Terraces in Phylogenetic Tree Space
Michael J. Sanderson, *et al.*
Science **333**, 448 (2011);
DOI: 10.1126/science.1206357

does not adequately describe the $M^{3/4}$ scaling of whole-organism metabolism for the species in our study because they span different physiological groups with different normalization constants (*4*, *16*) (fig. S1). Hence, the uniform abundance scaling documented here across all species indicates that, at any particular trophic level, populations of similarly sized species in different physiological groups flux different amounts of energy: endotherms > vertebrate ectotherms > parasitic or free-living invertebrates (fig. S1).

The uniform scaling of abundance found here has another general implication—that of "production equivalence." Specifically, species at the same trophic level produce biomass at the same average rate across all body sizes and functional groups. This occurs because, in contrast to metabolic rates, a single line can describe the $M^{3/4}$ scaling of individual biomass production, P_{ind}, for organisms of different physiological groups (*31*) (fig. S1). Consequently, the population production rate equals $P_{pop} = P_{ind}N$, which scales as $M^{3/4}M^{-3/4} = M^{0}$. Indeed, estimating population production for the species in the three estuaries supports the existence of this invariant biomass production with body size (Fig. 4 and fig. S1) (*11*). Thus, although population energy flux (and, consequently, demand on resources) may vary among physiological groups, opposing differences in production efficiency among these groups cause population biomass production to scale invariant of body size across all groups. Because production reflects biomass availability to consumers, production equivalence indicates a comparable ecological relevance for any single species within a trophic level, regardless of body size or functional group affiliation: invertebrate or vertebrate, ectotherm or endotherm, free-living or parasitic.

Accommodating parasitic and free-living species into a common framework highlights the utility of Eq. 3 to incorporate body size, temperature, and food-web information into ecological scaling theory in a simple and generally applicable way. Equations 3 and 4 may allow testing of the generality of the findings documented here for any ecosystem and any form of life.

References and Notes

1. P. W. Price, *Evolutionary Biology of Parasites* (Princeton Univ. Press, Princeton, NJ, 1980).
2. T. de Meeûs, F. Renaud, *Trends Parasitol.* **18**, 247 (2002).
3. A. P. Dobson, K. D. Lafferty, A. M. Kuris, R. F. Hechinger, W. Jetz, *Proc. Natl. Acad. Sci. U.S.A.* **105** (suppl. 1), 11482 (2008).
4. J. H. Brown, J. F. Gillooly, A. P. Allen, V. M. Savage, G. B. West, *Ecology* **85**, 1771 (2004).
5. R. H. Peters, *The Ecological Implications of Body Size* (Cambridge Univ. Press, Cambridge, 1983).
6. P. Arneberg, A. Skorping, A. F. Read, *Am. Nat.* **151**, 497 (1998).
7. S. Morand, R. Poulin, *Evol. Ecol. Res.* **4**, 951 (2002).
8. J. H. Brown, J. F. Gillooly, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1467 (2003).
9. S. Jennings, S. Mackinson, *Ecol. Lett.* **6**, 971 (2003).
10. D. C. Reuman, C. Mulder, D. Raffaelli, J. E. Cohen, *Ecol. Lett.* **11**, 1216 (2008).
11. See supporting material on *Science* Online.
12. R. L. Lindeman, *Ecology* **23**, 399 (1942).
13. D. G. Kozlovsky, *Ecology* **49**, 48 (1968).
14. P. Calow, *Parasitology* **86**, 197 (1983).
15. J. Damuth, *Nature* **290**, 699 (1981).
16. J. F. Gillooly, J. H. Brown, G. B. West, V. M. Savage, E. L. Charnov, *Science* **293**, 2248 (2001).
17. A. P. Allen, J. H. Brown, J. F. Gillooly, *Science* **297**, 1545 (2002).
18. W. R. Robinson, R. H. Peters, J. Zimmermann, *Can. J. Zool.* **61**, 281 (1983).
19. M. Kleiber, *Hilgardia* **6**, 315 (1932).
20. A. M. Hemmingsen, *Repts. Steno. Hosp. Copenhagen* **9**, 7 (1960).
21. H. Cyr, in *Scaling in Biology*, J. H. Brown, G. B. West, Eds. (Oxford Univ. Press, Oxford, 2000), pp. 267–295.
22. J. Damuth, *Biol. J. Linn. Soc. London* **31**, 193 (1987).
23. H. Cyr, J. A. Downing, R. H. Peters, *Oikos* **79**, 333 (1997).
24. J. E. Cohen, T. Jonsson, S. R. Carpenter, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1781 (2003).
25. R. F. Hechinger *et al.*, *Ecology* **92**, 791 (2011).
26. A. M. Kuris *et al.*, *Nature* **454**, 515 (2008).
27. T. D. Meehan, *Ecology* **87**, 1650 (2006).
28. B. J. McGill, *Am. Nat.* **172**, 88 (2008).
29. D. C. Reuman *et al.*, *Adv. Ecol. Res.* **41**, 1 (2009).
30. D. Baird, J. M. Mcglade, R. E. Ulanowicz, *Philos. Trans. R. Soc. B* **333**, 15 (1991).
31. S. K. M. Ernest *et al.*, *Ecol. Lett.* **6**, 990 (2003).
32. S. Nee, A. F. Read, J. J. D. Greenwood, P. H. Harvey, *Nature* **351**, 312 (1991).

Acknowledgments: We thank S. Sokolow, J. McLaughlin, J. Childress, and J. Damuth for discussion or comments on the manuscript. Supported by NSF/NIH EID grant DEB-0224565 and by CA Sea Grant R/OPCENV-01. The analyses in this manuscript used data published in Hechinger *et al.* (*25*), available at Ecological Archives (accession no. E092-066).

Supporting Online Material

www.sciencemag.org/cgi/content/full/333/6041/445/DC1
Materials and Methods
Figs. S1 and S2
Tables S1 to S11
References (*33*–*48*)

15 February 2011; accepted 27 May 2011
10.1126/science.1204337

Terraces in Phylogenetic Tree Space

Michael J. Sanderson,[1]* Michelle M. McMahon,[2] Mike Steel[3]

A key step in assembling the tree of life is the construction of species-rich phylogenies from multilocus—but often incomplete—sequence data sets. We describe previously unknown structure in the landscape of solutions to the tree reconstruction problem, comprising sometimes vast "terraces" of trees with identical quality, arranged on islands of phylogenetically similar trees. Phylogenetic ambiguity within a terrace can be characterized efficiently and then ameliorated by new algorithms for obtaining a terrace's maximum-agreement subtree or by identifying the smallest set of new targets for additional sequencing. Algorithms to find optimal trees or estimate Bayesian posterior tree distributions may need to navigate strategically in the neighborhood of large terraces in tree space.

Phylogenetic tree space, the collection of all possible trees for a set of taxa, grows exponentially with the number of taxa, creating computational challenges for phylogenetic inference (*1*). Nonetheless, phylogenetic trees and comparative analyses based on them are growing larger, with several exceeding 1000 species [e.g., (*2*)] and a recent one exceeding 50,000 (*3*). Understanding the landscape of tree space is important because heuristic algorithms for inferring trees using maximum likelihood (ML), maximum parsimony (MP), and Bayesian inference navigate through parts of this space guided by notions of its structure [e.g., (*4*)]. Moreover, analyses that use phylogenies to study evolutionary processes typically sample from tree space to obtain a good statistical "prior" distribution of phylogenetic relationships used in subsequent comparative analyses, but the design of sampling strategies hinges on the structure of tree space (*5*).

An important advance in understanding tree space was the formulation of the concept of "islands" of trees with similar MP or ML optimality scores (*6*, *7*). Trees belong to the same island if they are near each other in tree space and have optimality scores of L or better with respect to some data matrix. Distance in tree space can be measured by the number of rearrangements required to convert one tree to another. Nearest neighbor interchanges (NNIs), for example, are rearrangements obtained by swapping two subtrees around an internal branch of a tree. Conflicting signals or missing data can result in multiple large tree islands, separated by "seas" of lower-scoring trees, a landscape that can only be characterized by lengthy searches through tree space [e.g., (*8*)]. Empirical studies of phylogenetic tree islands flourished in the context of the single-locus data sets that were common in the 1990s. However, maintaining the same level of accuracy in the larger trees studied today requires combining multiple loci (*9*). The most widely used protocol for data combination is concatenation of multiple alignments of orthologous sequences, one next to another, analyzed as one "supermatrix," a procedure justified when gene tree discordance is low between loci (*10*). Notably, a hallmark of almost all large supermatrix studies is a sizable proportion of missing entries.

[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]Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, Christchurch, New Zealand.

*To whom correspondence should be addressed. E-mail: sanderm@email.arizona.edu

Consider a recent analysis (*11*) of deep arthropod phylogeny, which combined 129 alignments of separate loci obtained largely from expressed sequence tag libraries into a single supermatrix for 117 taxa. We represent such a collection of k multiple sequence alignments, D_i, which are concatenated, as a supermatrix, D, of k loci by n taxa. Loci for which fewer than n taxa have been sampled contain missing data (35% in the arthropod study). Let Y_i be the set of taxon labels that have been sampled for locus i, with the entire label set $X = \bigcup_{i=1}^{k} Y_i$, and $n = |X|$. A taxon coverage pattern, $S = \{Y_1,...,Y_k\}$, is a collection of subsets of X. Consider any binary tree T on X. Tree T displays a binary phylogenetic tree, T', if $T|Y = T'$, where the vertical bar means the subtree induced by restricting T to just the taxa in Y. If T displays the k subtrees, $T|Y_1,\ldots,T|Y_k$, then it is a parent tree of these subtrees. If T is the only such tree, the subtrees define T, and S is decisive for T (*12*). Let $\mathcal{L}(D, T)$ be a scoring function such as log likelihood, giving the score, ℓ_0, of tree T based on a sequence alignment D, and (implicitly) a model of evolution. Then

$$\mathcal{L}(D,T) = \sum_{i=1}^{k} \mathcal{L}(D_i,(T|Y_i)) \qquad (1)$$

This holds for MP because all sites are scored separately but also holds for partitioned models in ML [(*13*); e.g., RAxML (*14*); supporting online text] and Bayesian inference [e.g., MrBayes (*15*)]. It follows that any other tree that also displays $T|Y_1, ..., T|Y_k$ has the same score, ℓ_0. This leads to a fundamental observation:

The set of all parent trees of $T|Y_1, ..., T|Y_k$ has the same $\mathcal{L}$-score as tree T, namely, ℓ_0. We call this set a terrace.

All trees on a terrace are distinct from each other, but they are indistinguishable in two important respects: They display the same set of subtrees, and they have the same optimality score. Key properties of terraces can be understood with the theory of phylogenetic supertrees (trees constructed from collections of smaller trees). In the following we assume that each of the k induced subtrees can be rooted [for example, if there is at least one taxon, a reference taxon, sampled for all k loci (*10*)]. First, a terrace is part of a tree island. This follows from (*16*), which shows that trees in a terrace are all connected by NNI tree rearrangements in the same way that trees in an island are. Because they all have the same score, they must form at least a subset of some tree island whose threshold score, L, is worse than theirs.

Second, the trees in a terrace can be enumerated with an algorithm that generates all parent trees of a set of compatible subtrees (*17*). The latter are induced by any tree, T, from the terrace, together with the taxon coverage pattern, S. A search through tree space checking optimality scores is unnecessary, because the trees can be built directly with S and T. This is useful because the number of trees on a terrace can scale exponentially with the number of taxa in the displayed subtrees (*18*). Third, testing if two trees are on the same terrace can be done quickly because it merely requires tests of tree equality of the induced subtrees (*10*, *19*). Finally, the trees in a terrace can be summarized by a special consensus tree used in the supertree literature [the BUILD tree (*20*)] with three convenient properties: (i) It displays all the individual loci's induced subtrees; (ii) it is the Adams consensus tree of all trees on the terrace (*21*); and (iii) it can be constructed in polynomial time (*19*). Figure 1 illustrates these ideas with a small example.

We examined three recently published large supermatrix studies (*11*, *22*, *23*) (Table 1) that have typical levels of partial taxon coverage (52 to 66%), but differ with respect to fractional decisiveness, an index tied to the impact of missing data on tree construction (*10*, *12*). In an analysis of arthropods (*11*) with 129 loci and a very high fractional decisiveness (table S2), the 14 terraces found had just a single tree on each. However, in

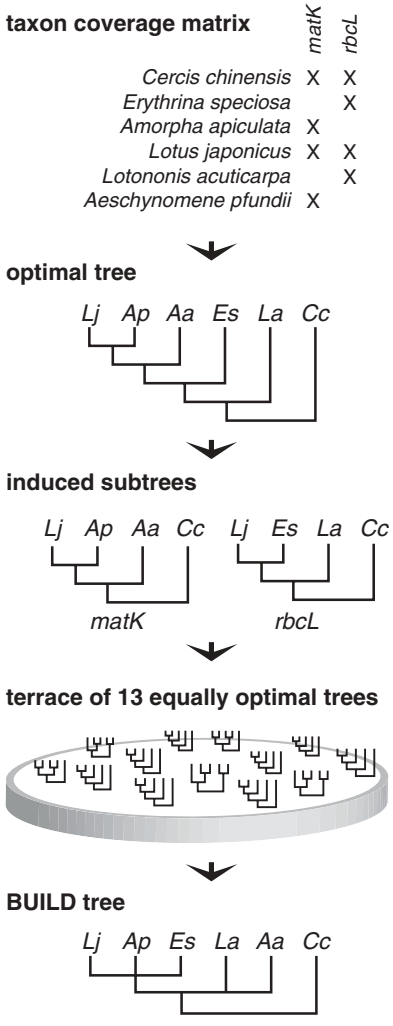

Fig. 1. Terrace in tree space for six species of the angiosperm clade Leguminosae and two loci, *matK* and *rbcL* (*10*). Taxon coverage is denoted by an "X" when sequence data are present. The optimal tree, an ML tree found using a partitioned model in RAxML (ln L = −6709.8), induces two locus-specific subtrees. Twelve additional trees for these six taxa also display these subtrees, together comprising a terrace of 13 equally optimal trees (labels and outgroup removed from trees on terrace). The BUILD tree (*20*) is a consensus of all trees on the terrace.

Table 1. Characteristics of data sets and their terraces.

Taxon/study	Arthropods (*11*)	Grasses (*22*)	Colubrid snakes (*23*)
Number of taxa	117	298	767
Number of loci	129	3	5
Number of sites	37,476	5074	5814
Coverage density	0.65	0.66	0.52
Terraces			
	ML optimal tree		
Terrace size	1	61.2 million	2205
	ML suboptimal trees		
Number found in 50 replicate searches	13	49	49
Smallest terrace size	1	893,025	315
Largest terrace size	1	>1 billion	33,075
	MP optimal trees		
Number found	1	8	8
Smallest terrace size	1	11,907	6615
Largest terrace size	1	4.1 million	6615

analyses with more taxa, fewer loci, lower decisiveness, but about the same fraction of missing data, terraces were much larger, ranging from hundreds to billions of trees in likelihood and parsimony searches (Table 1). Irrespective of terrace size, we could efficiently make the BUILD tree for each terrace without heuristic searches through tree space (e.g., running times of just seconds for terraces with ~100 million trees).

Exploring the position of terraces in tree islands is challenging because it involves searching tree space. However, a sense of the structure of an island in the immediate neighborhood of its peak can be obtained relatively easily by examining trees one rearrangement away, calculating their likelihood scores, and determining the size and number of terraces present. For the grass data (*22*), the ML tree is on a terrace of 61 million trees, and the tree itself is connected to 590 trees one NNI rearrangement away. Of these, 198 trees have a likelihood score within 5.0 log likelihood units of the ML tree, which we use as a cutoff for defining an island (*10*), and these comprise 168 distinct terraces the sizes of which range from 8.75 million to 428 million trees, or 1.1×10^{10} trees in all (Fig. 2). The island's structure is complicated by a broad plateau below the ML tree consisting of both large and small terraces with nearly equal likelihood scores.

The multiplicity of equally good trees in terraced landscapes poses obstacles to downstream comparative studies in ecology and evolutionary biology. However, a useful reduction in ambiguity can be obtained via a terrace's maximum-agreement subtree (MAST), which is a precise phylogenetic hypothesis on a smaller set of taxa. Although the MAST can be found in polynomial time when the input trees are binary (*24*), this may be infeasible in the present setting where there can be an exponentially large number of trees on a terrace.

However, a more appropriate variant of this problem can be solved efficiently (*10*), irrespective of the size of the terrace. Given a set of compatible rooted binary input trees, $T_1, \ldots, T_k$ with label sets $Y_1, \ldots, Y_k$; $X \equiv Y_1 \cup \ldots \cup Y_k$, the Maximum Defining Label Set problem seeks the largest label set $X^* \subseteq X$, such that $T_1|X^*,\ldots,T_k|X^*$ together define a parent tree T^* on X^*. For two loci (subtrees), this problem can be solved exactly in polynomial time (*10*). This could not be directly used for our data sets, the smallest of which (*22*) had $k = 3$ loci, so we used a heuristic strategy, solving the problem for all (three) pairs of loci (*10*). Removal of just 12 of 298 taxa reduced the terrace size of the ML tree from 61 million trees to one. Moreover, using a variant of this algorithm, we infer that completely sequencing all three loci for these 12 taxa could reduce the terrace size to one tree for the original larger set of taxa (*10*), a considerable savings over sequencing the entire 34% of the supermatrix that is empty.

The discovery of terraces has implications for search strategies for building large phylogenetic trees on the basis of ML, MP, and Bayesian methods that move through tree space. Each of these approaches spends substantial computational time evaluating scores on trees that are rearrangements of existing trees. Yet all trees within a terrace must have the same score, so it makes sense to direct tree search outside of known terraces. In Bayesian analysis, a better estimate of the posterior distribution might be obtained by quickly enumerating a sample of trees on a terrace once the first tree is visited. The extraordinarily large size of some terraces, however, makes exhaustive exploration of the islands in which they are found problematic because searching between terraces via tree rearrangements is still necessary. Progress may require engineering a compact data structure for the trees in a terrace to allow computing on what may be vast collections of reasonable trees in tree space. Otherwise, the boundaries of islands in complex data sets will likely remain shrouded.

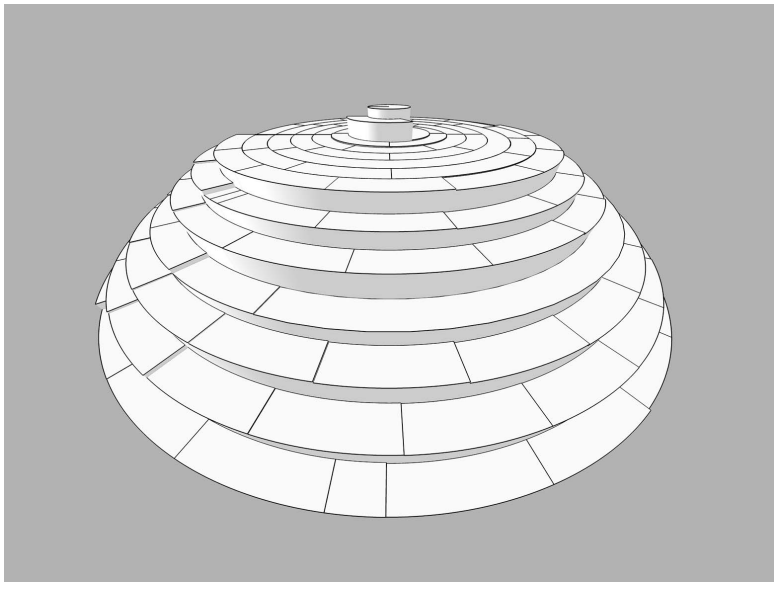

Fig. 2. Visualization of terraces in tree space near the ML tree for the grass data set (*22*). Areas of terraces are proportional to number of trees and height to likelihood score. Total number of trees on all terraces illustrated exceeds 10 billion.

References and Notes

1. J. Felsenstein, *Inferring Phylogenies* (Sinauer, Sunderland, MA, 2004).
2. O. R. P. Bininda-Emonds *et al.*, *Nature* **446**, 507 (2007).
3. S. A. Smith, J. M. Beaulieu, A. Stamatakis, M. J. Donoghue, *Am. J. Bot.* **98**, 404 (2011).
4. S. Whelan, D. Money, *Mol. Biol. Evol.* **27**, 2674 (2010).
5. A. Vanderpoorten, B. Goffinet, *Syst. Biol.* **55**, 957 (2006).
6. D. R. Maddison, *Syst. Zool.* **40**, 315 (1991).
7. L. A. Salter, *Syst. Biol.* **50**, 970 (2001).
8. L. A. McDade, T. F. Daniel, C. A. Kiel, *Am. J. Bot.* **95**, 1136 (2008).
9. E. Mossel, M. Steel, in *Mathematics of Evolution and Phylogeny* (Oxford Univ. Press, New York, 2005), pp. 384–412.
10. See supporting material on *Science* Online.
11. K. Meusemann *et al.*, *Mol. Biol. Evol.* **27**, 2451 (2010).
12. M. J. Sanderson, M. M. McMahon, M. Steel, *BMC Evol. Biol.* **10**, 155 (2010).
13. A. Stamatakis, M. Ott, *Philos. Trans. R. Soc. Lond B Biol. Sci.* **363**, 3977 (2008).
14. A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).
15. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* **19**, 1572 (2003).
16. M. Bordewich, thesis, University of Oxford (2003).
17. M. Constantinescu, D. Sankoff, *J. Classif.* **12**, 101 (1995).
18. C. Semple, *Discrete Appl. Math.* **127**, 489 (2003).
19. W. H. E. Day, *J. Classification* **2**, 7 (1985).
20. A. V. Aho, Y. Sagiv, T. G. Szymanski, J. D. Ullman, *SIAM J. Comput.* **10**, 405 (1981).
21. D. Bryant, in *BioConsensus*, M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, F. S. Roberts, Eds. (DIMACS ser. vol. 61, American Mathematical Society, Providence, RI, 2003), pp. 163–184.
22. Y. Bouchenak-Khelladi *et al.*, *Mol. Phylogenet. Evol.* **47**, 488 (2008).
23. R. A. Pyron *et al.*, *Mol. Phylogenet. Evol.* **58**, 329 (2011).
24. A. Amir, D. Keselman, *SIAM J. Comput.* **26**, 1656 (1997).

Acknowledgments: Thanks to C. Ané, M. Bordewich, D. Bryant, D. Fernández-Baca, M. Nachman, B. O'Meara, and C. Semple for helpful comments. This work was supported by NSF award 0829674 (to M.J.S. and M.M.M.). All data used in this paper were obtained from GenBank or TreeBASE and are detailed in the Supporting Online Material.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1206357/DC1
Materials and Methods
SOM Text
Tables S1 and S2
Fig. S1
References (*25*–*31*)

31 March 2011; accepted 7 June 2011
Published online 16 June 2011;
10.1126/science.1206357

www.sciencemag.org/cgi/content/full/science.1206357/DC1

Supporting Online Material for

Terraces in Phylogenetic Tree Space

Michael J. Sanderson,* Michelle M. McMahon, Mike Steel

*To whom correspondence should be addressed. E-mail: sanderm@email.arizona.edu

Published 16 June 2011 on *Science* Express
DOI: 10.1126/science.1206357

This PDF file includes:

Supporting Online Material

Materials and Methods

Sequence data. Data for the small example of Fig. 1 was obtained from GenBank (Table S1). Data matrices for the three published data sets were obtained from TreeBASE (*25*). Two of these included Nexus `charset` statements, which allowed determination of the identity of each locus. For the grass data matrix (*22*), locus boundaries were identified by inspection of the alignment to see where runs of question marks and/or dashes started or ended.

Phylogeny reconstruction. Phylogenetic trees for the three published data sets were built using MP with the program TNT (*26*) (with options xmult level=7 mxram 50). RAxML (*14*) was used for ML in all data sets (fully partitioned model; RAxML v. 7.2.7 HPC-SSE recompiled to allow 129 loci for the arthropod data set). We calculated likelihoods in RAxML under the model GTRGAMMA using a single tree per file, because in multi-tree files the program treats the first tree differently, optimizing some parameters only on it. Models were completely partitioned between loci (-M option), including separate branch lengths.

Nodes were not collapsed in MP searches; thus all trees obtained were binary. No non-binary trees were found in ML searches. All trees were rooted with a reference taxon selected arbitrarily from the available choices (for ref. *11*: *Tribolium*; *22*: *Zoysia*; *23*: *Xenopeltis*). A reference taxon is a taxon that has been sampled for all loci, and therefore can act formally as a root for the purposes of some of the rooted supertree algorithms described below. Trees can be rerooted after these analyses as desired.

Fractional decisiveness. We report fractional decisiveness values for each of the three published data sets (Table S2). A taxon coverage pattern, *S*, is decisive for a tree, *T*, if the induced subtrees define *T*. Sometimes *S* is decisive for all trees, but more commonly, it is only decisive for some or no trees. We report three aspects of fractional decisiveness. One is the fraction of all triples of taxa for which sequence data are present in a data set. A value of 100% is a necessary condition for decisiveness; it is also a sufficient condition when a reference taxon is present, as here (*12*). Another measure, $\angle_D$, is the proportion of all binary trees on the label set *X* for which *S* is decisive (*12*). A coverage pattern with high $\angle_D$ is more likely to define some specified tree found in a search. This is estimated by sampling from a uniform distribution of binary trees on *X* and evaluating decisiveness. Finally, $\angle_d$, is a measure of the average fraction of edges in trees on *X* that are distinguished by the taxon coverage pattern (*12*).

Terraces. First we provide a proof of the basic observation about terraces made in the main text.

Observation (definition of terrace): *The set of all parent trees of* $T \mid Y_1, \ldots, T \mid Y_k$ *has the same* $\mathcal{L}$*-score as tree T, namely,* ℓ_0 *. This set is a* **terrace**.

Proof: From Eq. (1),

$$\mathcal{L}(D,T) = \sum_{i=1}^{k} \mathcal{L}(D_i,(T \mid Y_i)) = \ell_0 \qquad (2)$$

and if T' is any other parent tree of $T \mid Y_1,...,T \mid Y_k$, then $T' \mid Y_i = T \mid Y_i$ (since T is binary) and so

$$\mathcal{L}(D,T') = \sum_{i=1}^{k} \mathcal{L}(D_i,(T' \mid Y_i)) = \sum_{i=1}^{k} \mathcal{L}(D_i,(T \mid Y_i))$$

which, by Eq. (2), implies that

$$\mathcal{L}(D,T') = \ell_0 .$$

Rooting and reference taxa. The basic observation and definition of terraces holds for rooted or unrooted trees. However, some algorithms used in the text assume rooted trees, including those used to enumerate all trees on a terrace (*17*) and to show that terraces are subsets of tree islands (*16*). To ensure rooted trees are available we make the sufficient (but not necessary) assumption that there exists in the data a reference taxon for which each locus has been sequenced. Such a reference taxon can serve as de facto root for algorithmic purposes whether it is the true phylogenetic root or not.

Analysis of terraces in data sets. For each data set, we determined *S*, the pattern of taxon coverage, using the boundaries of each of the *k* loci and scoring a locus as present (+) for a taxon if it contained any sequence data other than gaps or question marks. For any given binary tree, *T*, obtained in an MP or ML search, we generated the *k* induced subtrees with a PERL script, `displaysub.pl`. From these another PERL script, `maketriplets.pl`, generated a set of rooted triplets that define each of these trees. Neither of these involve significant computational or algorithmic issues. Finally, we implemented the algorithm described in Constantinescu and Sankoff (*17*) in a PERL script, `countParents.pl`, which counts the number of binary parent trees compatible with these rooted triplets. If this program exceeds a user-supplied upper bound (1 billion trees in the present analysis), it terminates and reports the upper bound has been reached.

Another script, `build.pl`, constructs the supertree displaying all the induced subtrees, using the BUILD algorithm (*20*; following *17*). This tree is also known to be the Adams consensus tree of all the parent trees of the induced subtrees (Theorem 2.10 of *21*). Thus, it as an extremely informative summary of the phylogenetic dimensions of a terrace, and it is obtained without the computationally infeasible step of enumerating and storing all parent trees.

A useful check on the calculation of terrace sizes and their consensus can be obtained for small to moderate input trees by exploiting another known result from the theory of supertrees. For a compatible set of input trees, as our induced subtrees are by definition, the set of all supertrees compatible with this set can be obtained by the MRP (matrix representation with parsimony) supertree method -- that is, by finding all the most parsimonious trees for the MRP matrix (*27*). The only problem in practice is that we must use an exact MP search algorithm, such as branch and bound, to guarantee finding all the optimal trees, which can be slow even though there is no conflict in this MRP matrix (because the inputs are compatible). Nonetheless, we found this to be a workable check for many of the trees examined in the colubrid data set, for which terrace sizes were often on the order of a few thousand trees. All checks indicated agreement between our direct implementations and this MRP workaround.

To determine how many different terraces are reflected in a sample of trees, we need to be able to check if two binary trees are on the same terrace. For this we again construct the *k* induced subtrees for each of the two trees as described above, and check whether the *i*th induced subtree for each is the same, for all *k* loci. This requires an efficient way to compare two phylogenetic trees. We exploited the data structure invented by Day (*19*) to build strict consensus trees, which has a running time of *O(n)*. Since there are at most *k* loci to check, the running time of this test for membership in the same terrace is *O(kn)*. We wrote a PERL module, `treesequal.pm`, for the equality test, called from a master script, `sameTerrace.pl`. For a set of input binary trees, the master script assigns the first tree to a new terrace, then checks if the second tree is on the same terrace as the first tree. Each new tree only needs to be checked against the unique terraces already found (rather than against all trees in the input), and then it is added to the list of terraces if it is new.

Visualization of a neighborhood in a large tree island. We examined the local neighborhood of the ML estimate for the grass data set (*22*; Fig. 2). Fifty replicate partitioned searches using the RAxML GTRCAT model were run, and the best tree under GTRGAMMA kept. PAUP* 4.0 was used to find the 590 nearest neighbor trees of this ML tree and their likelihoods were calculated using RAxML as described above. We noticed that one tree in this neighborhood had a slightly higher likelihood, and we therefore did another round of NNI exploration around that tree, using this final ML tree and its nearest neighbors.

To circumscribe the part of a tree island close to the ML tree, we considered all nearest neighbor trees within 5.0 log likelihood units of the maximum likelihood tree found in that second round. The choice of 5.0 is somewhat arbitrary (*6, 7, 28*). Let $L = L^* - \Delta$ be the minimum optimality score value required in the definition of a tree island, where L^* is the optimal score for the best tree found on the island. Small values of Δ induce smaller islands, at the cost of obscuring connectedness between trees with similar scores. On the other hand, larger values of Δ lead to inclusion of very large numbers of trees and increase the probability of multiple local optima per island (*28*). The number of neighbors connected to the optimal tree by NNI rearrangements grows exponentially with the NNI distance between the optimal tree and the included trees. Thus, Δ is a tuning

parameter providing different views on the landscape structure in tree space. However, numerical imprecision in likelihood calculations in large data sets also puts a lower limit on Δ (+1.0 log likelihood units is recommended in large data sets in the user manual of RAxML, *14*). We selected Δ = 5.0 log likelihood units for the grass data set, which was large enough to avoid numerical problems and include a large fraction of all the nearest neighbors of the optimal tree, but small enough that it discriminated against the even larger fraction of nearest neighbors, some with much lower likelihoods. Thus it provided a fine scale picture of the very close neighborhood of the peak and its terraces.

There were 198 nearest neighbor trees within Δ = 5.0 log likelihood units of the ML estimate, but there were only 168 distinct terraces, because some trees with adequate likelihood scores were found on the same terrace. The tree space around this ML estimate was visualized using the metaphor of agricultural terraces on a mountain side. Using the OpenGL graphics library we wrote software to draw horizontal partial ring shaped terraces, the area of which is proportional to the number of trees on a terrace, and the height of which is proportional to the likelihood score. The C program, `tree_space`, allows changes of viewpoint and zooming.

Data combination via concatenation. Our Eq. (1) and observations based on it assume phylogenetic inference is based on a concatenated supermatrix alignment, a very widely used protocol for the analysis of large phylogenetic data sets. Concatenation is theoretically proper when used for collections of loci on nonrecombining organellar genomes, such as metazoan mitochondrial (~16 kb in size) and plant plastid data sets (typically ~160 kb). It should also be appropriate in nuclear genomic regions in which recombination is low, such as mammalian haplotype blocks (often 10-100 kb regions between recombination hotspots). Concatenation has also been very widely used for putatively unlinked single copy orthologous nuclear markers, such as those derived from EST libraries (e.g., the arthropod data set), but in this setting it is well known that incomplete lineage sorting can induce considerable gene tree discordance that is "real"--that is the gene trees are correct for the genes even if incorrect for the species. However, discordance is expected to be low when the ratio of branch lengths to population size is large (*29*), as is often assumed to be the case in deep phylogenies, but phylogenomic evidence indicates it is high in recent radiations of closely related species. New methods of phylogenetic inference that capitalize on this discordance, by minimizing deep coalescence events or explicitly modeling the multilocus coalescent, appear to work best in the latter context but to be considerably less necessary in the former, where concatenation performs nearly as well (*29*).

Maximum Defining Label Set (MDLS) problem. For $k = 2$, the exact solution can be obtained in polynomial time by a straightforward application of Gordon's (*30*) strict consensus supertree method. If the strict consensus supertree from his algorithm is a binary tree, the input trees must define that tree. Assume the input trees are the induced subtrees from locus 1 and 2, T_1 and T_2, and the subtree on the overlapping taxa is t. For each edge, e, of t, there is a set R_1 of taxa unique to T_1 that attach to e, and R_2 of taxa unique to T_2 that attach to e. Let $r_1 = |R_1|$ and $r_2 = |R_2|$. In Gordon's algorithm, we do the following for each edge, e:

1. if $r_1 = r_2 = 0$, then we do nothing for edge e,
2. if $r_1 > 0$, and $r_2 = 0$, then we add R_1 to edge e with the same topology as is found on T_1,
3. if $r_1 = 0$, and $r_2 > 0$, then we add R_2 to edge e with the same topology as is found on T_2,
4. if $r_1 > 0$, and $r_2 > 0$, then we insert a node, z, in e that has a degree greater than 3, collapsing all the relationships in both R_1 and R_2 into a polytomy at z--implying the final output tree will not be binary

Figure S1A illustrates this. Case 4 is the only case that causes problems, and the obvious solution is to keep the larger set, R_1 or R_2, for edge e, and discard the other smaller set. With this in mind, our algorithm visits each edge of t and selects the set from T_1 or T_2 with larger size (Fig. S1B). This must produce the largest label set possible, because it gives the largest set at each branch of t. This algorithm is implemented in a PERL script, `mdls.pl`.

Properties of the MDLS solution. In our setting the subtrees forming the input of the MDLS problem are obtained by restricting some optimal tree obtained during tree search to the taxon coverage pattern for those loci. However, the choice of this tree does not matter: any tree on the same terrace leads to this solution:

Lemma: Suppose T_1 , ..., T_k are compatible rooted binary trees with label sets Y_1 ,..., Y_k; $X \alpha Y_1 * ... * Y_k$. Suppose $X^* \square X$ is a set for which $\{T_1 | X^*,..., T_k | X^*\}$ defines a tree t^*. Then for *any* tree T (binary or not) that displays $\{T_1, ..., T_k\}$ we must have $T | X^* = t^*$. In particular for any two trees T and T' that display $\{T_1, ..., T_k\}$ we have $T | X^* = T' | X^*$.

Proof. $T | X^*$ displays $T_1 | X^*,..., T_k | X^*$, since $(T | X^*) | Y_i = (T | Y_i)| X^* = T_i | X^*$, and since t^* is the only tree with label set X^* that displays $T_1 | X^*,..., T_k | X^*$ we must have $t^*=T | X^*$.

In addition, because the BUILD tree constructed from the induced subtrees, $T | Y_1,..., T | Y_k$ (and used to summarize the terrace) displays this set of subtrees, the MDLS solution, t, is the same as the BUILD tree restricted to X^*. Note, however, although the BUILD tree restricted to X^*, is binary, the BUILD tree itself may not be binary.

Heuristic solution for $k > 2$. No exact polynomial time solution for the MDLS problem is apparent for $k > 2$, and indeed its complexity is an open question. We can construct relatively brute force but slow solutions, but for our purposes we settle for a naive greedy heuristic. For the grass data set, we solved the exact MDLS problem for each of the three pairs of loci, took the pair that retained the largest taxon set (loci 1-2, discarding only 3 taxa), built its MDLS tree, and then used it with the subtree for locus 3 (discarding 9 additional taxa). We checked the terrace size of this tree using both the direct computation and the MRP solution and indeed only one tree was found. However, as this is heuristic, another solution might exist which discards fewer than 12 taxa.

Optimal new sequencing. The list of taxa discarded by the MDLS solution is an obvious candidate for additional sequencing. We address this here for $k = 2$ only. Visit each edge, *e*, of *t*, in which there are two sets of unique taxa attached, R_1 and R_2 (one set per locus) such that both $r_1 > 0$ and $r_2 > 0$, which is the only case in which taxa are discarded. Assume the discarded smaller set is R_1 without loss of generality. We sequence R_1 for locus 2 (Fig. S1C). Now we have a larger set of taxa, $R_2' \supset R_2$, in the induced subtree for locus 2. If the induced subtree for R_2 is not changed by the new data (admittedly a strong assumption), that is if the induced subtree on R_2 is a subtree of the induced subtree on R_2', then the only difference will be that the overlap tree, *t*, has more nodes in it, and those nodes will be within the original edge *e* or possibly attached to new edges that are attached to nodes within *e*. The taxa in R_1 will all be attached to new nodes of *t*, and hence will be removed from possible conflicts with taxa from R_2. Some of the taxa in R_2 will possibly be distributed between these new nodes, but there will no longer be any unique taxa from locus 1 in this vicinity since all have been accounted for in *t*, so case 3 above will apply. Since the MDLS problem discards the fewest unique taxa necessary to define a tree, restoring those taxa represents the optimal (fewest) number of taxa to sequence to define a tree.

Modification of MDLS for targeted species lists. Users of phylogenetic trees often have a set of phenotypic or other data for a list of species that does not necessarily match exactly the taxa in a supermatrix. Solution to the following slight variant of the MDLS problem will return a label set that defines a tree and contains the maximum number of species from an input list.

Problem: **Maximum defining targeted label set**

Given: Compatible rooted binary input trees, T_1 , ..., T_k with label sets Y_1,..., Y_k; $X \alpha Y_1 * ... * Y_k$; a set of taxa $Z \subseteq X$, comprising a list of targeted taxa to retain in the final tree.

Find: The largest label set $Z^* \subseteq Z$, such that $T_1 | Z^*$,..., $T_k | Z^*$ together define a parent tree T^* on Z^*.

This can be solved for $k = 2$ simply by modifying the algorithm described above to keep the contributions from whichever tree has the most targeted taxa rather than the most taxa overall. Another slight variant would maximize non-targeted taxa taxa once targeted taxa have been maximized to increase the overall size of the label set. This might be left to the discretion of the user to decide if non-target taxa add or detract from downstream comparative analyses.

Relationship to SMAST problem. A related problem has been discussed in the literature (*31*), the Supertree Maximum Agreement Subtree problem. This problem is designed to build a reasonable supertree from input trees that may conflict with respect to relationships among their shared taxa. The output of this problem is *any* parent tree that displays the input trees on X^*. It aims to eliminate the shared taxa that cause conflicts, not

the taxa unique to individual trees for which there is no "cross-talk" between input trees. Berry and Nicolas (*31*: sect. 3.2) show that any SMAST contains all the unique (their "specific") taxa. This follows because there is no information to contradict the specified relationships on the individual trees. In our case of compatible input trees induced by the taxon coverage pattern, SMAST would simply return any one tree on a terrace, complete with all its taxa.

Software availability. All software described above is free and open source and available for download via SourceForge at http://sourceforge.net/projects/phyloterraces/.

Supporting Text

Incompletely partitioned models. The terraces defined are inherent in MP analysis, but in ML and Bayesian analysis they emerge in the context of completely partitioned models. Although completely partitioned models seem increasingly warranted in response to widespread heterotachy in real data, it may not be required by any given data set. If the models for different loci share parameters in common, then the likelihood contributions of each locus will interact among loci, and Eq. 1 does not hold (exactly). This is the case, for example, if two loci have separate substitution rate models including rate variation across site parameters, but share a common set of branch lengths, or a common set of lengths differing by no more than a single scaling parameter between loci. Consider the sum of estimated branch lengths between two taxa that are both sequenced in the supermatrix. Each locus may have data missing for critical taxa that bear on some of the internal branches along that path, but since there is a common set of parameters bearing on the summed path length (and its branches), the likelihoods of these parameters will not be simply decomposable into terms for each locus.

We suspect that certain incompletely partitioned models will be associated with structures that are quite like terraces in the sense we have defined, but this requires further investigation. In the meantime, as users' preferences and the demands of data increasingly favor fully partitioned models, terraces are likely to be a fixture of tree space.

Data sets with little phylogenetic signal at decisive loci. The requirements for the existence of terraces are sufficient but not necessary in real data sets, because other factors can induce flat regions in the likelihood surface. To take an extreme example, consider a data set of 10 loci in which the first locus is completely sampled, but the other nine all have partial taxon coverage. Because even a single completely sampled locus causes a data set to be decisive (*12*), there would be no terraces with multiple trees for any tree found in a search. However, suppose that the partial taxon coverage in the nine loci would induce large terraces if the first locus were absent. Further, suppose that the first locus has very little variation, so that it contributes likelihood scores that are all very similar for different reasonable candidate trees in the island containing the ML tree. In that case, tree space might well look much like it did for the nine loci alone, perhaps modulated to be slightly non-planar by the first locus.

Supporting Tables

Table S1. GI numbers of GenBank accessions used in Fig. 1.

Species name	*matK*	*rbcL*
Cercis chinensis	183528996	148590346
Lotus japonicus	13518417	13518420
Amorpha apiculata	38373082	-
Aeschynomene pfundii	6466270	-
Erythrina speciosa	-	18157275
Lotononis acuticarpa	-	182411699

Table S2. Fractional decisiveness values. See SOM text for explanation.

	Arthropods (*11*)	Grasses (*22*)	Colubrids (*23*)
% triples	99.998	66.358	91.221
$\angle_D$	1.00	0.00	0.00
$\angle_d$	1.00	0.88	0.98

Supporting Figures

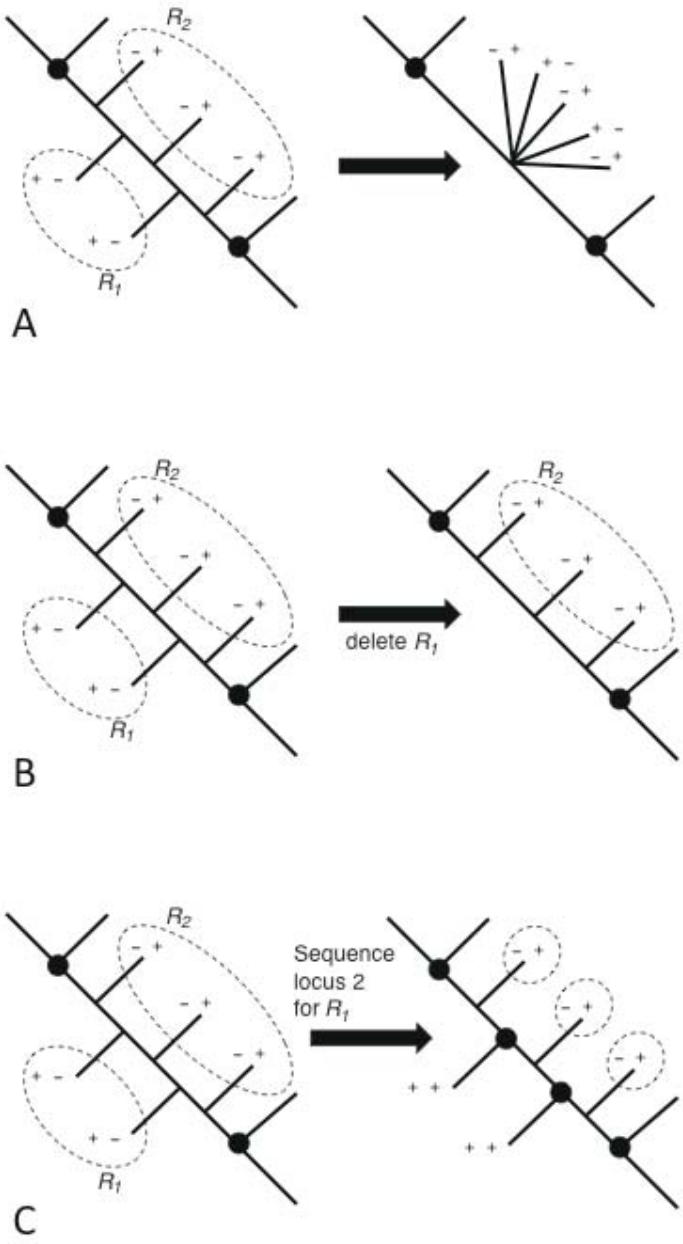

Fig. S1

Fig. S1. The MDLS and optimal new sequencing problems. Tree is the overlap tree, *t*, between two input trees representing the induced subtrees for two loci. Two nodes from *t* are shown as black filled circles. Taxa unique to tree 1 or 2 are labeled as "+-" or "-+" respectively to refer back to the taxon coverage pattern in the original supermatrix. In

other words, "+–", refers to a taxon sampled for locus 1 but not 2, and therefore present in induced subtree 1 but not subtree 2. Sets circled with dashed lines refer to the sets of unique taxa for each tree. A. The consequence of combining these two trees using Gordon's supertree algorithm is to induce a polytomy along the indicated edge, which implies that these two input trees do not define a tree. In the present context this means a terrace of size greater than 1 would be found. B. However, if the smaller set of taxa is simply deleted prior to combining the trees, the supertree is binary, defines a tree (and therefore would induce a terrace with only one tree), but is obviously missing r_1 taxa. C. This set of taxa can be rescued by simply sequencing the second locus. With the caveat that addition of the sequence does not change the topology of the induced subtree (a significant assumption), this would produce several new nodes in the overlap tree, and for each new edge, the unique taxa from tree 2 could also be added, because now they are the only taxa added to those edges (no unique taxa from tree 1 are added, because they have been accounted for in the overlap tree's new edges). Since these newly sequenced taxa are from the smaller set of R_1 or R_2, this leads to an obvious algorithm for sequencing the fewest additional taxa.

References and Notes

1. J. Felsenstein, *Inferring Phylogenies* (Sinauer Press, Sunderland, MA, 2004).
2. O. R. P. Bininda-Emonds *et al.*, *Nature* **446**, 507 (2007).
3. S. A. Smith, J. M. Beaulieu, A. Stamatakis, M. J. Donoghue, *Am. J. Bot.* **98**, 404 (2011).
4. S. Whelan, D. Money, *Mol. Biol. Evol.* **27**, 2674 (2010).
5. A. Vanderpoorten, B. Goffinet, *Syst. Biol.* **55**, 957 (2006).
6. D. R. Maddison, *Syst. Zool.* **40**, 315 (1991).
7. L. A. Salter, *Syst. Biol.* **50**, 970 (2001).
8. L. A. McDade, T. F. Daniel, C. A. Kiel, *Am. J. Bot.* **95**, 1136 (2008).
9. E. Mossel, M. Steel, in *Mathematics of Evolution and Phylogeny* (Oxford University Press, New York, 2005), pp. 384-412.
10 See Supporting Material on *Science* Online.
11. K. Meusemann *et al.*, *Mol. Biol. Evol.* **27**, 2451 (2010).
12. M. J. Sanderson, M. M. McMahon, M. Steel, *BMC Evol. Biol.* **10**, 155 (2010).
13. A. Stamatakis, M. Ott, *Phil. Trans. Roy. Soc. B* **363**, 3977 (2008).
14. A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).
15. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* **19**, 1572 (2003).
16. M. Bordewich, thesis. University of Oxford (2003).
17. M. Constantinescu, D. Sankoff, *J. Classif.* **12**, 101 (1995).
18. C. Semple, *Discrete Appl. Math.* **127**, 489 (2003).
19. W. H. E. Day, *J. Classification* **2**, 7 (1985).
20. A. V. Aho, Y. Sagiv, T. G. Szymanski, J. D. Ullman, *SIAM J. Computing* **10**, 405 (1981).
21. D. Bryant, in *BioConsensus*, M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, F. S. Roberts, Eds. (DIMACS. AMS., 2003), pp. 163-184.
22. Y. Bouchenak-Khelladi *et al.*, *Mol. Phylog. Evol.* **47**, 488 (2008).
23. R. A. Pyron *et al.*, *Mol. Phylog. Evol.* **58**, 329 (2011).
24. A. Amir, D. Keselman, *SIAM J. Computing* **26**, 1656 (1997).
25. TreeBASE (http://www.treebase.org)
26. P. A. Goloboff, J. S. Farris, K. C. Nixon, *Cladistics* **24**, 774 (2008).

27. C. Semple, M. Steel, *Phylogenetics* (Oxford University Press, New York, 2003).

28. D. A. Morrison, *Syst. Biol.* **56**, 988 (2007).

29. A. D. Leaché, B. Rannala, *Syst. Biol.* **60**, 126 (2011).

30. A. D. Gordon, *J. Classif.* **3**, 31 (1986).

31. V. Berry, F. Nicolas, *J. Discrete Algorithms* **5**, 564 (2007).