Syst. Biol. 64(5):709–726, 2015

DOI:10.1093/sysbio/syv024
Advance Access publication May 20, 2015

Impacts of Terraces on Phylogenetic Inference

MICHAEL J. SANDERSON[1,*], MICHELLE M. MCMAHON[2], ALEXANDROS STAMATAKIS[1,3,4], DERRICK J. ZWICKL[1],
AND MIKE STEEL[5]

[1]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA;* [2]*School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA;* [3]*Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg 69118, Germany;* [4]*Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany;* [5]*Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand;*

Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA; E-mail: sanderm@email.arizona.edu.

Michael J. Sanderson and Mike Steel contributed equally to this article.

Received 26 September 2014; reviews returned 2 December 2014; accepted 15 April 2015
Associate Editor: Laura Kubatko

Abstract.—Terraces are sets of trees with precisely the same likelihood or parsimony score, which can be induced by missing sequences in partitioned multi-locus phylogenetic data matrices. The potentially large set of trees on a terrace can be characterized by enumeration algorithms or consensus methods that exploit the pattern of partial taxon coverage in the data, independent of the sequence data themselves. Terraces can add ambiguity and complexity to phylogenetic inference, particularly in settings where inference is already challenging: data sets with many taxa and relatively few loci. In this article we present five new findings about terraces and their impacts on phylogenetic inference. First, we clarify assumptions about partitioning scheme model parameters that are necessary for the existence of terraces. Second, we explore the dependence of terrace size on partitioning scheme and indicate how to find the partitioning scheme associated with the largest terrace containing a given tree. Third, we highlight the impact of terrace size on bootstrap estimates of confidence limits in clades, and characterize the surprising result that the bootstrap proportion for a clade, as it is usually calculated, can be entirely determined by the frequency of bipartitions on a terrace, with some bipartitions receiving high support even when incorrect. Fourth, we dissect some effects of prior distributions of edge lengths on the computed posterior probabilities of clades on terraces, to understand an example in which long edges "attract" each other in Bayesian inference. Fifth, we describe how assuming relationships between edge-lengths of different loci, as an attempt to avoid terraces, can also be problematic when taxon coverage is partial, specifically when heterotachy is present. Finally, we discuss strategies for remediation of some of these problems. One promising approach finds a minimal set of taxa which, when deleted from the data matrix, reduces the size of a terrace to a single tree. [Bootstrap; partitioned model; phylogenetics; posterior probability; terrace.]

Inferred phylogenetic trees with thousands to tens of thousands of species are becoming increasingly commonplace (Rabosky et al. 2013; Zanne et al. 2014) and can serve at least two purposes: quantifying and conveying the scale and breadth of biodiversity, and providing statistical power to distinguish between alternative models of evolution (Wiens 2011; Boettiger et al. 2012; Chamberlain et al. 2012; Goldberg and Igic 2012; Marazzi et al. 2012; Christin et al. 2013; Davis et al. 2013). Reconstruction of large trees entails many challenges (Sanderson 2007; Izquierdo-Carrasco et al. 2011; Liu et al. 2012), including a recently discovered one: "terraces" (Sanderson et al. 2011). A terrace is a region in tree space in which all trees have precisely the same likelihood and parsimony score, which adds ambiguity to the "landscape" of trees (Fig. 1) and complexity to tree inference. Although "islands" in this landscape have been discussed for many years (Maddison 1991; Salter 2001), terraces may have been overlooked among the inevitable small numerical differences that arise in computing the likelihood score on different trees, especially in large data sets (e.g., roundoff errors, as when $(a+b)+c \neq a+(b+c)$). Indeed, the issue of whether two trees have *exactly* the same likelihood rarely arises in modern phylogenetic inference for this reason.

In trying to improve the efficiency of heuristic searches in RAxML, Stamatakis and Alachiotis (2010) noted a situation in which the likelihoods of different trees could be precisely identical. Suppose a sequence alignment is partitioned into two loci with separate model parameters for each, and sequences for some taxa for each locus are missing. For any tree T, there is an "induced subtree" for each locus, obtained simply by pruning the taxa for which no data are present, and the overall likelihood is the product of the likelihoods for these subtrees. During tree search, tree T might be rearranged to T', but the induced subtrees might stay the same, and consequently the overall likelihoods of T and T' are identical (Fig. 2). We suggested the term "terrace" for the set of trees emerging in this setting and exploited a variety of results concerning subtrees and supertrees to characterize these terraces, which in some data sets can be quite large (Sanderson et al. 2011). In general, trees with equal or nearly equal optimality scores can arise for many reasons, including lack of variable sites, homoplasy, and missing data (Wilkinson 1995). In parsimony analysis, terraces can arise when there are missing data irrespective of partitioning scheme, since the latter do not influence how trees are scored.

Data may be missing for many reasons, ranging from sampling biases inherent in studies that mine GenBank (Driskell et al. 2004; Sanderson 2008) to more biological causes, as in the loss of plastid genes transferred to the nuclear genomes of some plants (Sabir et al. 2014), or the differential expression of genes found in EST libraries or transcriptomes (Letsch et al. 2012). Missing data affect phylogenetic tree reconstruction in many ways (Wilkinson 1995, 2003; Kearney 2002; Burleigh et al. 2009; Lemmon et al. 2009; Cho et al. 2011; Simmons and Freudenstein 2011; Wiens and Morrill 2011; Crawley and Hilu 2012; Simmons 2012a,b, 2014; Simmons and Goloboff 2013, 2014; Hinchliff and Roalson 2013; Roure et al. 2013; Siu-Ting et al. 2014), some of which depend on the specific data that are present. However, many properties of terraces depend only on the overall "taxon coverage" (Fig. 2), which is just the set of taxon sets representing taxa for which *any* data are present in different elements of the data partition (Steel and Sanderson 2010; Sanderson et al. 2010, 2011). In particular, terraces are unproblematic, having only a single tree, any time the taxon coverage is "decisive" (Sanderson et al. 2010). Not surprisingly, this formalism is related to earlier results on ambiguity arising in supertree construction (cf. "groves" of phylogenetic trees: Driskell et al. 2004; Ane et al. 2009; and the impact of "effective" vs."ineffective" overlap in taxon sets: Wilkinson and Cotton 2006).

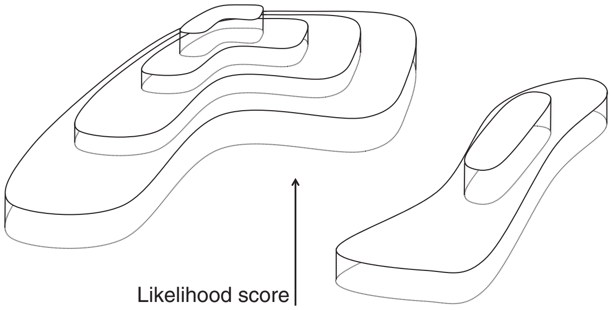
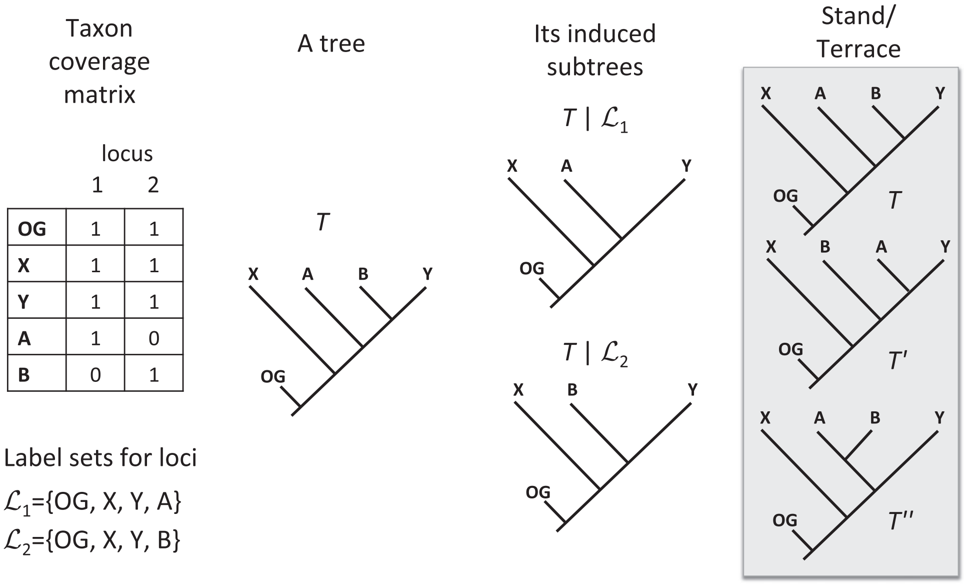


FIGURE 1. Schematic view of terraces in a likelihood surface of phylogenetic trees. The landscape shows two islands of trees separated by lower scores, but present on each island are regions of precisely equal score, terraces.

Although it can be computationally difficult to check for decisiveness, a model of randomly distributed missing data provides some clues about the likely properties of large data sets (Sanderson et al. 2010). Phylogenomic studies in which the number of loci greatly exceeds the number of taxa (like Hejnol et al. 2009: 94 taxa × 1487 loci; Salichos and Rokas 2013: 23 taxa × 1070 loci; Zwickl et al. 2014: 11 taxa × 473 loci) have a high probability of decisiveness, but when the number of taxa greatly exceeds the number of loci and missing data are common (e.g., Pyron and Wiens 2011: 2871 taxa × 12 loci; Smith et al. 2009: 55,473 taxa × 6 loci; Fabre et al. 2012: 1265 taxa × 11 loci; Rabosky et al. 2013: 7822 ray-finned fish taxa × 13 loci), the probability is low and large terraces are likely. The resulting increase in ambiguity poses challenges. Building on several basic mathematical properties of terraces we have characterized quantitatively (Sanderson et al. 2011), we extend our understanding of terraces in several directions, examining several new properties, the problems they induce, and strategies for overcoming them. We show that terraces can arise under more general conditions than we thought, and that they can be larger than believed before. We construct a method to identify the partitioning scheme that produces the maximal terrace for a given tree. Most importantly, we explore how terraces can affect confidence assessments and conventional views on how methods for characterizing ambiguity, such as consensus



	1	2
OG	1	1
X	1	1
Y	1	1
A	1	0
B	0	1

Label sets for loci

$\mathcal{L}_1$={OG, X, Y, A}

$\mathcal{L}_2$={OG, X, Y, B}

FIGURE 2. Stands and terraces. The two-locus taxon coverage matrix is at left (where '0' indicates missing data). Each locus has an associated "label set." For the given tree, T, two subtrees are induced by pruning taxa not in the respective label sets, denoted as $T|\mathcal{L}_1$ and $T|\mathcal{L}_2$. There are three trees at right that display them both. This collection of trees is a *stand*. The taxon coverage matrix in this example is thus not *decisive*. Because the parsimony scores (and likelihood scores, with an unlinked model), are identical on these three trees, the stand is also a *terrace*. All trees are assumed to be rooted with the aid of an outgroup, OG.

methods or bipartition support values, can be misled by them. Finally, we begin to extend results to the case in which the model partitioning scheme violates the sufficient conditions for terraces, but still generates data sets with patterns of ambiguity related to our other results on terraces.

Background

Definitions

Data.—Let D be a data matrix, generally assumed throughout to be a multiple sequence alignment, of n taxa ("rows") and l characters ("sites," "columns"), and let $\mathcal{P}$ be a "partitioning scheme" of the columns into m elements or blocks, referred to colloquially throughout as "loci," although blocks might be something like different sets of codon positions, etc. Let the *taxon coverage matrix*, $C=C(D,\mathcal{P})$, be an $n\times m$ matrix where the ij-th element is "1" if any sequence data are present for taxon i and locus j, and "0" if the data are entirely missing (Fig. 2). A "0" typically occurs because the locus was not sampled for that taxon. An entry of "1" does not imply that the data are *phylogenetically informative*—merely that they are not missing entirely. We use the phrase "partial taxon coverage" whenever C has at least one 0 in it.

The taxon *label set* for D, $\mathcal{L}$, is the set of all taxon names in D, and the label set, $\mathcal{L}_j=\mathcal{L}_j(D,\mathcal{P})$, for block j of D is the set of taxon names for which data are present for block j in partitioning scheme $\mathcal{P}$ (i.e., the set of taxon names for which $C_{ij}=1$) (Fig. 2).

Models.—Statistical models are used both to generate sequence data in simulations ("generating models") and to calculate likelihoods based on the data ("inference models"). Models can reflect a partitioning scheme, $\mathcal{P}$, in various ways. Suppose the model at locus i consists of two sets of free parameters $\{M_i,E_i\}$, where M_i is a set of free parameters associated with the substitution rate matrix alone, and E_i is a set of free edge-length parameters. An *edge-unlinked* (EUL) model has free edge-length parameters E_i for each locus (thus a total of km parameters, where m is the number of loci and k is the number of edges on the tree). This is the general "heterotachy" model of Pagel and Meade (2008). At the other extreme an *edge-linked* (EL) model has the same edge-length parameters for all loci (so exactly k edge length parameters). In between are *partially edge-linked* (PEL) models with an intermediate number of free parameters. For example a "proportional" model of the form $E_i\propto E_j$, for all i, j is PEL. We use the term "heterotachy" in a more general sense than Pagel and Meade (2008) to include both EUL and PEL models, excluding only the strictly EL model; that is, we include any models with different edge-length parameters at different loci.

The same scheme applies to the parameters of the substitution matrix: so we have rate-linked (RL), rate-unlinked (RUL), and partially rate-linked (PRL) models. In the literature, "linked" models generally refer to EL/RL models and unlinked models to EUL/RUL models. MrBayes (Ronquist et al. 2012) and RAxML (Stamatakis 2014) allow a decoupling of assumptions about the edge parameters and the substitution rate matrix parameters to enable EL/RUL (MrBayes) and EUL/RL models (MrBayes and RAxML). The "siterates" model in PAUP and GARLI (Zwickl 2006) is the "proportional" (PEL) model above. Empirical studies have entertained this entire range of models up to and including the most parameter-rich EUL models (Hess and Goldman 2011; Hedin et al. 2012; Xi et al. 2013).

Trees and subtrees.—For any tree, T, on the complete label set $\mathcal{L}$ (e.g., the best maximum likelihood (ML) tree found in a heuristic tree search based on D), the label set for any locus, j, $\mathcal{L}_j$ (which may have taxa missing) induces a subtree of T, which we write as $T|\mathcal{L}_j$. Let $\mathcal{Q}(\mathcal{P},C,T)=T|\mathcal{L}_1,T|\mathcal{L}_2,...,T|\mathcal{L}_m$, be the set of subtrees for this tree that are induced by the partitioning scheme and taxon coverage matrix—that is, $T|\mathcal{L}_j$ is the tree obtained from T by removing any taxa that have data that are entirely missing for block $\mathcal{L}_j$, for each block $\mathcal{L}_j$ of the partition (Fig. 2). Importantly, the subtrees induced from T in this way are *necessarily* compatible with each other; which is obviously not always true for subtrees that are actually inferred separately for each locus.

A tree T is a *resolution* of some other tree, say T^*, if T^* can be obtained from T by collapsing one or more edges of the tree, which transforms binary nodes to polytomies. A tree T on label set $\mathcal{L}(T)$ *displays* another tree T^* on $\mathcal{L}(T^*)$ (with $\mathcal{L}(T^*)\subseteq\mathcal{L}(T)$) if $T|\mathcal{L}(T^*)$ is equal to or is a resolution of T^*. Intuitively, this allows a larger tree to display a smaller tree even if that smaller tree is less resolved than the larger one. This extra generality is appropriate given the usual notion of polytomies as reflecting uncertainty (i.e., soft polytomies) rather than multiple speciation (hard polytomies: Maddison 1989; Semple and Steel 2003).

Decisiveness.—The taxon coverage matrix is said to be *decisive* for tree T and partitioning scheme $\mathcal{P}$ if T is the only tree that displays all subtrees in $\mathcal{Q}$ (Steel and Sanderson 2010; Sanderson et al. 2010). In the example in Figure 2, the taxon coverage matrix is not decisive for T because T, T', and T'' all display all subtrees in $\mathcal{Q}$. Some taxon coverage matrices are decisive for *every* tree, in which case we say the coverage matrix itself is decisive (Steel and Sanderson 2010). [We define decisiveness very differently than Goloboff (1991), who was referring to the relative strength of character signal in favor of topologies].

Stands.—The case of interest here is when a taxon coverage matrix is not decisive, and thus there exists somewhere in tree space a set of two or more trees that display the same subtrees in $\mathcal{Q}$. First, we define a new term for this not used in our previous work on terraces. A *stand* (of trees), $\mathcal{S}=\mathcal{S}(\mathcal{P},C,T)$, is the set of all binary phylogenetic trees on leaf set $\mathcal{L}$ that display every subtree in $\mathcal{Q}(\mathcal{P},C,T)$. In the example in Figure 2, there are three trees that display the subtrees in $\mathcal{Q}$, including the original T. Thus, when taxon coverage is not decisive there exist stand(s) with more than one tree.

The concept of a "stand" of trees is related to, but different from the concept of the "span" of a set of phylogenetic trees, which is more relevant to the supertree context (Semple and Steel 2003). Not only can a span of trees be empty, or contain non-binary trees (both of which are not possible for a "stand") but the input for a stand is not a set of trees, rather it is a single tree T and a pattern of taxon coverage. Nevertheless, the notions are related; a stand $\mathcal{S}=\mathcal{S}(\mathcal{P},C,T)$ is precisely the set of binary phylogenetic trees in the span of $\mathcal{Q}(\mathcal{P},C,T)$.

Terraces.—Stands and decisiveness depend only on the coverage pattern, partitioning scheme and tree, $\{\mathcal{P},C,T\}$; not on the actual sequence data, D, per se. However, under certain conditions, all trees in a stand, $\mathcal{S}$, have precisely the same optimality score with respect to D, in which case we call the stand, $\mathcal{S}$, a *terrace*, denoted $\mathcal{T}$ (Fig. 2). In particular, when maximum parsimony (MP) is used as the score, all trees in $\mathcal{S}$ have the same score for *any* partitioning scheme, $\mathcal{P}$, and $\mathcal{S}$ is always a terrace. Partitioning in parsimony analysis is usually aimed at discovery of different phylogenetic histories within the same data matrix, since different "models" are not typically defined. However, we will see below that irrespective of any partitioning (or not) imposed by the investigator, there exists an associated "maximal" partition that can help to identify a maximal set of equally optimal trees when such trees exist.

The set $\mathcal{S}$ is also a terrace when likelihood is the optimality criterion if the likelihood is determined by a model and partition, $\mathcal{P}$, which is RUL/EUL (see above) (Sanderson et al. 2011). Below we show this to be a sufficient but not necessary condition.

Trivially, both stands and terraces can have sizes of only one tree, but the case of interest throughout this paper is just that setting (lack of decisiveness) in which they have more than one element, and which thereby adds ambiguity to phylogenetic inference.

Properties of Terraces

Because all terraces are stands, a number of results derived for stands—based only on trees, subtrees, and coverage patterns—are helpful for characterizing terraces. Some of these results hold only for rooted trees, but many problems can be effectively "rooted" as long as there is one taxon in C that is sampled for all loci in the partition, in which case that taxon can serve as an "operational" root for the purposes of an algorithm. We assume trees are rooted unless stated otherwise. The number of trees on a stand (or terrace) can increase exponentially with the size of the tree (Semple 2003). Despite this, several properties of terraces make them more tractable than they might otherwise seem, mainly because several summary statistics can be obtained directly without any computation involving the data matrix, D (Sanderson et al. 2011). For example, all trees on a terrace can be enumerated without recalculation of optimality scores. This takes advantage of an algorithm due to Constantinescu (1995), with a running time that scales linearly with the size of the terrace (rather than, say, exponentially with the size of the tree, although in the worst case the size of the terrace can also grow exponentially).

The trees on a terrace can also be summarized by a strict consensus tree (Gordon 1986) or Adams consensus tree, either of which can be constructed in polynomial time. This last claim (that we can sidestep the enumeration of trees on the terrace again, or any further search using the data) is not obvious, but it holds in general for rooted trees. In the case of strict consensus, this was shown in Steel (1992) (using results from Aho et al. 1981), whereas for Adams consensus, it relies on a particularly elegant result due to (Bryant 1997) (Theorem 6.2) which states that the Adams consensus of a terrace is equal to the so-called BUILD supertree of the induced subtrees for each locus (i.e., the set $\mathcal{Q}(\mathcal{P},C,T)$), and this supertree can be computed quickly by the algorithm of Aho et al. (1981). Note that the Adams consensus tree displays each of those subtrees; that is, it is identical to (or resolves) them if extraneous taxa and edges are removed. Interpretation of Adams consensus trees in phylogenetics is somewhat fraught, however, as clusters do not necessarily represent clades (Wilkinson 1994).

Terraces are reminiscent of tree islands but are contained within them, at least for rooted trees. An island is a region of tree space with optimality score better than some threshold, separated from other such regions by regions of lower score (Maddison 1991; Salter 2001). Here a "region" is a set of trees that can be enumerated by a series of topological rearrangements that do not leave the region. Terraces are always wholly contained within a tree island, because all trees on a terrace of rooted trees can be reached by a series of nearest neighbor interchanges (NNIs) between trees of the same optimality score (Bordewich 2003; Sanderson et al. 2011). For unrooted trees this property does not necessarily hold. For rooted trees, then, tree space can be thought of as a collection of terraces (on islands), the size of which are determined entirely by the partition $\mathcal{P}$ and taxon coverage matrix C. Only the heights of the terraces depend on the data. To the extent that a tree "island" is a useful metaphor, it may be best to envision it as a rough landscape covered with terraces of different sizes and heights (Fig. 1).

New Results on Terraces

Terraces Occur in Likelihood Inference under a Less Restrictive Set of Assumptions

In our previous work, we showed that terraces can occur in ML inference whenever the inference models for separate loci in partition $\mathcal{P}$ are simultaneously EUL and RUL (Sanderson et al. 2011). This was a sufficient but not necessary condition. In fact, as we show now, terraces can arise even when models are EUL and the substitution rate matrices are the same across loci.

Proposition 1. *Let $\mathcal{S}(C,\mathcal{P},T_0)$ be a stand of trees containing tree T_0 for partitioning scheme, $\mathcal{P}$ and taxon coverage matrix, C. For any model for computing the likelihood score that is edge-unlinked (EUL), but where the model parameters (M_i) are constrained to be identical across the loci (i.e., across $\mathcal{P}$), all trees in $\mathcal{S}$ have the same maximum likelihood score, and hence $\mathcal{S}$ is a terrace.*

The proof of this result is presented in the Appendix. This finding implies that it is the lack of commonality between *edge-length parameters* across loci that is necessary for a stand of trees to be a terrace under likelihood. Rate matrices between loci need not have different parameter sets. Henceforth, we will refer to this kind of ML inference with just an EUL assumption as "ML-EUL."

Maximum Size of Stands and Terraces

Given a partitioning scheme, $\mathcal{P}$, and taxon coverage matrix, $C(\mathcal{P},D)$, we can find the stand, $\mathcal{S}$ containing some tree, T_0 and calculate its size. If the optimality criterion is MP or ML-EUL, then $\mathcal{S}$ is also a terrace. Generally, the presumption is that $\mathcal{P}$ is chosen to reflect meaningful biological aspects of the data, such as different loci, or disjoint sets of codon positions. Nonetheless, it is possible that T_0 is contained within a larger stand for a different partitioning scheme $\mathcal{P}'$. In this section we show that for any data matrix, D, and tree, T_0, there exists a partitioning scheme that corresponds to a stand of maximal size, and the stands associated with all other partitioning schemes will be contained within it. This has important implications for heuristic search strategies and for choice of partitioning schemes.

Consider any partition $\mathcal{P}=\{B_1,\ldots,B_l\}$ of the columns of the data matrix, D (i.e., the sets B_i are disjoint subsets of columns, which cover every column). For a given block B_i of $\mathcal{P}$, let $\mathcal{L}_i=\mathcal{L}(B_i)$ be the taxa that are present for at least one column in B_i.

Now suppose we have another partition $\mathcal{P}'=\{B'_1,\ldots,B'_{l'}\}$. We say that $\mathcal{P}'$ *refines* $\mathcal{P}$ if each block of $\mathcal{P}'$ is a subset of some block of $\mathcal{P}$ (equivalently, each block of $\mathcal{P}$ is either equal to a block of $\mathcal{P}'$ or is the union of two or more blocks of $\mathcal{P}'$).

However, a refinement that includes breaking B_j into say, K, smaller sets, $B'_{i_1},\ldots,B'_{i_K}$, should be disallowed if any of the labels sets, $\mathcal{L}(B'_{i_k})$, are duplicated, as otherwise this would allow a trivial refinement of any partition into one in which each block consists of just a single column in the data matrix, D.

Given this restriction, there will be a unique (and usually nontrivial) *maximal partition* for any data matrix, D, that cannot be refined further, which we denote as $\mathcal{P}_{\max}$. This maximal permitted partition $\mathcal{P}_{\max}$ can be built as follows. Let $\Omega=\{\hat{\mathcal{L}}_1,\ldots,\hat{\mathcal{L}}_m\}$, where $\hat{\mathcal{L}}_j$ is the set of taxa present in column j (note that Ω will generally have size less than m, since the sets $\hat{\mathcal{L}}_j$ will usually not all be different). Then $\mathcal{P}_{\max}=(B_A:A\in\Omega)$, where $B_A=\{j:\hat{\mathcal{L}}_j=A\}$.

Proposition 2. *If $\mathcal{P}'$ refines $\mathcal{P}$ then $\mathcal{S}(C(\mathcal{P},D),\mathcal{P},T_0)\subseteq\mathcal{S}(C(\mathcal{P}',D),\mathcal{P}',T_0)$; in particular, the former set is never larger than the latter.*

That is, the stand associated with $\mathcal{P}$ is a subset of the stand associated with $\mathcal{P}'$.

Proof. Let $\mathcal{L}'_j=\mathcal{L}(B'_j)$, the taxa present in block B'_j for $j=1,\ldots,l'$. Suppose that T is a tree in the stand of T_0 relative to $\mathcal{P}$. This means that $T|\mathcal{L}_i=T_0|\mathcal{L}_i$ for $i=1,\ldots,l$. Consider a block B'_j of $\mathcal{P}'$. Since $\mathcal{P}'$ refines $\mathcal{P}$, B'_j is a subset of some block of $\mathcal{P}$, say block B_i, and so $\mathcal{L}'_j\subseteq\mathcal{L}_i$. In that case:

$$T|\mathcal{L}'_j=(T|\mathcal{L}_i)|\mathcal{L}'_j=(T_0|\mathcal{L}_i)|\mathcal{L}'_j=T_0|\mathcal{L}'_j.$$

Thus, $T|\mathcal{L}'_j=T_0|\mathcal{L}'_j$ for each block $\mathcal{L}'_j$ of $\mathcal{P}'$, and so T is in the stand of T_0 relative to $\mathcal{P}'$. ■

Corollay 3 For any permitted partition $\mathcal{P}$ we have: $|\mathcal{S}((C(\mathcal{P}_{\max},D),T_0,\mathcal{P}_{\max})|\geq|\mathcal{S}(C(\mathcal{P},D),T_0,\mathcal{P})|$ for all $\mathcal{P}$.

Proof. This follows from Proposition 2, since $\mathcal{P}_{\max}$ is a refinement of every permitted partition of the loci. ■

Thus, the largest stand containing a given tree can be obtained by constructing the maximal partitioning scheme described above. Under MP or ML-EUL inference, this will also be the largest terrace containing this tree.

For example (Fig. 3a), suppose a multiple sequence alignment is first partitioned into two loci, $\mathcal{P}_I=\{I_1,I_2\}$, where I_j is the set of sites for the jth locus. This partition induces a stand of 13 trees containing the tree, T_0, shown (Fig. 3b). However, perhaps locus I_2 actually consists of three biologically meaningful blocks: two exons and an intron, and suppose further that the label sets for the exons are the same but differ from that for the intron. Our second partition is then described by $\mathcal{P}_{II}=\{II_1,II_2,II_3\}$, where II_j is the set of sites for the jth locus of this partition (Fig. 3a). Formally, $\mathcal{P}_{II}$ is a refinement of $\mathcal{P}_I$. The stand induced by this new three block partition, $\mathcal{P}_{II}$, has 23 trees (Fig. 3b). This is the MP for these data, and 23 is the size of the largest stand containing the given tree. Also, from Proposition 2, the smaller stand of trees is a subset of the larger.

Under MP and ML-EUL inference, these stands are also terraces, so the largest terrace containing the indicated tree has 23 trees on it. The sets of trees in these terraces arising from different partitioning schemes are the same for MP and ML-EUL inference, but there is an interesting distinction with respect to their optimality scores. For MP, the situation is particularly simple. For two partitioning schemes in which $\mathcal{P}_{II}$ is a refinement of $\mathcal{P}_I$, the two terraces have the same parsimony score,

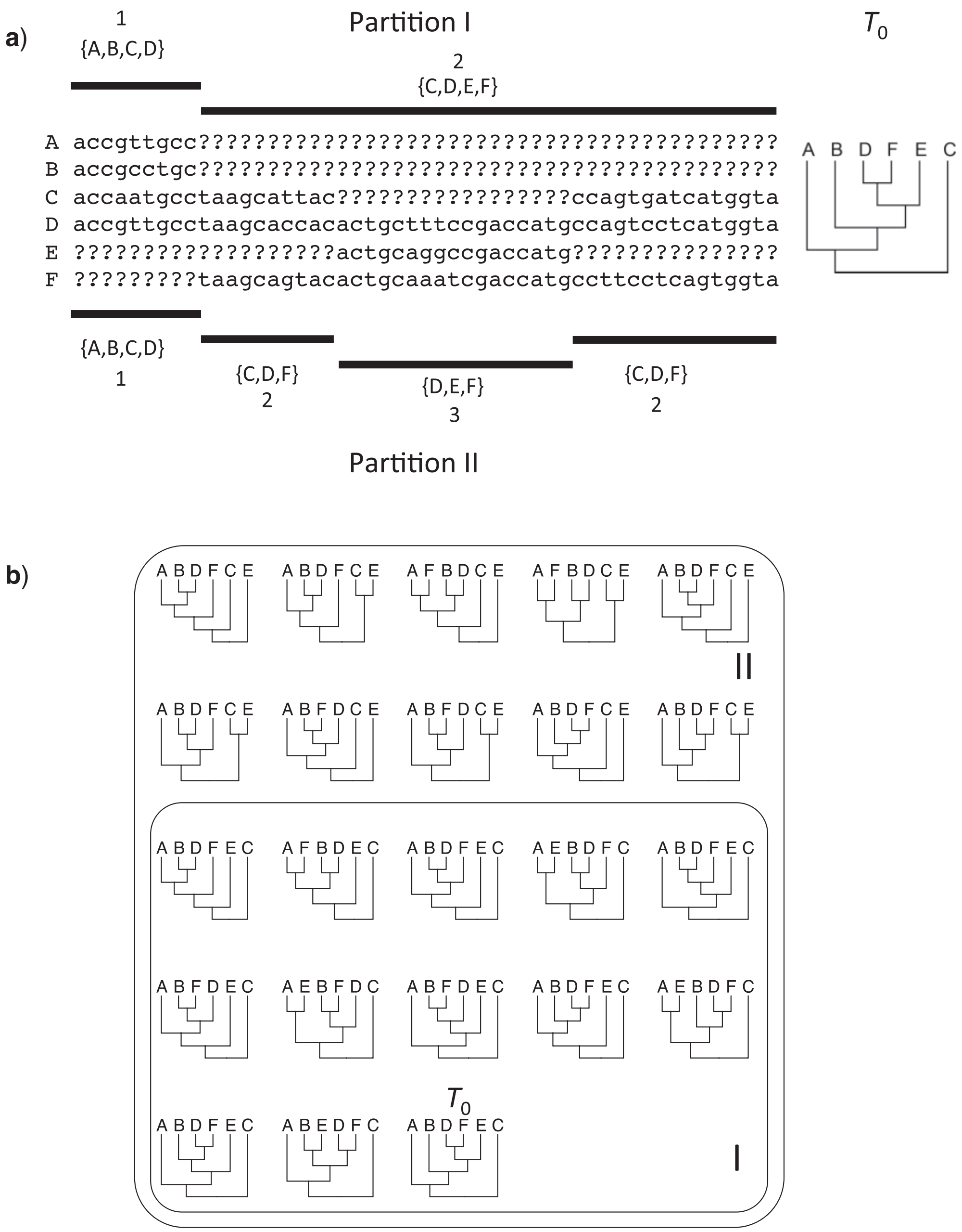


FIGURE 3. a) A multiple sequence alignment with different partitioning schemes, which can induce terraces of different sizes containing a given tree, T_0 (on right). Missing nucleotide data are indicated by '?'. Two partitioning schemes are labeled I and II. Partition elements ("loci", "blocks") are labeled with 1,2, and 3. Label sets for each locus are indicated in curly brackets. b) The terrace containing T_0 for Partition I has 13 trees, but has 23 trees for Partition II. Partition II is a "maximal partition" (see text).

so it is reasonable to visualize the smaller terrace as being literally imbedded in the larger one at the same "elevation" in the landscape of tree space. Consequently, for parsimony, when characterizing trees on a terrace based on a particular partition, $\mathcal{P}$, it is useful to check if $\mathcal{P}$ is the MP, and if not, also check $\mathcal{P}_{\max}$. The latter will be a more accurate representation of the actual extent of equally parsimonious trees around the given tree.

For ML-EUL, the situation is more complex. The likelihood scores of the trees in the terraces for partitioning schemes $\mathcal{P}_I$ and $\mathcal{P}_{II}$ could well be different, but all trees within either terrace will still have the same score once a partitioning scheme is fixed. This means changing a partitioning scheme to the maximal partitioning scheme, for example, not only expands the set of trees on the terrace, but potentially changes the optimality score as well, unlike in parsimony.

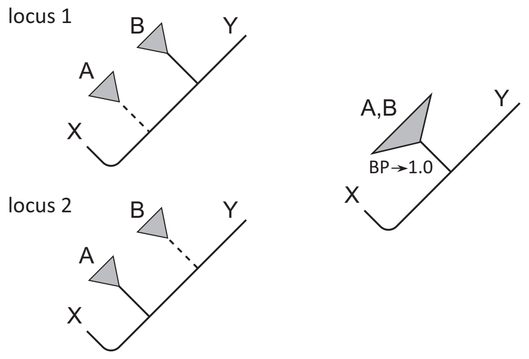


FIGURE 4. Impact of terraces on the bootstrap. True (rooted) tree and taxon coverage as in Figure 2 except outgroup removed and taxa A and B now contain n_{A} leaves (present only for locus 1) and n_{B} leaves (present for locus 2) respectively. As n_{A} and n_{B} get larger the bootstrap proportion for the incorrect bipartition at right goes to 100%.

Impact on Confidence Assessment: Bootstrap Proportions

Terraces with multiple trees are a consequence of missing data in phylogenetic tree inference. In the next two sections we discuss how the ambiguity terraces reflect can alter *estimates* of the quality of trees. The bootstrap (Felsenstein 1985) case is especially clear, as we demonstrate with detailed discussion of a simple example. For sequence data generated on trees of the form shown in Figure 4, with the indicated pattern of partial taxon coverage, the bootstrap proportion for an incorrect clade approaches 100% if the number of leaves in A and B is large. To show that this is a consequence of the partial taxon coverage, rather than the sequence data per se, we construct a hypothetical example in which we can isolate the two factors from each other.

Assume that the true tree, T, has four taxa, X,Y,A, and B, where X and Y are leaf taxa, but A and B are sets of leaves with n_A and n_B leaves respectively. Assume a specific tree having a bipartition given by XA|BY, as well as the two bipartitions A|BXY and B|AXY (Fig. 4). Let the data matrix, D, be partitioned into two blocks, D_1 and D_2, each of length k. To allow us to understand the impact of terraces in isolation from the sequence data, we make two simplifying assumptions. First, we assume that the sequence length, k, is sufficiently large that if there were no missing data, a bootstrap analysis of D, D_1, or D_2, would each return just a single tree, T, in every bootstrap replicate. This would imply perfect support for T. Second, we assume that if we delete some taxa from D_i, the tree reconstructed from the remaining sequences in D_i is just the subtree of T containing the remaining taxa; that is, taxon deletion does not affect inference for the remaining taxa (assuming k is large enough), and, again, each bootstrap replicate for D_i would return that subtree of T. Together these assumptions effectively remove the impact of noise and biases due to the sequence data proper. To make this a bit more concrete: simulations using Seq-Gen (Rambaut and Grassly 1997) showed that such ideal conditions were well approximated in small trees using an "F84" substitution model with equal base frequencies (equivalent to a Jukes–Cantor model); random root sequence; $k > 5000$; all edge lengths set to 0.1 substitutions/site; and constant rates across sites.

Now let there be partial taxon coverage, such that X and Y are sampled for both loci, but leaves in A are only sampled for locus 1, and leaves in B only for locus 2. Figure 4 is rooted with X but in general either X or Y can act as an "operational root" when needed in the following. With this taxon coverage matrix, the true tree, T, has two induced subtrees, $T|\mathcal{L}_1$ and $T|\mathcal{L}_2$, and T is in a stand, $\mathcal{S}$, with other trees. For example, if $n_A = 3$ and $n_B = 3$, then the stand has 107 binary trees, but importantly, only 2 of these are consistent with the XA|BY bipartition in the true tree, T. The remaining 105 trees conflict with T, specifically by interleaving leaves from A with leaves from B, such that the bipartition AB|XY is present instead of XA|BY. Assuming bootstrapping is implemented in a way that respects the partition boundaries (resampling from within each locus separately), then from our two assumptions above, each bootstrap replicate produces the same pair of induced subtrees, so in each replicate the tree found will be imbedded in this same stand of 107 trees. How this influences actual bootstrap proportions reported by software depends on whether parsimony or likelihood is used, and on technical choices about the handling of equally optimal trees.

In parsimony inference, standard branch rearrangement operations may be used to search for all equally parsimonious trees in the stand for each bootstrap replicate. This will work as long as memory is not exhausted because all trees in a stand can be reached by a series of NNIs, for example (Bordewich 2003; Sanderson et al. 2011). However, there are different options for tallying results across replicates. PAUP* (Swofford 1999) uses a "frequency within replicates" approach (Davis et al. 2004) to obtain the boostrap proportion of a bipartition, in which the frequency of a bipartition in the equally optimal trees found in each replicate is averaged across replicates. Because the same stand of 107 trees would be found for each bootstrap replicate, this procedure would return a bootstrap proportion of $105/107 = 0.98$ for the incorrect bipartition in Figure 4. If $1 - BP$ is regarded as a P-value for the null hypothesis of non-monophyly (DiCiccio and Efron 1996; Susko 2009), this significance level is greatly inflated. An alternative and much more conservative "strict consensus" approach would require all trees in a set of equally optimal trees contain the bipartition for it to be considered present in that replicate (Davis et al. 2004; Simmons and Freudenstein 2011). This procedure would result in a bootstrap proportion of zero for the incorrect AB|XY bipartition but would also result in zero for the correct XA|BY bipartition, which is less misleading but not altogether a more useful result.

Matters are different with current implementations of likelihood inference, which typically return a single tree. The very notion of "equality" of the likelihood score in likelihood calculations is problematic due to the imprecise nature of numerical optimization and floating point imprecision. ML tree search in RAxML (Stamatakis

TABLE 1. Scaling of terrace size and frequency of incorrect AB|XY bipartition in Figure 4

Number of taxa	Terrace size	Bipartition frequency
1	3	0.3333333
2	17	0.8823529
3	107	0.9813084
4	602.6	0.996681
5	3127.571	0.9993605
10	6906214	0.9999997
15	10867704749	1
20	1.501564e+13	1
25	1.935652e+16	1

Note: Number of taxa is the number of leaf taxa in both A and B of Figure 4.

2014) and GARLI (Zwickl 2006), for example, operates on the implicit assumption that equally optimal trees either do not exist or cannot be unambiguously identified during tree search. During a search in a single bootstrap replicate, small differences in likelihood scores arising simply from different addition orders of the per-site log likelihoods (depending on the compiler or number of threads being used), therefore, may result in different trees on the terrace being sampled. In addition, the order in which rearrangement moves are applied may also affect which tree on the terrace is being sampled. To the extent that score imprecision and distinct search paths are randomly distributed, repeated bootstrap replicates will sample trees on the terrace with equal probability, leading to the same bootstrap proportion of 0.98 for the incorrect clade of Figure 4, for example.

Surprisingly, the number of trees in $\mathcal{S}$, and support for the incorrect bipartition, grows with n_A and n_B (Table 1). We can show this analytically by deriving how many rooted binary trees correctly display the subtrees $T|\mathcal{L}_A$ and $T|\mathcal{L}_B$, where we have changed notation a bit and $\mathcal{L}_A$ refers to the subtree having just the leaf taxa in clade A, and likewise for B. Note that we can consider these subtrees as rooted by operationally rooting the entire tree with X or Y. Notice that the label sets $\mathcal{L}_A$ and $\mathcal{L}_B$ are disjoint. Let $n=n_A+n_B$, and $R(k)=1\times 3\times 5\times\cdots\times(2k-3)$ be the number of rooted binary trees for k leaves. If n_A or $n_B \leq 3$, then the number of trees on the terrace is

$$2+\frac{R(n)}{R(n_A)R(n_B)}, \tag{1}$$

a result that depends only on the two subtree sizes. The "2" term corresponds to the two trees on the terrace with bipartitions of XA|BY and XB|AY. The term on the right corresponds to the set of all trees that contain the bipartition AB|XY, and accounts for the fact that there are possibly many ways to interleave the A and B subtrees and still display the original two subtrees. We pause briefly to explain why this second term equals $\frac{R(n)}{R(n_A)R(n_B)}$ when (say) $n_B \leq 3$. Given two rooted binary trees T_A and T_B on disjoint label sets $\mathcal{L}_A$ and $\mathcal{L}_B$, respectively, let $N(T_A,T_B)$ denote the number of rooted binary trees on the total label set $X=\mathcal{L}_A\cup\mathcal{L}_B$ that restrict to T_A on $\mathcal{L}_A$ and T_B on $\mathcal{L}_B$. Then regardless of whether or not $n_B \leq 3$, the sum of $N(T_A,T_B)$ over all pairs T_A and T_B equals $R(n)$, and since $N(T_A,T_B)$ takes the same value for any of the trees T_A, we have the following identity (from Constantinescu and Sankoff 1986), namely:

$$\sum_{T_B} N(T_A,T_B)=R(n)/R(n_A). \tag{2}$$

Now, when $n_B \leq 3$ symmetry considerations show that $N(T_A,T_B)$ takes the same value for any of the (at most three) trees T_B, and so $N(T_A,T_B)=R(n)/[R(n_A)R(n_B)]$.

If both A and B are larger than three leaves, the expression depends on the topology of the subtrees (i.e., the entire tree), not just their sizes, an observation due to Constantinescu and Sankoff (1986), so a more complex calculation for Equation (1) would be necessary. On *average*, however, across a uniform random sample of subtree topologies, the mean number of trees converges to the expression in Equation (1)—this follows from Equation (2) upon dividing both sides by $R(n_B)$. Thus, for certain patterns of partial taxon coverage, the bootstrap proportion for a given clade may tend to zero or one depending not on the data but on the number of taxa (and possibly tree topology) and taxon coverage pattern in specific parts of the tree. It is the interaction of these factors that renders bootstrap proportions somewhat oddly at the whim of tree combinatorics, since in this case sequence data are assumed sufficient to eliminate all errors in the reconstruction of the two subtrees.

Impact on Confidence Assessment: Bayesian Posterior Probabilities

Given the issues raised by terraces for MP and ML-EUL, it is reasonable to expect some effects on calculation of Bayesian posterior probabilities, but in this case the impact is apparently determined by an interaction between the taxon coverage pattern, tree topology, and priors on edge lengths. Let us simplify the bootstrap example by setting $n_A=n_B=1$ on the same "true" tree, T (Fig. 4) and keep the same taxon coverage matrix and partitioning scheme. We also use the same simulation protocol and assume that likelihood calculations (in Bayesian inference) are carried out with an EUL inference model, so stands are also terraces with respect to the likelihood score. There is then a stand of three binary trees, all with the same maximized likelihood (in fact, these are all possible binary trees for four taxa).

Perhaps surprisingly, however, their posterior probabilities may not be equal. In particular, the posterior for the incorrect bipartition (AB|XY) increases as the lengths of the edges leading to A and B increase (calculated in MrBayes v. 3.1.2 [Ronquist et al. 2012]; 5 million generations; default burn in; Jukes–Cantor

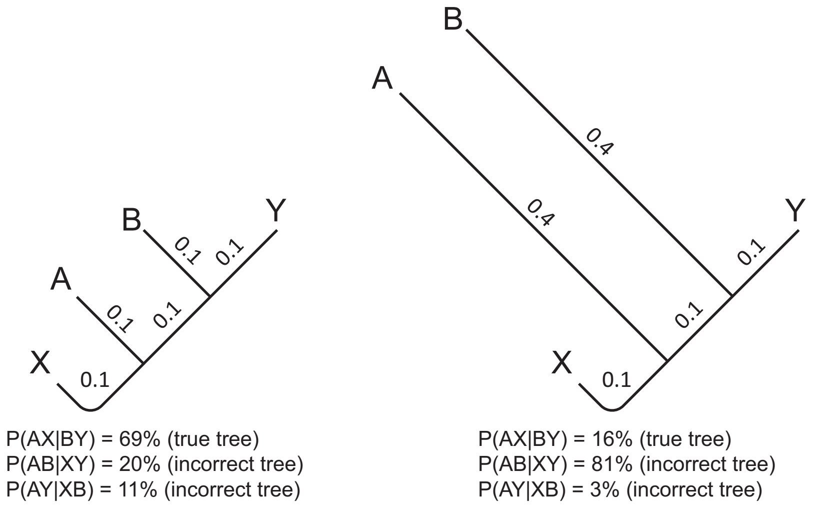

B
Y
A
0.1
0.1
0.1
0.1
X
0.1
P(AX|BY) = 69% (true tree)
P(AB|XY) = 20% (incorrect tree)
P(AY|XB) = 11% (incorrect tree)
B
A
0.4
0.4
Y
0.1
0.1
X
0.1
P(AX|BY) = 16% (true tree)
P(AB|XY) = 81% (incorrect tree)
P(AY|XB) = 3% (incorrect tree)

FIGURE 5. Impact of terraces on Bayesian posterior probabilities. True tree is on left. As terminal edge lengths for leaves A and B get longer, the posterior probability inferred by MrBayes (v. 3.1.2: 5 million generations; default burn in) for the incorrect bipartition, AB|XY increases. True edge lengths are indicated in substitutions/site.

model with edge length parameters "unlinked"). For example, if the two long edges are four times the length of the other edges in the tree, the posterior probability is 81% for an incorrect tree (Fig. 5), whereas if the edge lengths are all the same, the correct tree has highest posterior probability. This is evidently not a product of the "long branch attraction" possible under Bayesian inference in the absence of missing data (Susko 2008; Kolaczkowski and Thornton 2009), because if the missing data are returned to the matrix the correct tree has 99% posterior probability.

This phenomenon can be traced ultimately to how Bayesian inference handles missing data. Consider the simplest context of partial taxon coverage: with a trivial "partition" consisting of only a single block and having one leaf taxon, x, with missing data throughout that block. In other words, this is a data set in which one leaf, x, has no data at all. Consider a tree, T, with branch lengths, containing taxon x as a leaf, and the tree, T_{-x} (with its inherited branch lengths) obtained by deleting leaf x (and its incident branch) from T. If the sequence of x consists entirely of "?"s, the parsimony and likelihood scores of any tree, T', obtained by attaching x to some edge of T_{-x} is the same. In other words, a method such as maximum likelihood, which relies *only* on these scores is unable to decide the position of x.

However, in Bayesian inference the position of x may also be influenced by prior probabilities. In fact, we will now show that for the usual exponentially distributed prior on edge lengths used in phylogenetic inference, the posterior probability for different placements of x in the tree is determined entirely by the length of the edge to which x attaches; if these differ, then some placements of x in this tree will have higher posterior probabilities. Ultimately, this will affect posterior probabilities of trees on terraces in more complex partitioning schemes with partial taxon coverage.

Consider data generated by a reversible Markov process EL model on a binary phylogenetic tree T with branch lengths assigned, and analyzed under a Bayesian approach in which (i) all rooted binary phylogenetic trees have the same prior probability, and (ii) an exponential prior applies independently across the branch lengths.

Theorem 4 *Let T' be any binary phylogenetic tree that agrees with T up to the placement of taxon x, and consider a set of aligned sequences of length k generated on T with fixed branch lengths (under a standard reversible model) but thereafter with the sequence for taxon x replaced with all "?"s (missing data). Then under conditions (i) and (ii) above, the Bayesian posterior probability of T' for this data converges in probability to $l(e_x)/\sum_e l(e)$ as the sequence length k grows. Here, $l(e)$ is the length of edge e in T_{-x}, the summation is over all edges of T_{-x}, and e_x is the edge of T_{-x} to which x must attach to produce the tree topology T'.*

The proof of this result is presented in the Appendix. Intuitively, the theorem says that the relative chances of a leaf taxon having all missing data attaching to two alternative edges of a tree are (for long sequences) given by the ratio of the two edge lengths: the longer edge will "attract" the taxon with missing data to a greater extent. This result may seem like something of a curiosity for a trivial partitioning scheme consisting of just a single locus, and a leaf taxon with no data, but it explains why the posterior probability of trees on a terrace may be different even if their likelihoods are the same. In particular, it pinpoints the important role that edge lengths can play in influencing subsets of the trees on a terrace to have higher posterior probabilities, as seen in the example in Figure 5.

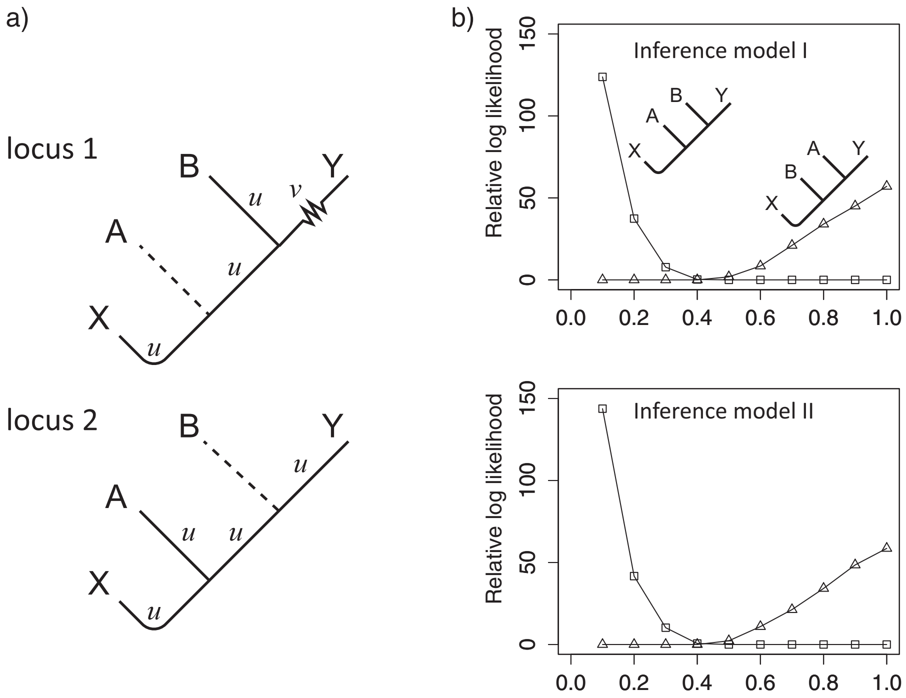


FIGURE 6. a) A simple generating model of heterotachy for two loci, G^{PEL}. Trees labeled 1 and 2 are annotated with edge length parameters for the two loci in the partitioning scheme. Partial taxon coverage pattern is indicated as a dotted edge for the leaf taxon having missing data. All edge lengths are u for loci 1 and 2, except the outlier edge labeled v for locus 1. b) Likelihood scores of the best and second best trees relative to the worst tree as a function of edge length v. Worst scoring tree is always the XY|AB tree. Inference model I assumes the same edge parameters for both loci. Inference model II assumes edge parameters are proportional between loci. Open squares refer to relative log likelihood of AX|BY tree to the worst scoring tree; open triangles refer to the relative log likelihood score of the BX|AY tree to the worst tree. The two best trees are drawn in the top panel for illustration.

Extensions to Partially Linked Models Used in ML Estimation

If the problem of terraces arises when using EUL models for inference, perhaps it is generally better to use EL or PEL models for inference. After all, the generating models used in our simulations have been simpler than EUL, so perhaps EUL models are "overparameterized." In this section, we discuss an example which would suggest that in fact overparameterizing may be better than mis-specification with a simpler inference model. In particular a small degree of heterotachy in the generating model can lead to a reordering of the relative ML scores when using EL or PEL inference models, so that an incorrect tree is favored, much as in the Bayesian case described above. Under these conditions, EUL inference, though overparameterized, is more conservative.

Consider the simple PEL generating model, G^{PEL}, having four leaves and two loci (Fig. 6a). Assume all edges for both loci have evolved with the same rate, $u = 0.1$ substitutions per site, across different simulation settings, except for a single terminal edge for one of the loci, which has rate v, ranging from 0.1 to 1.0 substitution/site. Each locus in the partition has 5000 sites and sequences are simulated on the model tree using Seq-Gen (Rambaut and Grassly 1997) with a Jukes–Cantor model, as described above in the bootstrap example. As before, the pattern of partial taxon coverage induces a stand consisting of all three of the binary trees possible for four taxa. Thus, with MP or ML-EUL inference, the optimality scores of all three trees are the same.

Now, for inference models, consider two models that are less general than an EUL model and have fewer parameters. Neither model therefore has the properties described earlier that are sufficient for terraces to emerge. Model I assumes the substitution rate matrix is the same for both loci and the edge-length parameters are the same for both loci (i.e., EL with five edge parameters); Model II assumes that the substitution rate matrix is the same for both loci but that edge lengths for all edges at the second locus are strictly proportional to the corresponding edge lengths for the first locus (i.e., a PEL model with five edge parameters plus a parameter for the proportionality constant). This allows rate variation between loci, but the edge parameters are still highly constrained. Likelihood calculations were carried out using PAUP* 4.0 (Model II implemented with the "siterates" command: Swofford 1999). Under these conditions the likelihood scores for the three possible binary trees are different for Models I and II (Fig. 6b). In particular, if v is more than about $4u$, then an incorrect tree has the highest likelihood (Fig. 6b). In other words, use of an EL or PEL model for inference does not just avoid the ambiguity of the terrace phenomenon that would have been seen using EUL inference, it is positively misleading. Under these conditions, using an EUL model for inference and dealing with the resulting ambiguity might be a preferable (if more conservative) procedure.

This case clearly involves model misspecification, since neither of the two inference models is the model actually generating the sequences. Interestingly, if we "fix" the data set by filling in the missing entries with data generated under our generating model, the correct ranking is restored under EL or PEL inference, even with this model misspecification. By the same token, using a simpler EL *generating* model, obtained by setting $u=v$ in G^{PEL}, but retaining missing data according to the specified partial coverage pattern, also restores the correct ranking of trees when using the EL or PEL inference model. Thus, the re-ranking of trees we observe in this example is not a product of either partial coverage alone or our particular generating model alone. Both are required, and the correct ranking of the three trees' likelihood scores can be rescued by eliminating one or the other.

Although these results imply strongly that ML inference can be statistically inconsistent (i.e., converging to the wrong tree) under the combination of missing data and data generated by G^{PEL}, we have not been able to formally prove this in general. However, some aspects of the converse are provable. For data generated by G^{PEL}, both parsimony and likelihood inference using Model I (EL) are statistically consistent when there are no missing data. The proof that parsimony will be consistent for G^{PEL} and the branch lengths indicated (for any u,v) follows from a straightforward application of Theorem 8.7.1 in Semple and Steel (2003), which shows that parsimony is consistent for the locus 1 and 2 trees with the indicated branch lengths (for any u and also any v in Fig. 6). Since parsimony is a linear scoring scheme, if it is consistent on each block of a partition it is consistent on all the data.

To prove the same for likelihood, consider first a single site. Notice that the locus 2 tree is obtained from the locus 1 tree by changing just one pendant branch length from u to v. So, if z is the state at this pendant leaf (Y), x the state at the other end (interior node) of this edge, and W is the collective set of states at the rest of the tree, then by the Markov property for the G^{PEL} generating model we have:

$$Pr(W,z|x)=\alpha Pr_u(W,z|x)+(1-\alpha)Pr_v(W,z|x), \qquad (3)$$

where proportion α of sites evolve on locus 1 tree (and $1-\alpha$ on locus 2 tree), $Pr_s(W,z|x)$ denotes the conditional probability of leaf states W and z given that interior node is in state x, and s is the length of the pendant edge in question. Now, consider the tree obtained from the locus 1 tree setting the length of the pendant edge incident with Y equal to θ, where (for the Jukes–Cantor model):

$$\exp\left(-\frac{4}{3}\theta\right)=\alpha\exp\left(-\frac{4}{3}u\right)+(1-\alpha)\exp\left(-\frac{4}{3}v\right) \qquad (4)$$

then, from Equation (3)

$$Pr(W,z|x)=Pr_\theta(W,z|x)$$

for all x,y,Z, and so

$$Pr(W,z)=Pr_\theta(W,z).$$

In particular, the site pattern distribution under this G^{PEL} generating model matches exactly the probability distribution of a Model I generating model in which the length of the pendant edge incident with leaf Y is the particular value θ chosen by Eqn. (4), and the remaining edges have length u. In other words, this G^{PEL} generating model produces site pattern probabilities the same as Model I on the same topology and with an intermediate branch length. These results are based on a single site, but it follows by Theorem 5.1 of Chang (1996), that if we now use ML (on sequences from the model) under the Model I inference model this will be a statistically consistent estimator of tree topology.

Discussion

The new results on terraces presented here are relevant to an increasingly common setting for phylogenetic analysis in which large multi-locus data sets with significant numbers of missing sequences are combined with models that partition the data in various ways during likelihood or Bayesian inference. They are also relevant to analyses using maximum parsimony irrespective of partitioning schemes, which remains a computationally attractive method for analyses of very large data sets, especially with the use of phylogenetic placement techniques (Matsen 2015). Previously we had shown that the terrace of equally optimal trees surrounding any specific tree could be quite large, although it was possible to exploit certain mathematical results on subtrees and supertrees to help characterize these sets of trees. In the present paper we found it useful to define "stands" of trees and distinguish between stands and terraces. Stands are collections of trees each of which displays all the subtrees associated with the separate blocks within a partitioned data set. Under some conditions the trees in a stand can all have the same optimality score, in which case they are called terraces, to reflect their constant "elevation" along a vertical axis corresponding to that score.

If the score is parsimony, a stand is always a terrace. The first new result we described specifies sufficient conditions under which this will also be true for the likelihood score. We showed that the parameters describing edge lengths in different partition blocks must be independent ("edge-unlinked" or "EUL"). Previously we had also assumed parameters of the rate matrix to be unlinked (Sanderson et al. 2011). This is sufficient but not necessary. Because practice in current phylogenetics of multi-locus supermatrices ranges from using completely linked models to highly partitioned ones that unlink both rate matrices and edge parameters, impacts of these divergent strategies should be assessed. Our results indicate that it is possible for EUL models to induce potentially large terraces around any given tree found during tree reconstruction. This raises the question of whether this ambiguity accurately reflects phylogenetic uncertainty or is a product of overparameterization of the EUL model. Overparameterization is expected to lead to higher variance in estimates of the parameters, including tree topology (Li et al. 2008). A deeper study of model selection in the context of terraces may be helpful.

The second new result indicates that the stand associated with one partitioning scheme can actually be part of a larger stand under a refined partitioning scheme. For parsimony, this leads to strong conclusions: it is possible to easily identify the extent of the largest terrace associated with any given partitioning scheme simply by constructing the so-called maximal partitioning scheme for the data set (a simple function of the distribution of missing data in the matrix), and then characterizing the terrace for that maximal partitioning scheme. This places an upper bound on the extent of the ambiguity associated with partial taxon coverage, and therefore should serve as the default terrace reported for any specific tree when parsimony is the optimality score.

These results are fundamentally based on the properties of stands. However, in considering the optimality scores of the trees on a stand, it is worth remembering that even a "maximal" terrace may be imbedded in a yet larger collection of trees with the same optimality score owing to homoplasy, constant sites, or possibly other factors in the data not related to partial taxon coverage. The maximal number of trees on a terrace is a lower bound on total this larger number of trees. The fact that it can be calculated sidestepping the sequence data proper may be relevant for tree search heuristics, which are clearly challenged by tree landscapes such as these (Goloboff 2014). For example, rather than doing tree rearrangements to find equally parsimonious trees, it should be possible to enumerate them directly. These can form "seeds" to continue searching using expensive tree score computations.

In likelihood inference the size of a terrace is also conditional on the partitioning scheme, but so is the likelihood score of the trees on the terrace. The terrace associated with the maximal partitioning scheme may reflect a worst case with respect to ambiguity but may or may not be close to a reasonable partitioning scheme for the data at hand. Selection of the best partitioning scheme is presumably guided by principles of model selection—not the size of the resulting terrace—but for some data sets it may be worth considering the consequences of model selection on terrace sizes. Moreover, although the terraces corresponding to different partitioning schemes may have different likelihood scores, the containment relations of the sets of trees in stands for different partitioning schemes may allow more efficient computational exploration of the impact of partitioning on the overall likelihood landscape.

The remaining new results address various aspects of the impact of terraces on assessing tree accuracy. The third and fourth results refer to situations in which bootstrap proportions and Bayesian posterior probabilities, respectively, are determined in large part not by the information content of the sequence data but by the pattern of partial taxon coverage. These happen for different reasons. In bootstrapping, the trees sampled from a terrace end up reflecting the frequencies of trees on that terrace, which in turn reflects the pattern of taxon coverage and tree shape. One consequence is that a clade can be highly supported in such an analysis because most trees on the terrace (all with equal likelihood scores) have that clade. Wilkinson and Benton (1996, p. 14) pointed out a similar finding for the impact of an unstable "rogue" taxon on support values derived from majority rule consensus frequencies. It is possible to view this as either a "feature" or a "bug," but either way one might not be expecting such sensitivity to arise from factors other than the sequence data. Simmons and colleagues (Simmons and Freudenstein 2011; Simmons b; Simmons and Goloboff 2014) have noted a number of impacts of missing data generally on bootstrap estimates and have been critical of its application in this context, suggesting several strategies to check well supported clades for spurious support. We suspect some but not all of their observations in real data sets (e.g., Simmons and Goloboff 2014) are due to issues related to terraces, but other factors involving homoplasy and phylogenetic signal proper are no doubt also involved.

Similarly, Bayesian inference can be influenced by posterior probabilities differing across trees on a terrace (despite the likelihood being the same). In this case, the result is largely explained by the influence of priors. With standard exponentially distributed prior probabilities on edge lengths, the probability that a single taxon with all its data missing will attach to a particular edge of the tree is determined by that edge length in relation to the length of the entire tree. In the absence of any other information, Bayesian inference will place a taxon on the longest edge of the tree (though not necessarily with high probability). Again, this may be viewed as desirable or not, but it has the downstream consequence that trees in a stand can have very different posterior probabilities, which are determined by a fairly opaque convolution of the priors, edge lengths, partial coverage pattern, and combinatorics of terraces and tree shapes.

These findings are superficially similar to the "long branch attraction" described elsewhere for Bayesian phylogenetic inference (Susko 2008; Kolaczkowski and Thornton 2009) in the absence of missing data. Although we do not yet fully understand the interaction between factors contributing to apparent long branch attraction artifacts in the small example we posed (Fig. 5), the effect of missing data was much stronger than edge lengths alone. Restoring the missing data to the matrix was sufficient to completely remove any long-branch artifacts in the posterior probability distribution.

A reasonable response to some of these concerns would be to avoid the possibly overparameterized EUL partitioning schemes in likelihood or Bayesian inference and use simpler and more homogeneous models for statistical inference. Our last new results suggest reasons to proceed carefully, however. When sequence data are generated on a tree with even a small amount of heterotachy between loci, an EL or PEL model is used for inference, and there is only partial taxon coverage, then the likelihood score for the correct tree can actually be worse than that of an incorrect tree. Simulations with long sequence lengths imply this is an instance of statistical inconsistency. As stated this might seem like simply another case in which model mis-specification causes problems, but in this instance, it is clearly a negative interaction between model mis-specification and missing data, because the problem can be avoided by supplying the missing data. If this is not possible, use of the (overparameterized) EUL inference model at least avoids selection of an incorrect tree, but it provides no evidence in favor of the correct one.

Much further work is needed to narrow down the precise effects of terraces on confidence estimation and accuracy of inference. In the meantime, however, there are options for remediation. Setting aside the strategy of acquiring the missing sequences to eliminate partial taxon coverage entirely, which may be expensive or impossible depending on availability of DNA samples, there are computationally promising approaches. Previously we posed the “maximum defining label set” (MDLS) problem (Sanderson et al. 2011), in which we seek the smallest number of leaf taxa to delete from the coverage matrix such that the stand for a given tree is reduced to a single tree. For example, in the two-locus partitioning scheme I of Figure 3, the removal of taxa E and F makes the taxon coverage matrix decisive, and therefore the indicated tree is alone on a terrace of size one (which is also true for every other tree). The cost, of course, in this simple example, is the significant loss of potential information about the deleted taxa.

The MDLS problem has an exact and efficient solution for two loci, and experiments with data sets indicate there are interesting instances in which elimination of relatively few taxa can solve the problem, even while leaving a significant amount of missing data. However, there is no known exact solution for the case of three or more loci. The good news is that simple heuristics in the case of three or more loci can find solutions that eliminate terraces (Sanderson et al. 2011); they just may not be optimal (it may have been possible to do the same thing and keep more taxa in the matrix).

It would also be interesting to explore how procedures designed to ameliorate the impact of unstable “rogue” taxa, characterized by having much missing data, apply to terraces. “Safe taxonomic reduction” (Wilkinson 1995) deletes a taxon with missing data if the character states that are present are a subset of another taxon’s, thus reducing the size of the set of equally optimal trees found in parsimony analysis. Solutions to our MDLS problem do not depend on the states observed in the data, so a better understanding of the relationship of the two approaches might lead to a more synthetic characterization of all factors producing equally optimal trees.

To highlight the impacts clearly, most of our results were in the context of sequence data sets so large that the only error was due to partial taxon coverage. In real data sets there is also error from the finite sample taken from the substitution process. This translates into a broadening of the bootstrap or posterior distribution of trees. In addition, there may be distinct terraces associated with each sample tree taken from these distributions; and there may be a distinct MDLS solution for deleting some set of taxa for each of these trees. How do we integrate across this information to make headway in reducing the overall impact of terraces? A simple but conservative fix might be to replace any sampled tree in a bootstrap replicate or an MCMC run with the strict consensus of the stand in which that tree is imbedded. Then any clade on that tree is present in all trees on the terrace. This would tend to reduce the false-positive clades uncovered in an analysis.

Another approach would be to rely on statistics other than those based on consensus. For example, a terrace could be characterized by the average dissimilarity among its trees, based on a measure of distance between trees, such as the Robinson–Foulds distance (RF: Robinson and Foulds 1981). This could allow a better assessment of the extent of the confidence set of trees sampled from bootstrap replicates or the posterior distributions, perhaps based on measures of “distance” between terraces. One approach would be to use a measure of distance between sets, such as the Hausdorff distance (Yu et al. 2014). The Hausdorff distance is small when each tree on one terrace is close (measured here by RF distance) to some tree on the other terrace. This might be true even if the average RF distance between trees *within* a terrace is much larger. Whether this approach ultimately proves promising or not, some means to characterize more fully the relationships between entire sets of trees seems to be a necessity when terraces are commonplace.

Finally, the existence of terraces poses an immediate practical challenge for numerical phylogenetic analysis using maximum likelihood—how to distinguish the termination of a heuristic search because of discovery of a local peak in the likelihood surface from termination due to “entrapment” on a terrace of equally optimal trees. Because termination criteria vary and are heavily affected by considerations like numerical precision, no one solution may be available, but minimally it should prove useful to canvas some space of nearby trees that are “off-terrace” to check if their likelihood scores are indeed lower (or at least not higher).

Funding

This work was supported by the US National Science Foundation (DEB-1353815 to MJS, DJZ, AS, MMM, and MS) and the New Zealand Marsden Fund (to MS).

Acknowledgments

We thank J. Charboneau and M. Simmons for discussion; and Laura Kubatko, Mark Wilkinson, and an anonymous reviewer for helpful comments.

Appendix: Mathematical proof of Proposition 1 and Theorem 4

Proof of Proposition 1

Consider an analysis where for each locus i we are free to select E_i-parameters, but the M_i parameters are constrained to be identical (i.e., $M_i = M$ for all i). Let $\varphi(T)$ denote the log-likelihood of tree T (having leaf set $\mathcal{L}$) for the data $D = (D_1, \ldots, D_k)$ for the $k = |\mathcal{P}|$ loci. Then assuming the loci evolve independently (conditional on the parameter choices) we have:

$$\varphi(T) = \sup_{(M,(E_i))} \sum_{i=1}^{k} \log \mathbb{P}(D_i | T, M, (E_i)),$$

where 'sup' refers to supremum (i.e., maximum if it is attained, else its limiting value) as we search over M and the E_i parameter spaces, and where (E_i) is short for $(E_1, \ldots, E_k)$. Now,

$$\mathbb{P}(D_i | T, M, (E_i)) = \mathbb{P}(D_i | (T|\mathcal{L}_i), M, E_i),$$

(notice that T and (E_i) on the left has been replaced by $(T|\mathcal{L}_i)$ and E_i on the right). Combining the above two equations gives:

$$\varphi(T) = \sup_{(M,(E_i))} \sum_{i=1}^{k} \log \mathbb{P}(D_i | (T|\mathcal{L}_i), M, E_i),$$

and so

$$\varphi(T) = \sup_{M} \sum_{i=1}^{k} \sup_{E_i} \log \mathbb{P}(D_i | (T|\mathcal{L}_i), M, E_i). \tag{A.1}$$

Now, suppose that T is a phylogenetic tree on the entire leaf set $\mathcal{L}$ and that T maximizes $\varphi(*)$. Let T' be any other phylogenetic tree on leaf set $\mathcal{L}$ for which $T|\mathcal{L}_i = T'|\mathcal{L}_i$ for all i (i.e., T' lies in the same terrace at T). Then from Equation (A.1) we have:

$$\varphi(T) = \sup_{M} \sum_{i=1}^{k} \sup_{E_i} \log \mathbb{P}(D_i | (T|\mathcal{L}_i), M, E_i) =$$

$$\sup_{M} \sum_{i=1}^{k} \sup_{E_i} \log \mathbb{P}(D_i | (T'|\mathcal{L}_i), M, E_i) = \varphi(T'),$$

so T' is an ML tree also. In other words, all trees on the same terrace as T are ML trees. This completes the proof. □

Proof of Theorem 4

We first begin by defining more formally some of the notions mentioned earlier.

- Given any binary phylogenetic X tree T', and any taxon x from X, consider the tree T'_{-x} that is obtained from T' by deleting leaf taxon x and its incident edge $e(x)$. Note that each edge of T'_{-x} corresponds to an edge of T', except for the edge e_{-x} of T'_{-x} which corresponds to the two edges (e_1, e_2) of T', that are incident with $e(x)$ in T'. In this way, if T' comes equipped with a branch length assignment l (so $l(e)$ is the length of edge e), then the induced branch length l_{-x} function for T'_{-x} is thus given by:

$$l_{-x}(e) = \begin{cases} l(e), & \text{if } e \neq e_{-x}; \\ l(e_1) + l'(e_2), & \text{if } e = e_{-x}. \end{cases}$$

- For any data set D that consists of a sequence of k aligned site patterns on X, and any taxon $x \in X$, let D_{-x} denote sequence of k aligned site patterns on $X - \{x\}$ obtained by deleting the sequence for x.

- Now suppose that the sequence sites in D have evolved i.i.d. on some fixed binary phylogenetic $X-$tree T with branch length assignment λ, under a reversible Markovian process. Thus the sites in D_{-x} evolve i.i.d. on T_{-x} with branch length assignment λ_{-x}.

We wish to apply a Bayesian approach to compare different placements of the taxon x into T_{-x} given the censored data D_{-x}. We assume a prior probability distribution on the set of binary phylogenetic X-trees with branch lengths for which:

(i) each binary tree has the same probability (i.e., the 'PDA distribution');

(ii) edge lengths are independent exponential random variables.

Without loss of generality (by rescaling) we may assume that the mean of the exponential distribution in (ii) is 1.

Suppose we have two binary phylogenetic X-trees T' and T'' that satisfy $T'_{-x} = T''_{-x} = T_{-x}$ (i.e., two different placements of leaf x in T_{-x}). One (or neither) of these trees might be T. We are interested in the ratio of posterior probabilities $\frac{\mathbb{P}(T'|D_{-x})}{\mathbb{P}(T''|D_{-x})}$. The following result states that for long sequences this ratio converges towards the ratio of the lengths of the two edges of T_{-x} to which the missing taxon (x) is attached. To establish Theorem 4, we prove the following result.

Theorem 5 *For data generated by a reversible Markov process on a phylogenetic X-tree T with branch length assignment λ, consider, for any $x \in X$, any two phylogenetic X-trees T' and T'' obtained by attaching x to edges of T_{-x} of length l' and l'' respectively. Then the ratio $\frac{\mathbb{P}(T'|D_{-x})}{\mathbb{P}(T''|D_{-x})}$ converges in probability to l'/l'', as the sequence length k grows.*

Notice that Theorem 5 implies Theorem 4 since the statistical consistency of Bayesian phylogenetics under identifiable models implies that any tree which is different from T_{-x} when x is deleted has a posterior probability that converges to zero as the sequences length k grows (thus $\mathbb{P}(T''|D_{-x})$ sums to one as we sum over all binary trees T'' that agree with T up to the placement of x). Thus, the remainder of our argument is tailored toward proving Theorem 5 under the same conditions stated for Theorem 4 (in particular, conditions (i) and (ii) in the preamble to that theorem).

From Bayes' identity we have:

$$\mathbb{P}(T'|D_{-x})=\frac{\mathbb{P}(D_{-x}|T')\mathbb{P}(T')}{\mathbb{P}(D_{-x})}. \tag{A.2}$$

Now, $\mathbb{P}(T)=\mathbb{P}(T')$ (by assumption (i)), and so, from Equation (A.2) and the analogous identity for $\mathbb{P}(T|D_{-x})$:

$$\frac{\mathbb{P}(T'|D_{-x})}{\mathbb{P}(T|D_{-x})}=\frac{\mathbb{P}(D_{-x}|T')}{\mathbb{P}(D_{-x}|T)}. \tag{A.3}$$

Moreover,

$$\mathbb{P}(D_{-x}|T)=\int_{\Gamma}\mathbb{P}(D_{-x}|T_{-x},l)f(l)dl \tag{A.4}$$

where l is the set of branch length assignment on T_{-x}, and where $f(l)$ refers to the density of the branch lengths on T_{-x} that is induced by independent exponential prior branch lengths on T (Γ is the set of possible branch lengths of T_{-x}).

Similarly,

$$\mathbb{P}(D_{-x}|T')=\int_{\Gamma}\mathbb{P}(D_{-x}|T'_{-x},l)f'(l)dl \tag{A.5}$$

where l is the branch length assignment on $T'_{-x}(=T_{-x})$, and where $f'(l)$ refers to the density of the branch lengths on T_{-x} induced by the independent exponential priors on T'.

Let $\hat{s}$ denote the empirical frequency distribution of site patterns on $X-x$, and for any assignment l of branch lengths to T_{-x} that is complete (i.e., l assigns a length to every edge of T_{-x}) let $p(l)$ denote the vector of site pattern probabilities generated by T_{-x} with these branch lengths. We then have the identity:

$$\mathbb{P}(D_{-x}|T_{-x},l)/\prod_i \hat{s}_i^{\hat{s}_i k}=\exp(-kd_{\mathrm{KL}}(\hat{s}||p(l)), \tag{A.6}$$

where i ranges over all site patterns, and where $d_{\mathrm{KL}}(P||Q)=\sum_i P_i\log(P_i/Q_i)$ refers to Kullback–Leibler separation of probability distributions P and Q. Notice that the branch lengths l are fixed (and completely specified for T_{-x}) in (A.6). Similarly, for any complete assignment l of branch lengths to T_{-x},

$$\mathbb{P}(D_{-x}|T'_{-x},l)/\prod_i \hat{s}_i^{\hat{s}_i k}=\exp(-kd_{\mathrm{KL}}(\hat{s}||p(l)), \tag{A.7}$$

Combining Equations (A.3), (A.4), (A.5), (A.6), and (A.7) we obtain:

$$\frac{\mathbb{P}(T'|D_{-x})}{\mathbb{P}(T|D_{-x})}=\frac{\int_{\Gamma}\exp(-kd_{KL}(\hat{s}||p(l))f'(l)dl}{\int_{\Gamma}\exp(-kd_{KL}(\hat{s}||p(l))f(l)dl}. \tag{A.8}$$

Now, let B_k denote the subspace of the branch length space Γ of T_{-x} that is within (l_∞) distance $k^{-1/4}$ of λ_{-x}. Then for $g=f$ or $g=f'$ we have the following convergence in probability as k grows:

$$R_k:=\frac{\int_{\Gamma-B_k}\exp(-kd(\hat{s}||p(l))g(l)dl}{\int_{B_k}\exp(-kd(\hat{s}||p(l))g(l)dl}\xrightarrow{p}0. \tag{A.9}$$

The proof of this last equation is given in a separate subsection below. We apply it as follows. Notice that for $g=f$ or $g=f'$, we have:

$$\int_{\Gamma}\exp(-kd(\hat{s}||p(l))g(l)dl=\int_{B_k}\exp(-kd(\hat{s}||p(l))g(l)dl$$
$$+\int_{\Gamma-B_k}\exp(-kd(\hat{s}||p(l))g(l)dl,$$

and so

$$\int_{\Gamma}\exp(-kd(\hat{s}||p(l))g(l)dl=$$
$$(1+R_k)\int_{B_k}\exp(-kd(\hat{s}||p(l))g(l)dl. \tag{A.10}$$

Moreover, since g is continuous, and the nested sequence of sets B_k convergences on the vector λ_x as $k\rightarrow\infty$ we have:

$$\frac{\int_{B_k}\exp(-kd(\hat{s}||p(l))g(l)dl}{\int_{B_k}\exp(-kd(\hat{s}||p(l))dl}\xrightarrow{p}g(\lambda_{-x}), \tag{A.11}$$

as k grows. Thus, combining Equations (A.8), (A.10), and (A.11) we obtain:

$$\frac{\mathbb{P}(T'|D_{-x})}{\mathbb{P}(T|D_{-x})}\xrightarrow{p}\frac{f'(\lambda_{-x})}{f(\lambda_{-x})}. \tag{A.12}$$

Now, by assumption (ii), the branch lengths in T_{-x} are independent exponentials (of mean 1) for all edges other than e_{-x}, and for this edge the branch length is the sum of two independent exponential(s) of mean 1, which has a gamma distribution with density $f(t)=t\exp(-t)$. Thus,

$$f(l)=\prod_{e'\neq e_{-x}}\exp(-l(e))\cdot[l(e_{-x})\exp(-l(e_{-x}))]$$
$$=l(e_{-x})\prod_{e}\exp(-l(e)), \tag{A.13}$$

where the last product term is over all edges of T_{-x}. Similarly, if e'_{-x} is the edge of $T_{-x}(=T'_{-x})$ that x is attached to in T' then

$$f'(l)=l(e'_{-x})\prod_{e}\exp(-l(e)). \tag{A.14}$$

From Equations (A.13) and (A.14) we have:

$$\frac{f'(\lambda_{-x})}{f(\lambda_{-x})} = \frac{\lambda_{-x}(e'_{-x})}{\lambda_{-x}(e_{-x})},$$

which, from (A.12), implies that for T' and T'' with $T'_{-x} = T''_{-x} = T_{-x}$, we have:

$$\frac{\mathbb{P}(T'|D_{-x})}{\mathbb{P}(T''|D_{-x})} \xrightarrow{p} \frac{l'}{l''},$$

where $l' = \lambda_{-x}(e'_{-x})$ and $l'' = \lambda_{-x}(e''_{-x})$, and where e'_{-x} and e''_{-x} are the corresponding edges of T_{-x} that x attaches to in T' and T'' respectively. This completes the proof of Theorem 5 and thereby Theorem 4, modulo the remaining step of establishing Equation (A.9) which we attend to below. □

Proof of Equation (A.9)

A classic result (e.g., Wilk's theorem) ensures that the following convergence in distribution holds:

$$2kd(\hat{s}||p(\lambda_{-x})) \xrightarrow{D} \chi^2_{N-1}$$

where χ^2_{N-1} is a chi-square distribution with $N-1$ degrees of freedom (here N is the number of possible site patterns). By the continuous mapping theorem, it now follows that:

$$\exp(-kd(\hat{s}||p(\lambda_{-x}))) \xrightarrow{D} W \tag{A.15}$$

where $W = \exp(-\chi^2_{N-1})$ is a continuous and non-negative random variable.

Moreover, if a sequences of branch length vectors l_k lies within (l_∞) distance $\frac{1}{k}$ of λ_{-x} then we also have:

$$\exp(-kd(\hat{s}||p(l_k))) \xrightarrow{D} W. \tag{A.16}$$

(For further details see Serfling (1980), esp. Section 3.5).

Next, Pinsker's inequality (see Cover and Thomas (2006)) gives for any $l \in \Gamma$:

$$d(\hat{s}||p(l)) \geq \frac{1}{2}||\hat{s} - p(l)||_1^2,$$

where $\|\cdot||_1$ refers to the l_1 metric. The triangle inequality for this metric then gives:

$$d(\hat{s}||p(l)) \geq \frac{1}{2}(\|p(\lambda_{-x}) - p(l)\|_1 - \|\hat{s} - p(\lambda_{-x})\|_1)^2. \tag{A.17}$$

Now, for any $l \in \Gamma - B_k$, we have $\|l - \lambda_{-x}\|_\infty \geq k^{-1/4}$, and so, by Theorem 2.1(2) of (Moulton and Steel 1999) there exists a pair of leaves i, j so that the difference in path length between these leaves under branch lengths l and λ_{-x} is at least $\frac{1}{2}k^{-1/4}$. Since the site substitution model is reversible, the probability that two leaves are in the same state is a monotone decreasing function of the path length between them (a positive mixture of exponential functions). This in turn implies that the event that leaves i and j are in the same state differs in probability under the branch lengths l and λ_{-x} by an amount that is at least $k^{-1/4}$ times some constant (dependent on the model, and λ_{-x}). In particular,

$$\|p(\lambda_{-x}) - p(l)\|_1 \geq ck^{-1/4}, \text{ for some constant } c > 0. \tag{A.18}$$

Also,

$$\|\hat{s} - p(\lambda_{-x})\| \leq k^{-1/3}, \tag{A.19}$$

with probability converging to 1 as k grows. Consequently, by combining Equations (A.17), (A.18) and (A.19), the following inequality holds for all $l \in \Gamma - B_k$ with probability converging to 1 as k grows:

$$\frac{\exp(-kd(\hat{s}||p(l)))}{\exp(-dk^{1/2})} \leq 1,$$

for some constant $d > 0$. Thus, with probability converging to 1 as k grows:

$$\frac{\int_{\Gamma - B_k}\exp(-kd(\hat{s}||p(l))g(l)dl}{\exp(-dk^{1/2})} \leq \lim_{k\to\infty}\int_{\Gamma - B_k} 1 \cdot g(l)dl = 1. \tag{A.20}$$

On the other hand, if we let $B_k^* \subset B_k$ be the set of branch length vectors that lie within (l_∞) distance at most $1/k$ from λ_{-x} then

$$\frac{\int_{B_k^*}\exp(-kd(\hat{s}||p(l))g(l)dl}{\int_{B_k^*}g(l)dl} \geq \sup_{l\in B_k^*}\{\exp(-kd(\hat{s}||p(l))\}. \tag{A.21}$$

Now, for a value $C > 0$ that is independent to k (but dependent on λ_{-x}), we have $\int_{B_k^*}g(l)dl \geq Ck^{-E}$, where E is the number of edges of T. Also, from Equation (A.16) $\sup_{l\in B_k^*}\{\exp(-kd(\hat{s}||p(l))\}$ converges in distribution to the random variable $W = \exp(-\chi^2_{N-1})$ as k grows. Thus, from Equation (A.21), the following inequality holds with probability converging to 1 as k grows: $\int_{B_k^*}\exp(-kd(\hat{s}||p(l))g(l)dl / Ck^{-E} \geq W$, and thus,

$$\frac{\int_{B_k}\exp(-kd(\hat{s}||p(l))g(l)dl}{k^{-E}} \geq W, \tag{A.22}$$

since $B_k^* \subset B_k$ and the integrand is non-negative. Combining Equations (A.20) and (A.22) the following inequality holds with probability converging to 1 with increasing k:

$$R_k \leq \frac{\exp(-dk^{1/2})}{Ck^{-E}} \cdot \frac{1}{W},$$

Notice that the second term on the right $(\frac{1}{W})$ is a continuous random variable, but since $\mathbb{P}(W > 0) = 1$ and since the first term on the right converges to 0 (absolutely) as k tends to infinity, this suffices to establish Equation (A.9).

References

Aho A., Sagiv Y., Szymanski T.G., Ullman J. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM J. Comput. 10:405–421.

Ane C., Eulenstein O., Piaggio-Talice R., Sanderson M.J. 2009. Groves of phylogenetic trees. Ann. Comb. 13:139–167.

Boettiger C., Coop G., Ralph P. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution 66: 2240–2251.

Bordewich M. 2003. The complexity of counting and randomized approximation [Thesis]. University of Oxford.

Bryant D. 1997. Building trees, hunting for trees, and comparing trees: theory and methods in phylogenetic analysis [Thesis]. University of Canterbury.

Burleigh J.G., Hilu K.W., Soltis D.E. 2009. Inferring phylogenies with incomplete data sets: a 5-gene, 567-taxon analysis of angiosperms. BMC Evol. Biol. 9. DOI: 10.1186/1471-2148-9-61.

Chamberlain S.A., Hovick S.M., Dibble C.J., Rasmussen N.L., Van Allen B.G., Maitner B.S., Ahern J.R., Bell-Dereske L.P., Roy C.L., Meza-Lopez M., Carrillo J., Siemann E., Lajeunesse M.J., Whitney K.D. 2012. Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta-analysis. Ecol. Lett. 15: 627–636.

Chang J. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. 137:51–73.

Cho S., Zwick A., Regier J.C., Mitter C., Cummings M.P., Yao J.X., Du Z.L., Zhao H., Kawahara A.Y., Weller S., Davis D.R., Baixeras J., Brown J.W., Parr C. 2011. Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? Syst. Biol. 60: 782–796.

Christin P.-A., Osborne C.P., Chatelet D.S., Columbus J.T., Besnard G., Hodkinson T.R., Garrison L.M., Vorontsova M.S., Edwards E.J. 2013. Anatomical enablers and the evolution of C-4 photosynthesis in grasses. P. Natl. Acad. Sci. USA 110:1381–1386.

Constantinescu M. 1995. An efficient algorithm for supertrees. J. Classif. 12:101–112.

Constantinescu M., Sankoff D. 1986. Tree enumeration modulo a consensus. Journal of Classif. 3:349–356.

Cover T.M., Thomas J.A. 2006. Elements of information theory, 2nd edn. New York: Wiley-Interscience.

Crawley S.S., Hilu K.W. 2012. Impact of missing data, gene choice, and taxon sampling on phylogenetic reconstruction: the Caryophyllales (Angiosperms). Plant Syst. Evol. 298:297–312.

Davis J.I., Stevenson D.W., Petersen G., Seberg O., Campbell L.M., Freudenstein J.V., Goldman D.H., Hardy C.R., Michelangeli F.A., Simmons M.P., Specht C.D., Vergara-Silva F., Gandolfo M. 2004. A phylogeny of the monocots, as inferred from rbcL and atpA sequence variation, and a comparison of methods for calculating jackknife and bootstrap values. Syst. Bot. 29:467–510.

Davis M.P., Midford P.E., Maddison W. 2013. Exploring power and parameter estimation of the BiSSE method for analyzing species diversification. BMC Evol. Biol. 13:11.

DiCiccio T., Efron B. 1996. Bootstrap confidence intervals. Stat. Sci. 11:189–228.

Driskell A.C., Ané C., Burleigh J.G., McMahon M.M., OMeara B., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172–1174.

Fabre P.-H., Hautier L., Dimitrov D., Douzery E.J.P. 2012. A glimpse on the pattern of rodent diversification: a phylogenetic approach. BMC Evol. Biol. 12.

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

Goldberg E.E., Igic B. 2012. Tempo and mode in plant breeding system evolution. Evolution 66:3701–3709.

Goloboff P. 2014. Hide and vanish: data sets where the most parsimonious tree is known but hard to find, and their implications for tree search methods. Mol. Phyl. Evol 79:118–131.

Goloboff P.A. 1991. Homoplasy and the choice among cladograms. Cladistics 7:215–232.

Gordon A.D. 1986. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. J. Classif. 3:31–39.

Hedin M., Starrett J., Akhter S., Schoenhofer A.L., Shultz J.W. 2012. Phylogenomic resolution of paleozoic divergences in Harvestmen (Arachnida, Opiliones) via analysis of next-generation transcriptome data. PLOS One 7.

Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., Martinez P., Baguna J., Bailly X., Jondelius U., Wiens M., Muller W.E.G., Seaver E., Wheeler W.C., Martindale M.Q., Giribet G., Dunn C.W. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. P. Roy. Soc. Lond. B Bio. 276:4261–4270.

Hess J., Goldman N. 2011. Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: yeasts revisited. PLOS One 6.

Hinchliff C.E., Roalson E.H. 2013. Using supermatrices for phylogenetic inquiry: an example using the sedges. Syst. Biol. 62:205–219.

Izquierdo-Carrasco F., Smith S.A., Stamatakis A. 2011. Algorithms, data structures, and numerics for likelihood-based phylogenetic inference of huge trees. BMC Bioinf. 12. DOI: 10.1186/1471-2105-12-470.

Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: Mistaken assumptions and conclusions. Syst. Biol. 51: 369–381.

Kolaczkowski B., Thornton J.W. 2009. Long-branch attraction bias and inconsistency in Bayesian phylogenetics. Plos One 4.

Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58: 130–145.

Letsch H.O., Meusemann K., Wipfler B., Schuette K., Beutel R., Misof B. 2012. Insect phylogenomics: results, problems and the impact of matrix composition. P. Roy. Soc. Lond. B Bio. 279:3282–3290.

Li C.H., Lu G.Q., Orti G. 2008. Optimal data partitioning and a test case for ray-finned fishes (actinopterygii) based on ten nuclear loci. Syst. Biol. 57:519–539.

Liu K., Linder C.R., Warnow T. 2012. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. PLOS One 6.

Maddison D.R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. 40:315–328.

Maddison W. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5:365–377.

Marazzi B., Ané C., Simon M.F., Delgado-Salinas A., Luckow M., Sanderson M.J. 2012. Locating evolutionary precursors on a phylogenetic tree. Evolution 66:3918–3930.

Matsen F.A. 2015. Phylogenetics and the human microbiome. Syst. Biol. 64:e26–e41.

Moulton V., Steel M.A. 1999. Retractions of finite distance functions onto tree metrics. Discr. Appl. Math. 91:215–233.

Pagel M., Meade A. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. Philos. T. R. Soc. B. 363:3955–3964.

Pyron R.A., Wiens J.J. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Mol. Phylogenet. Evol. 61: 543–583.

Rabosky D.L., Santini F., Eastman J., Smith S.A., Sidlauskas B., Chang J., Alfaro M.E. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. Nat. Commun. 4.

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

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Hohna 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 datasets. Mol. Biol. Evol. 30:197–214.

Sabir J., Schwarz E., Ellison N., Zhang J., Baeshen N.A., Mutwakil M., Jansen R., Ruhlman T. 2014. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotech. J. 12:743–754.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Salter L.A. 2001. Complexity of the likelihood surface for a large DNA dataset. Syst. Biol. 50:970–978.

Sanderson M.J. 2007. Construction and annotation of large phylogenetic trees. Aust. Syst. Bot. 20:287–301.

Sanderson M.J. 2008. Phylogenetic signal in the eukaryotic tree of life. Science 321:121–123.

Sanderson M.J., McMahon M.M., Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol. Biol. 10:155.

Sanderson M.J., McMahon M.M., Steel M. 2011. Terraces in phylogenetic tree space. Science 333:448–450.

Semple C. 2003. Reconstructing minimal rooted trees. Discr. Appl. Math. 127:489–503.

Semple C., Steel M. 2003. Phylogenetics. New York: Oxford University Press.

Serfling R.J. 1980. Approximation theorems of mathematical statistics. (Wiley Series in Probability and Mathematical Statistics). New York: John Wiley & Sons.

Simmons M.P. 2012a. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics 28:208–222.

Simmons M.P. 2012b. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. Mol. Phylogenet. Evol. 62:472–484.

Simmons M.P. 2014. Limitations of locally sampled characters in phylogenetic analyses of sparse supermatrices. Mol. Phylogenet. Evol. 74:1–14.

Simmons M.P., Freudenstein J.V. 2011. Spurious 99% bootstrap and jackknife support for unsupported clades. Mol. Phylogenet. Evol. 61:177–191.

Simmons M.P., Goloboff P.A. 2013. An artifact caused by undersampling optimal trees in supermatrix analyses of locally sampled characters. Mol. Phylogenet. Evol. 69:265–75.

Simmons M.P., Goloboff P.A. 2014. Dubious resolution and support from published sparse supermatrices: the importance of thorough tree searches. Mol. Phylogenet. Evol. 78:334–348.

Siu-Ting K., Pisani D., Creevey C.J., Wilkinson M. 2014. Concatabominations: identifying unstable taxa in morphological phylogenetics using a heuristic extension to safe taxonomic reduction. Syst. Biol. 64:137–143.

Smith S.A., Beaulieu J.M., Donoghue M.J. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evol. Biol. 9.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics.

Stamatakis A., Alachiotis N. 2010. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. Bioinformatics 26:i132–i139.

Steel M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. J. Classif. 9:91–116.

Steel M., Sanderson M.J. 2010. Characterizing phylogenetically decisive taxon coverage. Appl. Math. Lett. 23:82–86.

Susko E. 2008. On the distributions of bootstrap support and posterior distributions for a star tree. Syst. Biol. 57:602–612.

Susko E. 2009. Bootstrap support is not first order correct. Syst. Biol. 58:211–233.

Swofford D.L. 1999. PAUP *. Phylogenetic anaysis using parsimony and other methods version 4.

Wiens J.J. 2011. Re-evolution of lost mandibular teeth in frogs after more than 200 million years, and re-evaluating Dollo's law. Evolution 65:1283–1296.

Wiens J.J., Morrill M.C. 2011. Missing data in phylogenetic analysis: Reconciling results from simulations and empirical data. Syst. Biol. 60:719–731.

Wilkinson M. 1994. Common cladistic information and its consensus representation: Reduced adams and reduced cladistic consensus trees and profiles. Syst. Biol. 43:343–368.

Wilkinson M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst Biol. 44:501–514.

Wilkinson M. 2003. Missing entries and multiple trees: instability, relationships and support in parsimony analysis. J. Vert. Paleo. 23:311–323.

Wilkinson M., Benton M.J. 1996. Sphenodontid phylogeny and the problems of multiple trees. Philos. T. Roy. Soc. B. 351:1–16.

Wilkinson M., Cotton J.A. 2006. Supertree methods for building the tree of life: divide-and-conquer approaches to large phylogenetic problems. In: Hodkinson T.R., Parnell J.A.N., editors. Reconstructing the Tree of Life: Taxonomy and Systematics of Large and Species Rich Taxa. London: CRC Press (Systematics Association Special Volume), p. 61–75.

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

Yu C., He R.L., Yau S.S.T. 2014. Viral genome phylogeny based on Lempel-Ziv complexity and Hausdorff distance. J. Theor. Biol. 348:12–20.

Zanne A.E., Tank D.C., Cornwell W.K., Eastman J.M., Smith S.A., FitzJohn R.G., McGlinn D.J., O'Meara B.C., Moles A.T., Reich P.B., Royer D.L., Soltis D.E., Stevens P.F., Westoby M., Wright I.J., Aarssen L., Bertin R.I., Calaminus A., Govaerts R., Hemmings F., Leishman M.R., Oleksyn J., Soltis P.S., Swenson N.G., Warman L., Beaulieu J.M. 2014. Three keys to the radiation of angiosperms into freezing environments. Nature 506:89–92.

Zwickl D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [Thesis]. School of Biological Sciences, University of Texas at Austin.

Zwickl D.J., Wing R., Stein J., Ware D., Sanderson M.J. 2014. Disentangling methodological and biological sources of gene tree discordance on *Oryza* (Poaceae) chromosome 3. Syst. Biol. 63: 645–659.