Syst. Biol. 65(6):1009–1023, 2016

DOI:10.1093/sysbio/syw042
Advance Access publication May 6, 2016

Estimating Bayesian Phylogenetic Information Content

Paul O. Lewis[1,*], Ming-Hui Chen[2], Lynn Kuo[2], Louise A. Lewis[1], Karolina Fučíková[1], Suman Neupane[1], Yu-Bo Wang[2], and Daoyuan Shi[2]

[1]*Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, CT 06269, USA;* [2]*Department of Statistics, University of Connecticut, 215 Glenbrook Road, Unit 4120, Storrs, CT 06269, USA*

[*]*Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, CT 06269, USA; E-mail: paul.lewis@uconn.edu*

Received 4 January 2016; reviews returned 15 April 2016; accepted 1 May 2016
Associate Editor: Mark Holder

Abstract.—Measuring the phylogenetic information content of data has a long history in systematics. Here we explore a Bayesian approach to information content estimation. The entropy of the posterior distribution compared with the entropy of the prior distribution provides a natural way to measure information content. If the data have no information relevant to ranking tree topologies beyond the information supplied by the prior, the posterior and prior will be identical. Information in data discourages consideration of some hypotheses allowed by the prior, resulting in a posterior distribution that is more concentrated (has lower entropy) than the prior. We focus on measuring information about tree topology using marginal posterior distributions of tree topologies. We show that both the accuracy and the computational efficiency of topological information content estimation improve with use of the conditional clade distribution, which also allows topological information content to be partitioned by clade. We explore two important applications of our method: providing a compelling definition of saturation and detecting conflict among data partitions that can negatively affect analyses of concatenated data. [Bayesian; concatenation; conditional clade distribution; entropy; information; phylogenetics; saturation.]

Measuring the amount of phylogenetic information in data is of importance to systematists, and not surprisingly many methods have been proposed to measure the information content (or, equivalently, noise) of systematic data and how it is apportioned across clades: for example, the consistency index (Kluge and Farris 1969); bootstrapping (Felsenstein 1985); permutation tests (Archie 1989; Faith 1991); tests that account for nucleotide composition attraction (Steel et al. 1993; 1995); relative apparent synapomorphy analysis (Lyons-Weiler et al. 1996), tree length skew statistics (Hillis and Huelsenbeck 1992); and, most recently, entropy-based methods (Shpak and Churchill 2000; Xia et al. 2003; Shi et al. 2008; Lemey et al. 2009; Xia 2009; Tippery et al. 2012; Brown 2014). The main goal of another class of methods (e.g., phylogenetic informativeness profiling (Townsend 2007; Fischer and Steel 2009), use of Fisher information in phylogenetic experimental design (Goldman 1998; Massingham and Goldman 2000; Geuten et al. 2007; Mauro et al. 2009), and phylogenetic signal and noise analysis (Townsend et al. 2012)) is primarily concerned with predicting the value of additional data rather than measuring information in existing data.

Bayesian phylogenetic methods potentially allow more complex, biologically relevant models than do methods based only on the likelihood function because prior distributions can be used to constrain poorly identified parameters to reasonable values. This effect of the prior complicates assessment of information content because the prior adds information of its own to the analysis. The Bayesian paradigm is, however, ideal for quantifying the amount of information present in data and separating information in data from information provided by the prior (Lindley 1956; Bernardo and Smith 1994; Cover and Thomas 2006), yet few phylogenetic studies (e.g., Lemey et al. 2009; Tippery et al. 2012; Brown 2014) have taken advantage of the natural information content measures available within the Bayesian framework. This is partly due to the fact that until recently the major ingredients (e.g., marginal posterior tree topology distributions) were difficult to estimate accurately from Bayesian phylogenetic MCMC output. A recently published method (Larget 2013) for estimating distributions of tree topologies given conditional clade posterior distributions potentially allows much more accurate estimation of tree topology distributions than is possible using simple sample proportions. We show how Lindley's (1956) entropy-based Bayesian measure of information content can be computed directly from Larget's (2013) conditional clade distribution, making possible estimation of information content even when coverage (the proportion of the posterior distribution captured in a posterior sample of tree topologies) is less than 1.

One obvious application of information estimation is the determination of substitutional saturation, which is commonly assumed to be a source of systematic error in phylogeny estimation (Zhong et al. 2011; 2013; Parks et al. 2012; Xi et al. 2013; 2014; Fučíková et al. 2014a; Liu et al. 2014). Removing putatively fast-evolving sites from alignments has become common practice, and software tools have been provided to automate site-stripping (Goremykin et al. 2010; Cummins and McInerney 2011; Nguyen et al. 2011). Variation is, however, necessary for phylogenetic inference, and even the fastest evolving sites may provide much valuable information (Yang 1998). It is thus important to measure the amount of information provided by a subset of sites before deciding to ignore them.

Information provided by different data subsets is not necessarily concordant, and it is arguably just as important to measure information dissonance (conflicting phylogenetic signal in different subsets of sites) as it is to measure overall information content. There are many reasons why information in one data subset may conflict with information in a different subset. Incomplete lineage sorting and lateral transfer may result in different true phylogenies for different genes, and systematic error resulting from convergence in nucleotide composition, codon bias, or other factors may result in different estimated phylogenies even if the underlying tree topologies are identical. Ideally, concatenation of gene sequences combines compatible (but possibly weak) signal from individual genes to produce a well-resolved estimated phylogeny; however, it is clear that concatenation can also hide the effects of true discordance (Mossel and Vigoda 2005; Carstens and Knowles 2007; Edwards et al. 2007; Kubatko and Degnan 2007; Heled and Drummond 2010; Roch and Steel 2015). Analysis of the information content of individual sequence subsets and dissonance among these subsets is thus highly advisable prior to analyses involving concatenation.

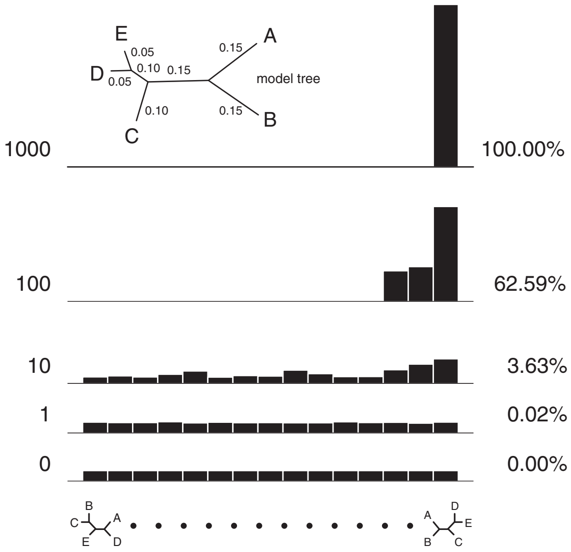


FIGURE 1. Posterior distributions estimated using a STANDARD analysis for simulated data sets with varying numbers of sites. All sites were simulated on the model tree shown using the K80 model with transition/transversion rate ratio 2. The 15 bins in each histogram each correspond to one of the 15 possible unrooted tree topologies for 5 taxa. The number of sites simulated is shown on the left. Information as defined in Equation (1) is shown on the right, expressed as a percentage of the maximum possible information.

Lindley's Information Measure

The posterior distribution of an unknown quantity θ can be viewed as the prior distribution of θ updated with information from data. If there is no information in the data relevant to θ, the posterior distribution exactly equals the prior distribution. Normally, prior distributions are made intentionally vague (high variance), in which case information present in data makes some possible values of θ less plausible than they are under the prior distribution, yielding a posterior that is concentrated compared with the prior.

In the context of phylogenetics, θ represents a (rooted or unrooted) tree topology. For example, θ represents one of 15 possible unrooted tree topologies in the case of 5 taxa. Assuming that the marginal prior distribution for tree topology is Discrete Uniform, the estimated posterior becomes increasingly concentrated over the true tree as the number of sites included grows from 0 to 10, 100, and 1000 sites, reflecting the increasing amount of information in the data relevant to tree topology estimation (Fig. 1).

The difference in entropy between the prior and posterior distributions was used as a measure of information content by Lindley (1956) and has been applied to tree topology distributions by Tippery et al. (2012) and Brown (2014). In the example above, the prior has maximum entropy (log15) because each of the 15 distinct tree topologies has equal probability (1/15), whereas the posterior based on 1000 sites has minimum entropy (0.0). The decrease in entropy as the sample size increases from 0 to 1000 provides a natural measure of the relevant information gained as sites are added. Note that information content does not grow linearly with the number of sites: comparing the 100 site cases with the 1000 site cases, 900% more sites yields only 60% more information. This decreasing return on investment is due to the fact that the information in different sites is partially redundant. It is clear that adding 1000 more sites (2000 sites total) could not reduce the posterior entropy any further even if the additional 1000 sites contained just as much information as the first 1000 sites.

Entropy has been used in other contexts in phylogenetics to measure information content. Shi et al. (2008) computed the entropy of posterior distributions from individual genes using likelihood weights to approximate posterior probabilities and used this measure as input to an automatic clustering algorithm that combined genes for which concatenation lowered entropy. Lemey et al. (2009) quantified information about the geographic location of the root of a phylogeny for several competing phylogeographic models using an entropy-based measure. Shpak and Churchill (2000) used entropy to measure the position of simulated data sets along the path from no variation to complete substitutional saturation as a function of tree topology and substitution rate using two-state Markov models. Xia et al. (2003) and Xia (2009) used a similar approach to measure the degree to which site patterns resemble those characteristic of complete substitutional saturation. Entropy has been used for a variety of other aspects of phylogenetics besides measuring information content, including estimation of pairwise distances (Bai et al. 2013), detection of heterotachy (Wang et al. 2011), and assessment of phylogenetic diversity (Allen et al. 2009).

Materials and Methods

Information from Prior and Posterior Entropy

Systematists are keenly interested in measuring information relevant to tree topology. For this the discrete version of Shannon's (1948) entropy measure is appropriate:

$$
\begin{aligned}
H &= H(\mathbf{p}) = -\sum_{\tau \in \mathbb{T}} p(\tau) \log p(\tau) \\
H^* &= H(\mathbf{p}^*) = -\sum_{\tau \in \mathbb{T}} p^*(\tau) \log p^*(\tau) \\
I &= H - H^*.
\end{aligned} \tag{1}
$$

$\mathbb{T}$ is the set of all distinct binary labeled tree topologies, τ represents a realization of the random variable T representing a tree topology, $|\mathbb{T}|$ represents the size of the set $\mathbb{T}$ (i.e., the number of possible binary labeled tree topologies), $\mathbf{p}^*$ is the marginal posterior distribution, with $p^*(\tau)$ the marginal posterior probability of tree topology τ, and $\mathbf{p}$ is the marginal prior distribution, with $p(\tau)$ the marginal prior probability of the specific tree topology τ. Throughout, $\log(\cdot)$ denotes the natural logarithm (base e). To simplify notation, we omit the conditional dependence of the posterior on data, instead using an asterisk ($*$) to denote quantities derived from the posterior and to distinguish them from quantities derived from the prior (which lack the asterisk). Lindley information, I, measures the degree to which the prior is concentrated into a posterior dominated by a smaller number of tree topologies. Assuming that the tree topology prior is Discrete Uniform on the set $\mathbb{T}$ of all distinct tree topologies, I ranges from 0.0 (no information about topology) to $\log|\mathbb{T}|$ (maximum information; all posterior probability concentrated in a single tree topology), and, when expressed as a percentage, it should be understood that 100% implies $I = \log|\mathbb{T}|$.

Relationship to Kullback–Leibler Divergence

If (and only if) the tree topology prior is Discrete Uniform over all distinct tree topologies (in which case $p(\tau) = 1/|\mathbb{T}|$ is a constant for all $\tau \in \mathbb{T}$), then Lindley information is equivalent to the Kullback–Leibler (KL) divergence measured from prior ($\mathbf{p}$) to posterior ($\mathbf{p}^*$):

$$
\begin{aligned}
\mathrm{KL}(\mathbf{p}^*, \mathbf{p}) &= \sum_{\tau \in \mathbb{T}} p^*(\tau) \log\left(\frac{p^*(\tau)}{p(\tau)}\right) \\
&= \left\{\sum_{\tau \in \mathbb{T}} p^*(\tau) \log p^*(\tau)\right\} - \sum_{\tau \in \mathbb{T}} p^*(\tau) \log p(\tau) \\
&= \left\{\sum_{\tau \in \mathbb{T}} p^*(\tau) \log p^*(\tau)\right\} - \log p(\tau) \cancelto{1}{\sum_{\tau \in \mathbb{T}} p^*(\tau)} \\
&= \left\{\sum_{\tau \in \mathbb{T}} p^*(\tau) \log p^*(\tau)\right\} - \log p(\tau) \sum_{\tau \in \mathbb{T}} p(\tau) \\
&= \left\{-\sum_{\tau \in \mathbb{T}} p(\tau) \log p(\tau)\right\} - \left\{-\sum_{\tau \in \mathbb{T}} p^*(\tau) \log p^*(\tau)\right\} \\
&= H - H^* \\
&= I.
\end{aligned} \tag{2}
$$

This KL interpretation is useful because, as we show later, the overall information content can be partitioned into additive clade-specific components that are themselves clade-specific KL divergences weighted by the posterior probability of the clade.

The KL interpretation also gives rise to a means of estimating Lindley information. Letting $\mathbf{y}$ denote the data on which the posterior distribution $\mathbf{p}^*$ is based,

$$
\begin{aligned}
\mathrm{KL}(\mathbf{p}^*, \mathbf{p}) &= \sum_{\tau \in \mathbb{T}} p^*(\tau) \log\left(\frac{p^*(\tau)}{p(\tau)}\right) \\
&= \sum_{\tau \in \mathbb{T}} p(\tau|\mathbf{y}) \log\left(\frac{p(\tau|\mathbf{y})}{p(\tau)}\right) \\
&= \sum_{\tau \in \mathbb{T}} p(\tau|\mathbf{y}) \log\left(\frac{p(\mathbf{y}|\tau)\cancel{p(\tau)}}{p(\mathbf{y})\cancel{p(\tau)}}\right) \\
&= \sum_{\tau \in \mathbb{T}} p(\tau|\mathbf{y}) \log p(\mathbf{y}|\tau) - \log p(\mathbf{y}) \cancelto{1}{\sum_{\tau \in \mathbb{T}} p(\tau|\mathbf{y})} \\
&= E_{T|\mathbf{y}}\left[\log p(\mathbf{y}|T)\right] - \log p(\mathbf{y}),
\end{aligned}
$$

where $p(\mathbf{y}|\tau)$ is the likelihood marginalized over all model parameters on a fixed tree topology τ, and $p(\mathbf{y})$ is the likelihood marginalized over all model parameters (including tree topology). This approach would clearly require considerable computation, as the total marginal likelihood as well as all tree-specific marginal likelihoods for trees having nonnegligible posterior probabilities must be estimated. We next consider a more tractable approach using conditional clade distribution estimated from a single posterior sample of trees.

Partitioning Information Using Conditional Clade Probabilities

Following on previous work by Höhna and Drummond (2011), Larget (2013) showed that much more accurate estimates of posterior probabilities of tree topologies are possible using posterior summaries of their component clades (or splits in the case of unrooted trees). For example, if only 1000 trees are sampled during an MCMC analysis, the rarest tree topology in the sample will necessarily have an estimated marginal

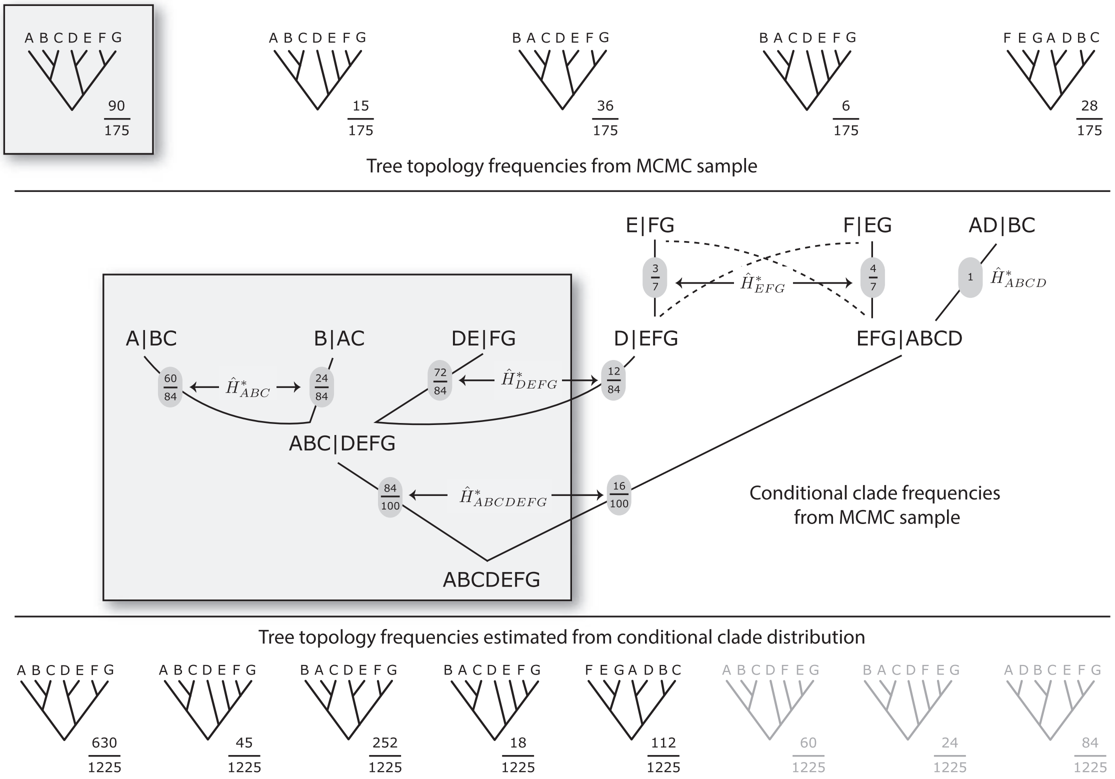


FIGURE 2. Relationship between the Markov chain Monte Carlo (MCMC) sample, the conditional clade distribution (CCD) derived from that sample, and the estimated posterior distribution derived from the CCD. Top: frequencies of the five distinct tree topologies in the MCMC sample. Middle: graph depicting the conditional clade distribution constructed from the MCMC sample (dotted lines indicate combinations that were not observed in the MCMC sample). $\hat{H}^*_{\mathrm{ABCDEFG}}$, $\hat{H}^*_{\mathrm{ABC}}$, $\hat{H}^*_{\mathrm{DEFG}}$, $\hat{H}^*_{\mathrm{EFG}}$, and $\hat{H}^*_{\mathrm{ABCD}}$ show which conditional clade probabilities are used in each component of the total posterior entropy computed in Equation (3). Bottom: posterior probabilities of tree topologies estimated using the CCD. Tree topologies shown in gray are implied by the CCD (see dashed lines in the CCD graph) but were not observed in the MCMC sample. Shaded boxes indicate the portion of this figure further explained in Figure 3.

posterior probability $\geq$ 0.001, which may represent a gross overestimate of its true marginal posterior probability. Using estimates of conditional clade probabilities (probabilities of child clades given their parent clade), it is possible to estimate the marginal posterior probability of a tree topology with much greater accuracy. Larget's method also allows one to estimate the posterior *coverage*, φ, defined to be the fraction of the total posterior probability represented in the sampled tree topologies, an important quantity that was previously unavailable.

Estimating entropy using simple frequencies of sampled tree topologies overestimates information content when $\varphi < 1$ because the posterior distribution appears to be more concentrated than it really is. Ideally, a measure of entropy used in computing I would take into account 100% rather than $(100 \times \varphi)$% of the posterior distribution. That is, it needs to include frequencies of tree topologies not sampled but which nevertheless have nontrivial posterior probabilities, and this need grows more important as the information content of the data decreases. For information-poor data sets, φ will be very low, and the apparent concentration of the posterior is correspondingly extreme, yielding a greater disparity between the actual and estimated information content.

Fortunately, it is possible to use the estimated conditional clade distribution to efficiently compute the entropy of the marginal posterior tree topology distribution, thus allowing tree topologies actually sampled *as well as those not sampled* to contribute, avoiding the apparent contraction of the posterior that accompanies $\varphi < 1$. Assume that the five tree topologies along the top of Figure 2 were sampled in an MCMC analysis. The frequency of each of these topologies in the posterior sample is shown to its right. The set of all distinct rooted tree topologies ($\mathbb{T}$) for seven taxa contains $|\mathbb{T}| = 10,395$ tree topologies, so the fact that there are only five tree topologies represented in the posterior sample implies that the data contain a substantial amount of information about tree topology. The prior entropy (H) and the posterior entropy (H^*) and Lindley information (I) estimated using these frequencies are:

$$\hat{H}^* = -\left[\frac{90}{175}\log\left(\frac{90}{175}\right) + \frac{15}{175}\log\left(\frac{15}{175}\right) + \frac{36}{175}\log\left(\frac{36}{175}\right)\right.$$
$$\left. + \frac{6}{175}\log\left(\frac{6}{175}\right) + \frac{28}{175}\log\left(\frac{28}{175}\right)\right] = 1.28671$$

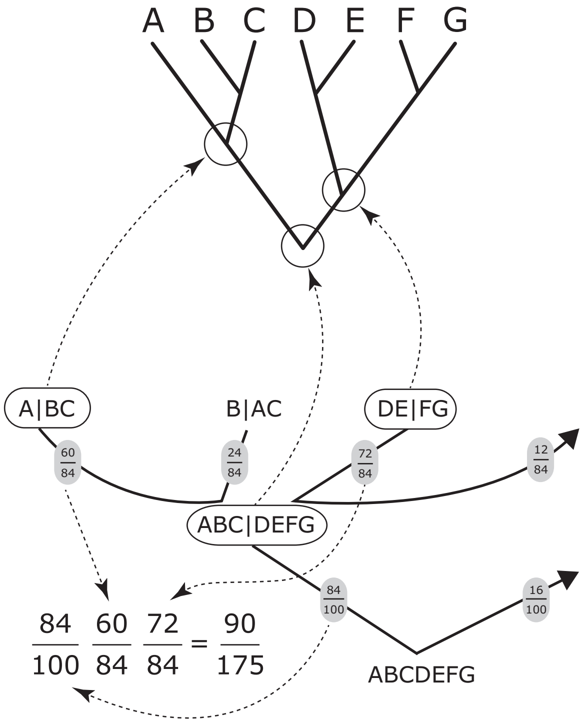


FIGURE 3. Portion of Figure 2 illustrating the calculation of the posterior probability of a given tree topology as the product of three nontrivial conditional clade probabilities, each corresponding to a distinct internal node of the tree.

$$
\begin{aligned}
H &= -\sum_{i=1}^{10395} \frac{1}{10395} \log\left(\frac{1}{10395}\right) = \log(10395) = 9.24908 \\
\hat{I} &= H - \hat{H}^* = 7.96237\,(86.1\%).
\end{aligned}
$$

The middle section of Figure 2 summarizes the conditional clade distribution derived from the sampled tree topologies. Relevant products of conditional clade probabilities yield estimates of marginal tree topology posterior probabilities. For example, the conditional clade estimate of the probability of the first tree topology (Fig. 3) is

$$
\left(\frac{84}{100}\right)\left[\left(\frac{60}{84}\right)\left(\frac{72}{84}\right)\right] = \frac{4320}{8400},
$$

which is identical to the value 90/175 shown to the right of this tree in Figure 2. The estimates of marginal posterior probabilities derived from the conditional clade distribution for all five sampled tree topologies are shown to the right of each topology in the bottom panel of Figure 2. The sum of these five conditional clade-based marginal tree topology probabilities yields the estimated coverage, $\hat{\varphi} = 0.863$, which is less than 1.0. The three additional tree topologies shown in gray in the bottom panel account for the remaining 13.7% of the posterior distribution. The conditional clade-based probabilities of these topologies involve combinations indicated by dashed lines in Figure 2. These combinations of clades did not occur in any of the tree topologies sampled, but presumably *would have been sampled* had the MCMC simulation continued.

Lindley information calculated from the probabilities of all eight tree topologies demonstrates that 84% of the maximum possible information is present in the data used for this analysis:

$$
\begin{aligned}
\hat{H}^* &= -\left[\frac{630}{1225}\log\left(\frac{630}{1225}\right) + \frac{45}{1225}\log\left(\frac{45}{1225}\right) + \frac{252}{1225}\log\left(\frac{252}{1225}\right)\right. \\
&\quad + \frac{18}{1225}\log\left(\frac{18}{1225}\right) + \frac{112}{1225}\log\left(\frac{112}{1225}\right) + \frac{60}{1225}\log\left(\frac{60}{1225}\right) \\
&\quad \left. + \frac{24}{1225}\log\left(\frac{24}{1225}\right) + \frac{84}{1225}\log\left(\frac{84}{1225}\right)\right] \\
&= 1.47793 \\
H &= \log(10395) = 9.24908 \\
\hat{I} &= H - \hat{H}^* = 7.77115\,(84.0\%).
\end{aligned}
$$

Note that using the simple sample frequencies of tree topologies results in an inflated estimate (86.1%) of information content compared with the estimate derived from the conditional clade distribution (84.0%). The same calculation can be carried out without enumerating all eight tree topologies using a traversal of the conditional clade probability graph (Algorithm 1, Appendix 1, and proof in Appendix 2):

$$
\begin{aligned}
\hat{H}^* &= \hat{H}^*_{\text{ABCDEFG}} + \frac{84}{100}\left[\hat{H}^*_{\text{ABC}} + \left(\hat{H}^*_{\text{DEFG}} + \frac{12}{84}\hat{H}^*_{\text{EFG}}\right)\right] \\
&\quad + \frac{16}{100}\left[\hat{H}^*_{\text{EFG}} + \hat{H}^*_{\text{ABCD}}\right],
\end{aligned} \tag{3}
$$

where

$$
\begin{aligned}
\hat{H}^*_{\text{ABCDEFG}} &= -\left[\frac{84}{100}\log\left(\frac{84}{100}\right) + \frac{16}{100}\log\left(\frac{16}{100}\right)\right] = 0.43967 \\
\hat{H}^*_{\text{ABC}} &= -\left[\frac{60}{84}\log\left(\frac{60}{84}\right) + \frac{24}{84}\log\left(\frac{24}{84}\right)\right] = 0.59827 \\
\hat{H}^*_{\text{DEFG}} &= -\left[\frac{72}{84}\log\left(\frac{72}{84}\right) + \frac{12}{84}\log\left(\frac{12}{84}\right)\right] = 0.41012 \\
\hat{H}^*_{\text{EFG}} &= -\left[\frac{3}{7}\log\left(\frac{3}{7}\right) + \frac{4}{7}\log\left(\frac{4}{7}\right)\right] = 0.68291 \\
\hat{H}^*_{\text{ABCD}} &= -\left[1.0\log(1.0)\right] = 0.0.
\end{aligned}
$$

Equation (3) can be rearranged into the following form, illustrating that the total entropy can be partitioned by clade:

$$
\begin{aligned}
\hat{H}^* &= \hat{H}^*_{\text{ABCDEFG}} + \frac{84}{100}\hat{H}^*_{\text{ABC}} + \frac{84}{100}\hat{H}^*_{\text{DEFG}} + \frac{28}{100}\hat{H}^*_{\text{EFG}} + \frac{16}{100}\hat{H}^*_{\text{ABCD}} \\
&= 1.47793.
\end{aligned}
$$

More generally, Lindley information itself can be expressed as a sum of clade-specific contributions,

$$I=\sum_{C\in S} p^*(C)\,\mathrm{KL}_C \tag{4}$$

$$\mathrm{KL}_C=\sum_{C^{(l)},C^{(r)}|C\in\mathbb{C}_C} p^*(C^{(l)},C^{(r)}|C)\log\left(\frac{p^*(C^{(l)},C^{(r)}|C)}{p(C^{(l)},C^{(r)}|C)}\right),$$

where clade C is a nonempty subset of the set S of all taxa, $p^*(C)$ is the marginal posterior probability of parent clade C, $p^*(C^{(l)},C^{(r)}|C)$ is the marginal posterior probability of left ($C^{(l)}$) and right ($C^{(r)}$) child clades given the parent clade (C) using the conditional clade distribution derived from the posterior, and $p(C^{(l)},C^{(r)}|C)$ is the corresponding conditional clade probability derived from the discrete uniform prior. Just as the overall Lindley information may be interpreted as the KL divergence from the (Discrete Uniform) prior to the posterior, the sum in (4) shows that the contribution of each clade C in the conditional clade hierarchy to the overall Lindley information may be interpreted as the conditional KL divergence, KL_C, weighted by the marginal posterior clade probability of the parent clade, $p^*(C)$. A proof of this equivalence is provided in Appendix 2.

Returning to the example in Figure 2, the weights and KL divergences for each clade in the conditional clade distribution that form the additive components of the total Lindley information are calculated below using Algorithm 2 (Appendix 1):

$$\begin{aligned}
p(\mathrm{ABCDEFG}) &= 1.0,\\
\mathrm{KL}_{\mathrm{ABCDEFG}} &= \frac{84}{100}\log\left(\frac{84}{100}\Big/\frac{45}{10395}\right)\\
&\quad+\frac{16}{100}\log\left(\frac{16}{100}\Big/\frac{45}{10395}\right)=5.00275,\\
p(\mathrm{ABC}) &= \frac{84}{100},\\
\mathrm{KL}_{\mathrm{ABC}} &= \frac{60}{84}\log\left(\frac{60}{84}\Big/\frac{1}{3}\right)\\
&\quad+\frac{24}{84}\log\left(\frac{24}{84}\Big/\frac{1}{3}\right)=0.50034,\\
p(\mathrm{DEFG}) &= \frac{84}{100},\\
\mathrm{KL}_{\mathrm{DEFG}} &= \frac{72}{84}\log\left(\frac{72}{84}\Big/\frac{1}{15}\right)\\
&\quad+\frac{12}{84}\log\left(\frac{12}{84}\Big/\frac{3}{15}\right)=2.14099,\\
p(\mathrm{ABCD}) &= \frac{16}{100},\\
\mathrm{KL}_{\mathrm{ABCD}} &= (1)\log\left(1\Big/\frac{1}{15}\right)=2.70805,\\
p(\mathrm{EFG}) &= \frac{28}{100}=\frac{16}{100}+\frac{84}{100}\frac{12}{84},\\
\mathrm{KL}_{\mathrm{EFG}} &= \frac{3}{7}\log\left(\frac{3}{7}\Big/\frac{1}{3}\right)\\
&\quad+\frac{4}{7}\log\left(\frac{4}{7}\Big/\frac{1}{3}\right)=0.41570.
\end{aligned}$$

The resulting estimated Lindley information,

$$\begin{aligned}
\hat{I} &= (1.0)(5.00275)+(0.84)(0.50034)+(0.84)(2.14099)\\
&\quad+(0.16)(2.70805)+(0.28)(0.41570)\\
&= 7.77115,
\end{aligned}$$

is identical to that calculated using the marginal posterior probabilities of tree topologies. The calculation is efficient because there is no enumeration of tree topologies, and the contribution of each clade to the overall information is itself of considerable interest to researchers.

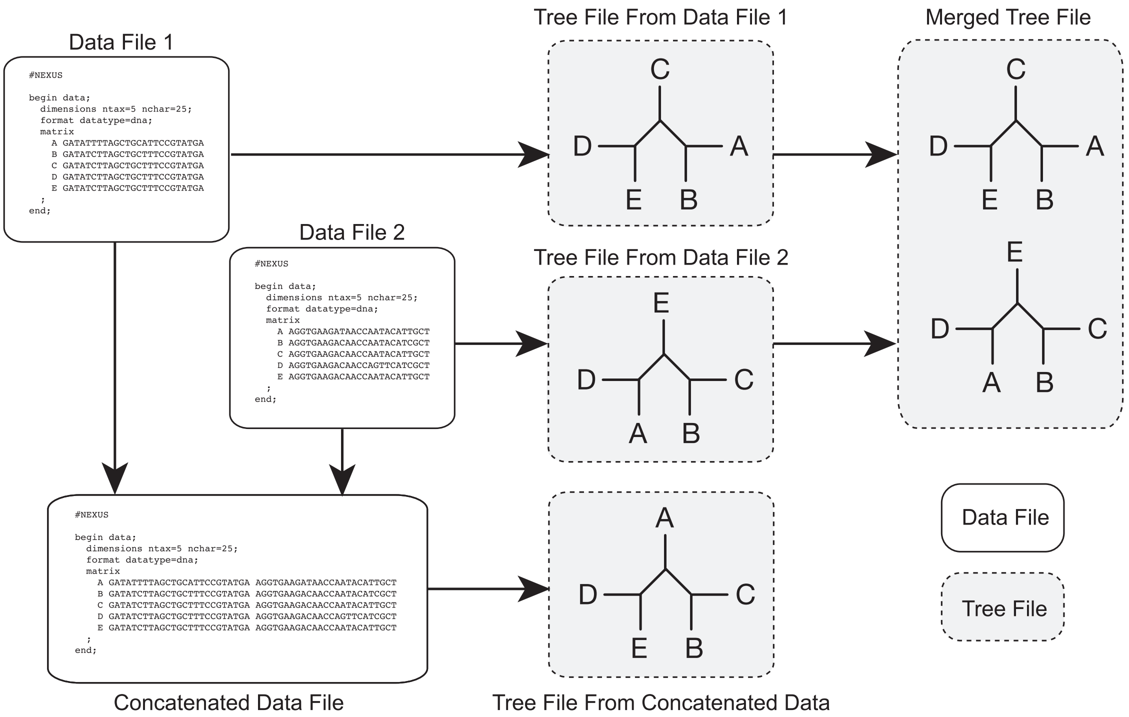
While the worked example above uses rooted trees, unrooted trees present no challenges. Rooted at an arbitrary leaf, unrooted trees may be treated as rooted trees with one fewer taxon. Because the conditional clade distribution is unaffected by rooting (Larget 2013), the choice of which leaf to use as the root does not affect the overall estimate of I but will change how the additive components of I are distributed across the tree.

Phylogenetic Dissonance

Consider a partitioned data set in which each site is assigned to one of K mutually exclusive subsets. Because entropy is a concave function, Jensen's inequality assures us that the average entropy of posterior samples from different subsets must be less than or equal to the entropy of an average tree sample obtained by merging tree files from separate analyses of each subset. To be specific, the average of the entropy computed from "Tree File From Data File 1" and the entropy computed from "Tree File From Data File 2" must be less than or equal to the entropy computed from "Merged Tree File" in Figure 4. Letting $\mathbf{p}_k^*$ be the posterior distribution for subset k, and w_k be the weight associated with subset k, Jensen's inequality says

$$\sum_{k=1}^{K} w_k H(\mathbf{p}_k^*)\le H\left(\textstyle\sum_{k=1}^{K} w_k\mathbf{p}_k^*\right)=H^*_{\text{merged}}. \tag{5}$$

We define *phylogenetic dissonance*, D, to be the difference between the entropy of the merged posterior tree sample and the average entropy of the posterior tree sample from



```
#NEXUS

begin data;
  dimensions ntax=5 nchar=25;
  format datatype=dna;
  matrix
    A GATATTTTAGCTGCATTCCGTATGA
    B GATATCTTAGCTGCTTTCCGTATGA
    C GATATCTTAGCTGCTTTCCGTATGA
    D GATATCTTAGCTGCTTTCCGTATGA
    E GATATCTTAGCTGCTTTCCGTATGA
  ;
end;
```

```
#NEXUS

begin data;
  dimensions ntax=5 nchar=25;
  format datatype=dna;
  matrix
    A AGGTGAAGATAACCAATACATTGCT
    B AGGTGAAGACAACCAATACATCGCT
    C AGGTGAAGACAACCAATACATTGCT
    D AGGTGAAGACAACCAGTTCATCGCT
    E AGGTGAAGACAACCAATACATTGCT
  ;
end;
```

```
#NEXUS

begin data;
  dimensions ntax=5 nchar=25;
  format datatype=dna;
  matrix
    A GATATTTTAGCTGCATTCCGTATGA AGGTGAAGATAACCAATACATTGCT
    B GATATCTTAGCTGCTTTCCGTATGA AGGTGAAGACAACCAATACATCGCT
    C GATATCTTAGCTGCTTTCCGTATGA AGGTGAAGACAACCAATACATTGCT
    D GATATCTTAGCTGCTTTCCGTATGA AGGTGAAGACAACCAGTTCATCGCT
    E GATATCTTAGCTGCTTTCCGTATGA AGGTGAAGACAACCAATACATTGCT
  ;
end;
```

FIGURE 4. Relationships between data files and the tree files generated from those data files. Concatenated data files are created by combining different data subsets into one matrix. Merged tree files are created by combining trees from different tree files. Note that a merged tree file is different than a tree file created by analyzing a concatenated data set. Tree files always contain multiple trees, but single tree icons are used here to make it easier to see that the *Merged Tree File* is composed of all trees from both *Tree File From Data File 1* and *Tree File From Data File 2*.

individual subsets,

$$D = H^*_{\text{merged}} - \sum_{k=1}^{K} w_k H(\mathbf{p}^*_k), \tag{6}$$

where $H(\mathbf{p}^*_k)$ is the entropy of the posterior tree sample from the k-th subset and $\sum_{k=1}^{K} w_k = 1$. Although (5) and (6) are general, in this manuscript we assume $w_k = 1/K$. Note that a tree file constructed by combining trees sampled from independent MCMC analyses of different data subsets is very different than a tree file resulting from a single MCMC analysis of the concatenated data (Fig. 4). It is clear from (5) and (6) that $0 \le D \le H^*_{\text{merged}}$. This upper bound allows D, like I, to be expressed as a percentage of its maximum value.

Assuming that the prior distribution $\mathbf{p}$ is identical for each subset, D may be defined equivalently in terms of Lindley information:

$$\sum_{k=1}^{K} w_k H(\mathbf{p}^*_k) \le H\left(\sum_{k=1}^{K} w_k \mathbf{p}^*_k\right)$$

$$\sum_{k=1}^{K} w_k \left(H(\mathbf{p}) - H(\mathbf{p}^*_k)\right) \ge H(\mathbf{p}) - H\left(\sum_{k=1}^{K} w_k \mathbf{p}^*_k\right)$$

$$\sum_{k=1}^{K} w_k I_k \ge I_{\text{merged}}$$

$$D = \left\{\sum_{k=1}^{K} w_k I_k\right\} - I_{\text{merged}} \ge 0.$$

Example Data Sets

Two previously published data sets provide empirical examples of information content estimation. The data set ALGAE comprises chloroplast *psa*B sequences from 33 taxa of green algae (phylum Chlorophyta, class Chlorophyceae, order Sphaeropleales) analyzed by Fučíková et al. (2014b). The original seven-gene data set may be downloaded from http://treebase.org/ using Study ID 13960. The alignments of just the *psa*B gene used in this study are available in the Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.1dn50. These data were chosen because of their deep divergence, which invites hasty judgements of saturation, especially of third codon position sites. We analyzed second and third codon position sites separately and used I to assess which subset has more phylogenetic information.

The five sequences of *rps*11 composing the data set BLOODROOT were extracted from the Supplementary Information available on Dryad for Figure 3 in Bergthorsson et al. (2003). The alignments used in this study are available in the Supplementary Material available on Dryad. These data were chosen because they represent a case in which horizontal transfer of

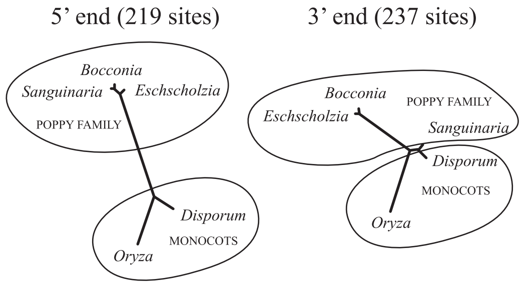

5’ end (219 sites) 3’ end (237 sites)

FIGURE 5. Maximum posterior probability trees for the 5′ and 3′ ends of the *rps*11 gene (BLOODROOT data set) from three taxa in the angiosperm family Papaveraceae (the Poppy Family) and two monocots. In *Sanguinaria* (a member of the Papaveraceae), the 3′ end of the gene was horizontally transferred from a monocot, whereas the 5′ end shows vertical inheritance.

half of the gene results in different true tree topologies for the 5′ (219 nucleotide sites) and 3′ (237 nucleotide sites) subsets, which allows investigation of information content estimation in the presence of true conflicting phylogenetic signal (Fig. 5). We analyzed each half of the data separately and measured phylogenetic dissonance, which is expected to be high in this case.

Phylogenetic Analyses

Simulated DNA data were generated using Seq-Gen version 1.3.3 (Rambaut and Grassly 1997). Corrected and uncorrected pairwise distances were computed using PAUP* version 4.0a146 for Macintosh OS X (Swofford 2003). Corrected distances used maximum likelihood under the GTR+I+G model with parameters estimated on a neighbor-joining tree. Bayesian MCMC analyses (hereafter referred to as the STANDARD analysis) were carried out using MRBAYES 3.2.4 (Ronquist et al. 2012) using the GTR+I+G model. Parameters of this model and the priors used were:

$$\begin{aligned}
\text{Tree topology } T &\sim \text{Discrete Uniform}(1, |\mathbb{T}|)\\
\text{Tree length } L &\sim \text{Exponential}(0.01)\\
\text{Edge length proportions } \mathbf{e} &\sim \text{Dirichlet}(1, \cdots, 1)\\
\text{Nucleotide frequencies } \boldsymbol{\pi} &\sim \text{Dirichlet}(1, 1, 1, 1)\\
\text{Exchangeabilities } \mathbf{r} &\sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)\\
\text{Discrete Gamma shape } \alpha &\sim \text{Exponential}(1)\\
\text{Proportion invariable sites } p &\sim \text{Uniform}(0, 1)
\end{aligned}$$

A compound Dirichlet prior (Rannala et al. 2012) was used to model tree length and edge length proportions, and the variance of the exponential tree length component (10,000) was increased by factor of 100 over the default variance (100) to avoid constraining tree length in analyses of fast-evolving third position sites. After removing the initial 10% as burn-in, MCMC samples contained 10,000 trees, saving every 1000 generations. Two independent analyses were performed, each comprising four Metropolis-coupled chains using the default heating schedule (chain powers 1.0, 0.909, 0.833, and 0.769). Information measures described in this article were computed using the program Galax version 1.0, which is available at http://phylogeny.uconn.edu/software.

Explicit instructions and scripts for performing all analyses described in this article are provided in the Supplementary Material available on Dryad.

RESULTS AND DISCUSSION

This article presents one approach to measuring the amount of information relevant to phylogenetic tree topology using the difference in entropy between the prior and posterior tree topology distributions. To reduce the impact of MCMC sample size in cases of low information, the conditional clade distribution is used to approximate the posterior entropy, allowing tree topologies that were not sampled, but nevertheless contain clades that were sampled, to smooth the posterior distribution. This smoothing leads to more accurate estimation of the posterior entropy, conditional on the accuracy of the approximation of marginal posterior tree probabilities by the conditional clade distribution.

Before discussing the behavior of our method in simulation and possible applications using empirical examples, it is important to make clear what, exactly, is being measured. First, Lindley information (I) defined in (1) measures the information contained in the data *that is not redundant with information provided in the prior*. An extremely informative prior that forces the posterior to place nearly all probability on one tree topology will result in a low estimated value of I because the posterior distribution has essentially the same entropy as the prior: the information in the data is redundant because the prior has already determined the posterior. The method for estimating I in this article assumes that the prior has zero information, but the issue of redundancy could nevertheless arise if I was estimated for both an informative prior (either analytically or using a tree sample from an analysis without data) and the associated posterior, and the difference used to measure information.

Second, it is possible that the information measured by I represents *misinformation*. Any model violation that causes a systematic bias that leads to an incorrect tree topology receiving more posterior probability than the correct tree topology will nevertheless increase the apparent amount of information in the data. We show below that concatenation of data subsets that have different underlying true tree topologies can result in maximum information but an incorrect tree topology. Phylogenetic dissonance (D) makes use of Lindley information to detect when different data subsets disagree, but caution must be exercised when

TABLE 1. Results of four-taxon simulations showing information content

	No. sites	$\hat{I}$
(a) Number of sites	1	0
	10	4
	100	90
	1000	100
	Rel. rate	$\hat{I}$
(b) Substitution rate	0.01	18
	0.1	99
	1	100
	10	64
	100	1.5
	missing (%)	$\hat{I}$
(c) Missing data	0	100
	50	98
	90	0
	100	0
	Rate variance	$\hat{I}$
(d) Rate heterogeneity	1	100
	10	97
	100	13
	1000	0

Notes: Each $\hat{I}$ value presented is expressed as percentage of maximum and is the mean of 100 replicate four-taxon simulations. Data generation: JC69 (Jukes–Cantor) model, 1000 sites (except (a)), all edge lengths 0.1 substitutions/site (except (b)), 0% missing (except (c)), and rate homogeneity (except (d)). Data analysis: STANDARD. Rel. rate in (b) is the factor by which all edge lengths were multiplied in the model tree, missing data in (c) was randomly distributed across taxa and sites after data were simulated, and rate variance in (d) is the variance of continuous-gamma distributed relative rates across sites used in the generating model.

interpreting any measure of information content for single, unpartitioned data sets.

Simulations

Four-taxon simulations illustrate some general expectations with respect to the information content of DNA sequence data (Table 1). The first 100 sites evaluated (of 1000 sites total) contribute 90% of the maximum possible information; any additional sites contribute mostly redundant information, mostly reinforcing exclusion of tree topologies already ruled out by the first 100 sites (Table 1a). As the rate of evolution increases or decreases away from the optimal rate, the information contributed per site decreases due to the noise associated with multiple hits, or the lack of variability, respectively (Table 1b). As the proportion of missing data increases, information content drops (Table 1c). Finally, as rate heterogeneity increases, information content drops (Table 1d) because of the combination of low substitution rate for most sites and high rates for the remaining sites, leaving few sites that evolve at an optimal rate for preserving history.

TABLE 2. Conditional clade ($\hat{I}$) and empirical frequency ($\hat{I}_{\text{freq}}$) estimates of topological information content (both expressed as percentage of maximum) when true information content is zero

Taxa	Trees	$\hat{\varphi}$	$\hat{I}$	$\hat{I}_{\text{freq}}$
4	3	1.000	0.01031	0.01001
5	15	1.000	0.02398	0.02400
6	105	1.000	0.07105	0.1093
7	945	1.000	0.1855	0.6934
8	10395	0.6539	0.4840	6.353
9	135135	0.09134	1.235	22.46
10	2027025	0.01010	3.404	36.60
11	34459425	0.00151	7.304	46.93
12	654729075	0.00040	13.30	54.63

Notes: Results from paired runs in a STANDARD analysis were averaged. Coverage ($\hat{\varphi}$) begins to drop for analyses involving more than seven taxa because the MCMC sample size was fixed at 10,000.

Zero Information

MCMC analyses exploring discrete uniform prior distributions for various numbers of taxa show how well topological information content can be estimated when there is zero information and thus the number of tree topologies that should be sampled grows much larger than the MCMC sample size (Table 2). The number of trees sampled was fixed at 10,000, so coverage begins to drop when the number of possible tree topologies grows larger than 10,000. As coverage drops, the empirical frequency estimate of information content increasingly overestimates the true information content, which is zero in every case. While estimates using the conditional clade method described here also begin to overestimate the true amount, $\hat{I}$ remains much lower than $\hat{I}_{\text{freq}}$ even when the sample accounts for only a tiny fraction (0.0004) of the posterior distribution. That said, note that even at 12 taxa coverage has dropped so low that a true information content of zero cannot be accurately measured, even using the conditional clade approach. Any estimate of information content should thus be accompanied by the estimated coverage, which speaks to the reliability of the information content estimate.

Saturation

A STANDARD analysis of the ALGAE data set demonstrates that commonly used methods for assessing saturation can be quite misleading (Table 3). Analyzing second codon position sites alone, 19,997 of the 20,000 tree topologies sampled from the posterior distribution were unique, and the entire sample captured only an estimated 0.026% of the total posterior probability ($\hat{\varphi} = 0.00026$). In addition to being used to measure dissonance among data subsets, D also serves as a sensitive measure of topological MCMC convergence if the tree files supplied are from replicate MCMC analyses that differ only in the pseudorandom number seed used. D is expected to be zero in this case as all replicate MCMC analyses are exploring exactly the same posterior distribution. Despite the extremely

TABLE 3. Results from analyses of the ALGAE data set

	Unique	$\hat{\varphi}$	H	$\hat{H}^*$	$\hat{I}$ %
(a) First codon positions only ($\hat{D}=0.7392\%$)					
Run 1	9627	0.03616	97.29	15.73	83.83
Run 2	9593	0.04004	97.29	15.66	83.91
Average	9610	0.03810	97.29	15.69	83.87
Merged	18705	0.05854	97.29	15.81	83.75
(b) Second codon positions only ($\hat{D}=1.374\%$)					
Run 1	9999	0.00012	97.29	23.45	75.89
Run 2	9998	0.00016	97.29	23.39	75.96
Average	9998.5	0.00014	97.29	23.42	75.92
Merged	19997	0.00026	97.29	23.75	75.59
(c) Third codon positions only ($\hat{D}=0.8486\%$)					
Run 1	8722	0.1839	97.29	13.21	86.42
Run 2	8639	0.1965	97.29	13.06	86.57
Average	8680.5	0.1902	97.29	13.14	86.50
Merged	16359	0.2462	97.29	13.25	86.38

Notes: $\hat{I}$ and $\hat{D}$ are expressed as % maximum. Note that here $\hat{D}$ measures dissonance among replicate MCMC analyses and low values are used as an indication that both runs have converged with respect to tree topology

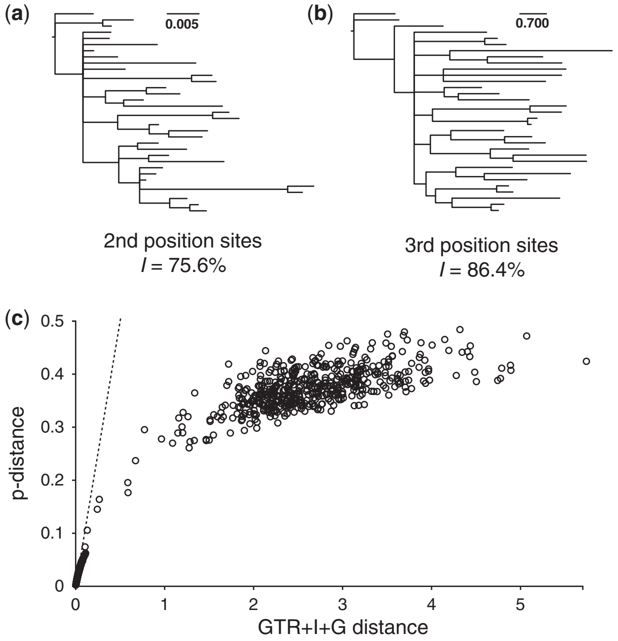


FIGURE 6. Saturation plots and majority rule consensus trees from analyses of *psa*B (ALGAE data set) second (a) and third (b) codon positions. In this example, fast evolving third position sites have more phylogenetic information than slow evolving second position sites despite appearing saturated in the plot (c) of pairwise distances corrected using the GTR+I+G (general time reversible with invariable sites and discrete Gamma rate heterogeneity) model (x-axis) against p-distance (proportion of sites different) (y-axis). Each point represents a single pair of taxa.

low coverage, D was only 1.374% of its maximum value, which indicates good agreement between the posterior distributions estimated from replicate MCMC simulations. The information content of these second position sites was $\hat{I}=75.59\%$ (estimated from the merged posterior sample). Coverage is very low, suggesting that this information content estimate may be unreliable, and this was confirmed using a comparable analysis of the prior alone, which yielded an information content of 82.8% and an estimate of D across the two replicates of 11.59. The greater consistency across replicates of the second position data (as indicated by the lower dissonance D) indicates that second position sites have some information, but this example illustrates the difficulty of estimating information content accurately when both information content and coverage are low.

Third position sites of *psa*B, on the other hand, exhibited lower D among replicates (0.8486% of maximum), higher $\hat{I}$ (86.38%), and three orders of magnitude greater coverage ($\hat{\varphi}=0.2462$). Although saturation plots make the third codon positions appear greatly saturated compared to second codon positions (Fig. 6c), I indicates that the third position sites contain more information about tree topology than do second position sites. The greater information of third positions is evident in the better resolution of the third position majority rule consensus (Fig. 6a,b).

Concatenation

The phylogenetic analysis of concatenated data continues to receive criticism because of the tendency for concatenated analyses to mask incongruence among individual data subsets (Edwards et al. 2007; Kubatko and Degnan 2007; Degnan and Rosenberg 2009; Edwards 2009; Heled and Drummond 2010; Leaché and Rannala 2011; Weisrock et al. 2012; Lemmon and Lemmon 2013; Roch and Steel 2015). We present here an empirical example in which concatenation of two data subsets containing conflicting phylogenetic signal is clearly misleading, strongly supporting a tree topology only weakly supported by each of the constituent data subsets. Coverage was 1.0 and phylogenetic dissonance was 0.01320% or less for replicate Bayesian MCMC analyses conducted using the BLOODROOT data set (Table 4), indicating that these runs all converged with respect to topology. The first analysis, based on 219 nucleotides from the 5′ end of the gene (Table 4a), yielded three unique tree topologies and an information content of 74.47%. The second analysis, based on 237 sites from the 3′ end of the gene (Table 4b), yielded an information content of 66.83% and also produced three distinct topologies. The third analysis, based on the entire (concatenated) data set (Table 4c), yielded just one tree topology and 100% information.

The 100% information content in the concatenated analysis is not surprising: information content is expected to increase with the number of sites included (Table 1a). If the information in the two halves of the *rps11* gene was concordant, then phylogenetic dissonance

TABLE 4. Results from analyses of the BLOODROOT data set

	Unique	$\hat{\varphi}$	H	$\hat{H}^*$	$\hat{I}$ (%)
(a) 5′ end (219 nucleotide sites, $\hat{D}=0.01320\%$)					
Run 1	3	1.000	2.70805	0.7023	74.07
Run 2	3	1.000	2.70805	0.6805	74.87
Average	3	1.000	2.70805	0.6914	74.47
Merged	3	1.000	2.70805	0.6915	74.47
(b) 3′ end (237 nucleotide sites, $\hat{D}=0.00769\%$)					
Run 1	3	1.000	2.70805	0.9054	66.56
Run 2	3	1.000	2.70805	0.8911	67.10
Average	3	1.000	2.70805	0.8983	66.83
Merged	3	1.000	2.70805	0.8983	66.83
(c) Concatenated (456 nucleotide sites, $\hat{D}=0.00000\%$)					
Run 1	1	1.000	2.70805	0.0000	100.00
Run 2	1	1.000	2.70805	0.0000	100.00
Average	1	1.000	2.70805	0.0000	100.00
Merged	1	1.000	2.70805	0.0000	100.00
(d) Run 1 across subsets ($\hat{D}=42.56\%$)					
5′ run 1	3	1.000	2.70805	0.7022	74.07
3′ run 1	3	1.000	2.70805	0.9054	66.56
Average	3	1.000	2.70805	0.8039	70.32
Merged	5	1.000	2.70805	1.3994	48.32
(e) Run 2 across subsets ($\hat{D}=43.27\%$)					
5′ run 2	3	1.000	2.70805	0.6805	74.87
3′ run 2	3	1.000	2.70805	0.8911	67.10
Average	3	1.000	2.70805	0.7858	70.98
Merged	5	1.000	2.70805	1.3853	48.85

Notes: $\hat{I}$ and $\hat{D}$ are expressed as % maximum. Note that in a-c $\hat{D}$ measures dissonance among replicate MCMC analyses and low values are used as an indication that both runs have converged with respect to tree topology, whereas in d-e $\hat{D}$ measures dissonance between data subsets and high values indicate incongruence.

should be low, yet $\hat{D}>42\%$ in two independent analyses comparing the 5′ subset to the 3′ subset (Table 1d,e), indicating that the information in the 5′ end of the gene disagrees strongly with the information in the 3′ end of the gene.

When D is high, the value of $\hat{I}$ from the merged tree sample will be lower than I estimated from the concatenated data set. Information content should go down, not up, when a data subset is added that conflicts with data already included. The more intuitive behavior of merged $\hat{I}$ (compared to concatenated $\hat{I}$) is evident in the results from the BLOODROOT data set: individual subsets have (averaging across two replicate analyses) $\hat{I}=74.5\%$ (5′) and $\hat{I}=66.8\%$ (3′), while $\hat{I}=48.6\%$ for the merged sample, contrasting sharply with $\hat{I}=100\%$ for the concatenated data.

The single tree topology obtained in the concatenated analysis is the only topology sampled by analyses of both 5′ and 3′ halves of the gene (Table 5). The best 5′ tree topology receives no posterior support in the 3′ analysis, and the best 3′ tree topology receives no posterior support in the 5′ analysis, so the concatenation tree topology is the only topology able to be tolerated by all sites, even though it is considered mediocre by every site.

TABLE 5. Marginal posterior probabilities of tree topologies for the 5′, 3′, and concatenated data sets

Topology	5′	3′	Merged	Concatenated
(D, O, (E, (B, S)))	0.7738	—	0.3869	—
(D, S, (O, (E, B)))	—	0.6436	0.3218	—
(D, O, (S, (B, E)))	0.1108	0.1760	0.1434	1.0000
(D, (S,O),(E, B))	—	0.1804	0.0902	—
(D, O, (B, (E, S)))	0.1155	—	0.0578	—

Notes: A dash (—) means that the marginal posterior probability was estimated to be zero. Taxon abbreviations: D = *Disporum*, O = *Oryza*, S = *Sanguinaria*, B = *Bocconia*, and E = *Eschscholzia*.

Alternative Tree Priors

There are two classes of models with respect to the marginal probability distribution of labeled rooted binary tree topologies: 1) equiprobable or proportional to distinguishable rearrangements (PDA) models; and 2) random-joining or equal rate Markov (ERM) models (Maddison and Slatkin 1991; Blum and François 2006). The methods described here assume a discrete uniform prior, which corresponds to the PDA model. Many Bayesian analyses use priors that jointly specify divergence times and topology in which the generating model for topology is an ERM model. For example, BEAST 2 (Bouckaert et al. 2014), MrBayes 3.2 (Ronquist et al. 2012), and RevBayes (Höhna et al. 2014) all allow birth–death (including the pure birth Yule model) and coalescent tree priors. Gernhard (2006) showed that the KL divergence from the discrete uniform model (probability distribution $\mathbf{p}$) over all rooted binary trees having n taxa to the pure-birth Yule model (probability distribution $\mathbf{p}_Y$) is

$$\mathrm{KL}(\mathbf{p}_{\mathrm{Y}},\mathbf{p}) = \log[(2n-3)!!]-n\sum_{k=2}^{n-1}\frac{g(k)}{k+1}=H(\mathbf{p})-H(\mathbf{p}_Y)$$
$$g(k) = \frac{1-k}{k}\log\frac{k-1}{2}+\log\frac{k}{2}+\log(k+1)-\frac{1}{k}\log(k!). \quad (7)$$

The Lindley information for a posterior distribution $\mathbf{p}^*$ relative to a Yule prior requires subtracting (7) from (2)

$$\begin{aligned} I &= H(\mathbf{p}_Y)-H(\mathbf{p}^*) \\ &= \big[H(\mathbf{p})-H(\mathbf{p}^*)\big]-\big[H(\mathbf{p})-H(\mathbf{p}_Y)\big] \\ &= \mathrm{KL}(\mathbf{p}^*,\mathbf{p})-\mathrm{KL}(\mathbf{p}_{\mathrm{Y}},\mathbf{p}). \end{aligned}$$

Coalescent tree priors and birth–death priors differ from the Yule model in their distribution of sojourn times between speciation events; however, the marginal distribution of tree topologies is identical, allowing (7) to be used for both. Lindley information can thus be easily computed for most tree priors in common use.

Challenges

Multimodal posteriors and conditional clade probabilities.—Whidden and Matsen (2015) showed that when the marginal posterior distribution of tree topologies is multimodal, the "mix-and-match" assumption made by Larget's conditional clade approximation is violated and the conditional clade approach will tend to underestimate the probability of tree topologies at the peaks and overestimate the probability of trees in the valleys. Such smoothing tends to overestimate the entropy of the marginal posterior distribution of tree topologies and thus underestimate information content. The degree to which this affects information content estimation depends on the details of a particular data set, but clearly multimodal posteriors are expected to bias information content estimation in a predictable direction when using the conditional clade distribution to estimate posterior probabilities. Software tools such as those provided by Whidden and Matsen (2015) can be used to assess the number and distinctness of islands in tree space if multimodality is a concern. An assumption made here is that the extent of the bias due to multimodal posteriors is less than the bias caused by using tree topology frequencies directly (i.e., assuming 100% coverage).

Another potential drawback of using the conditional clade distribution is that any tree topology containing a clade not sampled in the posterior will necessarily receive zero posterior probability using the conditional clade estimate. Thus, even though some tree topologies not present in the posterior sample are accounted for using the conditional clade distribution, there are potentially many tree topologies that are relevant but not considered. We note, however, that the simple frequency approach also ignores any tree topology containing a clade not sampled in the posterior.

Performance when information content is low.—The most challenging test case for any estimator of topological information content occurs when there is zero information (Table 2). In Bayesian phylogenetics, this situation occurs when the marginal posterior distribution of tree topologies is identical to the marginal prior distribution of tree topologies, which is assumed to be discrete uniform across all possible distinct tree topologies. For even moderately small problems, the space of tree topologies is sufficiently vast that an MCMC sample of any reasonable size fails to sample even a small fraction of possible tree topologies. For example, an MCMC sample of size 1 billion could, at most, include less than 1% of the prior when the study includes only 14 taxa!

Using the conditional clade distribution greatly improves estimation of I (Table 2); however, the case of low information content remains quite difficult, as illustrated by the psaB second codon position example. Information can be low for two very different reasons, similar to the way images can have low information content because of over- or underexposure: 1) sequences are saturated (overexposed) and thus noise from multiple substitutions obliterates any historical signal; and 2) sequences have very low variability (underexposed) due to short divergence time or low substitution rate. Polytomy priors (Lewis et al. 2005) may be helpful in both of these scenarios by allowing the posterior distribution to spread out over trees containing polytomies, including the "star tree" containing just one internal node. Dividing posterior probability assigned to polytomous trees among the fully-resolved tree topologies they subtend allows an MCMC sample of a given size to extend its reach beyond what could be achieved by restricting proposals to only fully-resolved tree topologies.

Future Directions

We have described an entropy-based measure of phylogenetic (topological) information content and suggested two possible applications (assessing degree of saturation and detecting topological conflict among data subsets), but we envision many more uses of I and D. For example, phylogenetic dissonance (D) serves as a sensitive measure of convergence in tree topology. Posterior samples from independent Bayesian MCMC analyses using the same model and data should ideally have zero phylogenetic dissonance because the information in the average tree sample should be indistinguishable from the average information from separate tree samples (see Tables 3a,b,c and 4a,b,c for examples of D used for this purpose). Bayesian analyses in which apparent convergence is assessed only by viewing trace plots of the log-likelihood may miss lack of convergence in tree topology (Nylander et al. 2008). Phylogenetic dissonance provides an alternative to the standard deviation of split frequencies (Ronquist et al. 2012), AWTY (Nylander et al. 2008), and the topological Gelman–Rubin-like measure proposed by Whidden and Matsen (2015) for assessing tree topology convergence in Bayesian analyses.

Information estimation (I) would also be useful in assessing the impact of missing data, a topic of recent concern in phylogenomics (Roure et al. 2013), and in assessing the information thrown away by site-stripping methods such as OV (Goremykin et al. 2010). The ability to partition I by clade allows for its use in phylogenetic profiling (assessing for which time period a particular data subset provides the most information), and the ability to partition D could help in identifying which parts of the phylogeny are responsible for incongruence among data subsets in multigene analyses. Clearly, all of these future directions depend on accurate information content estimation, and improvements in estimation accuracy for low-information data sets (using polytomy priors, for example) is thus our highest priority in the short term.

Software Availability

The software Galax can be used to estimate information content from tree files representing samples from the Bayesian phylogenetic marginal posterior distribution of trees. Galax can be obtained from http://phylogeny.uconn.edu/software/ or https://github.com/plewis/galax.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.1dn50.

Funding

This material is based upon work supported by the National Science Foundation [grant number DEB-1354146 to P.O.L., M.H.C., L.K., and L.A.L. and DEB-1036448 to L.A.L. and P.O.L.]. M.H.C.'s research was also partially supported by National Institutes of Health [grant numbers GM 70335 and P01 CA142538].

Acknowledgements

The authors wish to thank Frank Anderson, Tom Near, Mark Holder, Tanja Stadler, reviewers Erik Matsen, Jeremy Brown and Jeff Townsend, as well as one anonymous reviewer, for their very helpful comments. This study benefited from computing resources made available through the Bioinformatics Facility of the University of Connecticut Biotechnology Center and the Booth Engineering Center for Advanced Technology (School of Engineering, University of Connecticut).

Appendix 1

Algorithm 1 Computing entropy using conditional clade distribution.

```
procedure CalcEntropy(C)
    $H \leftarrow 0$
    for all $C_1, C_2 | C$ do    ▷ $C_1$ and $C_2$ are child clades in a bipartition of $C$
        $p \leftarrow \Pr(C_1, C_2 | C)$
        $H \leftarrow H - p\{\log p - \textsc{CalcEntropy}(C_1) - \textsc{CalcEntropy}(C_2)\}$
    end for
    return H
end procedure
```

Algorithm 2 Computing information specific to one clade

```
procedure CalcInfoForClade(C)
    $p_C \leftarrow p^*(C)$    ▷ $p^*(C)$ is marginal posterior for clade $C$
    $KL_C \leftarrow 0$    ▷ $KL_C$ = Kullback-Leibler divergence for clade $C$
    for all $C_1, C_2 | C$ do    ▷ $C_1$ and $C_2$ are child clades in a bipartition of $C$
        $KL_C \leftarrow p^*(C^{(l)}, C^{(r)}|C)\left(\log p^*(C^{(l)}, C^{(r)}|C) - \log p(C^{(l)}, C^{(r)}|C)\right)$
    end for
    return $p_C$ $KL_C$
end procedure
```

Appendix 2

Definitions

$$S = \text{the set of all taxa}$$

$$\tau = \text{a rooted, binary tree topology on } S$$

$$\mathbb{T} = \text{the set of all distinct rooted, binary tree topologies on } S$$

$$C = \text{a subset of } S \text{ (i.e., a clade)}$$

$$C^{(l)}, C^{(r)} = \text{a bipartition of clade } C \text{ such that } |C^{(l)}| > 0, |C^{(r)}| > 0, C^{(l)} \cup C^{(r)} = C, \text{ and } C^{(l)} \cap C^{(r)} = 0$$

$$\mathbb{C} = \text{the set of all possible triplets } C, C^{(l)}, C^{(r)}$$

$$\mathbb{C}_C = \text{the set of all possible combinations } C^{(l)}, C^{(r)} \text{ given parent clade } C$$

$$\mathbb{C}_t = \text{the subset of } \mathbb{C} \text{ associated with subtrees in } \tau$$

$$C^{(l)}, C^{(r)}|C = \text{the set of all possible } C^{(l)}, C^{(r)} \text{ pairs having parent clade } C$$

$$\mathbf{1}_x = \text{indicator variable: equals 1 if } x \text{ is true and 0 otherwise}$$

Partitioning Entropy Among Clades

Here we show that the entropy H of a probability distribution on tree topologies is approximated by a weighted sum of clade-specific entropies, H_C, where weights are marginal clade probabilities, $p(C)$:

$$H = \sum_{C \in S} p(C) H_C.$$

Proof. The proof makes use of the following result from Larget (2013), which shows that the probability of a tree, $p(\tau)$, is approximated by a product of conditional clade probabilities:

$$p(\tau) \approx \prod_{C, C^{(l)}, C^{(r)} \in \mathbb{C}_T} p(C^{(l)}, C^{(r)}|C)$$

$$
\begin{aligned}
H &= -\sum_{\tau\in\mathbb{T}} p(\tau)\log p(\tau)\\
&\approx -\sum_{\tau\in\mathbb{T}} p(\tau)\log\left[\prod_{C,C^{(l)},C^{(r)}\in\mathbb{C}_t} p(C^{(l)},C^{(r)}|C)\right]\\
&= -\sum_{\tau\in\mathbb{T}} p(\tau)\sum_{C,C^{(l)},C^{(r)}\in\mathbb{C}_t}\log p(C^{(l)},C^{(r)}|C)\\
&= -\sum_{\tau\in\mathbb{T}} p(\tau)\sum_{C,C^{(l)},C^{(r)}\in\mathbb{C}}\log p(C^{(l)},C^{(r)}|C)\,\mathbf{1}_{C,C^{(l)},C^{(r)}\in\mathbb{C}_t}\\
&= -\sum_{C,C^{(l)},C^{(r)}\in\mathbb{C}}\log p(C^{(l)},C^{(r)}|C)\sum_{\tau\in\mathbb{T}} p(\tau)\,\mathbf{1}_{C,C^{(l)},C^{(r)}\in\mathbb{C}_t}\\
&= -\sum_{C,C^{(l)},C^{(r)}\in\mathbb{C}} p(C,C^{(l)},C^{(r)})\log p(C^{(l)},C^{(r)}|C)
\end{aligned}
$$

$$
\begin{aligned}
&= -\sum_{C,C^{(l)},C^{(r)}\in\mathbb{C}} p(C)p(C^{(l)},C^{(r)}|C)\log p(C^{(l)},C^{(r)}|C)\\
&= -\sum_{C\in S} p(C)\sum_{C^{(l)},C^{(r)}|C\in\mathbb{C}_C} p(C^{(l)},C^{(r)}|C)\log p(C^{(l)},C^{(r)}|C)\\
&= \sum_{C\in S} p(C)H_C.
\end{aligned}
$$

■

Partitioning Information among Clades

Here we show that total information can be approximated using the conditional clade distribution as a weighted sum of clade-specific Kullback-Leibler divergences. Note that (A.1) assumes that the marginal prior probability of any given tree topology τ is a constant: $p(\tau)=1/|\mathbb{T}|$.

Proof.

$$
\begin{aligned}
I &= H-H^*\\
&= \left(-\sum_{\tau\in\mathbb{T}} p(\tau)\log p(\tau)\right)-\left(-\sum_{\tau\in\mathbb{T}} p^*(\tau)\log p^*(\tau)\right)\\
&= \left(-\log p(\tau)\cancelto{1}{\sum_{\tau\in\mathbb{T}} p(\tau)}\right)-\left(-\sum_{\tau\in\mathbb{T}} p^*(\tau)\log p^*(\tau)\right) \qquad \text{(A.1)}\\
&= \left(-\log p(\tau)\sum_{\tau\in\mathbb{T}} p^*(\tau)\right)-\left(-\sum_{\tau\in\mathbb{T}} p^*(\tau)\log p^*(\tau)\right)\\
&= \left(\sum_{\tau\in\mathbb{T}} p^*(\tau)\log p^*(\tau)\right)-\left(\sum_{\tau\in\mathbb{T}} p^*(\tau)\log p(\tau)\right)
\end{aligned}
$$

$$
\begin{aligned}
&\approx \left(\sum_{C\in S} p^*(C)\sum_{C^{(l)},C^{(r)}|C\in\mathbb{C}_C} p^*(C^{(l)},C^{(r)}|C)\log p^*(C^{(l)},C^{(r)}|C)\right)\\
&\quad -\left(\sum_{C\in S} p^*(C)\sum_{C^{(l)},C^{(r)}|C\in\mathbb{C}_C} p^*(C^{(l)},C^{(r)}|C)\log p(C^{(l)},C^{(r)}|C)\right)\\
&= \sum_{C\in S} p^*(C)\sum_{C^{(l)},C^{(r)}|C\in\mathbb{C}_C} p^*(C^{(l)},C^{(r)}|C)\log\left(\frac{p^*(C^{(l)},C^{(r)}|C)}{p(C^{(l)},C^{(r)}|C)}\right)\\
&= \sum_{C\in S} p^*(C)\,\mathrm{KL}_C
\end{aligned}
$$

■

References

Allen B., Kon M., Bar-Yam Y. 2009. A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. Am. Nat. 174:236–243.

Archie J.W. 1989. A randomization test for phylogenetic information in systematic data. Syst. Zool. 38:239–252.

Bai F., Xu J., Liu L. 2013. Weighted relative entropy for phylogenetic tree based on 2-step Markov model. Math. Biosci. 246:8–13.

Bergthorsson U., Adams K.L., Thomason B., Palmer J.D. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424:197–201.

Bernardo J.M., A.F.M. Smith. 1994. Bayesian theory. John Wiley & Sons.

Blum M.G.B., François O. 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst. Biol. 55:685–691.

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comp. Biol. 10:e1003537.

Brown J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.

Carstens B.C., Knowles L.L. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst. Biol. 56: 400–411.

Cover T.M., Thomas J.A. 2006. Elements of information theory. 2nd ed. John Wiley & Sons.

Cummins C.A., McInerney J.O. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst. Biol. 60:833–844.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl Acad. Sci. USA 104:5936–5941.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Faith D.P. 1991. Cladistic permutation tests for monophyly and nonmonophyly. Syst. Zool. 40:366–375.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evol. 39:783–791.

Fischer M., Steel M. 2009. Sequence length bounds for resolving a deep phylogenetic divergence. J. Theoret. Biol. 256:247–252.

Fučíková K., Leliaert F., Cooper E.D., Škaloud P., D'Hondt S., De Clerck O., Gurgel C.F.D., Lewis L.A., Lewis P.O., Lopez-Bautista J.M., Delwiche C.F., Verbruggen H. 2014a. New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. Frontiers Ecol. Evol. 2:1–12.

Fučíková K., Lewis P.O., Lewis L.A. 2014b. Putting *incertae sedis* taxa in their place: a proposal for ten new families and three new genera in sphaeropleales (chlorophyceae, chlorophyta). J. Phycol. 50:14–25.

Gernhard T. 2006. Stochastic models for speciation events in phylogenetic trees. [Ph.D. dissertation] Technische Universität München.

Geuten K., Massingham T., Darius P., Smets E., Goldman N. 2007. Experimental design criteria in phylogenetics: where to add taxa. Syst. Biol. 56:609–622.

Goldman N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. Roy. Soc. Lond. B 265:1779–1786.

Goremykin V.V., Nikiforova S.V., Bininda-Emonds O.R.P. 2010. Automated removal of noisy data in phylogenomic analyses. J. Mol. Evol. 71:319–331.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Hillis D.M., Huelsenbeck J.P. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered. 83:189–195.

Höhna S., Drummond A.J. 2011. Guided tree topology proposals for Bayesian phylogenetic inference. Syst. Biol. 61:1–11.

Höhna S., Heath T.A., Boussau B., Landis M.J., Ronquist F., Huelsenbeck J.P. 2014. Probabilistic graphical model representation in phylogenetics. Syst. Biol. 63:753–771.

Kluge A., Farris J. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1–32.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Larget B. 2013. The estimation of tree posterior probabilities using conditional clade probability distributions. Syst. Biol. 62: 501–511.

Leaché A.D., Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60:126–137.

Lemey P., Rambaut A., Drummond A.J., Suchard M.A. 2009. Bayesian phylogeography finds its roots. PLoS Comp. Biol. 5:e1000520.

Lemmon E.M., Lemmon A.R. 2013. High-throughput genomic data in systematics and phylogenetics. Ann. Rev. Ecol. Evol. Syst. 44: 99–121.

Lewis P.O., Holder M.T., Holsinger K.E. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253.

Lindley D.V. 1956. On a measure of the information provided by an experiment. Ann. Math. Stat. 27:986–1005.

Liu Y., Cox C.J., Wang W., Goffinet B. 2014. Mitochondrial phylogenomics of early land plants: mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. Syst. Biol. 63:862–878.

Lyons-Weiler J., Hoelzer G.A., Tausch R.J. 1996. Relative apparent synapomorphy analysis (RASA). I: the statistical measurement of phylogenetic signal. Mol. Biol. Evol. 13:749–757.

Maddison W.P., Slatkin M. 1991. Null models for the number of evolutionary steps in a character on a phylogenetic tree. Evolution 45:1184–1197.

Massingham T., Goldman N. 2000. EDIBLE: experimental design and information calculations in phylogenetics. Bioinformatics 16: 294–295.

Mauro D.S., Gower D.J., Massingham T., Wilkinson M., Zardoya R., Cotton J.A. 2009. Experimental design in caecilian systematics: phylogenetic information of mitochondrial genomes and nuclear rag1. Syst. Biol. 58:425–438.

Mossel E., Vigoda E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309:2207–2209.

Nguyen M.A.T., Klaere S., von Haeseler A. 2011. MISFITS: evaluating the goodness of fit between a phylogenetic model and an alignment. Mol. Biol. Evol. 28:143–152.

Nylander J., Wilgenbusch J., Warren D., Swofford D.L. 2008. AWTY(are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581–583.

Parks M., Cronn R., Liston A. 2012. Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). BMC Evol. Biol. 12:100.

Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13:235–238.

Rannala B., Zhu T., Yang Z. 2012. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. Mol. Biol. Evol. 29:325–335.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Pop. Biol. 100:56–62.

Ronquist F., Teslenko M., Van Der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61: 539–542.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30:197–214.

Shannon C.E. 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27:379–423, 623–656.

Shi X., Gu H., Field C. 2008. Pattern classification of phylogeny signals. Stat. Appl. Genet. Mol. Biol. 7:1–20.

Shpak M., Churchill G.A. 2000. The information content of a character under a Markov model of evolution. Mol. Phylogenet. Evol. 17: 231–243.

Steel M.A., Lockhart P.J., Penny D. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440–442.

Steel M., Lockhart P.J., Penny D. 1995. A frequency-dependent significance test for parsimony. Mol. Phylogenet. Evol. 4:64–71.

Swofford D.L. 2003. PAUP*. phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Tippery N.P., Fučíková K., Lewis P.O., Lewis L.A. 2012. Probing the monophyly of the sphaeropleales (chlorophyceae) using data from five genes. J. Phycol. 48:1482–1493.

Townsend J.P., Su Z., Tekle Y.I. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst. Biol. 61:835–849.

Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Wang H.-C., Susko E., Roger A.J. 2011. Fast statistical tests for detecting heterotachy in protein evolution. Mol. Biol. Evol. 28:2305–2315.

Weisrock D.W., Smith S.D., Chan L.M., Biebouw K., Kappeler P.M., Yoder A.D. 2012. Concatenation and concordance in the reconstruction of mouse lemur phylogeny: an empirical demonstration of the effect of allele sampling in phylogenetics. Mol. Biol. Evol. 29:1615–1630.

Whidden C., Matsen F.A. 2015. Quantifying MCMC exploration of phylogenetic tree space. Syst. Biol. 64:472–491.

Xia X., Xie Z., Salemi M., Chen L., Wang Y. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26:1–7.

Xia X. 2009. Assessing substitution saturation with DAMBE. In: Lemey P., Salemi M., Vandamme A.-M., editors. The phylogenetic handbook: a practical approach to phylogenetic analsysis and hypothesis testing. 2nd ed. Cambridge: Cambridge University Press, p. 613–629.

Xi Z., Liu L., Rest J.S., Davis C.C. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. 63:919–932.

Xi Z., Rest J.S., Davis C.C. 2013. Phylogenomics and coalescent analyses resolve extant seed plant relationships. PLoS One 8:e80870.

Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.

Zhong B., Deusch O., Goremykin V.V., Penny D., Biggs P.J., Atherton R.A., Nikiforova S.V., Lockhart P.J. 2011. Systematic error in seed plant phylogenomics. Genome Biol. Evol. 3:1340–1348.

Zhong B., Xi Z., Goremykin V.V., Fong R., McLenachan P.A., Novis P.M., Davis C.C., Penny D. 2013. Streptophyte algae and the origin of land plants revisited using heterogeneous models with three new algal chloroplast genomes. Mol. Biol. Evol. 31:177–183.