Terraces in Phylogenetic Tree Space

Michael J. Sanderson *et al.*

Science **333**, 448 (2011);
DOI: 10.1126/science.1206357

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

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

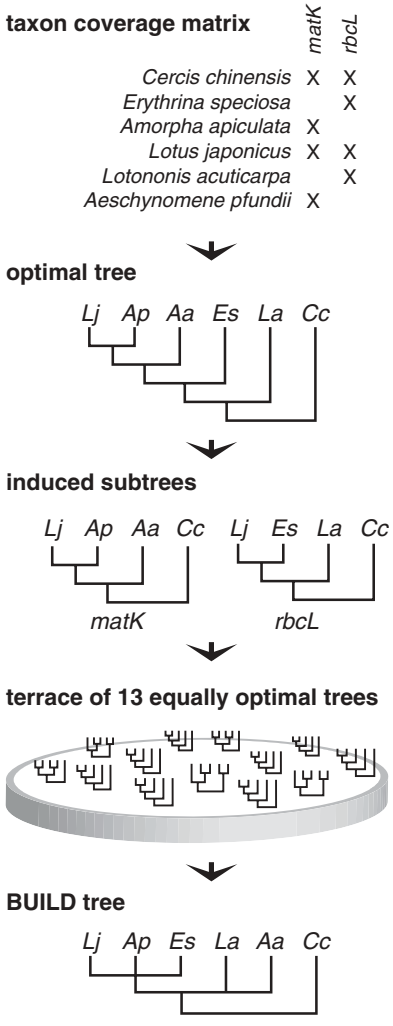

Fig. 1. Terrace in tree space for six species of the angiosperm clade Leguminosae and two loci, *matK* and *rbcL* (*10*). Taxon coverage is denoted by an "X" when sequence data are present. The optimal tree, an ML tree found using a partitioned model in RAxML (ln L = −6709.8), induces two locus-specific 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.

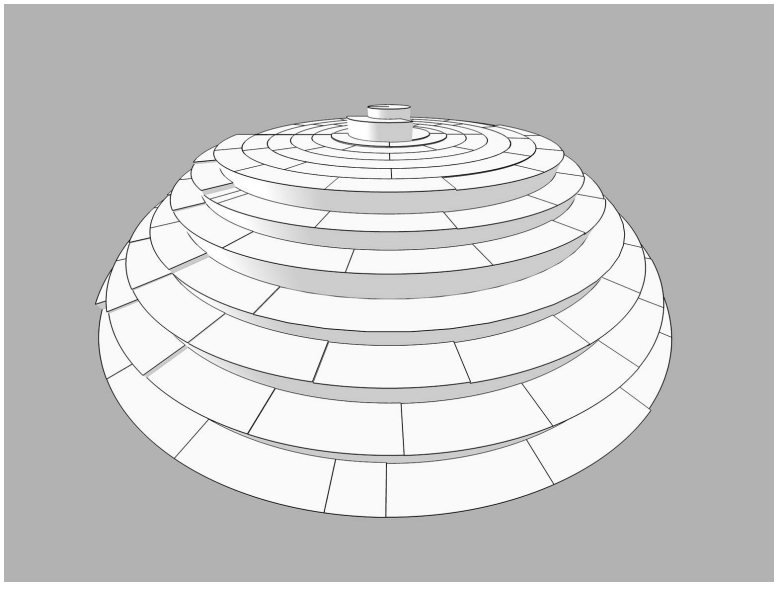

Fig. 2. Visualization of terraces in tree space near the ML tree for the grass data set (*22*). Areas of terraces are proportional to number of trees and height to likelihood score. Total number of trees on all terraces illustrated exceeds 10 billion.

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