Taxonomy and Temporal Diversity Patterns

Heidi E. Robeck; Carlo C. Maley; Michael J. Donoghue

Paleobiology, Vol. 26, No. 2 (Spring, 2000), 171-187.

Stable URL:
http://links.jstor.org/sici?sici=0094-8373%28200021%2926%3A2%3C171%3ATATDP%3E2.0.CO%3B2-7

Paleobiology is currently published by Paleontological Society.

Paleobiology, 26(2), 2000, pp. 171–187

Taxonomy and temporal diversity patterns

Heidi E. Robeck, Carlo C. Maley, and Michael J. Donoghue

Abstract.—Temporal diversity patterns have traditionally been analyzed by counting the number of families or genera present over a series of time periods. This approach has been criticized on the grounds that paraphyletic taxa might introduce artifacts. Sepkoski and Kendrick (1993) simulated phylogenetic trees and different classifications of those trees and concluded that paraphyletic taxa need not be rejected. We have reimplemented their model, extended it, and carried out statistical analyses under a variety of experimental conditions. Our results show that the focus on monophyly vs. paraphyly is misplaced. Instead, it appears that the number of groups in the classification and the distribution of the sizes of those groups have dramatic effects on the recovery of diversity information. Furthermore, the influence of these factors depends on whether the fossil record represents a low- or high-frequency sampling of lineages. When sampling is good, the best results are achieved by classifications with large numbers of small taxa. When sampling is poor, however, the best results are achieved by classifications that include some large and medium-sized groups as well as many smaller groups. This suggests that the best estimates of underlying diversity will be achieved by counting (in the same study) taxa assigned to different ranks, so as to best match the inferred quality of the paleontological sample. In practice this will mean abandoning the commitment to counting taxa at a single rank.

Heidi E. Robeck. Museum of Comparative Zoology, 26 Oxford Street, Cambridge, Massachusetts 02138. E-mail: hrobeck@oeb.harvard.edu

Carlo C. Maley. MIT NE43-937, 545 Technology Square, Cambridge, Massachusetts 02139. E-mail: cmaley@cs.unm.edu

Michael J. Donoghue. Harvard University Herbaria, 22 Divinity Avenue, Cambridge, Massachusetts 02138. E-mail: mdonoghue@oeb.harvard.edu

Accepted: 23 December 1999

Introduction

Temporal patterns of diversity have traditionally been analyzed by plotting the number of taxa at a particular rank (usually families or genera) present over a series of time periods (e.g., Valentine 1969; Raup 1972; Sepkoski 1984). In this endeavor paleobiologists have generally relied upon classifications in the literature. But is one classification just as good as another for this purpose? What attributes of a classification make it better or worse in reflecting the underlying pattern of lineage diversity?

Patterson and Smith (1987, 1988; Smith and Patterson 1988) argued that the inclusion of paraphyletic and monotypic groups could give misleading results. In the case of paraphyletic groups they reasoned that one might mistakenly count a lineage as having gone extinct when it actually has surviving members placed in another taxon at that same rank. Monotypic taxa were regarded as merely an expression of taxonomic opinion about phenetic distinctness or ignorance of relationships. When they eliminated ("culled") paraphyletic and monotypic families of fishes and echinoderms from the Raup and Sepkoski (1984, 1986) data set, and tallied only monophyletic families of two or more species, they failed to see approximately 20% of the supposed extinctions based on traditional classifications.

Motivated by these studies, Sepkoski and Kendrick (1993; hereafter S&K) conducted simulations to test how well classifications that include paraphyletic groups capture lineage diversity, particularly patterns of extinction, as compared to classifications from which paraphyletic groups are omitted. They found that when the fossil record represents a good sample of the underlying diversity, classifications with paraphyletic groups performed adequately, although generally not as well as classification with only monophyletic groups. However, under poor sampling of the fossil record, classifications with paraphyletic groups often performed better. They conclud-

 0094-8373/00/2602-0001/$1.00

ed from these experiments that "paraphyletic groups can adequately capture lineage information under a variety of conditions of diversification and mass extinctions" (p. 168). Although S&K were cautious in interpreting their results, they are widely cited as having vindicated the standard use of traditional classifications (Labandeira and Sepkoski 1993; Mayr 1994; Patzkowsky 1995; Wagner 1995; Foote 1996; Roy 1996; Roy et al. 1996).

Examination of S&K's Figure 1 (see our Fig. 1) suggests that their attempt to mimic Patterson and Smith's culling procedures introduced some confounding factors, some of which they noted (e.g., S&K: p. 178). In particular, when paraphyletic groups are omitted there may be many fewer taxa than in the original classification. The resulting classifications tended to have smaller taxa, since most of the large taxa were paraphyletic. Furthermore, the removal of paraphyletic groups, under either their "hard" or "soft" culls (see below), resulted in many lineages not being included in any taxon. That is, much of the lineage tree is no longer "covered" by the classification. In view of these differences, it is perhaps not surprising that monophyletic classifications produced by culling did not perform well under some circumstances. To investigate the influence of taxon number and taxonomic coverage of the underlying lineage tree, we reimplemented the S&K model, extended it, and carried out statistical analyses under a variety of experimental conditions.

Sepkoski and Kendrick's Analyses

S&K generated phylogenies of up to 700 lineages using a stochastic branching process based on either logistic or exponential growth. An example of a portion of one of their phylogenies is shown in Figure 1. They varied the probabilities of extinctions from 0.10 to 0.25 per time step, and the probabilities of speciation from 0.2 to 0.5 (mistakenly reported as 0.05 to 0.20 in S&K [J. Sepkoski personal communication 1996]). Boosting the extinction rate for a single time step and then resetting it to the original rate in the next time step created mass extinctions (S&K sometimes increased the probability of extinction over more than one time unit but did not report any resulting differences).

For each phylogeny, S&K simulated a traditional classification, which we refer to as the "mixed classification" because it contains both paraphyletic and monophyletic groups (Fig. 1B). This classification was created by randomly selecting an ancestor ("taxon-defining species") and designating all of its descendant species that did not already belong to another taxon as members of a new taxon. We will refer to any classified group as a taxon, whether monophyletic or paraphyletic; appropriate distinctions are made as necessary. If one of the descendant species had already been selected as a "taxon-defining species" then that species and all of its descendants were excluded from the new taxon, rendering it paraphyletic. The probability of any taxon's becoming an ancestor was set at 0.10; however, the root of the tree was always selected as a "taxon-defining species," which ensured that all species were included in some taxon. The default paraphyletic basal group tended to be quite large in most runs. In view of the predominance of small taxa produced by this procedure, S&K accepted only potential taxa with probability $(n + k - 1)/(n + k)$. Here k is a parameter of the model and n is whichever is smaller, the size of the new potential taxon or the size of the remaining taxon from which it would be segregated (Sepkoski 1978). If k is small, there is only a low probability of accepting a small group. If k is large, groups of all sizes will be readily accepted. The process of selecting taxon-defining species was repeated until 100 taxa were circumscribed.

Next, S&K attempted to emulate the culling procedure of Patterson and Smith (1987, 1988) by removing the paraphyletic and monotypic groups. They called this classification the "hard cull." Notice in the example shown in Figure 1C, where the dark gray groups are the only monophyletic groups identified from Figure 1B, that about half of the taxa are removed from consideration. Finally, and again following Patterson and Smith, they returned to the classification the largest monophyletic groups contained within the culled paraphyletic groups, the striped groups in Figure 1C. The classification including these internal

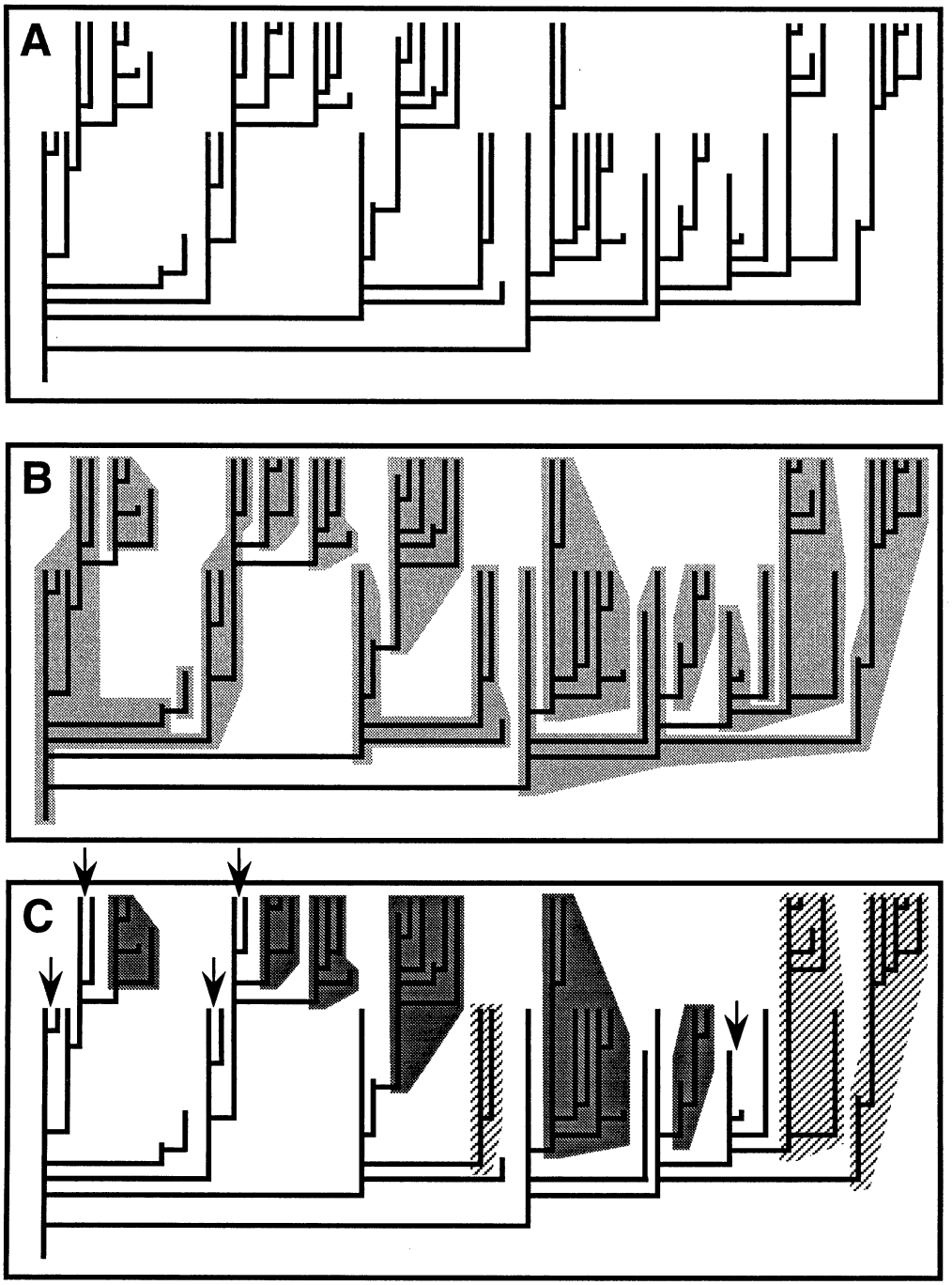

FIGURE 1. Redrawn from Sepkoski and Kendrick (1993: Fig. 1). A, An example of a portion of a simulated phylogeny. B, The "mixed" classification (the "paraphyletic" classification of S&K) is represented by shading. C, The "hard cull" is referenced by the dark-gray groups. Groups that would be added back into the classification in the "soft cull" are indicated by diagonal stripes. Arrows indicate groups that would be added back into the classification in "soft cull II."

monophyletic groups was termed the "soft cull." Although the small example shown in Figure 1C happens to have fewer taxa than the original mixed classification, soft-cull classifications contain more taxa in larger trees (see below).

The definition of speciation bears on the identification of monophyletic groups within culled paraphyletic groups. S&K assumed that each splitting event resulted in the creation of a single new species (J. Sepkoski personal communication 1996). Following the convention of Raup et al. (1973), the right-hand branch represents the new species and the left-hand line is the ancestral species. Alternatively, speciation can be viewed as producing two new species, neither one ancestral to the other (e.g., Hennig 1966). The practical consequence of this distinction can be seen in Figure 1C, where we have identified several groups (with arrows) that would be added back to the classification under the second view of speciation. These groups were not added back by S&K because if the entire ancestral species (on the left side) were included, they would be paraphyletic.

S&K treated all taxa created by their procedure as belonging to a single taxonomic rank, and counted the number of these taxa present through a series of times to estimate diversity in the underlying lineages. As a measure of performance they reported a standard squared product-moment correlation coefficient, r^2, between the taxa of the classification and the underlying diversity. The data were also detrended by examining the correlations of the residuals calculated from a linear regression on time. This procedure corrects for misleading results that might occur in time-series data with an increasing (or a decreasing) trend.

In addition to carrying out comparisons under the assumption that the fossil record provides perfect knowledge of the presence of taxa at each time, S&K also explored the effect of partial knowledge due to paleontological sampling. In one model, lineages across all time periods were assigned a constant probability of being found. They explored sampling probabilities from 0.05 to 0.2 with the additional assumption that in the final time step, all extant lineages are perfectly sampled. In a second model, motivated by Patterson and Smith (1987), they gave lineages in any given time period a probability of being sampled that was randomly chosen from an exponentially decaying distribution of probabilities. This distribution ranged from a non-zero minimum to a maximum value, which were additional parameters in the model. Thus, most time steps had a low sampling rate, punctuated by the occasional time step with a high sampling rate. Under both forms of sampling, if only a single lineage in a given taxon was sampled, then the taxon was assumed to be monotypic and, following Patterson and Smith, was removed in the culling procedure. If a lineage was sampled more than once, it was assumed to have been extant in all the time steps between the samples.

Our Analyses

S&K explored a wide variety of parameters (speciation rate, extinction rate, and the number and magnitude of mass extinctions) under two models of diversification (exponential and logistic). Because we wanted to examine specific attributes in enough detail to allow tests for statistical significance, we initially limited ourselves to varying a subset of the parameters under just one model, exponential diversification (but see below). S&K reported dramatic effects of paleontological sampling rate, and we have focused special attention on this parameter. We reimplemented the S&K algorithms for generating phylogenies (using exponential growth) and for creating mixed, soft-cull, and hard-cull classifications. Programming was carried out in LISP, as this language facilitates manipulation of the recursive structures of trees. Source code may be obtained from the Paleobiology Software Archive (http://geosci.uchicago.edu/paleo/csource) or by request from the second author.

We chose intermediate values in the ranges explored by S&K for the parameters that were not reported to have had an important effect on the dynamics. We set speciation rate = 0.4, extinction rate = 0.2, the number of taxa generated = 100, and $k = 0.05$. We invoked three mass extinctions, which eliminated approximately 70%, 95%, and 80% of the extant line-

ages and occurred after 20%, 45%, and 80% of the terminal nodes in the tree had been generated, respectively. For example, in the first mass extinction, every extant species had a 70% chance of going extinct during that one time step. Once that time step was over, the extinction rate was reset to the default of 0.2. We also increased the number of lineages per phylogeny to 1000 (S&K were limited to 700). This number was generally reached in the middle of a time step. Because the last time step of the simulation therefore differed from the rest we excluded data from it in our analyses.

Additional Classifications

In an attempt to uncouple not only the number and average size of taxa in a classification, but also the distribution of taxon sizes, from monophyly and paraphyly, we created five new classification types. For each of these, and for S&K's original three classifications, we kept track of the number and the size of taxa.

Soft Cull II.—Because we felt that the model of speciation used by S&K might affect the results, we implemented a version of the soft cull in which we viewed a speciation event as producing two new species. The effect is that more taxa are added back to the classification in the process of subdividing large paraphyletic groups into the largest monophyletic subgroups. Indeed, these classifications tended to have approximately two-thirds more taxa than the original mixed classification, even though monotypic groups were still excluded.

Full.—Because we saw no compelling reason to exclude monotypic taxa, and wanted to examine their impact on the analysis, we created a classification identical to soft cull II, but with monotypic groups added back in. These classifications typically contained over three times as many taxa as mixed classifications.

Random.—Because we suspected that the number of taxa might be an important factor, we created a classification that randomly selected nodes from the tree to form taxa (Fig. 2). This procedure has three useful properties: (1) the members of a taxon need not be, and often are not, convex; that is, they are usually neither monophyletic nor paraphyletic, but polyphyletic; (2) we can vary the number of taxa identified, up to a maximum of the number of terminal lineages (1000); and (3) like the mixed classification (with both paraphyletic and monophyletic groups), all of the terminal and ancestral nodes in the tree are included in some taxon. To generate a random classification we first designate the number of taxa, n, we wanted to produce. We randomly assigned lineages to the n taxa by uniformly selecting a number from 1 to n for each node in the tree. By chance, some potential taxa might never have had a terminal lineage assigned to them. Any "empty" taxon was given a terminal lineage from a "full" taxon that had two or more terminal lineages. This turned empty taxa into monotypic taxa. The leftmost "full" taxon was arbitrarily selected to be the "donor" taxon. We set n to be the midpoint between the number of terminal lineages in the tree (1000) and the number of taxa in our "full" classification, which resulted in the creation of many taxa (ca. 660). Random classifications have more taxa than any other type of classification but obviously have the least correspondence to patterns of relationship in the underlying tree.

Matched-Random.—Because we wanted to test the importance of factors other than the number of taxa, we implemented a form of random classification in which the number of taxa (n) was set equal to 100. Both the number of taxa and the coverage were matched to the mixed classification.

Matched-Distribution-Random.—Because the classifications described above differ not only in the average number of taxa, but also differ dramatically in the distribution of taxon sizes, we created a random classification that exactly matched the distribution of taxon sizes in the mixed classification. For example, if the mixed classification had a taxon with 432 terminal lineages, then the matched-distribution-random classification would also have a taxon of that size. This was done by creating taxa with the same number of ancestral and terminal nodes as taxa in the mixed classification, but with the particular ancestors and terminals selected randomly from across the entire phylogeny.

Table 1 provides a summary of all eight classifications. Note that comparisons involving the three kinds of random classifications

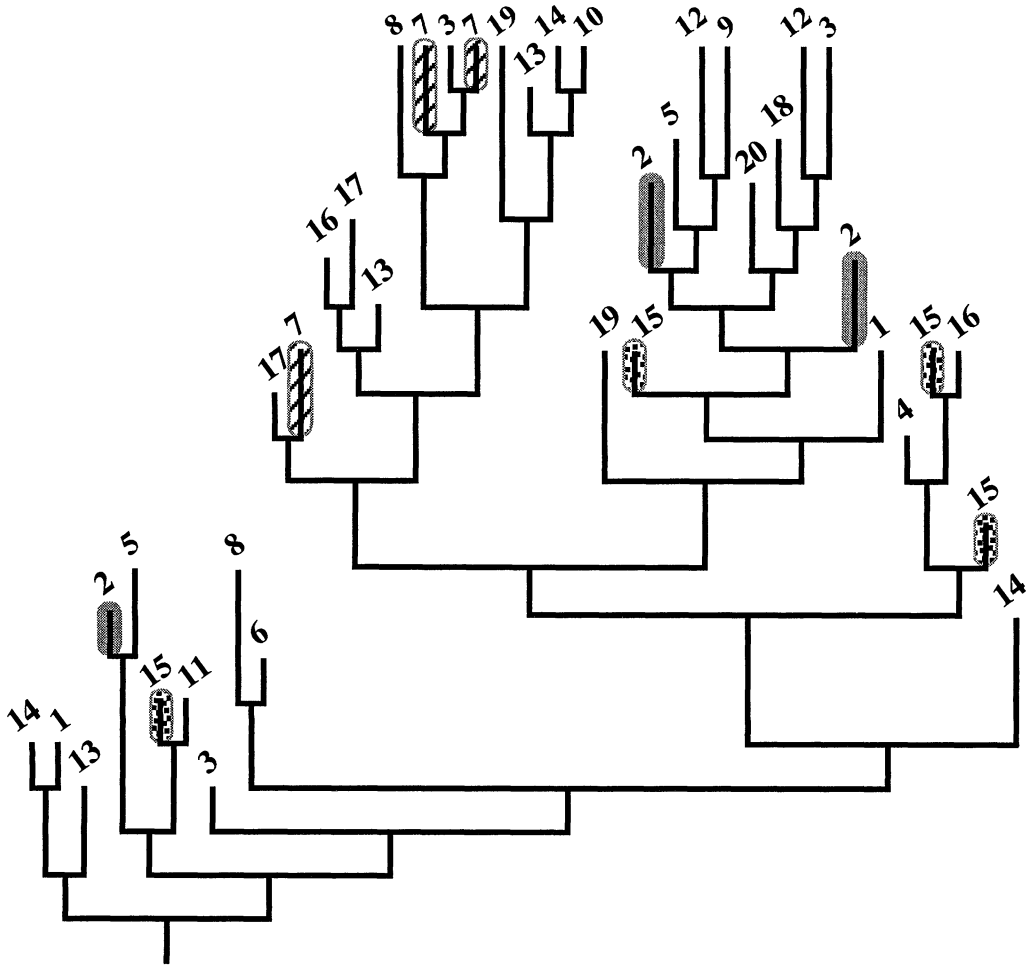

FIGURE 2. A portion of a simulated phylogeny showing the random classification. Lineages assigned to the same taxon are designated by the same number. Several taxa are highlighted to illustrate their polyphyly.

just described allow us to examine three potentially important factors: the number of taxa ("random" vs. "matched-random"), the distribution of taxon sizes ("matched-random" vs. "matched-distribution-random"), and the relationship between taxa and the underlying tree ("mixed" vs. "matched-distribution-random").

Sampling Procedures

S&K did not describe a substantial difference in the results of their two models of sampling: constant probability sampling vs. drawing sampling rates from an underlying distribution of probabilities. For simplicity, we therefore concentrated on a constant sampling rate. Furthermore, to maintain the consistency of a classification, we applied all of our classifications to the full tree. Thus, a taxon that was polytypic under perfect sampling but monotypic after sampling would not be removed in our implementation of the hard or soft cull.

S&K assumed that lineages spanned the time from their first to last sampled occurrences. In our analyses we extended this approach, using the phylogeny to infer the presence of lineages ("ghost lineages" [Norell 1992]) in the intervening time periods between samples. First, if a lineage is first sampled at time step n and last sampled at time step $n+k$, then we assume it was extant for all k time steps in between. Second, if a lineage's first sample is at time n and its sister lineage's first sample is at time $n+k$, then we assume

TABLE 1. Summary of classifications used.

S&K's original classifications (Fig. 1)	
Mixed	S&K's "traditional" classification, containing both paraphyletic and monophyletic groups
Hard cull	Same as mixed but paraphyletic and monotypic groups removed
Soft cull	Examines groups removed in hard cull and adds back the largest monophyletic groups contained within the culled paraphyletic groups; each speciation event yields one new species (see text)
Added classifications	
Soft cull II	Same as S&K's soft cull but each speciation event yields two new species (see text)
Full	Same as soft cull II but monotypic groups added back
Random	Randomly selects nodes from the tree to form taxa; taxa usually polyphyletic (Fig. 2)
Matched-random	Same as random but number of taxa and coverage matched to mixed classification
Matched-distribution-random	Same as matched-random but distribution of taxon sizes matched to mixed classification

that the sister lineage was also extant at time n. Third, if the first sample of a lineage is at time $n+k$ and the last sample of its ancestral lineage is at time n, then we assume that a speciation event occurred at some point in those k steps. We used the conservative heuristic that the speciation event happened just before the first sample of the descendent lineage, at time step $n+k$-1. In general, by taking advantage of the phylogeny, this approach should yield a somewhat higher correlation between sampled lineages and underlying diversity.

Figure 3 shows part of a phylogeny that has been sampled. In this example, lineages in each time step had a 0.05 probability of being sampled. We collected data for all eight classifications over sampling rates of 1.0 ("perfect sampling"), 0.25, 0.20, 0.15, 0.10, and 0.05. We conducted 500 runs for perfect sampling and 100 runs at the other rates, all with unique pseudo-random number seeds.

Measuring Performance

To compare diversity through time of the underlying lineages with the diversity of taxa in a particular classification, we followed S&K in using detrended data as a means of correcting for misleading results that can occur using the standard r^2 correlation with time-series data. An example of diversity curves for lineages and several classification types is given in Figure 4.

Additional Diversity Growth Models

Although S&K did not report major differences between exponential and logistic models of diversification, we ran the model under three additional conditions to determine whether our results generalized across different diversity growth models and in the absence of mass extinctions. Specifically, we examined S&K's logistic growth model with the same three mass extinctions used above and then both exponential and logistic growth in the absence of mass extinctions. All other parameters, classifications, sampling procedures, and measurements of performance were the same as above.

Results

Perfect Knowledge.—We examined the performance of the classifications under exponential growth with mass extinctions and under the conditions of perfect knowledge, when all the lineages were sampled every time step. The average squared correlations (r^2) for 500 runs with perfect knowledge are given in Table 2. Average numbers of taxa are also included in the table. We tested for significant differences between the means of the correlations in each of the classifications using a paired, two-tailed t-test. Because the distributions of these correlations were not normal, we used a square-root transformation on the correlations. There was a significant difference ($p < 0.05$) between all pairwise comparisons of the means.

Our results are similar to those of S&K. The soft cull performed the best, followed by the mixed classification and then the hard cull, with means (and standard deviations) of 0.884 (0.033), 0.775 (0.078), and 0.738 (0.097), respectively. The order of performance is strongly correlated with the number of taxa in the classification. The hard-cull procedure removed almost half of the groups from the analysis and performed the worst. The soft-

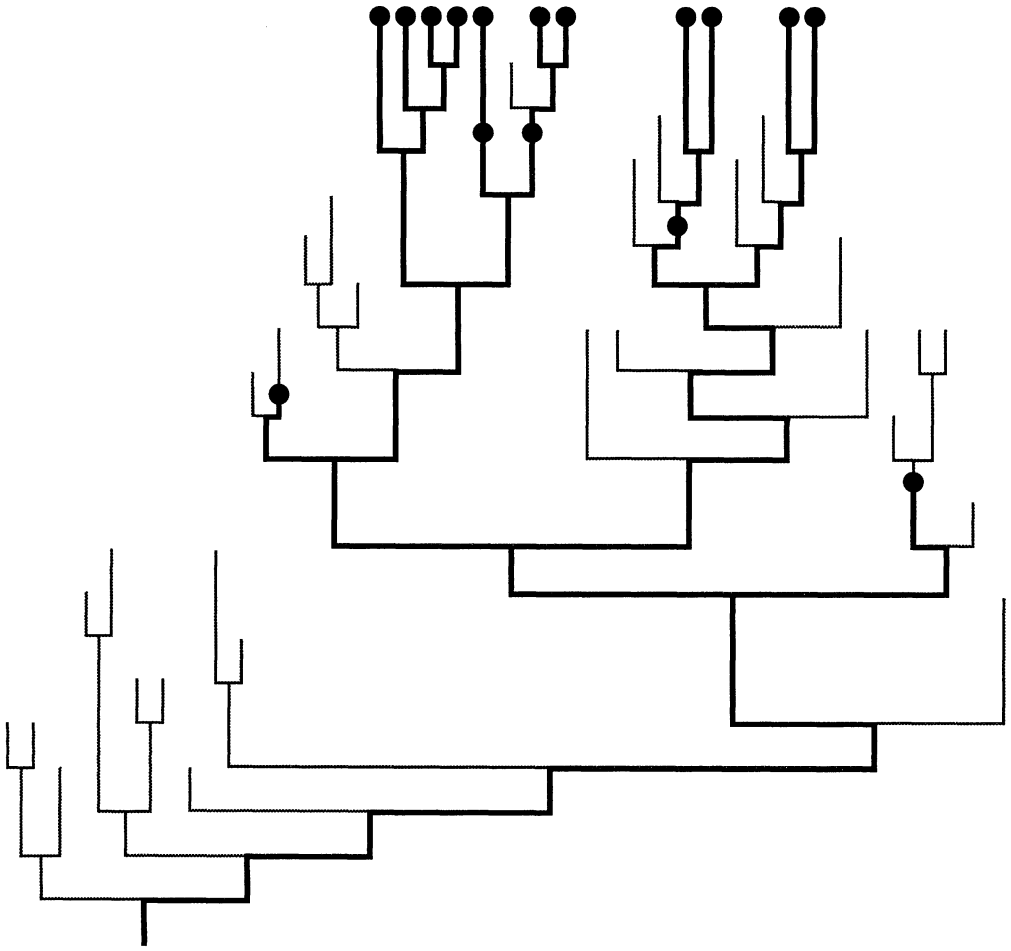

FIGURE 3. A portion of a simulated phylogeny showing an example of lineages sampled at the 0.05 rate. The dots represent sampling events; the existence of a lineage between events (indicated by dark branches) is inferred by reference to the phylogeny. The root of this tree assumes a known outgroup with a prior fossil record.

cull classification always had more taxa than the mixed classification. The random classification, which has the greatest number of taxa, outperforms all the other classifications with an r^2 of 0.995 (0.002). Specifically, it compares favorably to the matched-random classification, with 100 taxa, which had an average r^2 of 0.899 (0.030).

Comparison of the matched-random classification with the matched-distribution-random classification highlights the effect of the distribution of taxon sizes in the performance of a classification. The average r^2 of the matched-distribution-random is only 0.807 (0.092), now much closer to the mixed classification at 0.775. Figure 5 shows a natural log transformation of the cumulative distribution of taxon sizes for all the classifications. Although both the mixed and the matched-random classifications have the same number of taxa, the matched-random classification has many uniformly small groups while the mixed classification has a variety of small, medium, and very large groups.

Partial Knowledge.—We also examined the performance of the classifications under the same diversification model but under partial knowledge, with sampling rates of less than 1.0. An example of diversity curves for a sampling rate of 0.05 is shown in Figure 6. We calculated goodness-of-fit between all the curves representing different classifications and the curve representing the actual underlying lineages, as before. Table 3 shows the average r^2

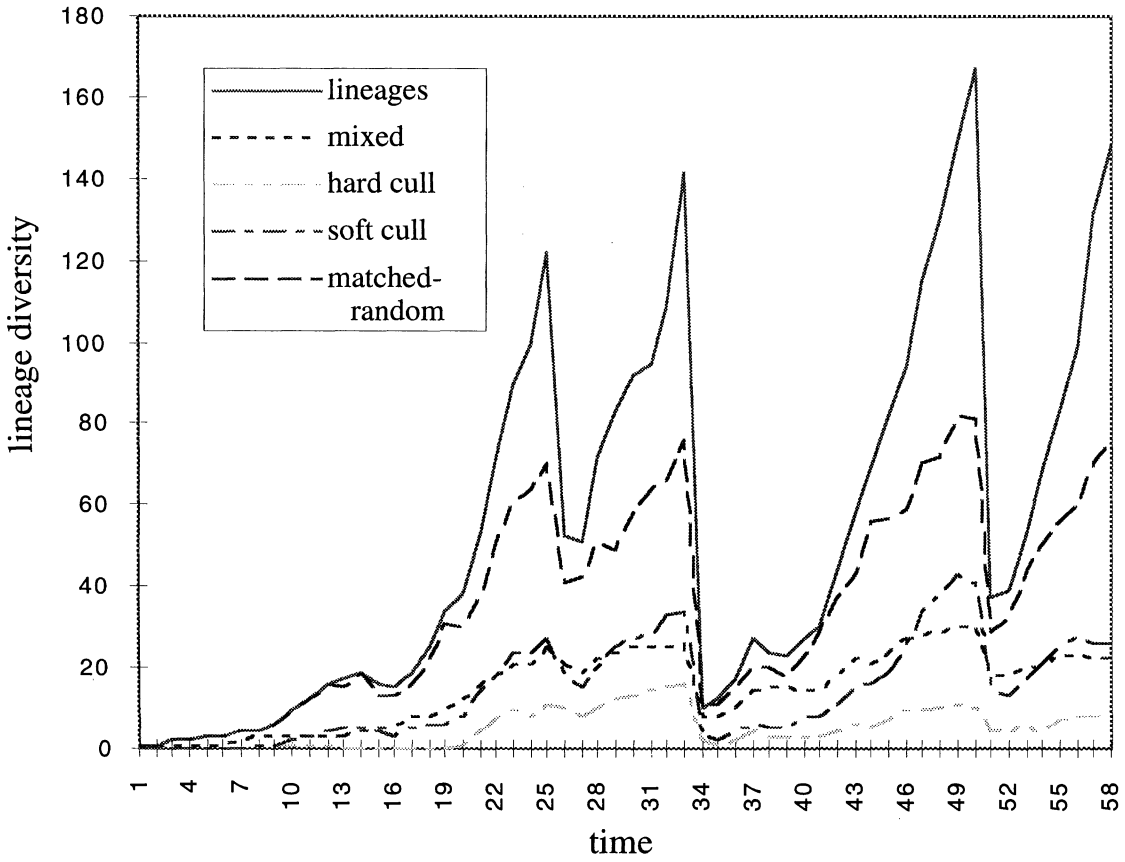

FIGURE 4. Example of diversity curves for four classifications under perfect knowledge. The line for each classification is compared with the line representing the underlying lineages.

based on detrended data for each of the classifications over 100 runs at sampling rates of 0.05, 0.10, 0.15, 0.20, and 0.25. For almost all pairwise comparisons between classifications, there is a distinct inflection point where one classification changes from being better to worse in relation to the other depending on the level of sampling (see Fig. 7). Except during this time, the paired, two-tailed t-test showed that there was a significant difference ($p < 0.05$) between most pairwise comparisons of the means for all classifications, despite the large variances and the small differences between some of the means. The two exceptions included the difference between soft cull II and matched-random at a sampling rate of 0.05 ($p = 0.069$) and the differences between the soft cull and the matched-random at sampling rates 0.15, 0.20, and 0.25 ($p = 0.437$, 0.357, and 0.436, respectively). However, all comparisons discussed below are significant. Probability values for pairwise comparisons that are not significant are given in Table 3.

TABLE 2. Results of 500 runs with perfect sampling. Standard deviations are given in parentheses.

Classification	Average detrended r^2	Average number of taxa
Mixed	0.775 (0.078)	100.00 (0.00)
Hard cull	0.738 (0.097)	54.31 (4.02)
Soft cull	0.884 (0.033)	145.04 (8.88)
Soft cull II	0.868 (0.036)	172.17 (11.42)
Full	0.878 (0.040)	332.95 (17.60)
Random	0.995 (0.002)	666.44 (8.80)
Matched-random	0.899 (0.030)	100.00 (0.00)
Matched-distribution-random	0.807 (0.092)	100.00 (0.00)

The correlations are much lower than under perfect knowledge, and the order of performance for the classifications is now inversely correlated with number of taxa (with the exception of the hard cull). The mixed classifi-

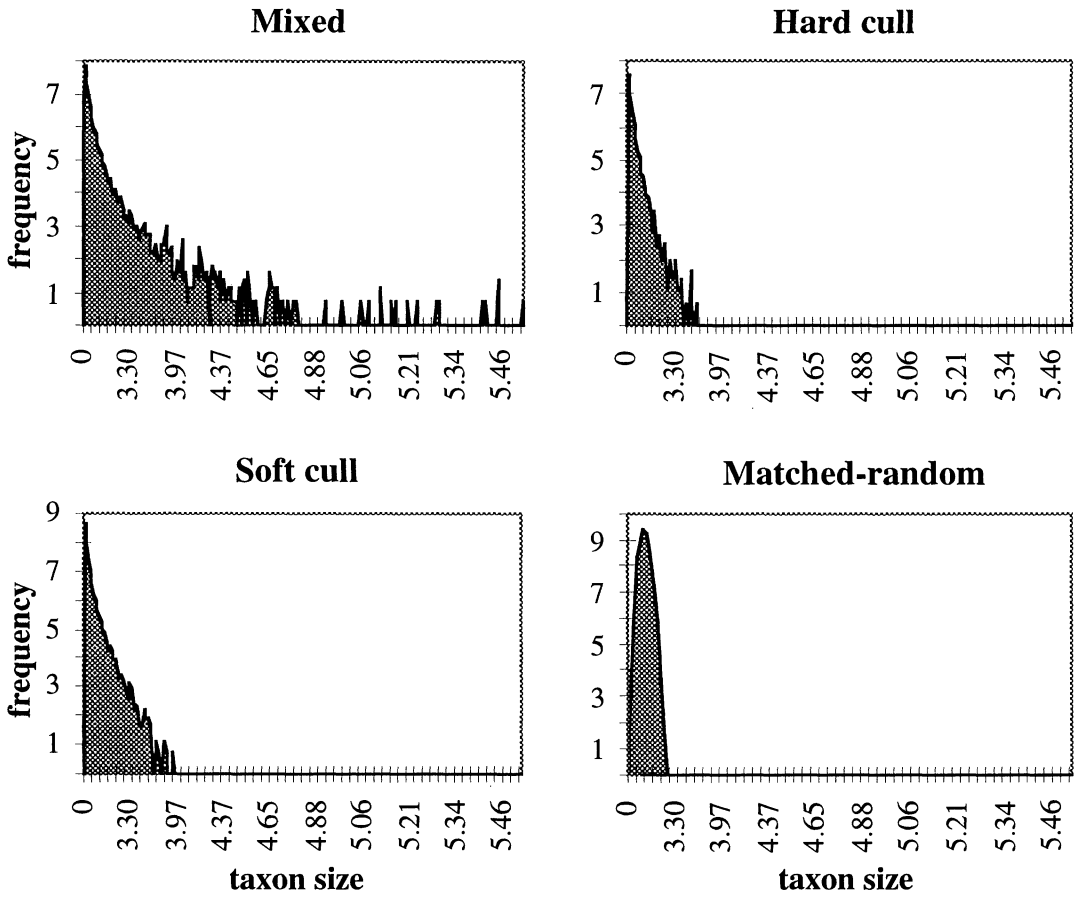

FIGURE 5. The distribution of taxon sizes for four classifications on a log-linear scale. The matched-distribution-random classification has the same distribution of taxon sizes as the mixed classification and is not shown separately.

cation is now doing better than the soft cull, with squared correlations of 0.278 (0.143) and 0.178 (0.112), respectively. The random classification's performance has "crashed," going from 0.995 (0.002) in perfect sampling down to 0.167 (0.095).

The matched-distribution-random classification has an average r^2 of 0.244 (0.123), as compared to the matched-random classification, which only has an average r^2 of 0.214 (0.114). The matched-distribution-random classification's performance is much closer to that of the mixed classification at 0.278, relative to the soft cull with an average r^2 of only 0.178 (0.112).

Additional Diversity Growth Models.—The Appendix shows the average detrended r^2 for S&K's logistic growth model with the same three mass extinctions used above and both exponential and logistic growth in the absence of mass extinctions. For each model, 250 runs were performed with perfect knowledge and 50 runs at sampling rates of 0.05, 0.10, 0.15, 0.20, and 0.25. Probability values from the paired, two-tailed *t*-test are available by request from the first author. The qualitative results noted above hold across these models, except for the case of exponential growth without mass extinctions. In this case, a rising curve of any shape will be highly correlated with the exponential curve, and so, even at the 0.05 sampling rates all the classifications have an r^2 greater than 0.84. In the other cases, the mixed and matched-distribution-random classifications do well under 0.05 sampling. At 0.25 sampling, the monophyletic (soft cull, soft cull II, and full) classifications tend to score best. The hard cull still performs abys-

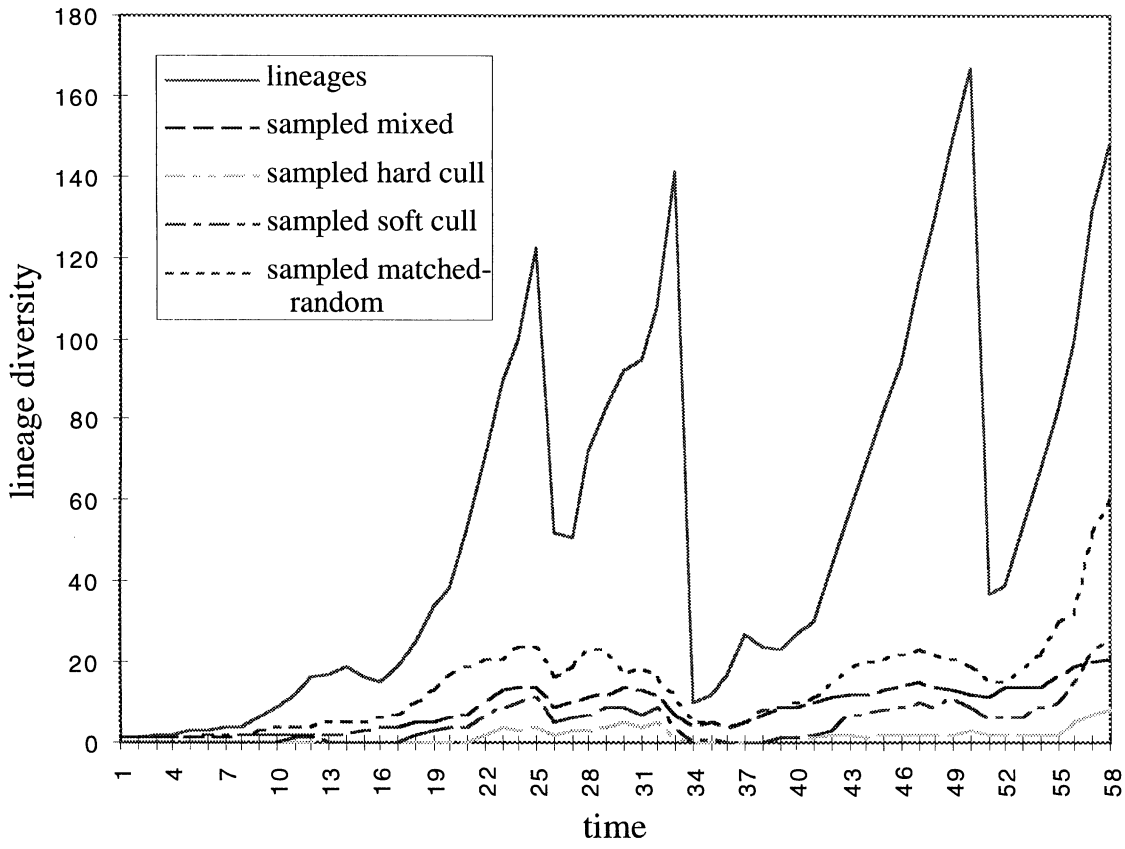

FIGURE 6. Example of diversity curves for four classifications at the 0.05 sampling rate. The line for each classification is compared with the line representing the underlying lineages.

mally. Finally, under perfect sampling the random classification does best.

Discussion

The difference in performance between the random (0.995) and the matched-random (0.899) classifications under perfect sampling shows that the sheer number of taxa in a classification is an important factor in tracking diversity. Of course, in a fully classified phylogeny, the number of taxa is inversely proportional to the average size of the taxa. So we might also say that performance is inversely correlated with the average size of the taxa under perfect sampling. In general, the results imply that when sampling is good small groups respond to fluctuations in diversity more readily than larger groups. When sampling is poor the performance of the classifications is inversely correlated with the number of taxa (and positively correlated with the average size of the taxa). The fact that the matched-random classification is still doing better than the mixed (0.775) classification under perfect sampling implies that there are other factors, beyond the number of taxa, that are affecting the performances of the classifications.

The distribution of taxon sizes is clearly a significant factor in the performance of a classification. The only difference between the matched-random (0.899) and the matched-distribution-random (0.807) classification is the distribution of taxon sizes. However, the remaining difference between the matched-distribution-random (0.807) classification and the mixed (0.775) classification suggests that there are other factors yet to be identified. One hypothesis for this difference focuses on the large basal group (often 30–50% of all lineages are included in this group) in the mixed classification. This is the default taxon into which

TABLE 3. Results of 100 runs at each of five sampling rates under an exponential growth model of diversification with three mass extinctions. Standard deviations are given in parentheses. All pairwise comparisons between classifications using the paired, two-tailed t-test showed a significant difference ($p < 0.05$) except SCII vs. MR at the 0.05 rate ($p = 0.069$); M vs. SCII and M vs. F at the 0.10 rate ($p = 0.822$ and 0.687, respectively); M vs. SC, M vs. MR, SC vs. MR, SCII vs. R, and SCII vs. F at the 0.15 rate ($p = 0.523$, 0.935, 0.437, 0.760, and 0.760, respectively); LS vs. M, R vs. M, SC vs. MR, SCII vs. R, and F vs. SCII at the 0.20 rate ($p = 0.779$, 0.802, 0.357, 0.139, and 0.139, respectively); and SC vs. MR and R vs. SC at the 0.25 rate ($p = 0.436$ and 0.065, respectively).

Classification	Average detrended r^2				
	0.05	0.10	0.15	0.2	0.25
Lineages sampled (LS)	0.159 (0.092)	0.323 (0.100)	0.464 (0.091)	0.569 (0.088)	0.642 (0.070)
Mixed (M)	0.278 (0.143)	0.430 (0.137)	0.521 (0.127)	0.576 (0.115)	0.618 (0.112)
Hard cull (HC)	0.117 (0.102)	0.231 (0.126)	0.336 (0.168)	0.424 (0.170)	0.457 (0.145)
Soft cull (SC)	0.178 (0.112)	0.367 (0.122)	0.526 (0.106)	0.619 (0.103)	0.668 (0.102)
Soft cull II (SCII)	0.206 (0.118)	0.430 (0.128)	0.575 (0.102)	0.666 (0.083)	0.709 (0.080)
Full (F)	0.200 (0.113)	0.424 (0.128)	0.575 (0.109)	0.670 (0.085)	0.716 (0.078)
Random (R)	0.167 (0.095)	0.339 (0.101)	0.475 (0.091)	0.575 (0.088)	0.648 (0.068)
Matched-random (MR)	0.214 (0.114)	0.391 (0.103)	0.518 (0.091)	0.609 (0.081)	0.658 (0.069)
Matched-distribution-random (MDR)	0.244 (0.123)	0.371 (0.135)	0.494 (0.128)	0.543 (0.111)	0.607 (0.113)

all unclassified lineages are put, and it typically dominates the early lineages in the tree. In contrast, while the matched-distribution-random classification has a taxon of equal size, that large taxon is scattered evenly about the tree. We suspected that the association of this large basal taxon with the early evolution of the tree has a significant effect on the mixed classification's goodness-of-fit to the underlying lineage diversity. To check this, we performed the detrended r^2 analysis on the last 30 time steps of each run. Since the length of the runs varied, this tended to capture the last half to three-quarters of the runs. Yet, the re-

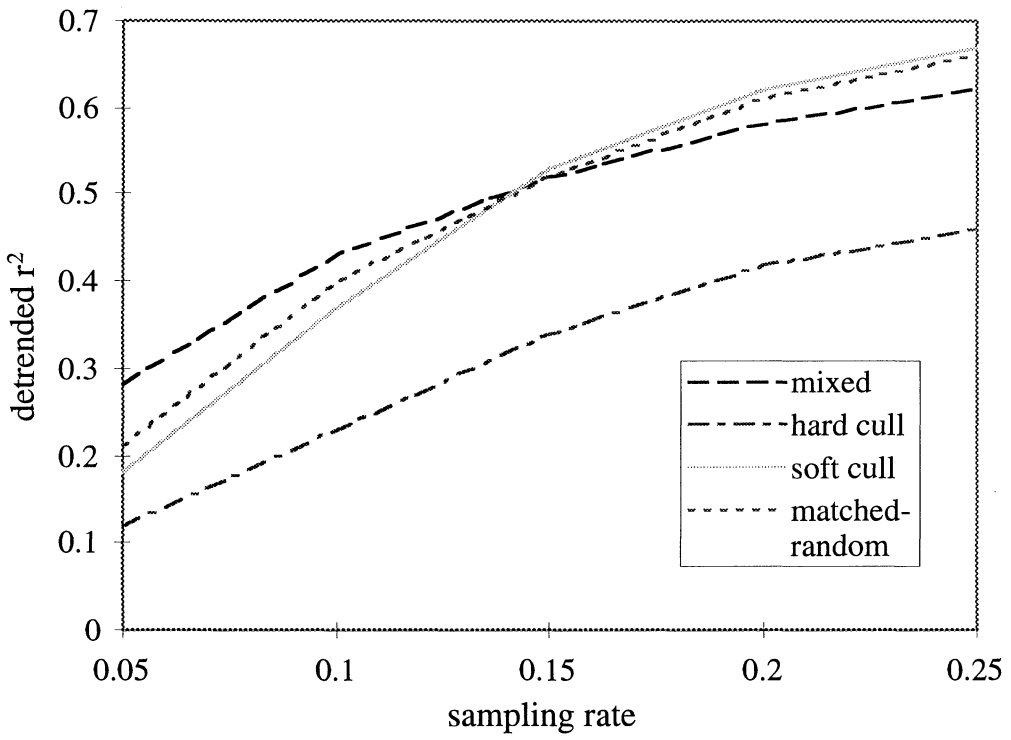

FIGURE 7. Effect of sampling in the fossil record on classification performance. Note inflection point at a sampling rate of 0.15, where the mixed classification changes from being better than the soft cull and random classifications, to being worse.

sults remained qualitatively the same, with the r^2 statistics rising for all the classifications. Under 0.05 sampling, the mixed classification r^2 "leaped" up to 0.334 while the matched-distribution-random increased to 0.264. The difference was exaggerated.

When sampling is poor, all of the classifications do a poor job of detecting anything at all. The relative increase in performance for the mixed classification is the same as that reported by S&K and is their result most often cited (Labandeira and Sepkoski 1993; Mayr 1994; Patzkowsky 1995; Wagner 1995; Foote 1996; Roy 1996; Roy et al. 1996; Lee 1997). It would be difficult to interpret any dips in the diversity curves as mass extinction events without a priori knowledge of those events. An unfortunate implication of this result is that if a particular group of interest occurs in the fossil record at a rate similar to this level of sampling (0.05 per million years per lineage), it is extremely difficult to detect overall trends in diversity. However, all of the classifications except the hard cull do better than the practice of simply counting up sampled lineages. S&K also noted the poor performance of relying on lineage-level data under sampling (p. 179). By grouping lineages into taxa, we can alleviate the fluctuations introduced by the poor sampling. Depending on the classification, the effects of the two critical factors—number of taxa and distribution of taxon sizes—invert at a sampling rate between 0.10 and 0.20 (Fig. 7).

In a sense, there is continuum across the different classifications, from lineages at one extreme, to matched-distribution-random, to mixed, and finally to monophyletic at the far extreme. To see this, it is helpful to think about the probability that a splitting event in a classified lineage will introduce a new taxon to the diversity count. If we are simply counting lineages, then every splitting event will increase the diversity count. However, consider the difference between the mixed classification and the matched-distribution-random classification. While both classifications have the same number of taxa and even the same sizes of taxa, a splitting event in a lineage of the matched-distribution-random classification will very often herald the origin of a new taxon because phylogenetic relationships were ignored in this classification. In contrast, some of the splitting events in a lineage of the mixed classification will introduce new taxa, but frequently splitting events will just lead to more lineages in the same taxon as the ancestral lineage. At the far extreme lie the monophyletic classifications. Splitting events within a monophyletic group at a particular rank will never result in a new taxon at that rank. Thus, if we hold the number of taxa and the taxon size distributions constant, we should expect a continuum of classification performance corresponding to this continuum of splitting behavior. Whether performance increases or decreases depends on the sampling rate.

Another way to see this is to consider the effect of the loss of a lineage due to poor sampling. This may incorrectly appear to herald an extinction. If we are simply tracking lineage diversity, we will always be misled by such a disappearance. If we are tracking a matched-distribution-random classification, we are likely to be misled by sampling error because the taxa are scattered about the tree. At any given time, there are just a few lineages—perhaps only one—extant. In contrast, a taxon from a mixed classification is convex and so is more likely to have more than one lineage extant. Such a clustering of lineages into taxa buffers the taxa from the effects of failure to sample individual lineages. Finally, monophyletic taxa are most likely to have multiple lineages extant at any given time. Thus, all other things being equal, monophyletic taxa should be the least susceptible to the diversity fluctuations introduced by infrequent sampling, and the random classifications should be the most susceptible to fluctuations.

In summary, the results indicate that when sampling is good, classifications with many small groups are optimal. When sampling is poor, large groups help to dampen the fluctuations injected by the sampling. This dampening effect seems to compensate for the inability of these larger groups to track the underlying lineage diversity. Our results indicate that the mixed classification is doing better under poor sampling, not because of any inherent property of the paraphyletic groups it contains, but simply because it has a variety of larger groups.

This is a property of the specific algorithm used in these simulations.

Our results suggest a number of questions for future study. While we have presented evidence that both the number of taxa and the distribution of the sizes of the taxa are important factors, there is clearly more to the story. If we compare the detrended r^2 of the mixed classification under perfect sampling (0.785) with the matched-distribution-random classification (0.812), we can see that even when we hold constant the number of taxa and the distribution of their sizes, we still get different results. Under the 0.05 sampling rate, there is also a difference: the mixed classification has a 0.278 correlation vs. 0.244 for the matched-distribution-random. What remaining factors account for these differences? We identified at least two major differences between these classifications. First, because the random classifications collect lineages from all over the tree under a single taxon name, a random taxon may appear and disappear many times over the course of time. In contrast, a taxon from the mixed classification always appears as a temporally contiguous set of lineages and so will appear and disappear only once under perfect sampling. Second, while there is an exact match between the sizes of the taxa in the two classifications, taxa of the same size may appear at different times in the two classifications. This is easiest to see when we consider the large basal group that typically appears in the mixed classification. This group generally includes the root of the tree as well as roughly 25% of the terminal nodes of the tree, with all the ancestral nodes in between. The large group in the matched-distribution-random classification includes the same number of nodes but they are scattered haphazardly across the tree.

It should also be noted that while we have identified the distribution of taxon sizes as an important factor in the performance of a classification, we have not isolated the aspects of a distribution that make it perform well under perfect or poor sampling. The mixed classification performs well under 0.05 sampling (0.278) with a variety of taxon sizes. However, when we created a monophyletic classification with a taxon size distribution that closely matched the mixed classification's distribution, it outperformed the mixed classification under the 0.05 sampling with an average detrended r^2 of 0.362. This classification was created by working down from the largest taxa in the mixed classification and attempting to find a yet unclassified monophyletic group of the same size in the tree. When an exact match could not be found, the next smaller monophyletic group was accepted. To compensate for the fact that the larger groups of this monophyletic classification were sometimes slightly smaller than the analogous groups in the mixed classification, the small groups of the monophyletic classification often had to be slightly larger than the small groups of the mixed classification. Interestingly, this matched-distribution-monophyletic classification performs abysmally under perfect sampling (0.652). These effects are probably due to the small differences in the taxon size distribution as well as additional unidentified factors hinted at above.

The remaining open questions focus on the validity of the simulations. We have followed the methodology of S&K, and so our results are vulnerable to many of the same sorts of criticisms and caveats as theirs. It is highly doubtful that the mixed classification, upon which most of the other classifications are based, is an accurate model of real taxonomies. For example, the large groups in the mixed classification were almost always paraphyletic. It is an open problem how one might best simulate a real taxonomy. Conversely, we do not know how a realistic taxonomy would perform under the different sampling regimes. It should also be noted that we have assumed the availability of an accurate phylogeny. Knowledge of phylogeny is still rudimentary in many groups, and it will be important to explore the effects of such uncertainty (see Donoghue and Ackerly 1996).

Clearly, our conclusions are based only on the diversification models and parameters that we have explored. The large number of parameters precluded an exhaustive exploration of parameter space. We chose to address this problem by focusing on the parameters that S&K reported to have had a significant impact. Parameters we did not explore include

the speciation and extinction rates, the proportion of taxa to lineages in the base mixed classification, the taxon size scaling parameter *k*, and the mass extinctions inflicted on the system. Preliminary results suggest that varying the severity of the mass extinctions can have dramatic effects on the performances of the different classifications. This deserves further exploration.

We have used a simplistic model of the sampling effects of fossilization and the recovery of those fossils. This was initially justified by the fact that S&K reported no significant difference between a constant sampling rate and the more sophisticated model where the sampling rate for a particular time step was drawn from a distribution of potential sampling rates. The presence of occasional time steps with good sampling may result in a much better resolution of the underlying tree. Given the dramatic effects of sampling rate on the performances of the classifications, other models of paleontological sampling ought to be examined. Development of methods to estimate the proportion of lineages sampled by the fossil record and by paleontologists is especially crucial (e.g., Foote and Raup 1996; Foote et al. 1999).

Conclusions and Implications

Our analyses have an important bearing on discussions surrounding the use of paraphyletic groups in paleobiology and in general. The Sepkoski and Kendrick paper has been widely cited as vindicating the use of traditional classifications including paraphyletic groups, but this is inappropriate. Sepkoski and Kendrick interpreted their analyses in terms of monophyly and paraphyly. This was, after all, the issue that had been raised by Patterson and Smith and which they meant to examine in their simulations. We have shown that several factors that they did not consider explicitly, and did not control for, are primarily responsible for their results. In particular, we have identified taxon number, taxon size, and the distribution of taxon sizes as critical factors influencing the ability of a classification to accurately recover information on temporal diversity patterns. We have also highlighted how these factors interact with sampling of the fossil record. Thus, under good sampling the best results are obtained by classifications with many small taxa, and this is true whether the groups are monophyletic or paraphyletic, or even random assemblages of species. When sampling is poor, better results can sometimes be obtained with a range of taxon sizes including some large groups. Again, this is true whether those groups are monophyletic or paraphyletic. Ultimately, then, the explanation for the performance of a particular classification has little to do with monophyly vs. paraphyly.

What about assertions to the effect that paraphyletic groups are actually superior in reflecting diversity patterns? According to Roy (1996: p. 438), for example, "recent simulation studies have suggested that, given the nature of the fossil record, paraphyletic groups may, in fact, be better at capturing large scale macro-evolutionary patterns than monophyletic groups." As we have demonstrated, these claims are not supported by the work of Sepkoski and Kendrick, who Roy cites, or by our own work. Wagner (1995: p. 434) made the even bolder claim, based on his analysis of early gastropods, that paraphyletic groups are *necessary* to achieve an accurate picture of diversity: "[F]or supraspecific taxonomy to reflect diversity patterns accurately, it is not simply permissible for the taxa to be paraphyletic, but it probably is *requisite* [emphasis in original] for many of the taxa to be paraphyletic." Our analyses imply that such conclusions are suspect until the factors that we have identified here as being of overwhelming importance (the number of taxa, taxon sizes, etc.) are taken fully into consideration.

Where, then, does this leave us? We should not be content with arguments for the *adequacy* of one kind of classification or another. We would like to obtain the best understanding of diversity patterns we can, and it is clear that a general theory is needed regarding how classifications reflect temporal diversity patterns. Our analyses can be viewed as a step in this direction, but, as discussed in the previous section, much work remains to be done to understand fully the factors that determine the ability of a classification to recover diversity patterns. There may be, for example, ways and times in which monophyly and paraphyly will

make a difference, but this will need to be addressed in very carefully designed simulations. These need not follow the details of S&K's design. In the meantime, it is not safe to assume that classifications with paraphyletic groups provide an acceptably accurate picture of diversity patterns through time. Whether they do or not simply depends on a variety of other factors, some of which we have emphasized here.

Unfortunately, classification biases cannot be avoided by simply counting species to estimate diversity patterns. When sampling is poor, counting species results in one of the worst correlations with underlying lineage diversity. In fact, the decision to focus on any single rank unnecessarily restricts one's ability to infer underlying diversity patterns. The distribution of taxon sizes at any single rank is highly variable (ranging from one to thousands of species) and will rarely, if ever, match the optimal distribution of taxon sizes for estimating lineage diversity. A better estimate of lineage diversity may often be obtained by counting (in the same study) taxa assigned to different ranks, so as to best match the inferred quality of the paleontological sample. If the sampling were deemed to be good, we would recommend counting as many taxa, containing as few lineages each, as possible. In practice this might include a mixture of genera, subgenera, species, etc., depending on the status of taxonomic knowledge in the group and the confidence one has in assigning specimens to taxa at different levels. In the long run, we imagine the development of methods that depend not on taxonomic ranks but directly on knowledge of the underlying phylogenetic tree.

Acknowledgments

We dedicate this work in memory of Jack Sepkoski, who provided us with many insights into his own analyses and very helpful comments on the manuscript. We truly admire his enormous contributions to paleobiology. We also thank D. Kendrick, J. Cadle, and E. Macklin for their helpful comments and assistance, and M. Patzkowsky and J. Sepkoski for their constructive reviews. This research was supported in part by National Defense Science and Engineering Graduate fellowship DAAH04-95-1-0557 awarded to C. C. Maley.

Literature Cited

Donoghue, M. J., and D. D. Ackerly. 1996. Phylogenetic uncertainties and sensitivity analyses in comparative biology. Philosophical Transactions of the Royal Society of London B 351: 1241–1249.

Foote, M. 1996. Perspective: evolutionary patterns in the fossil record. Evolution 50:1–11.

Foote, M., and D. M. Raup. 1996. Fossil preservation and the stratigraphic ranges of taxa. Paleobiology 22:121–140.

Foote, M., C. M. Janis, and J. J. Sepkoski Jr. 1999. Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. Science 283:1310–1314.

Hennig, W. 1966. Phylogenetic systematics. University of Illinois Press, Urbana.

Labandeira, C. C., and J. J. Sepkoski Jr. 1993. Insect diversity in the fossil record. Science 261:310–315.

Lee, M. S. Y. 1997. Documenting present and past biodiversity: conservation biology meets paleontology. Trends in Ecology and Evolution 12:132–133.

Mayr, E. 1994. Ordering systems. Science 266:715–716.

Norell, M. 1992. Taxic origin and temporal diversity: the effect of phylogeny. Pp. 89–118 *in* M. J. Novacek and Q. D. Wheeler, eds. Extinction and phylogeny. Columbia University Press, New York.

Patterson, C., and A. B. Smith. 1987. Is periodicity of mass extinctions a taxonomic artifact? Nature 330:248–251.

———. 1988. Periodicity in extinction: the role of systematics. Ecology 70:802–811.

Patzkowsky, M. E. 1995. A hierarchical branching model of evolutionary radiations. Paleobiology 21:440–460.

Raup, D. M. 1972. Taxonomic diversity during the Phanerozoic. Science 177:1065–1071.

Raup, D. M., and J. J. Sepkoski Jr. 1984. Periodicity of extinctions in the geologic past. Proceedings of the National Academy of Sciences USA 81:801–805.

———. 1986. Periodic extinctions of families and genera. Science 231:833–836.

Raup, D. M., S. J. Gould, T. J. M. Schopf, and D. S. Simberloff. 1973. Stochastic models of phylogeny and the evolution of diversity. Journal of Geology 81:525–542.

Roy, K. 1996. The roles of mass extinction and biotic interaction in large-scale replacements: a reexamination using the fossil record of stromboidean gastropods. Paleobiology 22:436–452.

Roy, K., D. Jablonski, and J. W. Valentine. 1996. Higher taxa in biodiversity studies: patterns from eastern Pacific marine molluscs. Philosophical Transactions of the Royal Society of London B 351:1605–1613.

Sepkoski, J. J., Jr. 1978. A kinetic model of Phanerozoic taxonomic diversity. I. Analysis of marine orders. Paleobiology 4:223–251.

———. 1984. A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. Paleobiology 10:246–267.

Sepkoski, J. J., Jr., and D. C. Kendrick. 1993. Numerical experiments with model monophyletic and paraphyletic taxa. Paleobiology 19:168–184.

Smith, A. B., and C. Patterson. 1988. The influence of taxonomic method on the perception of patterns of evolution. Evolutionary Biology 23:127–216.

Valentine, J. W. 1969. Patterns of taxonomic and ecological structure of the shelf benthos during Phanerozoic time. Paleontology 12:684–709.

Wagner, P. 1995. Diversity patterns among early gastropods: contrasting taxonomic and phylogenetic descriptions. Paleobiology 21:410–439.

Appendix

Results under three models of diversification: logistic growth with mass extinctions, logistic growth with no mass extinctions, and exponential growth with no mass extinctions. For each model, 250 runs were performed with perfect knowledge and 50 runs at sampling rates of 0.05, 0.10, 0.15, 0.20, and 0.25. Standard deviations are given in parentheses. Abbreviations for classifications are as follows: lineages sampled (LS), mixed (M), hard cull (HC), soft cull (SC), soft cull II (SCII), full (F), random (R), matched random (MR), and matched-distribution random (MDR).

	Logistic growth with mass extinctions					
Classification	0.05	0.1	0.15	0.2	0.25	1
LS	0.079 (0.033)	0.242 (0.049)	0.387 (0.059)	0.514 (0.051)	0.618 (0.052)	1.000 (0.000)
M	0.326 (0.146)	0.551 (0.135)	0.645 (0.121)	0.715 (0.088)	0.741 (0.079)	0.814 (0.067)
HC	0.121 (0.115)	0.295 (0.165)	0.398 (0.164)	0.517 (0.148)	0.592 (0.131)	0.781 (0.087)
SC	0.184 (0.103)	0.497 (0.103)	0.645 (0.096)	0.754 (0.065)	0.802 (0.049)	0.907 (0.033)
SCII	0.240 (0.114)	0.563 (0.103)	0.703 (0.096)	0.794 (0.055)	0.826 (0.046)	0.890 (0.034)
F	0.220 (0.109)	0.548 (0.107)	0.705 (0.103)	0.795 (0.058)	0.833 (0.048)	0.901 (0.037)
R	0.091 (0.036)	0.272 (0.052)	0.422 (0.061)	0.548 (0.053)	0.654 (0.050)	0.996 (0.001)
MR	0.171 (0.061)	0.426 (0.063)	0.596 (0.056)	0.697 (0.049)	0.775 (0.043)	0.947 (0.011)
MDR	0.228 (0.117)	0.468 (0.126)	0.561 (0.158)	0.674 (0.094)	0.702 (0.122)	0.873 (0.068)
	Logistic growth with no mass extinctions					
Classification	0.05	0.1	0.15	0.2	0.25	1
LS	0.043 (0.032)	0.268 (0.064)	0.498 (0.063)	0.611 (0.055)	0.698 (0.045)	1.000 (0.000)
M	0.310 (0.251)	0.628 (0.176)	0.723 (0.179)	0.730 (0.135)	0.731 (0.138)	0.792 (0.109)
HC	0.197 (0.149)	0.122 (0.140)	0.110 (0.126)	0.163 (0.156)	0.179 (0.173)	0.456 (0.232)
SC	0.078 (0.064)	0.186 (0.147)	0.489 (0.180)	0.583 (0.153)	0.629 (0.134)	0.858 (0.053)
SCII	0.050 (0.058)	0.389 (0.169)	0.673 (0.141)	0.729 (0.095)	0.786 (0.055)	0.887 (0.046)
F	0.045 (0.058)	0.333 (0.165)	0.647 (0.138)	0.718 (0.088)	0.776 (0.056)	0.926 (0.033)
R	0.067 (0.044)	0.328 (0.064)	0.553 (0.067)	0.660 (0.053)	0.733 (0.043)	0.991 (0.003)
MR	0.309 (0.099)	0.644 (0.062)	0.780 (0.049)	0.822 (0.042)	0.867 (0.034)	0.939 (0.013)
MDR	0.263 (0.156)	0.476 (0.165)	0.670 (0.112)	0.721 (0.092)	0.771 (0.105)	0.904 (0.036)
	Exponential growth with no mass extinctions					
Classification	0.05	0.1	0.15	0.2	0.25	1
LS	0.758 (0.053)	0.764 (0.044)	0.784 (0.052)	0.801 (0.042)	0.807 (0.046)	1.000 (0.000)
M	0.975 (0.016)	0.970 (0.025)	0.971 (0.020)	0.972 (0.017)	0.969 (0.019)	0.858 (0.073)
HC	0.908 (0.043)	0.922 (0.031)	0.936 (0.027)	0.949 (0.028)	0.953 (0.022)	0.927 (0.039)
SC	0.950 (0.021)	0.965 (0.016)	0.973 (0.011)	0.980 (0.010)	0.981 (0.009)	0.898 (0.043)
SCII	0.977 (0.010)	0.980 (0.012)	0.980 (0.008)	0.980 (0.011)	0.976 (0.016)	0.836 (0.067)
F	0.968 (0.013)	0.975 (0.013)	0.977 (0.011)	0.982 (0.010)	0.981 (0.011)	0.807 (0.086)
R	0.848 (0.044)	0.854 (0.041)	0.872 (0.046)	0.885 (0.036)	0.890 (0.033)	0.980 (0.008)
MR	0.911 (0.037)	0.870 (0.055)	0.836 (0.062)	0.830 (0.057)	0.778 (0.081)	0.293 (0.175)
MDR	0.905 (0.052)	0.893 (0.052)	0.903 (0.047)	0.914 (0.039)	0.909 (0.043)	0.758 (0.126)