_3)})\right) \\
&= \frac{1}{8}\left(1-e^{-2({}^4t_1+{}^4t_2)} + e^{-2({}^4t_2+{}^4t_3)} - e^{-2({}^4t_1+{}^4t_3)}\right) \qquad (23) \\
l_5({}^4t) &= \pi(\mathtt{R})P_{\mathtt{R},\mathtt{Y}}({}^4t_1)P_{\mathtt{R},\mathtt{R}}({}^4t_2)P_{\mathtt{R},\mathtt{R}}({}^4t_3) + \pi(\mathtt{Y})P_{\mathtt{Y},\mathtt{Y}}({}^4t_1)P_{\mathtt{Y},\mathtt{R}}({}^4t_2)P_{\mathtt{Y},\mathtt{R}}({}^4t_3) \\
&= \frac{1}{2}\left(a({}^4t_1)(1-a({}^4t_2))(1-a({}^4t_3)) + (1-a({}^4t_1))a({}^4t_2)a({}^4t_3)\right) \\
&= l_4({}^4t) \qquad (24)
\end{aligned}
$$

$$
\begin{aligned}
l_6({}^4t) &= \pi(\mathtt{R})P_{\mathtt{R},\mathtt{R}}({}^4t_1)P_{\mathtt{R},\mathtt{Y}}({}^4t_2)P_{\mathtt{R},\mathtt{R}}({}^4t_3) + \pi(\mathtt{Y})P_{\mathtt{Y},\mathtt{R}}({}^4t_1)P_{\mathtt{Y},\mathtt{Y}}({}^4t_2)P_{\mathtt{Y},\mathtt{R}}({}^4t_3) \\
&= \frac{1}{2}\left((1-a({}^4t_1))a({}^4t_2)(1-a({}^4t_3)) + a({}^4t_1)(1-a({}^4t_2))a({}^4t_3)\right) \\
&= \frac{1}{16}\left((1+e^{-2({}^4t_1)})(1-e^{-2({}^4t_2)})(1+e^{-2({}^4t_3)}) + (1-e^{-2({}^4t_1)})(1+e^{-2({}^4t_2)})(1-e^{-2({}^4t_3)})\right) \\
&= \frac{1}{8}\left(1-e^{-2({}^4t_1+{}^4t_2)} - e^{-2({}^4t_2+{}^4t_3)} + e^{-2({}^4t_1+{}^4t_3)}\right) \qquad (25) \\
l_7({}^4t) &= \pi(\mathtt{R})P_{\mathtt{R},\mathtt{Y}}({}^4t_1)P_{\mathtt{R},\mathtt{R}}({}^4t_2)P_{\mathtt{R},\mathtt{Y}}({}^4t_3) + \pi(\mathtt{Y})P_{\mathtt{Y},\mathtt{Y}}({}^4t_1)P_{\mathtt{Y},\mathtt{R}}({}^4t_2)P_{\mathtt{Y},\mathtt{Y}}({}^4t_3) \\
&= \frac{1}{2}\left(a({}^4t_1)(1-a({}^4t_2))a({}^4t_3) + (1-a({}^4t_1))a({}^4t_2)(1-a({}^4t_3))\right) \\
&= l_6({}^4t) \qquad (26)
\end{aligned}
$$

Proof of Theorem 1 (cf. [37])

Since any real arithmetic operation $x \star y$, where $\star \in \{+, -, \times, /\}$ and $x, y \in \mathbb{R}$, is a continuous function $x \star y := \star(x, y) : \mathbb{R} \otimes \mathbb{R} \mapsto \mathbb{R}$, except when $y = 0$ under / operation. Since $\boldsymbol{x}$ and $\boldsymbol{y}$ are simply connected compact intervals, so is their Cartesian product $\boldsymbol{x} \otimes \boldsymbol{y}$. On such a domain $\boldsymbol{x} \otimes \boldsymbol{y}$, the continuity of $\star(x, y)$

(except when $\star = /$ and $0 \in \boldsymbol{y}$) ensures the attainment of a minimum, a maximum and all intermediate values. Therefore, with the exception of the case when $\star = /$ and $0 \in \boldsymbol{y}$, the range $\boldsymbol{x} \star \boldsymbol{y}$ has an interval form $[\min(x \star y), \max(x \star y)]$, where the min and max are taken over all pairs $(x, y) \in \boldsymbol{x} \otimes \boldsymbol{y}$. Fortunately, we do not have to evaluate $x \star y$ over every $(x, y) \in \boldsymbol{x} \otimes \boldsymbol{y}$ to find the global min and global max of $\star(x, y)$ over $\boldsymbol{x} \otimes \boldsymbol{y}$, because the monotonicity of the $\star(x, y^*)$ in terms of $x \in \boldsymbol{x}$ for any fixed $y^* \in \boldsymbol{y}$ implies that the extremal values are attained on the boundary of $\boldsymbol{x} \otimes \boldsymbol{y}$, i.e. the set $\{\underline{x}, \underline{y}, \overline{x}, \text{ and } \overline{y}\}$. Thus the theorem can be verified by examining the finitely many boundary cases. □

Proof of Theorem 2

$$\boldsymbol{x} \star \boldsymbol{y} = \{x \star y : x \in \boldsymbol{x}, y \in \boldsymbol{y}\} \subseteq \{x \star y : x \in \boldsymbol{x}', y \in \boldsymbol{y}'\} = \boldsymbol{x}' \star \boldsymbol{y}'. \square$$

Proof of Theorem 3 (cf. [37])

Since $\mathbf{f}(\boldsymbol{y})$ is well-defined, we will not run into division by zero, and therefore (i) follows from the repeated invocation of Theorem 2. We can prove (ii) by contradiction. Suppose $\text{range}(f; \boldsymbol{x}) \nsubseteq \mathbf{f}(\boldsymbol{x})$. Then there exists $x \in \boldsymbol{x}$, such that $f(x) \in \text{range}(f; \boldsymbol{x})$ but $f(x) \notin \mathbf{f}(\boldsymbol{x})$. This in turn implies that $f(x) = \mathbf{f}([x, x]) \notin \mathbf{f}(\boldsymbol{x})$, which contradicts (i). Therefore, our supposition cannot be true and we have proved (ii) $\text{range}(f; \boldsymbol{x}) \subseteq \mathbf{f}(\boldsymbol{x})$. □

Proof of Theorem 4 (cf. [37])

Any elementary function $f \in \mathfrak{E}$ with expression f is defined by the recursion 9 on its sub-expressions f_i where $i \in \{1, \ldots, n\}$ according to its DAG. If $f(x) = p(x)/q(x)$ is a rational function, then the theorem already holds by Theorem 3, and if $f \in \mathfrak{S}$ then the theorem holds because the range enclosure is exact for standard functions. Thus it suffices to show that if the theorem holds for $\mathsf{f}_1, \mathsf{f}_2 \in \mathfrak{E}$, then the theorem also holds for $\mathsf{f}_1 \star \mathsf{f}_2$, where $\star \in \{+, -, /, \times, \circ\}$. By $\circ$ we mean the composition operator. Since the proof is analogous for all five operators, we only focus on the $\circ$ operator. Since f is well-defined on its domain $\boldsymbol{y}$, neither the real-valued f nor any of its sub-expressions f_i has singularities in its respective domain $\boldsymbol{y}_i$ induced by $\boldsymbol{y}$. In particular f_2 is continuous on any $\boldsymbol{x}_2$ and $\boldsymbol{x}_2'$ such that $\boldsymbol{x}_2 \subseteq \boldsymbol{x}_2' \subseteq \boldsymbol{y}_2$ implying the compactness of $\mathbf{f}_2(\boldsymbol{x}_2) =: \boldsymbol{w}_2$ and $\mathbf{f}_2(\boldsymbol{x}_2') =: \boldsymbol{w}_2'$, respectively. By our assumption that $\mathbf{f}_1$ and $\mathbf{f}_2$ are inclusion isotonic we have that $\boldsymbol{w}_2 \subseteq \boldsymbol{w}_2'$ and also that

$$\mathbf{f}_1 \circ \mathbf{f}_2(\boldsymbol{x}_2) = \mathbf{f}_1(\mathbf{f}_2(\boldsymbol{x}_2)) = \mathbf{f}_1(\boldsymbol{w}_2) \subseteq \mathbf{f}_1(\boldsymbol{w}_2') = \mathbf{f}_1(\mathbf{f}_2(\boldsymbol{x}_2')) = \mathbf{f}_1 \circ \mathbf{f}_2(\boldsymbol{x}_2')$$

The range enclosure is a consequence of inclusion isotony by an argument identical to that given in the proof for Theorem 3. □

Proof of Theorem 5 (cf. [37])

The proof is given by an induction on the DAG for f similar to the proof of Theorem 4 (See [37]). □

Proof of Theorem 6

Let the domain $\mathbb{T}$ of the target $f^{\cdot}$ be an element of $\mathbb{IR}^n$. From (15) and (14) observe that $\widehat{g}^{\mathfrak{T}}(t) = g^{\mathfrak{T}}(t)N_{\widehat{g}^{\mathfrak{T}}}$. Let us define the following two subsets of $\mathbb{R}^{n+1}$,

$$\mathcal{B}(\widehat{g}^{\mathfrak{T}}) = \{(v,u) : v \in \mathbb{T}, 0 \le u \le \widehat{g}^{\mathfrak{T}}(v)\}, \text{ and } \mathcal{B}(f) = \{(v,u) : v \in \mathbb{T}, 0 \le u \le f(v)\} \ .$$

Algorithm 1 first produces a sample from the random vector (V,U) that is uniformly distributed in $\mathcal{B}(\widehat{g}^{\mathfrak{T}})$. We can see this by letting $h(v,u)$ denote the joint density of (V,U) and $h(u|v)$ denote the conditional density of U given $V = v$. Then,

$$h(v,u) = \begin{cases} g^{\mathfrak{T}}(v)\, h(u|v) & \text{if } (v,u) \in \mathcal{B}(\widehat{g}^{\mathfrak{T}}) \\ 0 & \text{otherwise .} \end{cases}$$

Since we sample a height u for a given v from the Uniform$[0, \widehat{g}^{\mathfrak{T}}(v)]$ distribution,

$$h(u|v) = \begin{cases} (\widehat{g}^{\mathfrak{T}}(v))^{-1} = (g^{\mathfrak{T}}(v)N_{\widehat{g}^{\mathfrak{T}}})^{-1} & \text{if } u \in [0, \widehat{g}^{\mathfrak{T}}(v)] \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$h(v,u) = \begin{cases} g^{\mathfrak{T}}(v)\, h(u|v) = g^{\mathfrak{T}}(v)(g^{\mathfrak{T}}(v)\, N_{\widehat{g}^{\mathfrak{T}}})^{-1} = (N_{\widehat{g}^{\mathfrak{T}}})^{-1} & \text{if } (v,u) \in \mathcal{B}(\widehat{g}^{\mathfrak{T}}) \\ 0 & \text{otherwise .} \end{cases}$$

Thus, we have shown that the joint density of the random vector (V,U) initially produced by Algorithm 1 is a uniformly distributed on $\mathcal{B}(\widehat{g}^{\mathfrak{T}})$. The above relationship also makes geometric sense since the volume of $\mathcal{B}(\widehat{g}^{\mathfrak{T}})$ is exactly $N_{\widehat{g}^{\mathfrak{T}}}$.

Now, let (T,S) be the accepted random vector during the accept/reject step of Algorithm 1, i.e.

$$(T,S) = (V,U) \iff (V,U) \in \mathcal{B}(f) \subseteq \mathcal{B}(\widehat{g}^{\mathfrak{T}}) \ .$$

Then, the uniform distribution of (V,U) on $\mathcal{B}(\widehat{g}^{\mathfrak{T}})$ implies the uniform distribution of (T,S) on $\mathcal{B}(f)$. Since the volume of $\mathcal{B}(f)$ is N_f, the density of (T,S) is identically $1/N_f$ on $\mathcal{B}(f)$ and 0 elsewhere. Hence, the marginal density of T on $\mathbb{T}$ is

$$\begin{aligned} \int_0^{f(t)} 1/N_f \, dh &= 1/N_f \int_0^{f(t)} 1 \, dh \\ &= 1/N_f \int_0^{N_f f^{\cdot}(t)} 1 \, dh, \quad \because f^{\cdot}(t) = f(t)/N_f \\ &= f^{\cdot}(t) \ . \end{aligned}$$

Thus, we have shown that the accepted random vector T has the desired density $f^{\cdot}$. □

Proof of Theorem 7

Due to Theorem 5,

$$\begin{aligned} \mathrm{wid}\,(\boldsymbol{t}_W^{(i)}) = \mathcal{O}(1/W) \quad &\Longrightarrow \quad \mathrm{dist}_\infty(\mathrm{range}(f;\boldsymbol{t}_W^{(i)}), \mathbf{f}(\boldsymbol{t}_W^{(i)})\ ) = \mathcal{O}(1/W) \\ &\Longrightarrow \quad \mathrm{wid}\,(\mathbf{f}(\boldsymbol{t}_W^{(i)})) = \mathcal{O}(1/W), \qquad \because f \in \mathfrak{E}_{\mathfrak{L}}\ . \end{aligned}$$

Therefore

$$\sum_{i=1}^{|\mathfrak{U}_W|} \left(\mathrm{wid}\,(\boldsymbol{t}_W^{(i)}) \cdot \mathbf{f}(\boldsymbol{t}_W^{(i)}) \right) = w \sum_{i=1}^{W} \mathbf{f}\left([\ \underline{t} + (i-1)w,\ \underline{t} + iw\ ]\right)\ ,$$

and we have

$$\mathrm{wid}\,(w \textstyle\sum_{i=1}^{W} \mathbf{f}(\boldsymbol{t}_W^i)) = \mathcal{O}(1/W) \quad \Longrightarrow \quad \boldsymbol{A}(\mathfrak{U}_W) = 1 - \mathcal{O}(1/W)\ .$$

Therefore the lower bound for the acceptance probability $\boldsymbol{A}(\mathfrak{U}_W)$ of MRS approaches 1 no slower than linearly with the refinement of $\mathbb{T}$ by $\mathfrak{U}_W$. Note that this should hold for a general nonuniform partition with w replaced by the mesh.□

Acknowledgments

R.S. is a Research Fellow of the Royal Commission for the Exhibition of 1851. This was partly supported by a joint NSF/NIGMS grant DMS-02-01037. Many thanks to Rob Strawderman, Warwick Tucker and Stephane Aris-Brosou for constructive comments, Joe Felsenstein for clarifying the transition probabilities under the HKY model and Ziheng Yang for various clarifications and encouragement.

References

1. Jones G, Hobert J: **Honest exploration of intractable probability distributions via Markov chain Monte Carlo**. *Statistical Science* 2001, **16**(4):312–334.
2. Mossel E, Vigoda E: **Phylogenetic MCMC algorithms are misleading on mixtures of trees**. *Science* 2005, **309**:2207–2209.
3. von Neumann J: **Various techniques used in connection with random digits**. In *John Von Neumann, Collected Works, Volume V*, Oxford University Press 1963.
4. Walker A: **An efficient method for generating discrete random variables with general distributions**. *ACM Trans on Mathematical Software* 1977, **3**:253–256.
5. Sainudiin R: **Machine interval experiments**. *pHd dissertation*, Cornell University, Ithaca, New York 2005.
6. Moore R: *Interval analysis*. Prentice-Hall 1967.
7. Semple C, Steel M: *Phylogenetics*. Oxford University Press 2003.
8. Felsenstein J: *Inferring phylogenies*. Sunderland, MA: Sinauer Associates 2003.
9. Yang Z: *Computational molecular evolution*. UK: Oxford University Press 2006.
10. Moore R: *Methods and applications of interval analysis*. Philadelphia, Pennsylvania: SIAM 1979.
11. Alefeld G, Herzberger J: *An introduction to interval computations*. Academic press 1983.
12. Hammer R, Hocks M, Kulisch U, Ratz D: *C++ toolbox for verified computing: basic numerical problems*. Springer-Verlag 1995.

13. Kulisch U, Lohner R, Facius A (Eds): *Perspectives on encolsure methods*. Springer-Verlag 2001.

14. Matsumoto M, Nishimura T: **Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator.** *ACM Trans. Model. Comput. Simul.* 1998, **8**:3–30.

15. Williams D: *Weighing the Odds: A Course in Probability and Statistics*. Cambridge University Press 2001.

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

17. Yang Z: **Complexity of the simplest phylogenetic estimation problem**. *Proceedings Royal Soc. London B Biol. Sci.* 2000, **267**:109–119.

18. Evans W, Kenyon C, Peres Y, Schulman L: **Broadcasting on trees and the Ising model**. *Advances in Applied Probability* 2000, **10**:410–433.

19. Neyman J: **Molecular studies of evolution: a source of novel statistical problems**. In *Statistical decision theory and related topics*. Edited by Gupta S, Yackel J, New York Academy Press 1971:1–27.

20. Jukes T, Cantor C: **Evolution of protein molecules**. In *Mammalian Protein Metabolism*. Edited by Munro H, New York Academic Press 1969:21–32.

21. Saitou N: **Property and efficiency of the maximum likelihood method for molecular phylogeny**. *Jnl. Mol. Evol.* 1988, **27**:261–273.

22. Yang Z: **Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods**. *Syst. Biol.* 1994, **43**:329–342.

23. Hosten S, Khetan A, Sturmfels B: **Solving the likelihood equations**. *Found. Comput. Math.* 2005, **5**(4):389–407.

24. Casanellas M, Garcia L, Sullivant S: **Catalog of small trees**. In *Algebraic statistics for computational biology*. Edited by Pachter L, Sturmfels B, Cambridge University Press 2005:291–304.

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

26. Sainudiin R, Yoshida R: **Applications of interval methods to phylogenetic trees**. In *Algebraic statistics for computational biology*. Edited by Pachter L, Sturmfels B, Cambridge University Press 2005:359–374.

27. Brown W, Prager E, Wang A, Wilson A: **Mitochondrial DNA sequences of primates, tempo and mode of evolution**. *Jnl. Mol. Evol.* 1982, **18**:225–239.

28. Marsaglia G: **Generating discrete random numbers in a computer**. *Comm ACM* 1963, **6**:37–38.

29. Galassi M, Davies J, Theiler J, Gough B, Jungman G, Booth M, Rossi F: *GNU Scientific Library Reference Manual - 2nd Ed.* Network Theory Ltd. 2003, [http://www.gnu.org/software/gsl/].

30. Hofschuster, Krämer: **C-XSC 2.0: A `C++` library for extended scientific computing**. In *Numerical software with result verification*, *Volume 2991 of* Lecture notes in computer science. Edited by Alt R, Frommer A, Kearfott R, Luther W, Springer-Verlag 2004:15–35.

31. Rannala B, Yang Z: **Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference**. *Jnl. Mol. Evol.* 1996, **43**:304–311.

32. Yang Z, Rannala B: **Branch-length prior influences Bayesian posterior probability of phylogeny**. *Syst. Biol.* 2005, **54**:455–470.

33. Green R, Krause J, Ptak S, Briggs A, Ronan M, Simons J, Du L, Egholm M, Rothberg J, Paunovic M, Pääbo S: **Analysis of one million base pairs of Neandertal DNA**. *Nature* 2006, **444**:330–336.

34. Efron B, Halloran E, Holmes S: **Bootstrap confidence levels for phylogenetic trees**. *Proc. Natl. Acad. Sci.* 1996, **93**:13429–13429.

35. Gilks W, Wild P: **Adaptive rejection sampling for Gibbs sampling**. *Applied Statistics* 1992, **41**:337–348.

36. Kulisch U: **Advanced arithmetic for the digital computer, interval arithmetic revisited**. In *Perspectives on encolsure methods*. Edited by Kulisch U, Lohner R, Facius A, Springer-Verlag 2001:50–70.

37. Tucker W: **Auto-validating numerical methods**. Lecture notes, Uppsala University 2004.

38. Jaulin L, Kieffer M, Didrit O, Walter E: *Applied interval analysis: with examples in parameter and state estimation, robust control and robotics*. Springer-Verlag 2004.

39. Strimmer K, von Haeseler A: **Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies**. *Mol. Biol. Evol.* 1996, **13**:964–969.

40. Levy D, Yoshida R, Pachter L: **Beyond pairwise distances: neighbor joining with phylogenetic diversity estimates**. *Mol. Biol. Evol.* 2006, **23**:491–498.