BIOINFORMATICS

Vol. 30 ECCB 2014, pages i534–i540
doi:10.1093/bioinformatics/btu461

Point estimates in phylogenetic reconstructions

Philipp Benner[1,*], Miroslav Bačák[1] and Pierre-Yves Bourguignon[1,2]

[1]Max-Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany and [2]Isthmus SARL, 75002 Paris, France

ABSTRACT

Motivation: The construction of statistics for summarizing posterior samples returned by a Bayesian phylogenetic study has so far been hindered by the poor geometric insights available into the space of phylogenetic trees, and *ad hoc* methods such as the derivation of a consensus tree makeup for the ill-definition of the usual concepts of posterior mean, while bootstrap methods mitigate the absence of a sound concept of variance. Yielding satisfactory results with sufficiently concentrated posterior distributions, such methods fall short of providing a faithful summary of posterior distributions if the data do not offer compelling evidence for a single topology.

Results: Building upon previous work of Billera *et al*., summary statistics such as sample mean, median and variance are defined as the geometric median, Fréchet mean and variance, respectively. Their computation is enabled by recently published works, and embeds an algorithm for computing shortest paths in the space of trees. Studying the phylogeny of a set of plants, where several tree topologies occur in the posterior sample, the posterior mean balances correctly the contributions from the different topologies, where a consensus tree would be biased. Comparisons of the posterior mean, median and consensus trees with the ground truth using simulated data also reveals the benefits of a sound averaging method when reconstructing phylogenetic trees.

Availability and implementation: We provide two independent implementations of the algorithm for computing Fréchet means, geometric medians and variances in the space of phylogenetic trees. TFBayes: https://github.com/pbenner/tfbayes, TrAP: https://github.com/bacak/TrAP.

Contact: philipp.benner@mis.mpg.de

1 INTRODUCTION

Phylogenetic trees are central to the study of evolution, so much that the sketch of a tree of species by Sir Charles Darwin has become the icon of this theory. Nowadays, trees relating units of selection (be it functional domains, genes or species) are structures of primary interest for systematists, but also instrumental to a wealth of other studies where evolutionary correlations need to be accounted for [see, for instance, McCue *et al*. (2001)]. Various statistical models pertaining to diverse types of observables can be found in the literature, as well as methods, for estimating their parameters and reconstructing a phylogenetic tree (Gascuel, 2005). Some estimation methods proceed by maximizing a posterior distribution or a likelihood function, and are amenable to an exact reconstruction of the optimal tree, but Bayesian phylogenetic analyses generally produce posterior distributions that are best explored by generating posterior samples. While a large enough posterior sample offers a faithful representation of the posterior knowledge, it is of little scientific interest unless summarized by some statistics (Robert, 2001). A summary can balance contributions from the different tree topologies occurring in the sample, resulting in a legit phylogenetic tree, or combine them within a phylogenetic network. Here we focus on the former, showing how to build a phylogenetic tree that faithfully represents the sample in its entirety, despite competing topologies occur.

Provided a unique topology with n edges occurs in the sample, each tree including its edge lengths can be identified by a point in the positive orthant of the Euclidean space $\mathbb{R}^n$. Performing an average of the sample in this linear representation is a straightforward operation, which produces a legit posterior mean tree. If more than one tree topology occurs, the trees are no longer mapped all to the same linear space, and the posterior mean is ill-defined. Selecting the a posteriori most probable tree topology may seem a sound alternative, however, with the unpleasant consequence of neglecting all the sampled trees of different topology, and therefore would not provide a satisfactory representation of the posterior. The construction of a consensus tree, using an absolute majority rule (Margush and McMorris, 1981) to decide which one among competing edges should be retained, has been widely adopted by the interested community as the method of choice to summarize posterior samples of phylogenetic trees. On the theoretical side, decision-theoretic justifications of this construction have been proposed (Holder *et al*., 2003; Huggins *et al*., 2011). However, they are built upon loss functions that neglect edge lengths, focusing only on the tree topology. Besides, from the authors' point of view, it is also a rather conservative approach, as the absence of an absolute majority among edges results in the inclusion of none of them, thereby producing unresolved branching points. The extended majority-rule consensus method (also known as greedy consensus method) has been introduced to remedy this drawback by adding edges with $<50\%$ support (Bryant, 2003). Despite this improvement, the consensus methods neglect much of the available information in a sample by ignoring the context in which an edge occurs (i.e. the remaining topology of the tree as well as all other edge lengths). Reporting a posterior mean that balances the contributions from each topology including edge lengths rather than isolated edges would therefore be of utmost interest.

Building upon the work published by Billera *et al*. (2001), who deciphered the geometric structure of the space of phygenetic trees and first proposed a construction of the *tree space* (sometimes also called *BHV tree space*, where BHV is an acronym of Billera, Holmes and Vogtmann), the purpose of this article is to show how the computation of the posterior mean of a sample of phylogenetic trees can be achieved by simply reaching out for the appropriate geometry. The BHV space is obtained by gluing together the positive orthants of the linear space associated to

*To whom correspondence should be addressed

each topology, so that a point in this space identifies both a tree topology and the lengths of the corresponding edges. The adjacency structure between any two orthants reflects the edges shared by the two corresponding topologies and permits the definition of paths visiting several orthants. Any two trees are connected by at least one path, and the one with minimal length is called a *geodesic*. Therefore, the length of a geodesic qualifies as a distance function between phylogenetic trees, and offers a theoretically and practically appealing alternative to existing distances (e.g. NNI or the Robinson–Foulds distance). Furthermore, using implicit characterizations of the posterior mean and median as minimizers of appropriate loss functions, algorithms developed by Bačák (2013, 2014); Miller *et al.* (2012); Sturm (2003) compute an approximation of these statistics by walking along geodesics. Here, the determination of the geodesics is done in polynomial time, thanks to an algorithm owing to Owen and Provan (2011).

This article gathers and combines technical results scattered across multiple mathematics papers, into a general statistical framework for analyzing posterior distributions over phylogenetic trees easily accessible to the target readership, namely practitioners in Bayesian phylogenetics. Certainly, many other applications exist where computing an average phylogenetic tree is of great importance. The methods presented here are not restricted to Bayesian statistics, yet in this context, they allow to recover basic statistical concepts such as the posterior mean and median, as well as a notion of variance to measure the posterior uncertainty.

After a gentle introduction to the geometry of the tree space in Section 2.1, the geometric median and Fréchet mean over this space are constructed in Section 2.2. The algorithms computing those quantities are outlined in Section 2.3, while Section 3 shows the method in action, and illustrates how it compares with the majority-rule consensus method.

2 METHODS

Given a generative model and a prior distribution over its parameter space, a Bayesian analysis of observations carried across species related by evolution produces a posterior distribution over the space of all possible phylogenetic trees for this set of species (Gascuel, 2005; Robert and Casella, 1999). The size of this space grows super-exponentially with the number of species, and it is often intractable to compute the normalization constant of this distribution. In such cases, sampling methods offer a way to explore the posterior distribution via an arbitrarily large sample drawn from it without requiring any further knowledge. However, although a posterior sample offers a representation of the full posterior distribution, it is of little scientific interest in absence of a method to summarize it. Building upon previous works by Billera *et al.* (2001), Miller *et al.* (2012), Owen and Provan (2011) and Bačák (2013, 2014), we propose here to define and compute posterior mean in a sound manner, an approach so far hindered by the poor geometrical insights into the space of phylogenetic trees (see also Gascuel, 2005).

2.1 The geometry of the tree space

The elucidation of the geometric structure of the space of phylogenetic trees is because of Billera, Holmes and Vogtmann (2001). For any integer $n \geq 3$, a *phylogenetic n-tree* is a connected graph without cycle that has $n + 1$ terminal vertices called *leaves*, which are labeled from 0 to n and associated with the $n + 1$ species considered in the phylogenetic study.

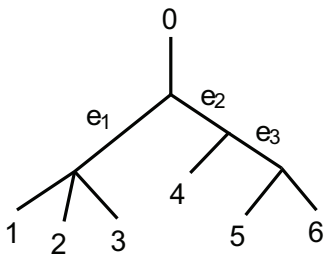

Fig. 1. An example of a 6-tree

The non-terminal vertices of a tree bear no label, as they are seen as sheer 'branching points'. An example of a 6-tree is shown in Figure 1.

The construction of the tree space relies fundamentally on the identification of edges of the tree as *splits*: an edge is uniquely associated to the partition of the set of leaves $L = \{0, \ldots, n\}$ into two disjoint and non-empty subsets $L_1 \cup L_2 = L$ that would be disconnected in the graph structure by its removal. Such a split is denoted by $L_1|L_2$. For instance, the edges labeled e_1, e_2 and e_3 of the tree in Figure 1 correspond to the splits $(0, 4, 5, 6|1, 2, 3)$, $(0, 1, 2, 3|4, 5, 6)$ and $(0, 1, 2, 3, 4|5, 6)$, respectively. Conversely, given a set of leaves and splits subjects to certain conditions, a unique tree is specified. Namely, it is required that any two splits $L_1|L_2$ and $L_1'|L_2'$ are *compatible*, that is, one of the sets

$$L_1 \cap L_2',\ L_1' \cap L_2,\ L_1 \cap L_1',\ L_2 \cap L_2'$$

must be empty.

The topology of a phylogenetic tree t can therefore be uniquely specified via the associated set of compatible splits, which is denoted by $S(t)$. Yet, a phylogenetic tree is more than a sheer topology, as its edges also have (positive) length. Writing $|e|_t$ for the length of the edge $e = L_1|L_2$ of t, one obtains a canonical mapping of the phylogenetic tree t onto $\mathbb{R}_+^{2^n-1}$ by further setting $|e'|_t = 0$ whenever $e' = L_1'|L_2' \notin S(t)$.

While any phylogenetic tree over a given set of species can be represented in this way in the linear cone $\mathbb{R}_+^{2^n-1}$, the converse is obviously not true: Assume that e and e' are two incompatible edges, any coordinates in the linear cone that have positive entries for e and e' do not correspond to any legit phylogenetic tree. In other words, the set of phylogenetic trees forms a manifold in the linear cone $\mathbb{R}_+^{2^n-1}$. All the phylogenetic trees t sharing a given topology also exhibit the same split set S, and are such that $|e|_t = 0$ whenever $e \notin S$. The BHV tree space can therefore be understood as a collection of smaller dimensional positive orthants embedded jointly in $\mathbb{R}^{2^n-1}$, each associated to a particular tree topology. Among these orthants, those of maximal dimension play a special role: a tree whose representation lies in their interior is binary, as such trees have the maximal possible number of edges. Shrinking the length of any edge down to zero results in the formation of a triple branching-point, so that the orthants associated to non-binary tree topologies appear as the faces of larger dimensional orthants.

More interestingly, non-maximal orthants (those associated to non-binary tree topologies) are faces of several orthants simultaneously. The simplest instance is a triple-branching point, from which three different edges can be grown depending on which pair of species diverged first. As these three edges cannot be compatible, these three departures from the triple-branching point lie in different orthants, which are therefore all adjacent. Figure 2 shows a section of $\mathcal{T}_4$, where every tree in the interior of the orthant is a binary tree. For instance, the point $(0, 1/2)$ may be reached by taking a tree from the interior and shrinking the length of the edge e_1 to zero. This location corresponds to a non-binary tree that lies at the face of three maximal orthants.

As an example, take the space $\mathcal{T}_3$, which consists of all trees with only four leaves. Binary trees in this space have a single inner edge identified as one of the splits $(0, 1|2, 3)$, $(0, 2|1, 3)$ or $(0, 3|1, 2)$. An orthant $[0, \infty)$ is associated with each of the splits. The origin 0 is a face of each orthant,

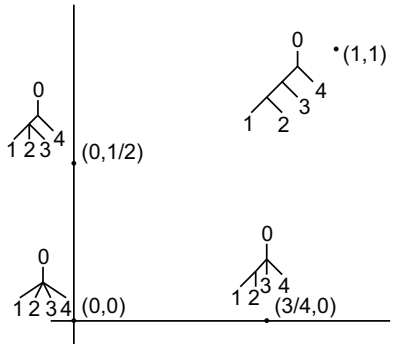

Fig. 2. 4-trees of a given combinatorial structure. The horizontal direction shows the length of $e_1 : (0, 3, 4|1, 2)$, whereas the vertical direction shows $e_2 : (0, 4|1, 2, 3)$

which represents the same non-binary tree. A piece of $\mathcal{T}_3$ is constructed by gluing all three orthants together at this common point.

The tree space, as a compound of orthants, is not a convex subset of $\mathbb{R}_+^{2^n-1}$: while one can form a linear interpolation between two trees of different topologies in the embedding space $\mathbb{R}_+^{2^n-1}$, by simply shrinking the edges to be removed and simultaneously growing those to be created, the cooccurrence of incompatible edges along this path places it outside of the tree space. However, as seen above, the orthants composing the tree space have a rich adjacency structure, which guarantees at least the path-connectedness of the space: any two phylogenetic trees can be connected by at least one path that remains in this space, although this path might not be a straight line as in $\mathbb{R}_+^{2^n-1}$ whenever the topologies differ. Yet, these paths have a length, and one can consider the shortest path connecting two trees. In geometric terms, such a path is called a *geodesic*. Defining a distance over the tree space as the length of the shortest path connecting any two points equips the space with a metric $d(\cdot, \cdot)$ which, as an alternative to the NNI or Robinson–Foulds distances, allows to measure the discrepancy between any two trees, whatever their topologies. Although the technical details are not important for this article, it should be noted that the BHV tree space is a Hadamard space (Billera *et al.*, 2001), which allows to use tools from this mathematical field.

2.2 Mean, median and variance of a sample of phylogenetic trees

Let us now exploit the above geometric properties of the tree space to summarize a sample of phylogenetic trees by a single point. Following the rationale of decision theory (e.g. Robert, 2001), the construction begins with the definition of a *loss function*, which measures how faithful a representation of the sample would be achieved by a given point in the tree space. A loss function is defined as the cost $\mathcal{L}(t, t')$ of choosing a phylogenetic tree t' instead of some other t, and the decision theory literature advocates strongly to summarize a posterior distribution by choosing $\hat{t}$ as the minimizer of the expected loss function

$$\hat{t} = \operatorname*{arg\,min}_{t' \in \mathcal{T}_n} \int_{\mathcal{T}_n} \mathcal{L}(t, t') \mu_{t|X} \mathrm{d}t \,, \tag{1}$$

where $\mu_{t|X}$ denotes the posterior distribution over phylogenetic trees given the data X. Approximating the latter via a posterior sample of phylogenetic trees $t_1, \ldots, t_K$, the above formula becomes

$$\hat{t} = \operatorname*{arg\,min}_{t' \in \mathcal{T}_n} \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(t_i, t') \,.$$

Two very typical choices for the loss function are the distance and the squared distance. When the parameters to be estimated lie in a Euclidean space, it is well-known that the resulting estimates coincide respectively with the median and mean of the posterior distribution. Although the tree space is not Euclidean, distances between pairs of trees are well defined, and a minimizer of (Equation 1) can be sought, respectively yielding the so-called *geometric median* and *Fréchet mean*.

In contrast to other approaches that provide a decision-theoretic argument for point estimates of phylogenetic trees (e.g. Holder *et al.*, 2003; Huggins *et al.*, 2011), the loss function considered here derives the *intrinsic metric* of the underlying space. In particular, the loss function considers both the topology and the branch lengths of phylogenetic trees, as opposed to those supporting the consensus method, and thereby considers all available information in a sample. Unfortunately, in tree space, a simple gradient search is not a practical method to solve such optimization problems (see Miller *et al.*, 2012). Therefore, appropriate algorithms are required and will be presented in Section 2.3.

A side benefit of the method presented here is the sound definition of the sample variance, also called the Fréchet variance, which is simply given as the value of the minimization problem with the squared distance. In complement to the point estimate, this quantity provides the modeler with insight onto the reliability of the point estimate. It is noteworthy that existing phylogenetic reconstruction methods are not tied to a notion of variance, and often retort to bootstrapping methods for reporting similar information.

2.3 Computing the sample mean, median and variance

The question of how to *compute* medians and means of a given set of trees will be addressed in the following. It turns out that efficient approximation methods from optimization theory can be extended into Hadamard spaces and applied to median and mean computations. For the reader's convenience, the simple version for unweighted medians and means is outlined here (see also Bačák, 2013).

2.3.1 Algorithm for computing medians Let us first describe the algorithm for computing a median of a given set of trees $\bar{t} = \{t_1, \ldots, t_K\}$ (i.e. the set of all tree samples).

Set $x_0 = t_1$ and suppose that at the i-th iteration, an approximation $x_i \in \mathcal{T}_n$ of the median of $\bar{t}$ is available. To find x_{i+1}, a tree t_k is selected from the set of trees $\bar{t}$ at random and x_{i+1} is defined as a point on the geodesic between x_i and t_k. (In other words, x_{i+1} is a convex combination of x_i and t_k, which will be denoted $(1 - \lambda)x_i + \lambda t_k$ for some $\lambda \in [0, 1]$.) The position of x_{i+1} on this geodesic is determined by a parameter $\eta_i \in [0, 1]$, which is computed at each iteration. By this procedure, we obtain a sequence of trees $x_1, x_2, \ldots$, which is known to converge to a median of $\bar{t}$.

Algorithm 2.1 (Computing median). *Let $x_0 = t_1$. At each step $i \in \mathbb{N}_0$, choose randomly $r_i \in \{1, \ldots, K\}$ according to the uniform distribution and put*

$$x_{i+1} = (1 - \eta_i)x_i + \eta_i t_{r_i},$$

with η_i defined by $\eta_i = \min\left\{1, \frac{1}{(i+1)d(t_{r_i}, x_i)}\right\}$, for each $i \in \mathbb{N}_0$, where d is the metric on $\mathcal{T}_n$.

2.3.2 Algorithm for computing means Computing the mean is similar to the computation of the median. As a matter of fact, it only differs in the coefficients determining the position of x_{i+1} on the geodesic from x_i to t_k.

Algorithm 2.2 (Computing mean). *Let $x_0 = t_1$ and, at each step, $i \in \mathbb{N}_0$, choose randomly $r_i \in \{1, \ldots, K\}$ according to the uniform distribution and put*

$$x_{i+1} = \frac{1}{i+1} x_i + \frac{i}{i+1} t_{r_i}.$$

The approximation algorithms for computing medians and means use (at each step) the algorithm for finding a geodesic in tree space by Owen and Provan (2011).

3 RESULTS

The content of this section is intended to illustrate the behavior of the posterior mean and median, in comparison with the majority-rule consensus tree, which is, for instance, computed by MrBayes (Huelsenbeck and Ronquist, 2001). Following a formal argument that relates these two estimates when extremely large or little information is available, the posterior distributions derived from real and simulated datasets are investigated in a way that best illustrates the different outcomes yielded by the existing and the proposed approaches. A prerequisite is the adoption of a specific statistical model, which is outlined first.

3.1 Consensus versus posterior mean

The majority-rule consensus method is a reference method to summarize samples from a posterior distribution. There, the consensus tree consists of those splits that occur in >50% of the samples. The average length of a retained edge is computed using the subsample where the corresponding split does occur, thereby neglecting a fraction of the posterior mass, but also the context in which the split occurs. In contrast, the Fréchet mean and geometric median account for the full posterior, and are expected to provide a more meaningful summary. However, both estimates have a property called stickiness (see Miller *et al.*, 2012): If there is a high posterior uncertainty on the topology, the Frechet mean and geometric median result in non-binary trees, a behavior that parallels the multiple branching points reconstructed by the consensus tree when no absolute majority occurs.

Take for instance the space $\mathcal{T}_3$ that consists of three orthants $[0, \infty)$ glued together at $\mathbf{0}$, as discussed earlier, and place a tree on each orthant. If all three trees are equally far apart, say at a distance r to the origin, then obviously the Fréchet mean lies at the origin. The term stickiness refers to the fact that the mean stays at the origin if one of the trees is moved further away. In fact, one tree may be located at a distance anywhere between r and $2r$ away from the origin without affecting the mean. Instead of moving one tree further away from the origin, one may similarly add another tree somewhere between the three trees, and the Fréchet mean would again stay at the origin.

A probabilistic counterpart of this phenomenon can be observed in the same setting. Equip $\mathcal{T}_3$ with a distribution whose trace in each orthant is a normalized Gaussian distribution, centered at identical distance from the origin, and truncated at $\mathbf{0}$. By symmetry the Fréchet mean is at the origin, and one can ask how far the location parameter m of one component can be perturbed without affecting the mean. In $\mathcal{T}_3$, m is just a scalar, and one can study the distance of the Fréchet mean $\hat{t}$ to the origin $\mathbf{0}$ as a function of m. An analytic but complicated solution of the distance $d(\mathbf{0}, \hat{t})$ exists; however, a fairly good answer is provided by the following approximation:

$$d(\mathbf{0}, \hat{t}) \approx \max\left\{0, \frac{m - \sqrt{2/(e\pi)} - 2\Phi(1)}{1 + 2\Phi(1)}\right\}, m \geq 1$$

where Φ is the standard normal distribution function. The Fréchet mean stays at the origin until m reaches ~2.16, which roughly matches the case of only three trees. Also in this case one may similarly increase the probability mass on one orthant and the Fréchet mean would stay at the origin until a certain threshold is reached.

3.2 Statistical model

The statistical model we use in this study is often used for the analyses of motifs of transcription factor binding sites. It became popular because of its analytic tractability (cf. Siddharthan *et al.*, 2005), and it is simple enough so that the marginal likelihood of an alignment, given a phylogenetic tree can be computed analytically. In particular, it permits Bayesian model selection, or to gain some insights on how well an estimate generalizes to new data, and therefore qualifies perfectly for the purpose of comparing different downstream methods for summarizing the posterior samples.

A dataset consists of an alignment of $n + 1$ homologous sequences. The observations within each column of the alignment are assumed to evolve according to a phylogenetic tree t. The topology of the phylogenetic tree is a priori uniformly distributed, and the edge lengths of t are drawn from a gamma distribution. The process of evolution is specified by a mutation model, which is parameterized by a column-specific stationary distribution Θ. It follows a priori a Dirichlet distributed with pseudocounts α. It is sufficient to discuss the model for a single column of the alignment. Binary trees have $2n - 1$ nodes, each of which is associated with a random variable $X^{(k)}$ that takes values in an alphabet $\mathbb{A}$. Assume that node k is the parent of node i. The mutation process along the edge between the two nodes is defined as

$$X^{(i)} | X^{(k)} = x, \Theta = \vartheta \sim p_{M_i}\text{Categorical}(\vartheta) + p_{\overline{M}_i}\delta_x,$$

which was introduced by Felsenstein (1981). The probability of a mutation from node k to node i is denoted p_{M_i} and depends on the length l of the edge that connects the two nodes, i.e. $p_{M_i} = 1 - \exp(-l)$. Furthermore, $p_{\overline{M}_i}$ denotes the probability of no mutation, given as $p_{\overline{M}_i} = 1 - p_{M_i}$.

To obtain samples from the posterior distribution, we implemented a Metropolis coupled MCMC algorithm (Geyer and Thompson, 1995) for the given model. In the sequel, these samples will serve as input to the reconstrution of the posterior mean, median, and consensus trees.

3.3 Estimation results

Using a multiple sequence alignment from a study by Karol *et al.* (2001), which was slightly modified by Yang and Rannala (2005), the phylogeny of the small subunit rRNA gene (SSU rRNA) from the nuclear genome of eight land plants and six charales (see Fig. 3) has been reconstructed. It appears that the edge that separates *Psilotum nudum* and *Dicksonia antarctica* from the remaining tree has a very short length of ~0.0027. Figure 4 shows the marginal posterior distribution of this edge (e_1) and a competing one (e_2) that groups *P.nudum* with *Taxus baccata* and *Arabidopsis thaliana*. There remains a high posterior uncertainty about the exact topology of the tree at this very

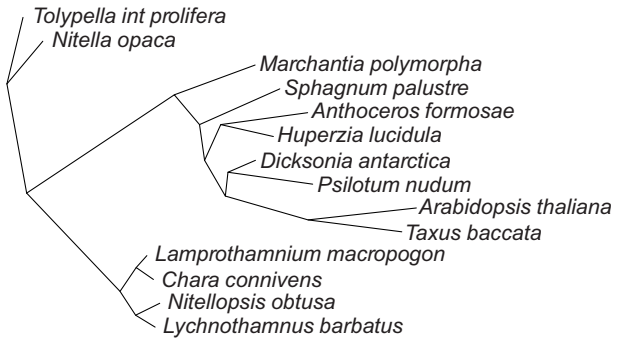

Fig. 3. Fréchet mean estimated from the small subunit rRNA gene (SSU rRNA) from the nuclear genome of eight land plants and six charales. Edge lengths are plotted in horizontal direction only

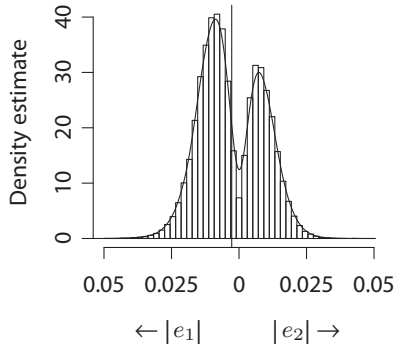

Fig. 4. Marginal posterior density estimate of two edges e_1 and e_2. The edge e_1 groups *P.nudum* with *D.antarctica*, while e_2 groups *P.nudum* with *T.baccata* and *A.thaliana*. The Fréchet mean is shown as a vertical line

branching. The posterior mass on e_1 is, however, sufficient for the majority-rule consensus tree to include this edge. However, there, it has a much longer length than in the posterior mean tree (by ~0.011) because this length is obtained by averaging only the lengths of this edge when it occurs in the sample, neglecting the contributions of the alternative edges. Clearly, the shorter edge length of the posterior mean tree better accounts for the uncertainty, and the consensus tree appears, in contrast, to have overestimated branch lengths.

The assessment of reconstruction methods for phylogenetic trees is notoriously hindered by the ignorance of the true evolutionary history to be uncovered, as the latter is never observed. Instead, the estimated tree of Figure 3 has been used to generate 50 alignments of length $m = 50$, 100, 250 and 500. For each generated dataset, 210 000 posterior samples were obtained using one cold Markov chain and three heated chains. The Fréchet mean, geometric median and consensus tree of the last 200 000 samples were computed. In the whole study, tree topologies are uniformly distributed a priori, while branch lengths are distributed according to a Gamma(1, 0.4).

Figure 5 shows the distances of the computed estimates to the generating tree. For alignments of the lengths considered, the Fréchet mean and geometric median are generally closer to the generating tree. A trend appears, from the greatest discrepancy observed for the shortest alignments, to an almost systematic agreement for the longest alignments. It should be noted that even shorter alignments generally result in very broad a posterior distributions so that all three estimates coincide with a star tree. At the opposite extreme, large datasets support a clear decision about the topology of the tree, placing most of the mass of the posterior distribution in a single orthant, and resulting in mostly agreeing estimates. One also observes that the geometric median is in most cases closer to the generating tree than the Fréchet mean. This comes at no surprise given the skewness of the gamma prior on the branch lengths.

The model permits an analytical computation of the marginal likelihood of an alignment given a phylogenetic tree, thereby offering an evaluation of how the estimated model generalizes to novel observations. Using a leave-one-out approach, the average (unnormalized) posterior value achieved by the estimators was computed on the remaining 49 datasets of the same length (see Table 1). For all alignment lengths, the Fréchet mean and

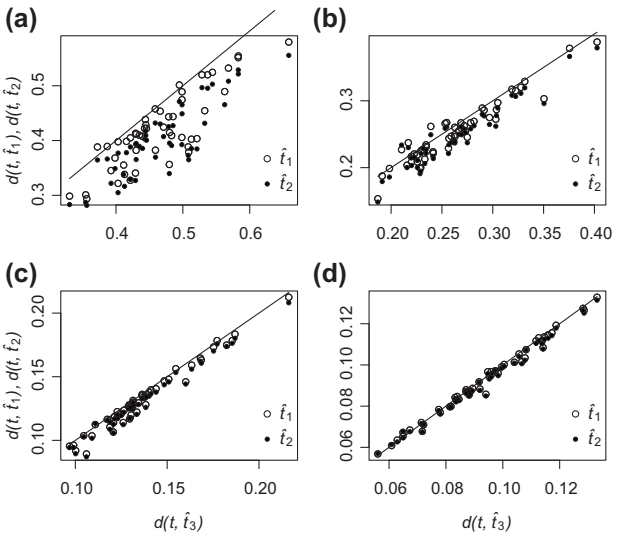

Fig. 5. Distances $d(\cdot, \cdot)$ of the Fréchet mean $\hat{t}_1$, geometric median $\hat{t}_2$ and consensus tree $\hat{t}_3$ to the generating tree t for alignments of length 50 (**a**), 100 (**b**), 250 (**c**) and 500 (**d**). The straight line shows the main diagonal

geometric median show a slightly higher average posterior value compared with the majority-rule consensus tree. But the difference is too minor to make any definite statements, as also shown by the variance of the estimates. Interestingly, however, while the variance of the mean and median steadily decreases with data length, the consensus estimate shows an increase in variance for a data length of 250. A much clearer picture is gained by considering how often the mean and median have a higher posterior value than the consensus (see Table 1). The results show that the consensus tree clearly perfomes worse.

Another quantity of interest is the Fréchet variance of the posterior distribution, which provides us with a measure of uncertainty. The mean variance is shown in Figure 6 separately for all four dataset lengths. Similar to the case of normal distributed i.i.d. random variables, the variance decreases approximately with $1/m$. Another, maybe more intuitive statistic, is to compute a credibility region around the Fréchet mean $\hat{t}$ that contains a given proportion c of the posterior mass. More precisely, consider the set of trees $B = \{t \in \mathcal{T}_n \mid d(\hat{t}, t) \leq d^*\}$ for some d^*

Table 1. (a) Mean posterior values for the Fréchet mean, geometric median and consensus tree (the variance is shown in brackets) and (b) percentage of times the mean and median show a higher posterior value on the remaining (joined) 49 datasets

	(a) Posterior			(b) Performance	
m	Mean	Median	Consensus	Mean	Median
50	–202.37 (11.23)	–201.76 (11.29)	–203.22 (10.10)	0.76	0.78
100	–409.30 (5.97)	–409.04 (5.88)	–409.39 (3.66)	0.66	0.64
250	–1035.66 (5.12)	–1035.62 (5.32)	–1036.72 (5.02)	0.90	0.90
500	–2074.11 (3.75)	–2074.01 (3.75)	–2074.96 (3.99)	0.78	0.80

Note: Both statistics were evaluated separately on datasets of length $m = 50$, 100, 250 and 500.

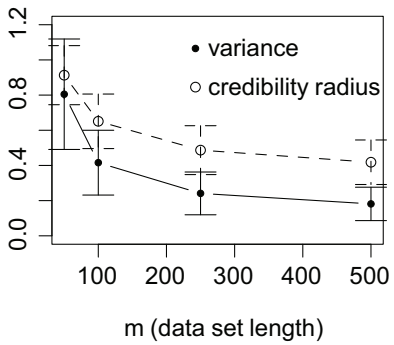

Fig. 6. Average Fréchet variance and credibility radius d^* for datasets of length 50, 100, 250 and 500. The error bars show 1 SD

such that $\int_B \mu_{t|X} \mathrm{d}t = c$. The bound d^* may be called the credibility radius. Figure 6 shows the results for $c = 0.68$.

4 DISCUSSION

By recognizing the global geometric nature of the space of phylogenetic trees, this article shows that the fundamental statistical notions defined over linear spaces, such as sample mean, median and variance, can be generalized to more complex spaces such as the tree space. Besides the sheer recovery of well-defined fundamental statistical quantities in the particular setting of phylogenetic studies, this study also demonstrates critical differences in the behavior of the posterior mean and the consensus tree.

The reconstruction of a consensus tree retains splits that occur in at least half of the samples. This absolute majority rule prevents the introduction of splits favored by sheer fluctuations, but also aims to maximize the information extracted from the sample. The length of the retained edges is indeed often simply set to the average length of their occurrences in the sample, so that the lengths of the discarded edges never enter the determination of the consensus tree. As illustrated on a real dataset by Figure 4, the neglection of a fraction of the sample results in biased estimates, where edge lengths are systematically overestimated from the perspective of the geometry of the tree space.

The extent of the bias born by the consensus tree is however tightly related to the concentration of the posterior distribution, which decreases the amount of information dropped in the reconstruction process, and the simulation-based study shown above confirms that the consensus tree and the posterior average disagree mostly when there exists no compelling evidence for a single topology. Illustrated on small datasets, the consensus tree appears dramatically further of the generating tree than the posterior mean, as a result of its neglection of a fraction of the information brought by the sample.

The proper definition of a variance for a sample of phylogenetic trees has consequences that should not be overlooked, and is believed by the authors to bear even more potential for applications. Not only is the reporting of the credibility of the Bayesian estimate made simple by this quantity, but it also opens the way to the generalization of variance-based studies of samples of phylogenetic trees, including principal components analysis, a task already tackled by Nye (2011). Measuring the spread of a set of trees is a useful tool not only to quantify the posterior uncertainty. For instance, in a recent study Salichos *et al.* (2014) developed an information theoretic measure to quantify the incongruence of gene trees.

Funding: This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 267087.

Conflict of Interest: none declared.

REFERENCES

Bačák,M. (2013) Computing medians and means in Hadamard spaces. *arXiv*, 1210.2145.

Bačák,M. (2014) *Convex Analysis and Optimization in Hadamard Spaces, volume 22 of De Gruyter Series in Nonlinear Analysis and Applications*. Walter de Gruyter & Co., Berlin.

Billera,L. *et al.* (2001) Geometry of the space of phylogenetic trees. *Adv. Appl. Math.*, **27**, 733–767.

Bryant,D. (2003) A classification of consensus methods for phylogenetics. *DIMACS Ser. Discrete Math. Theor. Comput. Sci.*, **61**, 163–184.

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

Gascuel,O. (2005) *Mathematics of Evolution and Phylogeny*. Oxford University Press, New York.

Geyer,C.J. and Thompson,E.A. (1995) Annealing markov chain monte carlo with applications to ancestral inference. *J. Am. Stat. Assoc.*, **90**, 909–920.

Holder,M. *et al.* (2003) A justification for reporting the majority-rule consensus tree in Bayesian phylogenetics. *Syst. Biol.*, **57**, 814–821.

Huelsenbeck,J. and Ronquist,F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Huggins,P.M. *et al.* (2011) Bayes estimators for phylogenetic reconstruction. *Syst. Biol.*, **60**, 528–540.

Karol,K.G. *et al.* (2001) The closest living relatives of land plants. *Science*, **294**, 2351–2353.

Margush,T. and McMorris,F.R. (1981) Consensus n-trees. *Bull. Math. Biol.*, **43**, 239–244.

McCue,L.A. *et al.* (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **29**, 774–782.

Miller,E. *et al.* (2012) Averaging metric phylogenetic trees. *arXiv*, 1211.7046.

Nye,T.M.W. (2011) Principal components analysis in the space of phylogenetic trees. *Ann. Statist.*, **39**, 2716–2739.

Owen,M. and Provan,S. (2011) A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Trans. Computat. Biol. Bioinform.*, **8**, 2–13.

Robert,C. (2001) *The Bayesian Choice. Springer Texts in Statistics*. 2nd edn. Springer-Verlag, New York. From decision-theoretic foundations to computational implementation, Translated and revised from the French original by the author.

Robert,C. and Casella,G. (1999) *Monte Carlo Statistical Methods*. 1st edn. Springer-Verlag, New York.

Salichos,L. *et al.* (2014) Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.*, **31**, 1500–1513.

Siddharthan,R. *et al.* (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Computat. Biol.*, **1**, e67.

Sturm,K.T. (2003) Probability measures on metric spaces of nonpositive curvature. In: *Heat kernels and Analysis on Manifolds, Graphs, and Metric Spaces (Paris, 2002), volume 338 of Contemporary Mathematics*. American Mathematics Society, Providence, RI, pp. 357–390.

Yang,Z. and Rannala,B. (2005) Branch-length prior influences bayesian posterior probability of phylogeny. *Syst. Biol.*, **54**, 455–470.