Estimating Bayesian Phylogenetic Information Content

PAUL O. LEWIS1,*, MING-HUI CHEN2, LYNN KUO2, LOUISE A. LEWIS1, KAROLINA FUCHKOVÁ1, SUMANT NEUPANE1, YU-BO WANG2, AND DAOYUAN SHI2

1Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, CT 06269, USA; 2Department 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

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 prior distribution compared with the entropy of the posterior 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 2003; 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; Fučková 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. 2001). 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 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.

Lindley’s Information Measure

The posterior distribution of an unknown quantity $\theta$ can be viewed as the prior distribution of $\theta$ updated with information from data. If there is no information in the data relevant to $\theta$, 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 $\theta$ less plausible than they are under the prior distribution, yielding a posterior that is concentrated compared with the prior.

In the context of phylogenetics, $\theta$ represents a (rooted or unrooted) tree topology. For example, $\theta$ 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 (log 15) 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:

\[
H = H(p) = - \sum_{\tau \in T} p(\tau) \log p(\tau)
\]

\[
H^* = H(p^*) = - \sum_{\tau \in T} p^*(\tau) \log p^*(\tau)
\]

\[
I = H - H^*.
\]

(1)

\(T\) is the set of all distinct binary labeled tree topologies, \(\tau\) represents a realization of the random variable \(T\) representing a tree topology, \(|T|\) represents the size of the set \(T\) (i.e., the number of possible binary labeled tree topologies), \(p^*\) is the marginal posterior distribution, with \(p^*(\tau)\) the marginal posterior probability of tree topology \(\tau\), and \(p\) is the marginal prior distribution, with \(p(\tau)\) the marginal prior probability of the specific tree topology \(\tau\). 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 \(T\) of all distinct tree topologies, \(I\) ranges from 0.0 (no information about topology) to \(\log|T|\) (maximum information; all posterior probability concentrated in a single tree topology), and, when expressed as a percentage, it should be understood that \(100\%\) implies \(I = \log|T|\).

Relationship to Kullback–Leibler Divergence

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

\[
KL(p^*; p) = \sum_{\tau \in T} p^*(\tau) \log \frac{p^*(\tau)}{p(\tau)}
\]

(2)

\[
= \sum_{\tau \in T} p^*(\tau) \log p^*(\tau) - \log p(\tau) \sum_{\tau \in T} p^*(\tau)
\]

\[
= \sum_{\tau \in T} p^*(\tau) \log p^*(\tau) - \sum_{\tau \in T} p^*(\tau) \log p^*(\tau)
\]

\[
= H - H^*.
\]

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 \(y\) denote the data on which the posterior distribution \(p^*\) is based,

\[
KL(p^*; p) = \sum_{\tau \in T} p^*(\tau) \log \frac{p^*(\tau)}{p(\tau)}
\]

(2)

\[
= \sum_{\tau \in T} p^*(\tau) \log \frac{p^*(\tau)}{p(\tau)}
\]

\[
= \sum_{\tau \in T} p^*(\tau) \log \frac{p^*(\tau)}{p(\tau)}
\]

\[
= \sum_{\tau \in T} p(\tau) \log p(\tau) - \log p(\tau) \sum_{\tau \in T} p(\tau)
\]

\[
= E_T \{\log p(\tau)\} - \log p(\tau).
\]

where \(p(\tau)\) is the likelihood marginalized over all model parameters on a fixed tree topology \(\tau\), and \(p(\tau)\) 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
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, $\psi$, 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 $\psi < 1$ because the posterior distribution appears to be more concentrated than it really is. Ideally, a measure of entropy used in computing $I$ should take into account 100% rather than (100 × $\psi$)% 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, $\psi$ 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 $\psi < 1$. 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 ($T$) for seven taxa contains $|T| = 10,395$ 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:

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

\[
H = -\sum_{i=1}^{100395} \frac{1}{10395} \log \left( \frac{1}{10395} \right) = \log(10395) = 9.24908
\]

\[
\hat{H}^* = H - \hat{H}^* = 7.96237 (86.1\%)
\]

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

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

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, \( \hat{\phi} = 0.863 \), 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:

\[
\hat{H}^* = -\frac{630}{1225} \log \left( \frac{630}{1225} \right) + \frac{45}{1225} \log \left( \frac{45}{1225} \right) + \frac{252}{1225} \log \left( \frac{252}{1225} \right) + \frac{18}{1225} \log \left( \frac{18}{1225} \right) + \frac{112}{1225} \log \left( \frac{112}{1225} \right) + \frac{60}{1225} \log \left( \frac{60}{1225} \right)
\]

\[
= 1.47793
\]

\[H = \log(10395) = 9.24908\]

\[
\hat{I} = H - \hat{H}^* = 7.77115 (84.0\%)
\]

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):

\[
\hat{H}^* = \hat{H}_{\text{arc}} + \frac{84}{100} \left[ \hat{H}_{\text{arc}} + \left( \frac{12}{84} \hat{H}_{\text{arc}} \right) \right] + \frac{16}{100} \left[ \hat{H}_{\text{arc}} + \hat{H}_{\text{arc}} \right]
\]

(3)

where

\[
\hat{H}_{\text{arc}} = -\left[ \frac{84}{100} \log \left( \frac{84}{100} \right) + \frac{16}{100} \log \left( \frac{16}{100} \right) \right] = 0.43967
\]

\[
\hat{H}_{\text{arc}} = -\left[ \frac{60}{84} \log \left( \frac{60}{84} \right) + \frac{24}{84} \log \left( \frac{24}{84} \right) \right] = 0.59827
\]

\[
\hat{H}_{\text{arc}} = -\left[ \frac{72}{84} \log \left( \frac{72}{84} \right) + \frac{12}{84} \log \left( \frac{12}{84} \right) \right] = 0.41012
\]

\[
\hat{H}_{\text{arc}} = -\left[ \frac{3}{7} \log \left( \frac{3}{7} \right) + \frac{4}{7} \log \left( \frac{4}{7} \right) \right] = 0.68291
\]

\[
\hat{H}_{\text{arc}} = -\left[ 1.0 \log (1.0) \right] = 0.0.
\]

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

\[
\hat{H}^* = \hat{H}_{\text{arc}} + \frac{84}{100} \hat{H}_{\text{arc}} + \frac{84}{100} \hat{H}_{\text{arc}} + \frac{28}{100} \hat{H}_{\text{arc}} + \frac{16}{100} \hat{H}_{\text{arc}}
\]

\[
= 1.47793.
\]
More generally, Lindley information itself can be expressed as a sum of clade-specific contributions:

\[ I = \sum_{C \in S} p^*(C) KL_C \]  
\[ KL_C = \sum_{C^0, C^0 C \in S} p^*(C^0, C^0|C) \log \left( \frac{p^*(C^0, C^0|C)}{p(C^0, C^0|C)} \right). \]

where clade \( C \) is a nonempty subset of the set \( S \) of all taxa, \( p^*(C) \) is the marginal posterior probability of parent clade \( C \), \( p^*(C^0, C^0|C) \) is the marginal posterior probability of left \( (C^0) \) and right \( (C^0) \) child clades given the parent clade \( (C) \) using the conditional clade distribution derived from the posterior, and \( p(C^0, C^0|C) \) 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):

\[ p(ABCDEF) = 1.0. \]
\[ KL_{ABCDEF} = 84 \log \left( \frac{84}{100} \right) + 16 \log \left( \frac{16}{100} \right) = 5.00275, \]
\[ p(ABC) = \frac{84}{100}. \]
\[ KL_{ABC} = 60 \log \left( \frac{60}{84} \right) + 24 \log \left( \frac{24}{84} \right) = 0.50034, \]
\[ p(DEF) = \frac{84}{100}. \]
\[ KL_{DEF} = 70 \log \left( \frac{72}{84} \right) + 12 \log \left( \frac{12}{84} \right) = 2.14099, \]
\[ p(ABCD) = \frac{16}{100}. \]
\[ KL_{ABCD} = (1) \log \left( \frac{16}{100} \right) = 2.70805, \]
\[ p(EFG) = \frac{28}{100} = \frac{16}{100} + \frac{84}{100} \]
\[ KL_{EFG} = 3 \log \left( \frac{3}{7} \right) + \frac{4}{7} \log \left( \frac{4}{7} \right) = 0.41570. \]

The resulting estimated Lindley information,
\[ \hat{i} = (1.0)(5.00275) + (0.84)(0.50034) + (0.84)(2.14099) + (0.16)(2.70805) + (0.28)(0.41570) \]
\[ = 7.77115, \]

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 \( p_k^* \) be the posterior distribution for subset \( k \), and \( w_k \) be the weight associated with subset \( k \), Jensen’s inequality says

\[ \sum_{k=1}^{K} w_k H(p_k^*) \leq H \left( \sum_{k=1}^{K} w_k p_k^* \right) = H_{\text{merged}}. \]

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
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,

$$D = H^* - \sum_{k=1}^{K} w_k H(p_k^*)$$,

where $H(p_k^*)$ 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 $0 \leq D \leq H^*$.

This upper bound allows $D$, like $I$, to be expressed as a percentage of its maximum value. Assuming that the prior distribution $p$ is identical for each subset, $D$ may be defined equivalently in terms of Lindley information:

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

Example Data Sets

Two previously published data sets provide empirical examples of information content estimation. The data set ALGAE comprises chloroplast psaB 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 psaB 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 rps11 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
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:

- Tree topology $T \sim \text{Discrete Uniform}(1, |T|)$
- Tree length $L \sim \text{Exponential}(0.01)$
- Edge length proportions $e \sim \text{Dirichlet}(1, \ldots, 1)$
- Nucleotide frequencies $\pi \sim \text{Dirichlet}(1, 1, 1, 1)$
- Exchangeabilities $r \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)$
- Discrete Gamma shape $\alpha \sim \text{Exponential}(1)$
- Proportion invariable sites $p \sim \text{Uniform}(0, 1)$

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 Metropoliscoupled 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](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 with the prior's information. 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
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.

| No. sites | \( \hat{I} \) |
|-----------|-------|
| 1        | 0     |
| 10       | 4     |
| 100      | 90    |
| 1000     | 100   |

(a) Number of sites

| Rel. rate | \( \hat{I} \) |
|-----------|-------|
| 0.01      | 18    |
| 0.1       | 99    |
| 1         | 100   |
| 10        | 64    |
| 100       | 1.5   |

(b) Substitution rate

| missing (%) | \( \hat{I} \) |
|-------------|-------|
| 0           | 100   |
| 50          | 98    |
| 90          | 0     |
| 100         | 0     |

(c) Missing data

| Rate variance | \( \hat{I} \) |
|---------------|-------|
| 1             | 100   |
| 10            | 97    |
| 100           | 13    |
| 1000          | 0     |

(d) Rate heterogeneity

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.

**Table 1.** Results of four-taxon simulations showing information content.

| Taxa | Trees | \( \hat{I} \) | \( \hat{I}_{true} \) |
|------|-------|-------|------------------|
| 4    | 3     | 1.00  | 0.0301           |
| 5    | 15    | 1.00  | 0.02398          |
| 6    | 105   | 1.00  | 0.07105          |
| 7    | 945   | 1.00  | 0.1835           |
| 8    | 10395 | 0.6539| 0.4840           |
| 9    | 13513 | 0.9934| 1.235            |
| 10   | 202705| 0.0010| 3.404            |
| 11   | 34459425| 0.00135| 7.304          |
| 12   | 654729075| 0.00040| 13.30           |

Notes: Results from paired runs in a STANDARD analysis were averaged. Coverage \( \hat{I} \) 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 \( \hat{I}_{true} \) 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 (\( \hat{\phi} = 0.00026 \)). 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
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 Ramalha 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 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 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 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 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 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 three unique tree topologies and an information content of 74.47%.

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

| TABLE 3. Results from analyses of the ALGAE data set |
|-----------------------------------------------|
| Unique | $\hat{q}$ | $H$ | $\hat{I}$ | $\hat{I}$ % |
|-------|--------|----|--------|----------|
| (a) First codon positions only ($\hat{D} = 0.7392\%$) |
| Run 1 | 9627 | 0.03646 | 97.29 | 15.73 | 83.83 |
| Run 2 | 5993 | 0.04004 | 97.29 | 15.66 | 83.91 |
| Average | 9610 | 0.03810 | 97.29 | 15.69 | 83.87 |
| Merged | 18510 | 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 | 75.99 |
| (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 |

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 Ramalha 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 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 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 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 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 three unique tree topologies and an information content of 74.47%.

The 100% information content in the concatenated data set (Fig. 6c) 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 Ramalha 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 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 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 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 three unique tree topologies and an information content of 74.47%.

The 100% information content in the concatenated data set (Fig. 6c) is evident in the better resolution of the third position majority rule consensus (Fig. 6a,b).
merged sample, contrasting sharply with analysis is the only topology sampled by analyses of concatenated data.

\[ I = H(p_Y) \]

\[ g(k) = \frac{1-k}{k} \log \left( \frac{k-1}{2} + \log \left( k+1 \right) - \frac{1}{k} \log(k!) \right). \]

(7)

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.

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 model, 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 (Hohna et al. 2014) all allow birth–death (including the pure-birth Yule model) and coalescent tree priors. Gerнhard (2006) showed that the KL divergence from the discrete uniform model (probability distribution \( p_Y \)) over all rooted binary trees having \( n \) taxa to the pure-birth Yule model (probability distribution \( p_Y \)) is

\[ KL(p_Y, p^*) = \log[(2n-3)!] - n \sum_{k=2}^{n-1} \frac{g(k)}{k+1} = H(p_Y) - H(p^*) \]

\[ g(k) = \frac{1-k}{k} \log \left( \frac{k-1}{2} + \log \left( k+1 \right) - \frac{1}{k} \log(k!) \right). \]

(7)

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

\[ I = H(p_Y) - H(p^*) = \frac{1}{2} \log \left( \frac{k-1}{2} + \log \left( k+1 \right) - \frac{1}{k} \log(k!) \right) \]

(7)

The single tree topology obtained in the concatenated analysis is the only topology sampled by analyses of

### Table 4. Results from analyses of the **BLOODROOT** data set

| Topology | 5' | 3' | Merged | Concatenated |
|----------|----|----|--------|--------------|
| 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 |

### Table 5. Marginal posterior probabilities of tree topologies for the 5', 3', and concatenated data sets

| Topology | 5' | 3' | Merged | Concatenated |
|----------|----|----|--------|--------------|
| (D, (O, (B, E))) | 0.7728 | — | 0.3869 | — |
| (D, (S, (O, (E, B)))) | — | 0.6436 | 0.3218 | — |
| (D, (O, (B, (E, S)))) | 0.1108 | 0.1780 | 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 = Disparum, O = Oryza, S = Sanguinaria, B = Bocconia, and E = Escholzia.
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 approach. For example, 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 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 polytymous trees among the fully-resolved tree topologies they subdivide 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 OVL (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 estimation.
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 information specific to one clade

```plaintext
1: procedure CALCULATE_CLAUDE(C)
2: \( H \leftarrow 0 \)
3: for all \( C_l, C_r \) do if \( C_l \) and \( C_r \) are child clades in a bipartition of \( C \) then
4: \( p \leftarrow p(C_l, C_r) \)
5: \( H \leftarrow H - p \cdot \left( \log p - \text{CALCULATE_ENTROPY}(C_l) - \text{CALCULATE_ENTROPY}(C_r) \right) \)
6: end for
7: return \( H \)
8: end procedure
```

Algorithm 2 Computing information specific to one clade

```plaintext
1: procedure CALCULATE_CLAUDE(C)
2: \( p_{C} \leftarrow p^* (C) \) \( \triangleright \) \( p^* (C) \) is marginal posterior for clade \( C \)
3: \( KL_C \leftarrow 0 \) \( \triangleright \) \( KL_C \) is Kullback-Leibler divergence for clade \( C \)
4: for all \( C_l, C_r \) do if \( C_l \) and \( C_r \) are child clades in a bipartition of \( C \) then
5: \( KL_C \leftarrow KL_C + p(C_l, C_r) \cdot \log \left( \frac{p(C_l, C_r)}{p(C(l)) \cdot p(C(r))} \right) \)
6: end for
7: return \( p_{C} \cdot KL_C \)
8: end procedure
```

APPENDIX 2

Definitions

\( S \) = the set of all taxa

\( \tau \) = a rooted, binary tree topology on \( S \)

\( \mathcal{T} \) = the set of all distinct rooted, binary tree topologies on \( S \)

\( C \) = a subset of \( S \) (i.e., a clade)

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

\( \mathcal{C} \) = the set of all possible triplets \( C^0, C^1 \)

\( \mathcal{C}_C \) = the set of all possible combinations \( C^0, C^1 \)

\( C(l) \) = the set of all possible \( C^0, C^1 \) pairs having parent clade \( C \)

\( 1_x \) = indicator variable: equals 1 if \( x \) 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, \( p(C) \):

\[
H = \sum_{C \in \mathcal{C}} p(C) H_C.
\]

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

\[
p(t) \approx \prod_{C \in \mathcal{C}^0, C^0 \in \mathcal{C}_C} p(C_0, C^0 | C)
\]
\[ H = - \sum_{t \in \mathcal{T}} p(t) \log p(t) \]

\[
\approx - \sum_{t \in \mathcal{T}} p(t) \log \left[ \prod_{C \in \mathcal{S}, C \subset C} p(C^{(0)}, C^{(t)} | C) \right] \\
= - \sum_{t \in \mathcal{T}} p(t) \sum_{C \in \mathcal{S}, C \subset C} \log p(C^{(0)}, C^{(t)} | C) \\
= - \sum_{t \in \mathcal{T}} p(t) \sum_{C \in \mathcal{S}, C \subset C} \log p(C^{(0)}, C^{(t)} | C) \sum_{t \in \mathcal{T}} 1_{C \in \mathcal{S}, C \subset C} \\
= - \sum_{t \in \mathcal{T}} p(t) \sum_{C \in \mathcal{S}, C \subset C} \log p(C^{(0)}, C^{(t)} | C) \\
= - \sum_{C \in \mathcal{S}} p(C) \sum_{C \subset C} \sum_{t \in \mathcal{T}} p(C^{(0)}, C^{(t)} | C) \log p(C^{(0)}, C^{(t)} | C) \\
= - \sum_{C \in \mathcal{S}} p(C) \sum_{C \subset C} p(C^{(0)}, C^{(t)} | C) \log p(C^{(0)}, C^{(t)} | C) \\
= - \sum_{C \in \mathcal{S}} p(C) \sum_{C \subset C} p(C^{(0)}, C^{(t)} | C) \log p(C^{(0)}, C^{(t)} | C) \\
= \sum_{C \in \mathcal{S}} p(C) H_{C}. \]

**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 approximations using the conditional clade distribution \( H_{C} \).

**Proof.**

\[ I = H - H^* \]

\[
= - \sum_{t \in \mathcal{T}} p(t) \log p(t) - \left( - \sum_{t \in \mathcal{T}} p^*(t) \log p^*(t) \right) \\
= - \log p(t) \sum_{t \in \mathcal{T}} p^*(t) - \left( - \sum_{t \in \mathcal{T}} p^*(t) \log p^*(t) \right) \tag{A.1} \\
= - \log p(t) \sum_{t \in \mathcal{T}} p^*(t) - \left( - \sum_{t \in \mathcal{T}} p^*(t) \log p^*(t) \right) \\
= \left( \sum_{t \in \mathcal{T}} p^*(t) \log p^*(t) \right) - \left( \sum_{t \in \mathcal{T}} p^*(t) \log p(t) \right) \\
= \sum_{C \in \mathcal{S}} p^*(C) \sum_{C \subset C} p(C^{(0)}, C^{(t)} | C) \log p(C^{(0)}, C^{(t)} | C) \\
= \sum_{C \in \mathcal{S}} p^*(C) \sum_{C \subset C} p(C^{(0)}, C^{(t)} | C) \log p(C^{(0)}, C^{(t)} | C) \\
= \sum_{C \in \mathcal{S}} p^*(C) \sum_{C \subset C} p(C^{(0)}, C^{(t)} | C) \log p(C^{(0)}, C^{(t)} | C) \\
= \sum_{C \in \mathcal{S}} p^*(C) K_{C}. \]

**REFERENCES**

Allen B., Kon M., Bar-Yam Y. 2009. A new phylogenetic diversity measure generalizing the Shannon index and its application to phylostomous 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.

Berghorson 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.

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.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:302–310.

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.

Fužíková K., Lielaitė P., Cooper E.D., Skaloud P., D’Hondt S., De Clerck O., Gurfel 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.
Nylander J., Wilgenbusch J., Warren D., Swofford D.L. 2008. AWTY (are
Nguyen M.A.T., Klaere S., von Haeseler A. 2011. MISFITS: evaluating
Mossel E., Vigoda E. 2005. Phylogenetic MCMC algorithms are
Mauro D.S., Gower D.J., Massingham T., Wilkinson M., Zardoya R.,
Liu Y., Cox C.J., Wang W., Goffinet B. 2014. Mitochondrial
Lindley D.V. 1956. On a measure of the information provided by an
Lewis P.O., Holder M.T., Holsinger K.E. 2005. Polytomies and Bayesian
Lemmon E.M., Lemmon A.R. 2013. High-throughput genomic data
Larget B. 2013. The estimation of tree posterior probabilities using conditional cladage probability distributions. Syst. Biol. 62: 301–311.
Leche 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, Suchard M.A. 2009. Bayesian phylogeny analysis 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.
Larget B. 2008. The estimation of tree posterior probabilities using molecular phylogenetic analyses. J. Hered. 89:189–195.
Höhna S., Drummond A.J. 2014. Probabilistic graphical model representation in phylogenetics. Syst. Biol. 63:753–771.
Kluge A., Farris J. 1969. Quantitative phyletics and the evolution of morphological characters. 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.
Kluge A., Farris J., Hedges S. 1972. Quantitative systematic and historical studies on the Anura. Syst. Zool. 21:9–14.
Kluge A., Farris J., Hedges S. 1973. Phylogenetic analysis of higher vertebrates. Science 179:401–410.
Kluge A., Farris J., Hedges S. 1974. Phylogenetic analysis of higher vertebrates. Syst. Zool. 23:401–410.
Kluge A., Farris J., Hedges S. 1975. Phylogenetic analysis of higher vertebrates. Syst. Zool. 24:401–410.
Kluge A., Farris J., Hedges S. 1976. Phylogenetic analysis of higher vertebrates. Syst. Zool. 25:401–410.
Kluge A., Farris J., Hedges S. 1977. Phylogenetic analysis of higher vertebrates. Syst. Zool. 26:401–410.
Kluge A., Farris J., Hedges S. 1978. Phylogenetic analysis of higher vertebrates. Syst. Zool. 27:401–410.
Kluge A., Farris J., Hedges S. 1979. Phylogenetic analysis of higher vertebrates. Syst. Zool. 28:401–410.
Kluge A., Farris J., Hedges S. 1980. Phylogenetic analysis of higher vertebrates. Syst. Zool. 29:401–410.
Kluge A., Farris J., Hedges S. 1981. Phylogenetic analysis of higher vertebrates. Syst. Zool. 30:401–410.
Kluge A., Farris J., Hedges S. 1982. Phylogenetic analysis of higher vertebrates. Syst. Zool. 31:401–410.
Kluge A., Farris J., Hedges S. 1983. Phylogenetic analysis of higher vertebrates. Syst. Zool. 32:401–410.
Kluge A., Farris J., Hedges S. 1984. Phylogenetic analysis of higher vertebrates. Syst. Zool. 33:401–410.
Kluge A., Farris J., Hedges S. 1985. Phylogenetic analysis of higher vertebrates. Syst. Zool. 34:401–410.
Kluge A., Farris J., Hedges S. 1986. Phylogenetic analysis of higher vertebrates. Syst. Zool. 35:401–410.
Kluge A., Farris J., Hedges S. 1987. Phylogenetic analysis of higher vertebrates. Syst. Zool. 36:401–410.
Kluge A., Farris J., Hedges S. 1988. Phylogenetic analysis of higher vertebrates. Syst. Zool. 37:401–410.
Kluge A., Farris J., Hedges S. 1989. Phylogenetic analysis of higher vertebrates. Syst. Zool. 38:401–410.
Kluge A., Farris J., Hedges S. 1990. Phylogenetic analysis of higher vertebrates. Syst. Zool. 39:401–410.
Kluge A., Farris J., Hedges S. 1991. Phylogenetic analysis of higher vertebrates. Syst. Zool. 40:401–410.
Kluge A., Farris J., Hedges S. 1992. Phylogenetic analysis of higher vertebrates. Syst. Zool. 41:401–410.
Kluge A., Farris J., Hedges S. 1993. Phylogenetic analysis of higher vertebrates. Syst. Zool. 42:401–410.
Kluge A., Farris J., Hedges S. 1994. Phylogenetic analysis of higher vertebrates. Syst. Zool. 43:401–410.
Kluge A., Farris J., Hedges S. 1995. Phylogenetic analysis of higher vertebrates. Syst. Zool. 44:401–410.
Kluge A., Farris J., Hedges S. 1996. Phylogenetic analysis of higher vertebrates. Syst. Zool. 45:401–410.
Kluge A., Farris J., Hedges S. 1997. Phylogenetic analysis of higher vertebrates. Syst. Zool. 46:401–410.
Kluge A., Farris J., Hedges S. 1998. Phylogenetic analysis of higher vertebrates. Syst. Zool. 47:401–410.
Kluge A., Farris J., Hedges S. 1999. Phylogenetic analysis of higher vertebrates. Syst. Zool. 48:401–410.
Kluge A., Farris J., Hedges S. 2000. Phylogenetic analysis of higher vertebrates. Syst. Zool. 49:401–410.
Kluge A., Farris J., Hedges S. 2001. Phylogenetic analysis of higher vertebrates. Syst. Zool. 50:401–410.
Kluge A., Farris J., Hedges S. 2002. Phylogenetic analysis of higher vertebrates. Syst. Zool. 51:401–410.
Kluge A., Farris J., Hedges S. 2003. Phylogenetic analysis of higher vertebrates. Syst. Zool. 52:401–410.
Kluge A., Farris J., Hedges S. 2004. Phylogenetic analysis of higher vertebrates. Syst. Zool. 53:401–410.
Kluge A., Farris J., Hedges S. 2005. Phylogenetic analysis of higher vertebrates. Syst. Zool. 54:401–410.
Kluge A., Farris J., Hedges S. 2006. Phylogenetic analysis of higher vertebrates. Syst. Zool. 55:401–410.
Kluge A., Farris J., Hedges S. 2007. Phylogenetic analysis of higher vertebrates. Syst. Zool. 56:401–410.
Kluge A., Farris J., Hedges S. 2008. Phylogenetic analysis of higher vertebrates. Syst. Zool. 57:401–410.
Kluge A., Farris J., Hedges S. 2009. Phylogenetic analysis of higher vertebrates. Syst. Zool. 58:401–410.
Kluge A., Farris J., Hedges S. 2010. Phylogenetic analysis of higher vertebrates. Syst. Zool. 59:401–410.
Kluge A., Farris J., Hedges S. 2011. Phylogenetic analysis of higher vertebrates. Syst. Zool. 60:401–410.
Kluge A., Farris J., Hedges S. 2012. Phylogenetic analysis of higher vertebrates. Syst. Zool. 61:401–410.
Kluge A., Farris J., Hedges S. 2013. Phylogenetic analysis of higher vertebrates. Syst. Zool. 62:401–410.
Kluge A., Farris J., Hedges S. 2014. Phylogenetic analysis of higher vertebrates. Syst. Zool. 63:401–410.
Kluge A., Farris J., Hedges S. 2015. Phylogenetic analysis of higher vertebrates. Syst. Zool. 64:401–410.
Kluge A., Farris J., Hedges S. 2016. Phylogenetic analysis of higher vertebrates. Syst. Zool. 65:401–410.