Paleobiology, 26(4), 2000, pp. 707–733

New methods for quantifying macroevolutionary patterns and processes

John Alroy

Abstract.—This paper documents a series of methodological innovations that are relevant to macroevolutionary studies. The new methods are applied to updated faunal and body mass data sets for North American fossil mammals, documenting several key trends across the late Cretaceous and Cenozoic. The methods are (1) A maximum likelihood formulation of appearance event ordination. The reformulated criterion involves generating a maximally likely hypothesized relative ordering of first and last appearances (i.e., an age range chart). The criterion takes faunal occurrences, stratigraphic relationships, and the sampling probability of individual genera and species into account. (2) A nonparametric temporal interpolation method called ''shrink-wrapping'' that makes it possible to employ the greatest possible number of tie points without violating monotonicity or allowing abrupt changes in slopes. The new calibration method is used in computing provisional definitions of boundaries among North American land mammal ages. (3) Additional methods for randomized subsampling of faunal lists, one weighting the number of lists that have been drawn by the sum of the square of the number of occurrences in each list, and one further modifying this approach to account for long-term changes in average local species richness. (4) Foote's new equations for instantaneous speciation and extinction rates. The equations are rederived and used to generate time series, confirm that logistic dynamics result from the diversity dependence of speciation but not extinction, and define the median duration of species (i.e., 2.6 m.y. for Eocene–Pleistocene mammals). (5) A method employing the *G* likelihood ratio statistic that is used to quantify the volatility of changes in the relative proportion of species falling in each of several major taxonomic groups. (6) Univariate measures of body mass distributions based on ordinary moment statistics (mean, standard deviation, skewness, kurtosis). These measures are favored over the method of cenogram analysis. Data are presented showing that even diverse individual fossil collections merely yield a noisy version of the same pattern seen in the overall continental data set. Peaks in speciation rates, extinction rates, proportional volatility, and shifts in body mass distributions occur at different times, suggesting that environmental perturbations do not have simple effects on the biota.

John Alroy. National Center for Ecological Analysis and Synthesis, University of California, 735 State Street, Santa Barbara, California 93101. E-mail: alroy@nceas.ucsb.edu

Accepted: 22 June 2000

Introduction

The study of large-scale patterns in the evolution of biodiversity has been the domain of paleontology for two centuries. Over the past three decades, however, this research program has expanded and strengthened dramatically with the introduction of quantitative methodologies. For example, well-established quantitative topics such as linear equilibrial models of taxonomic diversity dynamics (Sepkoski 1978), secular trends in turnover rates (Raup and Sepkoski 1982, 1984), nonlinear dynamics (Carr and Kitchell 1980), and scale-dependence of diversity patterns (Sepkoski 1988) all continue to attract attention.

However, basic issues concerning the preparation of the data used in these studies still remain. This paper, a companion to Alroy et al. 2000, builds on earlier ones (Alroy 1996, 1998d) in highlighting two major methodological themes. First, defining the temporal ranges of fossil taxa—a necessary prerequisite to any paleobiological analysis—is a serious, difficult problem of quantitative inference, and not just a matter of tradition and expert opinion. Despite steady progress in this area, and despite the growing use of maximum likelihood areas in cognate subdisciplines such as phylogenetic theory (Felsenstein 1981), this paper is the first to introduce a proper maximum likelihood method of inferring global taxonomic age ranges from faunal and stratigraphic data. Furthermore, it introduces a new, nonparametric method for calibrating age ranges to numerical time, one that is intended to satisfy workers who wish to see calibrations employ as much data as possible.

 0094-8373/00/2604-0010/$1.00

Second, although paleobiologists have long seen simple summations of age ranges as direct depictions of historically meaningful diversity patterns, these raw data are better seen as nonrandom statistical subsamples that might or might not carry a reliable signal (Raup 1976). Thus, in order to make them meaningful representations of historical trends, we must correct them somehow to render the sampling less idiosyncratic. Much work on this topic remains to be done. This paper discusses two new variations on a method of randomly drawing sets of fossil locality-based taxonomic lists within temporal intervals (Alroy 1996). The new algorithms are intended to deal with two problems: variation among localities in the number of individual fossils they represent, and variation through time in species richness at the locality level (i.e., alpha diversity).

Three other issues not directly related to the estimation of diversity per se also are addressed. First, computing taxonomic turnover rates turns out to be a difficult problem, with many diffferent indices having been proposed over the years (Foote 2000). Here I advocate the new equations of Foote (1999), explaining how these equations can be derived from common-sense assumptions. The new equations are favored not just because they avoid computational problems such as the existence of arbitrary upper bounds, but because they are the best possible estimates of instantaneous turnover rates—the true focus of interest in the study of diversity dynamics.

Second, patterns of replacement among major taxonomic groups are a long-standing topic of interest. To date, paleobiologists have mostly taken a somewhat typological approach to this problem, using quantiative methods to categorize groups into still larger categories such as ''evolutionary faunas'' (Flessa and Imbrie 1973; Sepkoski 1981). However, taxonomic replacement is really a matter of dynamics, not categorization, as recognized by Sepkoski (1978). Thus, on the one hand we might ask whether diversification patterns in sets of individual groups or evolutionary faunas can be modeled with dynamic equations (Sepkoski 1984; Miller and Sepkoski 1988). On the other, we might ask whether the overall tempo of taxonomic replacement—not just taxonomic turnover—is steady through time, or perhaps instead spurred by environmental or intrinsic perturbations. Remarkably, this simple question of quantification has not been addressed in the literature. Here I define two simple statistics that summarize the tempo of replacement in different time intervals, showing how these indices can be applied to arbitrarily large numbers of taxonomic groups.

Finally, perhaps the most notable development in paleobiology over the past decade has been the explosion of interest in quantitatively defined morphospaces (Raup 1966; Foote 1991). Morphospaces describing marine invertebrates have not always made use of paleoecologically significant measurements, but this is a common approach in the literature on fossil mammals (Van Valkenburgh 1985, 1988; Janis and Wilhem 1993; Hunter and Jernvall 1995; Jernvall et al. 1996). This paper grapples with a relatively minor debate in mammalian paleoecology concerning the study of body mass distributions. Although mostly confined to that body of literature, the discussion is important because the resulting statistical time series represents not just abstract morphometric statistics, but evolving community properties of real paleoecological interest. Furthermore, body mass distributions are one of just a few key variables studied by macroecologists (Brown 1995). Thus, more intense focus on variables like this one will foster synergy between macroecology and macroevolution.

Raw Data

The methods described here all are demonstrated using mammalian fossil data compiled from a set of 2828 publications. Most of the analyses hinge on the latest version of the North American Mammalian Paleofaunal Database (http://www.nceas.ucsb.edu/~alroy/nampfd.html). Currently, the data set includes 4978 faunal lists that total 30,951 taxonomic occurrences and include 1241 different genera and 3243 different species. Because 1089 of the 4484 taxa are singletons (i.e., are found only in one fossil collection) and 315 other taxa continue into the latest Pleistocene, 6475 first and last appearance events of genera

and species need to be arranged into a single best sequence. The ordination analysis of the next section is based on 217,673 demonstrated temporal overlaps of pairs of taxa (conjunctions [Alroy 1992]) and 289,141 additional first-appearance-before-last-appearance (F/L) statements (Alroy 1994).

The paper's next section deals with calibrating the event sequence to numerical time. This analysis involves a set of 186 geochronological age estimates for 434 fossil assemblages (9% of the total), many of which are tied to the same estimate because they are stratigraphically and geographically proximate (supplementary material, Table 1 available at http://www.psjournals.org). Two K-Ar estimates pertain to assemblages with no taxa determinate at the genus or species level; four Ur series estimates and one fission track estimate include only latest Pleistocene and Recent taxa; and two $^{40}Ar/^{39}Ar$ estimates, three paleomagnetic estimates, and one Ur series estimate are for small assemblages that are fully duplicated by other assemblages tied to different age estimates. These 13 dates are discarded because they provide no unique information on the numerical age of events prior to the very end of the sequence. Additionally, in contrast to earlier analyses the set of calibration points is restricted to the remaining 95 high-precision estimates (64 $^{40}Ar/^{39}Ar$, four Ur series, and 27 paleomagnetic). The additional, excluded estimates are based on low-precision methods (63 K-Ar, 11 fission-track, and four Sr isotope). The usable data points are augmented by a 0-Ma tie point at the end of the sequence. The use of high-precision age estimates makes a small but consistent difference in terms of improving the variance explained by the calibration.

The section on body mass distributions relies upon a set of 23,125 published lower first-molar measurements classed into 3398 population samples of 1969 species. The measurements are transformed into mass estimates using standard equations for each major mammalian order (Alroy 1998b). The use of such measurements in this context is common in the paleoecological literature because (1) nearly all mammals with teeth do have lower first molars (e.g., carnivoran carnassials), (2) population-level variability in lower first-molar measurements is small (Gingerich 1974), and (3) published regression equations are easily available and relatively robust (e.g., Damuth and MacFadden 1990). The current data set includes 51% more measurements and 28% more species than the one used by Alroy (1998b).

Maximum Likelihood Appearance Event Ordination

Basic Concepts.—The new method described here is an extension of appearance event ordination (AEO) (Alroy 1992, 1994, 1996, 1998a,c,d; Wing et al. 1995), an algorithm that infers age-ranges by quantitatively analyzing locality-specific faunal lists. AEO avoids the traditional system of North American land mammal ages, a series of qualitatively defined temporal bins of uneven duration that are loosely tied to first appearances of individual immigrant genera (Woodburne and Swisher 1995). First appearances of mammalian genera are extremely diachronous (Alroy 1998a), and independent geochronologic data show that correlations based on the traditional scale are three times less precise than those based on the AEO-derived age ranges (Alroy 1998c). The AEO method's basic steps have been reviewed previously (Alroy 1996, 1998d), and are summarized as follows:

1. Singleton taxa are excluded from the data set.

2. F/L statements are computed for all remaining pairs of taxa (species or genera). If two taxa i and j are found in the same faunal list, they are ''conjunct'': the statement ''Fi comes before [<] Lj'' is true and vice versa. If an occurrence of i is found below one of j in any stratigraphic section, Fi < Lj but the converse is not necessarily true. Li < Fj is tentatively assumed if no list includes both taxa and no section shows i occurring below j. Fi < Lj statements are assumed to be known with certainty, but Li < Fj statements are treated as hypotheses to be tested against candidate age ranges. Fi < Lj statements are generated automatically for all pairs of taxa for which either (a) $i = j$, because a taxon's first appearance must come before its own last appearance; or

(b) *j* is a living taxon (in this paper extinct latest Pleistocene taxa are treated as ''living'').

3. The square, pairwise F/L matrix is augmented by adding ''virtual'' conjunctions using the square graph algorithm (Alroy 1998d), which compensates for biogeographic effects that keep coeval taxa from ever being found in the same locality or section. The virtual conjunctions are used in the next step and then discarded.

4. As a starting point, a candidate linear sequence of F/L statements is computed by (a) using a variant of reciprocal averaging to derive scores for taxa from the F/L matrix, (b) using these scores to compute mean scores for faunal lists, (c) ordering the lists by their scores, and (d) computing first and last appearances by scanning across the sequence of lists. The event sequence is identical to an age range chart in which each taxon is represented by one F statement and one L statement occurring later on.

5. The initial appearance event sequence is optimized by a swapping algorithm. Earlier papers used a simple parsimony criterion to perform this optimization; a maximum likelihood approach to the problem is discussed below.

6. The appearance event sequence is numbered from oldest to youngest, and event positions are computed for the faunal lists. An event position is a minimal span of events going across the sequence that includes all of the taxa in a list; so if a list's position is 222–224, then all first appearances of the taxa occur by event 222 and all last appearances by event 224. In contrast to earlier studies, here the numbering is based on consecutive runs of like events (e.g., first appearances) instead of simple counts of events. For example, a stretch of seven events like F-F-F-L-F-L-L would count as just four runs. The new practice of counting event runs instead of events makes only a tiny difference to the calibration. However, by removing some small-scale distortions in the calibration the new numbering scheme decreases apparent variation among sampling bins in counts of lists and taxonomic occurrences.

7. Geochronologic age estimates are matched to the event positions using a new linear interpolation algorithm detailed in a later section. The algorithm seeks to find the largest set of ''hinge'' calibration points that implies a monotonic and reasonably steady relationship between time and the event sequence. In contrast, earlier studies used interpolation methods that employed small sets of statistically significant hinge points (Alroy 1996, 1998d).

8. The interpolation is used to estimate the age of each event in Ma, and these estimates in turn define numerical values for the age ranges of each taxon and the maximum/minimum ages of each list.

Justification.—The optimization algorithm has been improved by employing an explicitly formulated maximum likelihood criterion in deciding amongst alternative event sequences. Likelihood criteria are widely used in phylogenetics (Felsenstein 1981; Huelsenbeck and Crandall 1997; Wagner 1998) and ecology (Hilborn and Mangel 1997), but apparently have not been used in quantitative biochronology, even though there are some probabilistic biochronological methods that are related to graphic correlation (Agterberg and Gradstein 1999). Although maximum likelihood methods are often disputed on philosophical grounds, detailed expositions and defenses of the basic logic of likelihood already are available elsewhere (e.g., Hilborn and Mangel 1997).

The new algorithm is called maximum likelihood appearance event ordination (ML-AEO). The basic idea is to compute the probability of obtaining the observed F/L data given a candidate event sequence, a probabilistic model of sampling, and some set of nuisance parameters. Swaps of the event sequence and recalculations of the nuisance parameters are alternated until a stable solution (meaning a local optimum) is found. The maximum likelihood criterion of ML-AEO is justified by the following considerations:

1. The individual probabilities of observing each cell of the F/L matrix are what need to be explained by an event sequence. The overall log likelihood is just the natural log of the product of these cell-by-cell likelihoods, i.e., the sum of the logs of the likelihoods. As elsewhere in the scientific literature, likelihoods

are logged for computational reasons and for ease of representation; untransformed numbers may be so small as to cause memory-handling problems.

2. If two taxa have disjunct (nonoverlapping) age ranges in the event sequence (i.e., either Fi < Lj and Li < Fj, or Fj < Li and Lj < Fi), the likelihood is 1.0 because any reasonable sampling model should imply that when taxa are disjunct, they cannot be found together. The log of 1.0 being zero, these cases can be ignored in computing the overall log likelihood.

3. Cases of apparent overlap (hypothesized Fi < Lj and Fj < Li) are more complicated. The original AEO parsimony criterion (Alroy 1992, 1994) concerned itself only with minimizing cases where overlap is implied but the raw data do not demonstrate that an overlap must occur. However, demonstrated overlap is a probabilistic outcome that may or may not follow from overlap of age ranges in the real world. Therefore, likelihoods of matrix cell values have to be computed if the data either show that Fi < Lj and Fj < Li (matching the hypothesis) or fail to show this (implying age range disjunction).

4. The probability model should take three factors into account: (a) the amount of hypothesized overlap between pairs of taxa, measured in appearance events; (b) a ''nuisance'' parameter specifying the relative sampling probability of each taxon; and (c) the fact that demonstrating conjunction of a pair of taxa is much more likely if at some point one end of each taxon's age range overlaps with one end of the other taxon's age range, because each taxon then must appear in at least one faunal list that equates with such a dual range-ending event.

5. It is reasonable to think of the sampling process as involving a single, discrete sampling opportunity occurring at each appearance event. Of course, sampling opportunities correspond with collections of specimens, which constitute faunal lists; and not only may multiple fossil lists correlate with a single appearance event, but many lists will have broad event positions including multiple events. Therefore, the model assumption that one appearance event = one sampling event is a simplification.

Algorithm.—The likelihood expression can be formulated as follows: Assume that the probability of a taxon not being sampled at an event is κ (''crypsis''), and that of being sampled is $1 - \kappa$. The probability that two of N taxa, i and j, will both be found at an event is therefore $(1 - \kappa_i)(1 - \kappa_j) = 1 - \kappa_i - \kappa_j + \kappa_i\kappa_j$. The probability that this will not happen, so that no conjunction will be demonstrated, is $\kappa_i + \kappa_j - \kappa_i\kappa_j$. If the probability of never demonstrating a conjunction is $P_{\text{disj}(i,j)}$, then

$$P_{\text{disj}(i,j)} = (\kappa_i^d + \kappa_j^d - \kappa_i^d\kappa_j^d)^e, \tag{1}$$

where e = the number of events by which i and j overlap, i.e., the minimum of Li − Fi, Lj − Fj, Lj − Fi, and Li − Fj; and $d = 1$ at most events, but some other real number at dual range-ending events (i.e., events that equal both Fi and Lj, or Fj and Li, or Fi and Fj, or Li and Lj). Note that this additional nuisance parameter d is held constant across all such cases. If the overall log likelihood of the matrix is L($\mathbf{M}|\mathbf{E}, \mathbf{K}, d$), where $\mathbf{M}$ = the matrix, $\mathbf{E}$ = the event sequence, $\mathbf{K}$ = the vector of crypsis parameters, and d = the dual-event parameter, then

$$\text{L}(\mathbf{M}|\mathbf{E}, \mathbf{K}, d) = \sum_{i=1}^{N}\sum_{j=1}^{N} \log(\kappa_i{}^d + \kappa_j{}^d - \kappa_i{}^d\kappa_j{}^d)^e, \tag{2}$$

where the summation is computed only over cases in which Fi < Lj and Fj < Li and the two taxa are disjunct in the raw data set. An analogous summation is computed over cases where the taxa are conjunct and have overlapping hypothesized age ranges.

Before proceeding, it should be noted that the likelihood model ignores certain factors. (1) Relative time (counted in events) is employed instead of numerical time (counted in years) because the geochronological calibration is computed after the event sequence. It is conceivable that both things could be computed at once, but this would require some extremely intensive and complex computations. (2) As in earlier work on AEO, thicknesses of stratigraphic sections are ignored in all computations. No straightforward method of incorporating this information is apparent. (3) The model includes no information that relates directly to individual fossil localities (e.g., taphonomic regimes, sizes of fossil col-

lections, and biogeography). The effects of such attributes are not easily modeled because the likelihood model is expressed entirely as a relationship between a taxon-by-taxon hypothesis (the event sequence) and a taxon-by-taxon data set (the F/L matrix). Incorporating most of these locality-specific factors would require fundamentally reformulating the likelihood model, and it is not clear that such an approach would either be tractable or make much of a difference to the likelihood scores.

However, at least the biogeographic patterns could be dealt with by modeling the geographic range of each taxon with four parameters to encode the latitudinal and longitudinal range limits. Taxa with nonoverlapping geographic ranges would always be predicted to be disjunct. This method would require seven parameters per taxon instead of the three used in the current model (κ_i, Fi, and Li), so it might not significantly improve the results.

The general algorithmic problem is to find a combination of one event sequence, a vector of κ values, and one d value that maximizes the log likelihood. The appropriate algorithm therefore involves three alternating steps: (1) the event sequence is swapped using the current κ and d values; (2) the d parameter is recomputed using observed overlaps of age ranges and counts of conjunctions and disjunctions (but ignoring the **K** vector); and (3) the **K** vector is recomputed using the new event sequence and d value. On the first pass no κ and d values are available, so swapping is based on the parsimony criterion.

Swapping the event sequence involves considerable bookkeeping and is computationally intensive, but these inevitable programming details are of little interest. Obtaining nuisance parameters is a more general methodological problem. The computation of d involves a recursive equation. Deriving the equation requires first making the assumption that on average, most pairs of taxa have identical κ values equal to some overall mean ($\bar{\kappa}$). Furthermore, let the average number of events showing overlap between any pair of taxa ($\bar{e}$) equal E/O, where E = the sum of overlaps across all pairs and O is the number of overlapping pairs. For cases where $d = 1$ (i.e., anywhere but at dual range-ending events), this allows simplying equation (1) to yield

$$P_{\text{disj}(i,j)} = (2\bar{\kappa} - \bar{\kappa}^2)^{E/O}. \quad (3)$$

The summation of this average value over all of the O cases in which pairs of taxa overlap must equal the number of disjunct overlapping pairs (D), so $P_{\text{disj}(i,j)} = D/O$. It follows that

$$2\bar{\kappa} - \bar{\kappa}^2 = (D/O)^{O/E}. \quad (4)$$

An estimate of $\bar{\kappa}$ can be obtained recursively by taking advantage of two facts: (1) $\bar{\kappa} = \bar{\kappa}^1$, and (2) $1 = (2\bar{\kappa} - \bar{\kappa}^2)/(D/O)^{O/E}$. Therefore:

$$\bar{\kappa} = \bar{\kappa}^{([2\bar{\kappa} - \bar{\kappa}^2]/[D/O]^{O/E})}. \quad (5)$$

The exact value is obtained by first replacing $\bar{\kappa}$ on the right-hand side of the equation with $(D/O)^{O/E}$ and then iterating. Next, one recomputes D, O, and E for cases of dual range-ending events (i.e., where d is not equal to 1) and uses these numbers to recursively compute a second, independent estimate of the average κ value ($\bar{\kappa}'$). By definition, $\bar{\kappa}' = \bar{\kappa}^d$, so:

$$d = \ln \bar{\kappa}' / \ln \bar{\kappa}. \quad (6)$$

A different recursive computation is used to obtain the κ values. For each taxon i, one sums the observed number of conjunctions, c_i, then sums the probabilities of conjunction across all pairs that overlap with i in the hypothesized event sequence to obtain a predicted number of conjunctions, $\hat{c}_i$:

$$\hat{c}_i = \sum_{j=1}^{N} 1 - P_{\text{disj}(i,j)}$$

$$= \sum_{j=1}^{N} 1 - (\kappa_i{}^d + \kappa_j{}^d - \kappa_i{}^d \kappa_j{}^d)^e. \quad (7)$$

If this is a good estimate, then $\hat{c}_i = c_i$, so to make the estimate precise one can recursively compute:

$$\kappa_i = \kappa^{c_i/\hat{c}_i}. \quad (8)$$

Performance.—Interestingly, the ML-AEO analysis produces an event sequence that differs only in detail from sequences produced by parsimony swapping. The unswapped AEO sequence implies 580,205 conjunctions,

the sequence based on parsimony swapping implies 509,359 (12.2% fewer), and the ML sequence implies 514,088 (11.4% fewer); the overall log likelihood scores are 274,322.0 (unswapped), 256,202.6 (parsimony swapping: 6.6% less), and 248,250.5 (ML swapping: 9.5% less). In other words, both swapping algorithms produce sequences that are much better than the raw sequence, regardless of how fit is measured. Although the two yield sequences that are very similar, the ML-AEO sequence is far more impressive in terms of its log likelihood score (3.1% better) than is the parsimony sequence in terms of parsimony (0.9% fewer implied conjunctions).

The two units may seem to be incommensurate, but in fact a crude interconversion is possible. Assume for a moment that there are no differences between taxa in sampling probabilities, so actual conjunction and disjunction between temporally overlapping pairs is predicted randomly by the parsimony method. Then the prediction probabilities to be used in equation (2) are just the proportion of demonstrably conjunct or disjunct pairs across all overlapping pairs. If this proportion is roughly 50%, then each implied conjunction has a likelihood ''cost'' (see eq. 2) of about $-\log(0.5) = 0.69$. Thus, on this maximally simplistic model one would expect the parsimony sequence, which implies 509,359 conjunctions, to equate roughly with a likelihood score of 351,458.

If one instead improves the estimate by using the observed proportion of demonstrated conjunctions (43%) instead of 50%, one obtains $217{,}673[-\log(217{,}673/509{,}359)] + 291{,}686[-\log(291{,}686/509{,}359)] \cong 347{,}665$. Of course, the value reported above is 256,202.6—much lower because the prediction makes use of thousands of κ parameters. The important point is that the two kinds of scores are fundamentally interrelated even if they are not exactly interconvertible; one could use similar logic to make a rough estimate of the parsimony score from the likelihood score. Thus, the parsimony method is nothing more or less than a maximum likelihood method employing a very simplistic probability model.

One side benefit of ML-AEO is the κ or ''taxonomic crypsis'' values it generates for each taxon. These values, which indicate how likely it is that taxa with overlapping age ranges will fail to have demonstrated conjunctions, are remarkably intuitive. The worst κ value is 0.99653 for *Osbornoceros osborni,* an artiodactyl found at just two Miocene localities of substantially different ages: Gabaldon Badlands Level B (12 Ma) and Osbornoceros Quarry (10 Ma). Other species with very poor scores also tend to be found in two or three lists that are widely separated in time and span a well-sampled part of the timescale. Several taxa have scores of almost zero, indicating great abundance. An example is the common Late Cretaceous multituberculate *Meniscoessus,* which is found in 92 faunal lists—57% of all the lists that fall within its age range (this latter statistic is the list-wise sampling probability [Alroy 1998a]).

The κ scores are in general inversely related to the list-wise sampling probabilities, with the rank-order correlation being -0.737 for the 958 genera that occur in at least three lists. The rank-order correlation between the κ score and the number of lists including each genus (the abundance measure suggested by Walsh [1998]) is only -0.306, which makes sense because very long-ranging but rare genera may occur in many lists. For example, the Eocene insectivoran genus *Batodonoides* includes the smallest known mammal species (Bloch et al. 1998) and is very rare throughout its 11-m.y.-long range, so it has a high κ score of 0.97804 even though it occurs in 26 lists.

Calibration of the Event Sequence

Alroy (1996, 1998d) described a ''hinge'' interpolation method that fits a series of abutting lines to a plot of numeric time against numbered event runs for a set of faunal assemblages that have been tied to geochronological age estimates. This algorithm has much to recommend it; for example, it avoids assuming that transitions in the underlying ''faunal turnover clock'' follow some smooth, gradual function, and it generates only as many interpolation lines as are justified by a standard *F*-test. However, that second strength is also its greatest weakness: the original hinge interpolation method discards an enormous number of data points because it always errs on the side of accepting fewer rather than more. By doing this,

it implicitly assumes that the faunal turnover clock is constant until proven otherwise.

It might seem more intuitive to impose as many hinge points as one can, subject to the relationship must always be monotonic. In practice, however, doing so creates two major problems, both of which are attacked with the new ''shrink-wrap'' algorithm described here. First, it is difficult to find a large set of hinge points that both explains much of the variance and maintains monotonicity. The new algorithm attempts to do this by finding not one, but two monotonic lines that bracket the data, and then interpolating between them.

Second, solutions with many hinge points tend to posit many abrupt and dramatic shifts in the turnover clock, with slopes of neighboring line segments often differing by two orders of magnitude or more. These enormous shifts relate only to small analytical errors concerning faunas with very similar age estimates and event run numbers. Such errors may lead to fortuitously tied or nearly tied values for either variable. The new algorithm systematically rejects hinge points by examining the distribution of changes in the rate of the clock across the interpolation, discarding points that create unusually abrupt rate shifts. The exact algorithm is as follows:

1. Starting from the oldest data point to the lower left side of the event sequence (Fig. 1), a lower boundary line is drawn to connect a series of younger points in such a way that all other data points are younger than (above) the line. At each step, monotonicity is imposed by searching for the next data point to the right in the event sequence, making sure this data point actually is paired to a younger age estimate. The algorithm also pays attention to the stratigraphic relationship between the dated and fossiliferous horizons. If a dated horizon underlies a faunal locality, then it is a maximum estimate, so it is ignored in ''shrinking up'' the lower boundary line. Ignoring these points causes them to fall below the line, which is acceptable because maxima always may be older than the actual age of the fossil assemblage.

2. Starting from the youngest point, a monotonic upper boundary line is drawn so that all points either are on the line or are older. Again, all that is required is to add points in the order of the event sequence (now being read right to left). Minimum estimates (as determined by stratigraphy, see preceding step) are ignored.

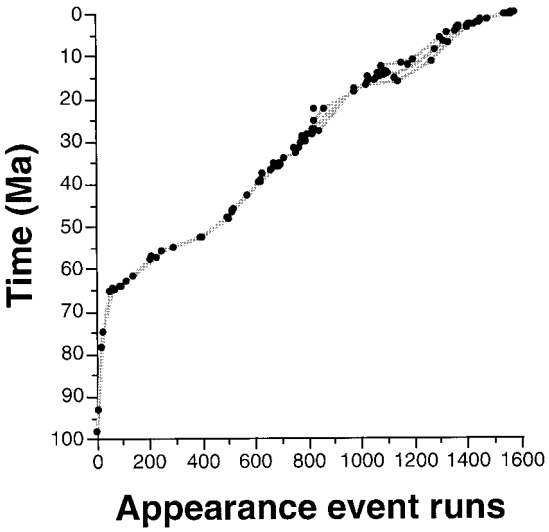

FIGURE 1. Shrink-wrap calibration of the appearance event ordination. Data points are based on concurrent range zones of faunal assemblages tied to geochronological age estimates (see supplementary material, Table 1). Gray lines show upper and lower boundaries produced by shrink-wrapping, as well as the midpoint line between them that is used in the calibration. A few points that would introduce abrupt shifts in the slopes of the boundary lines are left outside of the polygon they define.

3. The slope of each line segment (i.e., line connecting consecutive data points) in each boundary line is computed.

4. Logs of ratios of slopes of adjacent line segments (i.e., rate changes) are found.

5. For each line, differences of adjacent rate changes (not rates) are found. The resulting values describe changes across three consecutive segments, say, segments i, $i + 1$, and $i + 2$. High values often mean that there is a large change in slope followed by another large change in the opposite direction; i and $i + 2$ may be shallow while $i + 1$ is steep.

6. The mean and standard deviation of this double-difference distribution are computed.

7. The most extreme value in the double-difference distribution is found. If it is more than 1.96 standard deviations away from the mean, the original data point creating the larger of the two slope differences (not double differ-

TABLE 1. Age estimates for Cenozoic North American land mammal ages (NALMAs). Locations of boundaries between NALMAs are set using the break algorithm (Alroy 1992), which iteratively splits up the appearance event sequence in a way that separates as many pairs of age ranges as possible with the addition of each boundary. Two-standard-deviation analytical errors in age estimates are about 1.35 m.y. (see text). Cretaceous NALMAs are omitted. Age estimate for base of Puercan is set at the date of the Cretaceous/Tertiary boundary given by Berggren et al. (1995). Geringian, Monroecreekian, and Harrisonian are lumped by many authors as the ''Arikareean,'' but unlike any other NALMA this unit would encompass 10.2 m.y., an epoch boundary (Oligocene/Miocene), and an extraordinary amount of faunal turnover. Base (Ma) = age estimate for base of NALMA in Ma; Event run = identity of run beginning the NALMA (which is always a run of FAEs); Rank = order in which boundary was selected by the break algorithm (lower-ranked boundaries have weaker support); Reference locality = name-bearing faunal assemblage of biochron, i.e., assemblage whose placement defines the location of the land mammal age in the event sequence.

NALMA	Base (Ma)	Event run	Rank	Reference locality
Irvingtonian	1.8	1463	10	Irvington
Blancan	4.9	1357	7	Red Quarry
Hemphillian	10.3	1253	3	Coffee Ranch
Clarendonian	13.6	1159	12	MacAdams Quarry
Barstovian	16.3	1049	28	Hemicyon Quarry
Hemingfordian	20.6	937	21	Thomson Quarry
Harrisonian	24.8	861	40	Pine Ridge Escarpment
Monroecreekian	26.3	839	22	Monroe Creek (SDSM V-6229)
Geringian	30.8	777	1	Durnal Ranch Quarry
Whitneyan	33.3	739	27	Indian Stronghold (Protoceras Channel)
Orellan	33.9	717	39	Sage Creek Basin (West)
Chadronian	38.0	651	26	Chadronia Pocket
Duchesnean	42.0	591	32	Titanothere Quarry
Uintan	46.2	525	11	Myton Pocket
Bridgerian	50.3	453	2	Grizzly Buttes
Wasatchian	55.4	275	19	Reservoir Creek Bonanza
Clarkforkian	56.8	235	4	Holly's Microsite
Tiffanian	60.2	171	16	Mason Pocket
Torrejonian	63.3	115	8	West Flank Torreon Wash (Pantolambda Zone)
Puercan	(65.0)	61	15	Mammalon Hill

ences) is discarded. If not, the pruning algorithm (steps 3–7) halts.

8. A midpoint line between the two boundaries is computed by (a) finding the set of data points that are included in one or both boundary lines, (b) computing the value predicted by each boundary line at each of these points, and (c) averaging the two values to create a hinge point.

Because the algorithm is strictly heuristic, a shrink-wrap interpolation line is not guaranteed to explain more of the variance, or include more data points, than any other monotonic solution seeking to include a large number of data points. The method seems to perform well with the data set at hand, however, explaining 99.856% of the variance in the age estimates. The variance explained increases to 99.912% if one excludes estimates that are maxima falling below the interpolation line, or minima falling above it (see steps 1 and 2 above). This performance would be hard to match. Nonetheless, still more sophisticated approaches should be explored in the future.

The large number of interpolation segments produced by this algorithm makes it impossible to compute error terms for each and every segment. However, a rough idea of the error in the entire analysis can be determined by noting that the sum of squares of the age estimates is 54,819.3; if 99.912% of this variance is explained, then the expected departure of any one data point from the interpolation line is $[54{,}819.3/107 \times (1 - 0.99912)]^{0.5} = 0.674$ m.y. This value represents only one standard deviation, so a general error estimate of 1.35 m.y. is preferable. In contrast, Alroy (1998d) employed a calibration that explained 99.70% of the variance. A similar calculation implies a two-standard-deviation error of 2.17 m.y. The difference is partially due to improvements in the data set and ordination algorithm, but mostly attributable to the use of higher-precision radioisotopic dates. Including low-precision dates would increase the error from 1.35 to 2.04 m.y.

To facilitate evaluation of the hypothesized appearance event sequence by other workers,

Table 1 gives estimated boundaries among the conventionally recognized North American land mammal ages (NALMAs) (Woodburne and Swisher 1995). The boundaries were determined by using the break algorithm (Alroy 1992) to find 50 points that split the event sequence in a way that summarizes as many observed taxonomic disjunctions as possible. Matches between the breaks and conventional NALMAs then were determined by examining the positions of classic faunal assemblages (i.e., reference localities) in the event sequence. Further details of a revised North American land mammal timescale will be given in a later paper.

Randomized Subsampling of Faunal Lists

In this section I discuss three distinct methods for correcting diversity curves by basing the age ranges used to construct them on randomly drawn faunal lists (Alroy 1996). Previously just one algorithm has been employed for this purpose; the two new, more complex algorithms introduced here involve more realistic assumptions about the nature of species richness versus sampling relationships within individual faunal assemblages. A comparison of the diversity curves generated by these three methods shows consistent but small differences, with the third and most realistic method generating almost exactly the same result as the original algorithm. Therefore, I conclude that although sampling effects are important, the exact choice of a subsampling method is not a major problem for the North American mammal data set.

Basic Procedure.—The temporally calibrated sequence of first and last appearance events produced by the ML-AEO method could be used directly to infer a diversity curve and turnover rate data, because diversity is just the sum of overlapping age ranges at any one point in time. However, sampling intensity is known to vary by an order of magnitude through the Cenozoic (Alroy 1996, 1998d), so the total diversity curve presents a mixed signal of sampling artifacts and real trends. Therefore, this study follows earlier ones (Alroy 1996, 1998d, 1999a,b) by using randomized subsampling of entire faunal lists to generate sampling-standardized diversity data. The procedure, termed here by-list occurrences-weighted subsampling, is as follows:

1. The 191 lists from eastern North America (less than 4% of the total) are discarded to minimize the biogeographic spread of sampling through time.

2. The event sequence is broken up into uniformly spaced, 1.0-m.y.-long sampling bins (the bin length is conservative relative to the best possible trade-off of precision and accuracy in correlation [Alroy 1996]).

3. Faunal lists are assigned to bins, and within each bin entire faunal lists are drawn at random until a uniform quota is met. The quota can be expressed as a count of lists, but instead it is set to a count of taxonomic occurrences (called ''records'' by Alroy [1996, 1998d]) across all lists. So if three lists respectively include 5, 8, and 12 distinct taxa, the total is 25 occurrences. Distinct taxa in each list include (a) all identifiable species, plus (b) all genera that include no determinate species. In an important departure from earlier studies, here 2090 additional lists (42%) are excluded from the analysis because they do not fit in a single interval. This is done because of the combinatory difficulty of guaranteeing that all intervals will closely approach the correct quota when many individual lists (typically very short ones) fit into multiple intervals. Another approach, which has not yet been implemented, would be to randomly assign each of these lists to one of the several bins that could include it during each iteration of the subsampling algorithm.

4. After obtaining the full quota once for each bin, the age range of each taxon is computed across all the bins. So if a species was found in bins 10, 13, 15, 16, and 17 in the raw data but only occurs in lists that are sampled in bins 13 and 15 in a subsampling trial, then its range is considered to span only bins 13, 14, and 15 for the purposes of that trial.

5. Counts of species that cross the boundary between each neighboring pair of bins are computed; the series of counts defines a diversity curve. To prevent cases where genera are implied to be polyphyletic, diversity counts are incremented by one for each genus that occurs before and after a boundary even though no named species in that genus crosses

the boundary (hence the counts are actually of ''species lineages,'' and the pseudoextinction of the last species before such a gap and pseudo-origination of the first species after such a gap are discounted in the following step).

6. Counts of originations and extinctions within each bin also are totaled, but taxa occurring only in one bin (''singletons'') are discarded because these taxa create artifactual patterns in the turnover rate data (Alroy 1998d). The species in the previous example would be considered (a) present at the boundaries between bins 13 and 14, and 14 and 15; (b) to appear in bin 13; and (c) to disappear in bin 15.

7. Steps 3 through 6 are iterated (here 200 times), and the resulting boundary-crossing diversity counts and within-bin turnover rates are averaged.

A somewhat similar procedure was employed by Miller and Foote (1996) and Markwick (1998). However, these authors used a variant of the classical method of rarefaction (Sanders 1968; Raup 1975; Tipper 1979) in which individual taxonomic occurrences are drawn independently regardless of the composition of faunal lists. Additionally, they computed total levels of sampled diversity within bins, instead of ranging the data through and averaging counts of taxa that crossed boundaries between bins. These distinctions are important because (a) unlike drawing faunal lists, independently drawing taxonomic occurrences does not mimic any real-world sampling process (although the two methods should converge on the same result with large sample sizes [Smith et al. 1985]); (b) failing to range through the data yields diversity curves that are not comparable to the traditional, ranged-through diversity curves employed elsewhere in the paleobiological literature; (c) counting all taxa that range into a bin can lead to difficulties with timescale effects that relate to variation in the scale of time-averaging (Foote 1994)—in contrast, all of the species that cross a boundary must have been coeval at that time, so there is effectively no time-averaging; and (d) failing to compute age ranges for species across individual trials makes it impossible to identify and remove singletons, which can create additional statistical artifacts that even subsampling cannot remove (Alroy 1998d).

Randomized subsampling of entire lists was first employed by Shinozaki (1963), and some of the method's statistical properties were explored by Smith et al. (1985). However, the algorithm discussed by these authors differs crucially from the one introduced by Alroy (1996) in the way that the lists are weighted. Shinozaki's method simpy tallies the number of lists drawn, whereas occurrences-weighted subsampling targets a quota of taxonomic occurrences, not lists, and therefore keeps track of the number of occurrences encountered as lists are drawn.

Assumptions.—Unfortunately, each of the preceding subsampling methods makes unrealistic assumptions about the nature of sampling and species richness within collection localities. Shinozaki's unweighted by-list method assumes that lists are taphonomically comparable in different time intervals, so any systematic variation in the average species richness of lists reflects real biological patterns. The by-list occurrences-weighted method (Alroy 1996, 1998d) assumes instead that trends through time in average species richness are partially artifactual, and the artifacts can be ameliorated if one weights lists by occurrence counts.

Effectively, this assumption makes sense only if there is a roughly linear relationship between the number of specimens sampled and number of species found at a locality—i.e., if there is a linear collection curve. However, subsampling methods like rarefaction were justified in the first place in ecology by the observation that collection curves are nonlinear, and specifically asymptotic (Sanders 1968).

There are two cases in which the linearity assumption might make sense. First, most collections might represent comparable numbers of specimens. If so, then the lists all would fall in the same short segment of the collection curve, and short segments of such curves are indeed reasonably linear. However, if individual fossil collections really didn't vary much in size, then it would make even more sense to use Shinozaki's method, tallying the lists themselves and not the taxonomic occurrenc-

es. Furthermore, mean species richness of fossil localities in the raw data does vary greatly through time (Alroy 1998d), suggesting that some intervals are represented on average by much larger modal collections of fossils, and therefore that there must be considerable variation even within temporal bins.

A second possibility is that occurrence counts represent a composite signal of variation among fossil collections in specimen counts and true species richnesses. Although that alone would not justify assuming a linear relationship between occurrence and specimen counts, as discussed below it seems to be a fair compromise between more extreme assumptions.

The implicit assumptions of the classical rarefaction method (Miller and Foote 1996) with respect to alpha diversity and collection size are not obvious in this context. However, they may prove similar to the assumptions made by occurrences-weighted subsampling of lists.

Occurrences-Squared Weighted Method.—One way to improve the realism of subsampling methods is to assume a more realistic relationship between sample size and richness. Many collecting curves are approximately linear in a log-log space (May 1975). However, even if this is true the slope and intercept of the relationship may vary substantially among localities. The occurrences-tallied method implicitly assumes that the slope is one and the intercept is zero, but this is the maximum possible value for the slope, and the mathematical minimum is some small number just greater than zero—so real collecting curves most likely do not look anything like this.

Fortunately, it seems that most mammalian fossil assemblages can be characterized by a slope of about 0.5. This claim needs to be documented in detail, but the data on hand seem to support it. For example, a slope of just about 0.5 is seen for four very different samples of widely different ages (Fig. 2). Big Multi Quarry (late Paleocene of Wyoming [Wilf et al. 1998]) and Swift Current Creek (late Eocene of Saskatchewan [Storer 1984]) both have reasonable sampling across the body mass spectrum and large numbers of specimens (1665 and 997), but they differ in the proportion of species (2/37 and 9/41) that are each represented only by one specimen (i.e., unique species [Colwell and Coddington 1994]). Additionally, Big Multi Quarry displays the unimodal body mass distribution characteristic of early Tertiary assemblages, whereas Swift Current Creek has a markedly bimodal distribution. The remarkably small number of unique species at Big Multi suggests that very few species remain to be collected (Colwell and Coddington 1994). A rarefaction analysis demonstrates an asympotic relationship between sampling and richness in both cases, but flattening of the curve is more pronounced for Big Multi Quarry (Fig. 2A).

The two younger faunas are quite different (SDSM V-6229, late Oligocene of South Dakota [Macdonald 1972]; Achilles Quarry, middle Miocene of Nebraska [Voorhies 1990]). Like most middle and late Cenozoic assemblages, both of them suffer from profound size bias: respectively, just 30 of 712 and 18 of 816 of their generically identifiable specimens represent ungulates and carnivorans. Both faunas also include moderate numbers of unique species (3/25 and 6/28). The rarefaction curves, particularly for Achilles Quarry (Fig. 2D), are closer to linear in a log-log space for these faunas than for the others.

A slope of 0.5 describes all of these relationships fairly well despite the great differences among the assemblages in geological age, body mass distribution, and completeness of sampling. Reduced major-axis regression yields slopes of 0.534, 0.590, 0.525, and 0.512 for the four quarries, with r^2 values of 0.965, 0.977, 0.980, and 0.998. Similar observations were made by Gunnell (1998) in a study of mid-Eocene faunas.

Obviously, the estimated slope for asymptotic curves will be too low close to the origin and too high close to the asymptote. For example, a quadratic fit for the Big Multi data implies that the slope begins at around 1.02, but falls to 0.68 when there are 100 specimens and to 0.51 when there are 1000. Hence, if this kind of a relationship is typical for the early Paleogene, then assuming a slope of 0.5–0.7 makes sense for the kind of sample sizes typically encountered in ''good'' to ''excellent''

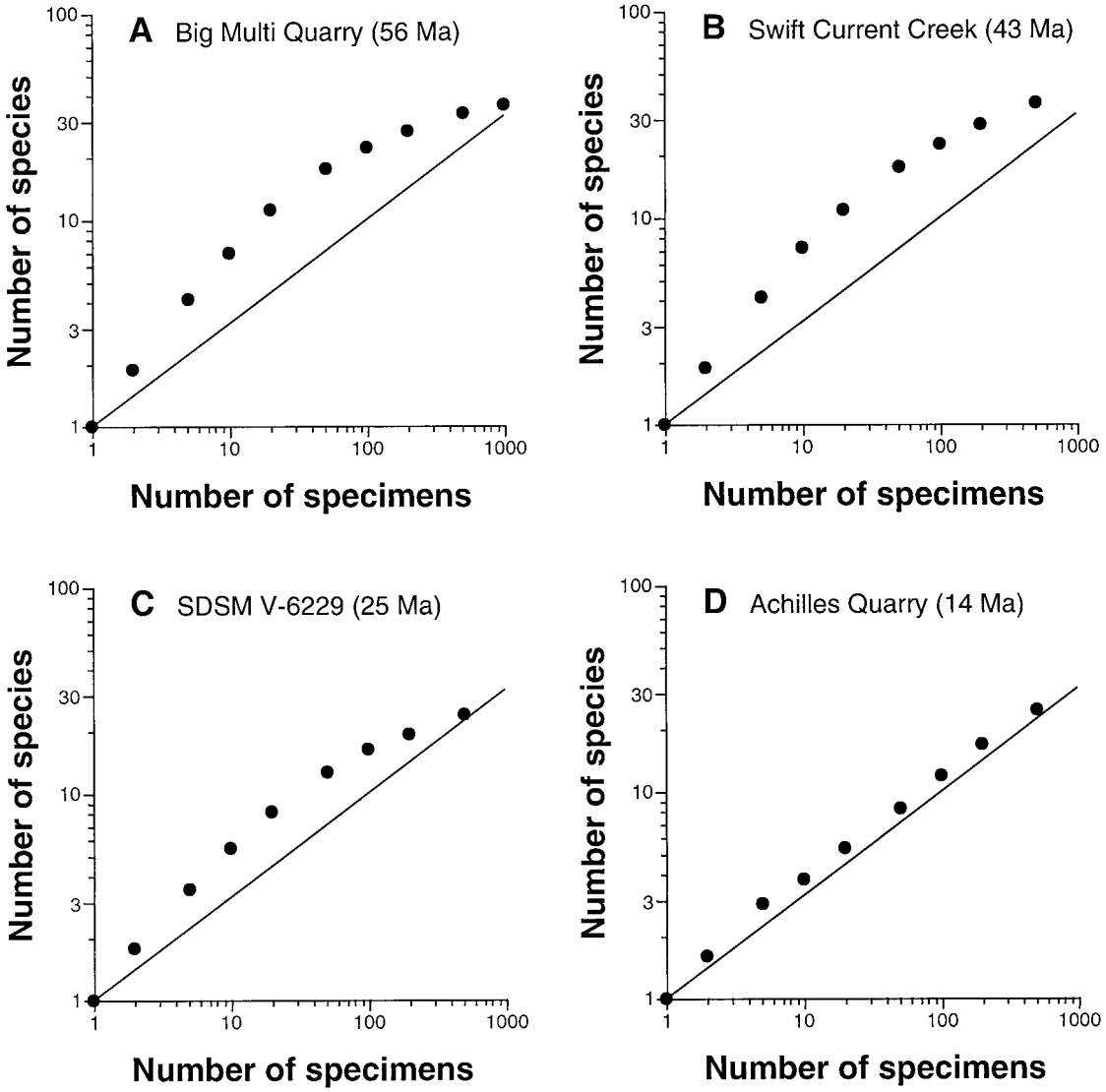

FIGURE 2. Rarefaction curves for selected mammalian faunas. Axes are log-transformed; points fall at 1, 2, 5, 10, 20, 50, 100, 200, 500, and 1000 specimens. Lines are not regression lines but instead have a slope of 0.5 and an intercept of one specimen and one species, showing the relationship assumed by the occurrences-squared subsampling method. A, Big Multi Quarry (Clarkforkian or late Paleocene, 56 Ma). B, Swift Current Creek (Uintan or late Eocene, 43 Ma). C, SDSM V-6229 (Monroecreekian or late Oligocene, 25 Ma). D, Achilles Quarry (Barstovian or middle Miocene, 14 Ma).

assemblages. Analyses of additional assemblages to be detailed elsewhere show similar patterns.

The important point is not whether a linear regression is appropriate, because the relationships are clearly not exactly linear; instead, the point of the exercise is to show that a simple linear approximation is operationally useful. In fact, using a ''square root rule'' to estimate the richness/sampling relationship for fossil mammals does work well. Given the linear log-log scaling, the expected number of species in a sample of N specimens can be approximated as the square root of N. For the four quarries, this rule predicts 41, 32, 27, and 29 species when there are 37, 41, 25, and 28. Only the low estimate for Swift Current Creek is off by more than 10%. Conversely, the num-

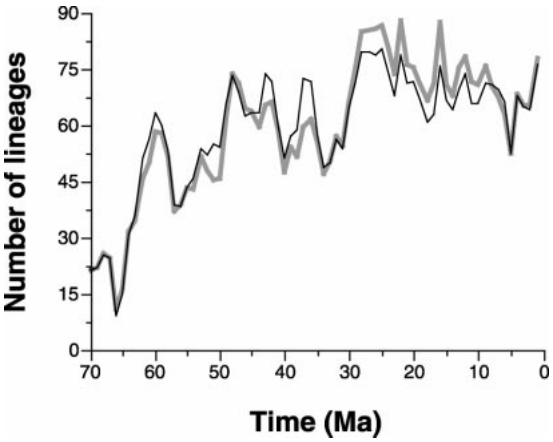

FIGURE 3. Comparison of sampling-standardized diversity curves based on occurrences weighted and occurrences-squared weighted sampling. Methods assume no change through time in alpha diversity. Thin black line: curve based on randomly subsampling lists that total 100 taxonomic occurrences per 1.0-m.y. temporal interval. Thick gray line: curve based on randomly subsampling lists that total 1800 occurrences-squared per interval.

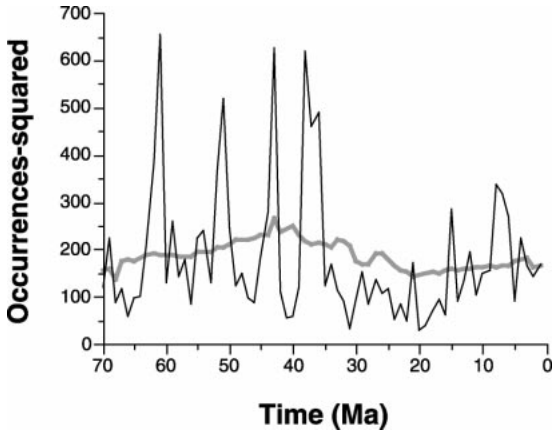

FIGURE 4. Mean number of occurrences-squared per faunal list. Thin black line: means for individual 1.0-m.y.-long bins. Thick gray line: smoothed means based on a 25-m.y.-long moving window.

ber of specimens for the vast majority of assemblages that have no known specimen counts can be estimated as the number of taxonomic occurrences squared.

The occurrences-squared rule of thumb implies that subsampling of lists should employ quotas based on sums of occurrences-squared instead of sums of occurrences. Fig. 3 contrasts an occurrences-tallied analysis using a quota of 100 occurrences per 1.0-m.y.-long temporal bin with an analysis using a quota of 1800 occurrences-squared per bin. Both quotas are set at the lowest point that is practical; any lower quota would cause a large number of bins to fall far short. The first analysis employs the same occurrences-per-bin quota as in Alroy 1998d; with a smaller data set Alroy (1996) employed a quota of 85 occurrences per bin. Both of the earlier analyses yielded curves that closely resemble this new occurrences-tallied curve.

The occurrences-squared weighted curve is substantially different (Fig. 3). The occurrences-weighted curve often has higher or lower peaks, but during intervals of rapid change the curves overlap closely. Because of this occasional close tracking, the differences seem to be fundamental and not merely attributable to difficulties in equating sampling quotas that have different units. In particular, one pattern seems to be important: the occurrences-squared weighted curve is consistently low during the Paleocene and Eocene and high afterwards.

Smoothing Method.—The lack of variation in the occurrences-squared weighted curve may reflect better responsiveness of the method to artifactual variation through time in the per-bin ratio of occurrences to lists. However, the low Paleocene–Eocene values seem to reflect failure to account for bona fide variation through time in alpha diversity. This inference is suggested by long-term trends in the occurrences/lists ratio (Fig. 4): despite considerable short-term variation, it does seem that Paleocene and Eocene lists are consistently longer than Oligocene and Neogene lists. Work in progress on the relationship between observed richness, total counts of specimens, and body mass distributions supports earlier evidence (Stucky 1990) that alpha diversity falls through the Cenozoic.

The simplest way to deal with this effect would be to ignore counts of occurrences or occurrences-squared per list and peg subsampling quotas to counts of lists—in other words, to fall back on the assumption that all variation in apparent richness is real. A more reasonable approach is to assume that short-term variation is in fact controlled by local sampling artifacts (e.g., strong size bias or small average counts of specimens in lists in a

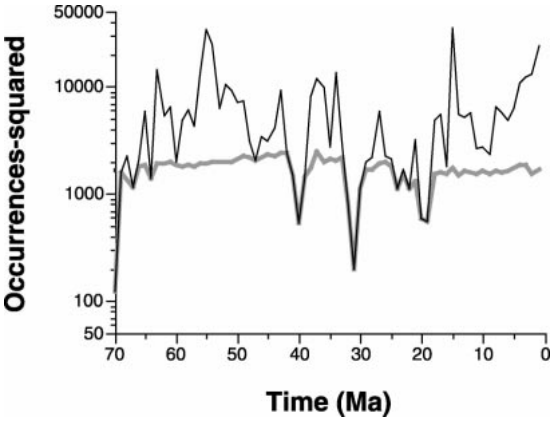

FIGURE 5. Occurrences-squared totals for 1.0-m.y. bins. Thin black line: total occurrences-squared counts across all faunal lists. Thick gray line: average number of occurrences-squared drawn in each of 200 randomized subsampling trials.

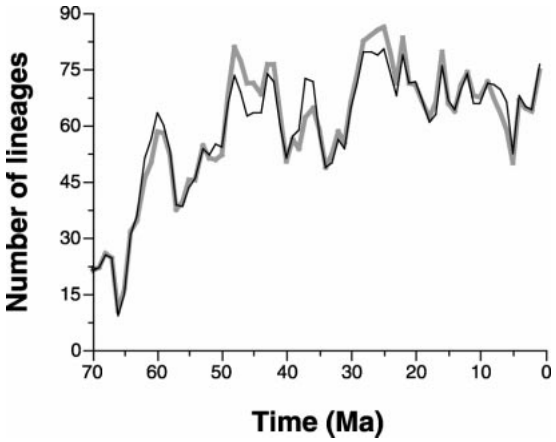

FIGURE 6. Comparison of sampling-standardized diversity curves based on occurrences weighted and smoothed occurrences-squared weighted sampling. Thin black line: curve based on 100 taxonomic occurrences per interval. Thick gray line: curve based on randomly subsampling lists that total occurrences-squared quotas shown in Figure 5. First method assumes no change through time in alpha diversity, second method assumes a gradual, long-term secular trend in alpha diversity like the one shown in Figure 4.

particular interval), but that long-term variation reflects biological trends.

A straightforward algorithm is as follows. First, one computes a running average of the ratio of occurrences-squared tallies to list tallies across a moving window of 25 bins:

$$\bar{o}_t^2 = \left(\sum_{i=t-12}^{t+12} \sum_{j=1}^{L_i} o_j^2 \right) \Big/ \sum_{i=t-12}^{t+12} L_i, \qquad (9)$$

where L_i = the number of lists in the ith interval, o_j = the number of occurrences in the jth list falling within the ith interval, and $\bar{o}_t^2$ = the mean occurrences-squared count for some interval t. The 25-bin width of the moving window is used to guarantee that the smoothed curve will reflect genuine trends on the scale of geological epochs. When the window falls partially outside of the range of the data, the summation is carried out over fewer intervals and the values in both the numerator and the denominator fall proportionately. The next step is to assume a list-based sampling quota (in this case nine lists per interval). Finally, for each interval the list quota is multiplied by $\bar{o}_t^2$ to obtain a tailor-made quota for each sampling bin that is expressed in units of occurrences-squared.

The resulting curve of sampled occurrences-squared (Fig. 5) is generally high in the Paleocene–Eocene and low thereafter. More occurrences-squared are sampled in the early Tertiary because the method assumes that the same number of specimens will yield more occurrences, or in other words that the slope of richness/sampling curves like those in Figure 2 should be slightly steeper in the early Tertiary (as suggested by Figure 4).

Because the artifactual Paleocene–Eocene low is now removed, the diversity curve produced by this smoothing algorithm is even more similar to the one originally obtained just by the simple occurrences-weighted method (Fig. 6). The two curves often overlap completely, and for the entire Cenozoic the curves have comparable geometric means (62.5 smoothed, 61.9 occurrences-weighted), standard deviations (14.0, 12.7), and most importantly serial correlations (+0.840, +0.808). By contrast, the original occurrences-squared weighted curve has a stronger serial correlation of +0.858 and also shows a stronger temporal trend (time vs. diversity: $r = -0.680$, vs. -0.545 [smoothed, occurrences-squared weighted] and -0.604 [occurrences-weighted]). As a result, the points in this curve fall into clusters that are strongly separated along the time axis.

In summary, there are substantial concerns about underlying assumptions in the original by-lists occurrences-weighted subsampling method—but these concerns turn out to be

misplaced. Almost exactly the same results are obtained using a much more sophisticated algorithm that is based on more concrete and well-justified assumptions, that varies sampling using a direct proxy for specimen counts, and that makes allowances for long-term secular trends in alpha diversity. However, this fortunate outcome may not hold for all paleontological data sets, so the more detailed smoothing algorithms are recommended for general use. In addition, varying the exponent used to translate counts of occurrences in lists into weights is recommended. For example, if richness scales roughly as the cube (not square) root of specimen counts in a particular data set, then occurrence counts should be raised to the third (not second) power to generate weights.

Speciation and Extinction Rates

The ordination and faunal subsampling methods provide three raw variables that describe the basic pattern of diversification: species richness values for boundaries between sampling bins (D), and counts of originations (O) and extinctions (E) of boundary-crossing taxa within each bin (which means that singletons are excluded). However, the raw origination and extinction data need to be transformed into rates before they can be employed in time series analyses. Raup (1985) and Foote (1994, 2000) surveyed some of the many equations that are normally used for this purpose in paleobiology. Alroy (1996, 1998d) advocated using per-taxon turnover rates in the form O/D and E/D. Although some authors also employed boundary-crosser counts in their denominators (e.g., Carr and Kitchell 1980), and others employed boundary crossers in their numerators (e.g., Harper 1996), none previously had combined the two.

However, Foote (1999, 2000) has shown that the use of raw per-taxon rates presents certain difficulties even if the rates only pertain to boundary crossers. Here I outline a slightly different argument intended to justify the use of the new turnover rate equations presented by Foote (1999). To begin with, I note that much of the modeling literature on diversity dynamics is premised on continuous-time exponential decay and growth equations (Raup 1985). Alternative models such as those assuming discrete ''generations'' (Gilinsky and Good 1991) have not received wide support. Raup's exponential equations take the form

$$D_t = D_0 e^{(\lambda - \mu)t}, \tag{10}$$

where t = the number of time intervals, λ = the per-taxon instantaneous rate of origination, and μ = the per-taxon instantaneous rate of extinction; also, let the intrinsic rate of increase $r = \lambda - \mu$.

If turnover is truly exponential then direct ratios of counts and diversity levels will be inaccurate estimates of these instantaneous rates. For example, suppose that $D_0 = 100$, $t = 1$, $O = 20$, and $E = 10$. If one assumes that $\lambda = O/D = 0.2$ and $\mu = E/D = 0.1$, then by equation (10) one actually expects O to be 20.03 (not 20) and E to be 9.52 (not 10) after one time interval. The discrepancies could be extreme: if $O = 100$ and $E = 50$, then using this to infer that $\lambda = 1.0$ and $\mu = 0.5$ leads us to back-predict O and E counts of 104.2 and 39.3.

The reason for discrepancies between true and back-predicted counts is compounding. Origination probabilities are not imposed just once at the end of an interval but continuously during an interval, so new species may themselves give rise to other new species before an interval ends. Likewise, extinction probabilities are not imposed suddenly, so even if extinction is initially rapid enough to imply an ultimate rate of 0.5 over one full interval, enough species are soon lost and therefore become ''immune'' to further extinction that the actualized rate is far lower.

Thus, analyzing simple ratios like O/D is just not consistent with modeling diversity dynamics in continuous time. Instead, one must determine the underlying, instantaneous rates by solving directly for r, λ, and μ. For net diversification one obtains

$$r = \ln(D_t/D_0)/t, \tag{11}$$

which was erroneously called an ''origination'' rate by Stucky (1990) (the statistic implicitly incorporates both origination and extinction counts, so the term is a misnomer). Now because

$$E = D_0 - D_0 e^{\mu t}, \quad (12)$$

the instantaneous extinction rate is

$$\mu = -\ln[(D_0 - E)/D_0]/t, \quad (13)$$

which is just the natural log of the fraction of species in the original cohort at time 0 that still survive to time t. The resulting expression for the instantaneous origination rate is not so intuitive. Because $r = \lambda - \mu$, $\lambda = r + \mu$, so one can combine equations (11) and (13) to yield

$$\lambda = \ln[D_t/(D_0 - E)]/t. \quad (14)$$

In other words, the origination rate is just the natural log of the number of species at the end of an interval divided by the number at the start minus the number in this cohort that went extinct. If there is no origination, then these numbers are equal and λ is zero; if there is no extinction, then equations (11) and (14) are equivalent, so $\lambda = r$.

It is important to repeat that the counts O and E must exclude singletons; otherwise the equations would not be valid. The equations are easily derived from those involved in cohort analysis (Raup 1978), which is such a well-established method that the failure of any author to employ these equations prior to Foote (1999) may come as a surprise. Indeed, there seems to be no earlier case in which instantaneous rate equations were used to study diversity dynamics in the fossil record. Likewise, it appears that none of the standard equations used by population ecologists (Murray 1997) correspond directly to equations (11) or (14).

In practice, the newly defined instantaneous rates yield virtually the same values as the singletons-excluded per-taxon rates of Alroy (1996, 1998d). The parametric correlation between the new λ and the old O/D is +0.983 (n = 70; Fig. 7A); for μ and E/D, it is +0.982 (Fig. 7B). Visually, the rates compare very closely, the only consistent difference being higher values yielded by the new equations during times of intense turnover (e.g., around the Cretaceous/Tertiary boundary at 65 Ma). Thus, the earlier equations turn out to have been excellent approximations of instantaneous rates. This result is predictable given that the bin lengths in this study are uniform and short relative to average turnover rates (Foote 2000).

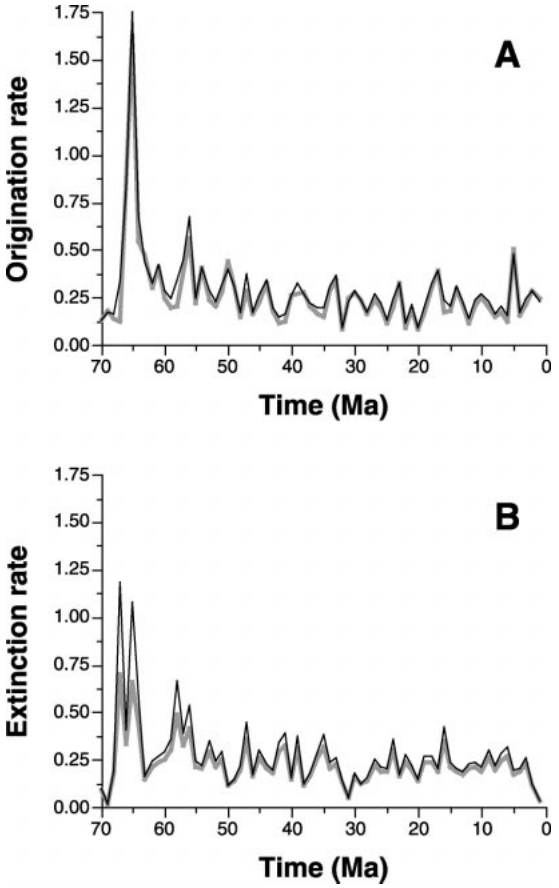

FIGURE 7. Cenozoic trends in taxonomic turnover statistics. All rates are instantaneous and apply to a single 1.0-m.y.-long sampling interval. Thin dark lines: instantaneous rates based on equations of Foote (1999). Thick gray lines: per-taxon rates based on equations of Alroy (1996, 1998d). A, Per-lineage origination rates. B, Per-lineage extinction rates.

Another minor note is that whereas dividing the rates by the amount of time within a bin may be necessary in some studies, here the use of uniform-length 1.0-m.y. bins renders any such correction superfluous.

A point of some empirical interest is the implied median duration of mammals implied by these data. The average μ value for the whole Cenozoic is 0.2672; for the 55 post-Paleocene data points, it is 0.2326. These figures respectively imply median durations of 2.14 and 2.62 m.y., both being much greater than (for example) an estimate of 1.7 m.y. based on a very different analysis of an earlier version

of this data set (Foote and Raup 1996). The difference has to do with the use in this study of species-lineage computations (Alroy 1996, 1998d), a heuristic method for removing the effects of pseudoextinction from the data.

In a companion study, Alroy et al. (2000) compare the instantaneous rates with marine oxygen isotope data for the last 60 m.y. This would be dangerous if both time series showed strong trends, because pairs of autocorrelated time series typically show correlations even when there are no causal relationships between them. However, the fact that Alroy et al. (2000) did not examine the first five of the ten highly volatile Paleocene data points means that the turnover rate data do not show strong trends or autocorrelation (Fig. 7A,B). For example, for λ plotted against time over the last 55 m.y., $r = +0.178$ (n.s.); for μ vs. time, $r = +0.152$ (n.s.).

This lack of correlation is unexpected, because mammalian origination rates are known to be negatively correlated with standing diversity levels (Alroy 1996, 1998d). Thus, one would expect to see origination rates fall not just slightly through time but dramatically as diversity increased. The fact that they do not makes it clear that the origination/diversity correlation is really not some side effect of secular trends in both variables.

Indeed, one does find a very strong correlation between λ and log standing species richness regardless of whether one examines the last 65, 60, or 55 m.y. of the Cenozoic ($r = -0.809, -0.602, -0.456$; $p < 0.001$). In contrast, μ seems to bear a relationship to diversity only if one includes Paleocene data points in the regression (same temporal intervals: $r = -0.570, -0.224, +0.075$; $p < 0.001$, n.s., n.s.). Thus, the new data confirm earlier findings (Alroy 1996, 1998d) that logistic growth in North American mammals is governed by the diversity dependence of origination rates, but not of extinction rates, which are essentially stochastic in the post-Paleocene interval.

The surprising thing about this result is that all other things being equal, a simple logistic curve should correlate with time, and hence time should correlate with origination—but in this case the time/diversity correlation quickly breaks down. The reason is that the first few data points include the initial climb; after that, diversity merely fluctuates around an equilibrium (Fig. 6).

The important thing to keep in mind is that before we even look to extrinsic variables like climate to explain trends in turnover (e.g., Alroy et al. 2000), we already have a powerful explanation for up to 65% of the variation in origination rates that is related entirely to intrinsic factors—i.e., biotic interactions like competition. Biotic interactions must be invoked here because no abiotic mechanism could predict a strong correlation between standing species richness and an origination-rate time series with low autocorrelation and no net temporal trend. By contrast, biotic interaction models that make just such a prediction are numerous (Rosenzweig 1975; Sepkoski 1978; Walker and Valentine 1984; Maurer 1989; Nee et al. 1992).

In addition to the λ and μ values (Fig. 7A,B; supplementary material, Table 2), Alroy et al. (2000) analyze three combined versions of these two statistics that represent different aspects of turnover: net diversification, here based on the difference of the instantaneous rates; diversification volatility, which is the absolute value of net diversification and represents the degree of net change in diversity in an interval regardless of its direction; and total turnover, which is the sum of the instantaneous λ and μ values and represents the amount of turnover in an interval regardless of how much this turnover causes diversity to change. None of these variables show strong temporal trends or autocorrelation.

Relative Diversity of Major Taxonomic Groups

Regional taxonomic richness is far from the only relevant indicator of evolutionary trends. It seems intuitive that ecological aspects of biotas such as alpha diversity and trophic diversity should be influenced by climate change, and much of the earlier literature on fossil mammals reflects this assumption (e.g., Webb 1977; Stucky 1990; Janis 1993, 1997; Gunnell et al. 1995; Morgan et al. 1995; Clyde and Gingerich 1998). Quantifying these features is far from easy. The most impressive study regarding alpha diversity was that of

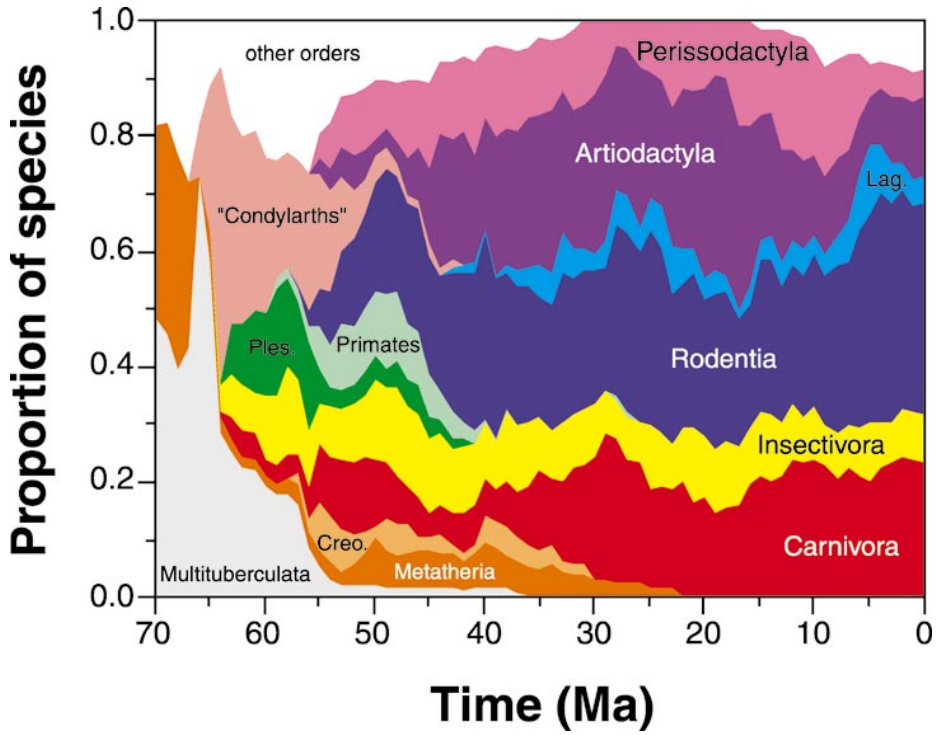

FIGURE 8. Proportionate diversity histories of twelve major taxonomic groups. All orders with at least 20 known genera are included, plus the paraphyletic stem group of ''condylarths'' that encompasses primitive ungulates. Abbreviations: ''Creo.'' = Creodonta; ''Lag.'' = Lagomorpha; ''Ples.'' = Plesiadapiformes.

Stucky (1990), who nonetheless only was able to compare the entire Paleocene–Eocene to the entire Oligocene–Neogene. Based on the difficulties he encountered, working out a full time series of alpha diversity values at 1.0-m.y. resolution would probably take years of effort. Likewise, earlier authors like Gunnell et al. (1995) used categorical data in studying trophic levels. Unfortunately, categorical data are not amenable to time series analysis, and obtaining good quantitative proxies of trophic data might require assembling a very large morphometric data set of the kind that Van Valkenburgh (1988) assembled for carnivores in selected faunas.

Because of these difficulties, here I will focus on two proxies of ecological disparity: the relative taxonomic diversity of major, ecologically distinct orders, and the general shape of the among-species body mass distribution. Both of these indicators have been used by other authors for similar purposes (e.g., Gunnell et al. 1995; Morgan et al. 1995).

At least in the North American Cenozoic fossil record, only a handful of orders have been relatively diverse at any one time (Alroy 1996). Most of these orders can easily be equated with one of four important trophic strategies distinguished by size (small vs. large) and diet (faunivorous vs. herbivorous). For example, mid- and late Cenozoic faunas are dominated by the small-sized and faunivorous Insectivora; the large and mostly faunivorous Carnivora; the small and mostly herbivorous Rodentia; the slightly larger and herbivorous Lagomorpha; and the large and herbivorous Artiodactyla and Perissodactyla. The situation is not so simple in the Cretaceous and Paleocene, with a larger number of orders each having moderate diversity levels (Fig. 8). Most of these groups had evolved only intermediate body sizes and do not show the extreme adaptations for specialized diets of Recent taxa. Because of the difficulty of assigning members of these primitive groups to ecological categories, and because of the likelihood that ecological roles evolved very quickly in the early Tertiary, it is presumably

conservative to treat each of the major extinct orders as a separate ecomorphologic entity.

Following an earlier study (Alroy 1996), this analysis will focus on orders including at least 20 genera. The ordinal-level taxonomy of marsupials and primitive ungulates (''condylarths'') is in flux, so each of these superordinal categories is treated as a unit. In practice, lumping the condylarth groups has little effect because only one of the potential orders (the Arctocyonia *sensu* McKenna and Bell 1997) would qualify as a ''major'' order by itself. The 12 major groups include four that are entirely extinct (Multituberculata, Creodonta, Plesiadapiformes, ''condylarths''), two that went extinct in North America before the Holocene (Metatheria exclusive of the late-Pleistocene immigrant *Didelphis,* Primates), and six that were extant throughout most of the Cenozoic (Carnivora, Insectivora, Rodentia, Lagomorpha, Artiodactyla, Perissodactyla).

Because absolute taxonomic richness is not of concern here, the analysis focuses on proportionate diversity curves for each of the groups (Fig. 8). These are computed by dividing the standing number of lineages in each order at the beginning of each 1.0-m.y. time interval by the total number of all mammalian lineages at that time. To avoid sampling artifacts, the same randomly subsampled data are used here as in the analysis of overall diversity and turnover.

Two methods are proposed for collapsing these relative diversity data. First, the data can be reduced to a single variable here termed the ''proportional volatility index.'' This is simply the sum of the absolute values of the differences between time slices in proportionate diversity for each order. For example, if the proportionate diversity of three groups is 0.1, 0.3, and 0.5 at time T and 0.1, 0.1, and 0.8 at time $T + 1$, then proportional volatility is $0 + 0.2 + 0.3 = 0.5$. The index ranges between 0.0 and 1.0.

The resulting time series is shown in Fig. 9A and in the supplementary material, Table 2. The main point established by this statistic is that large changes in relative taxonomic composition strongly track overall species-level turnover rates; periods of rapid extinction and/or speciation often witness certain

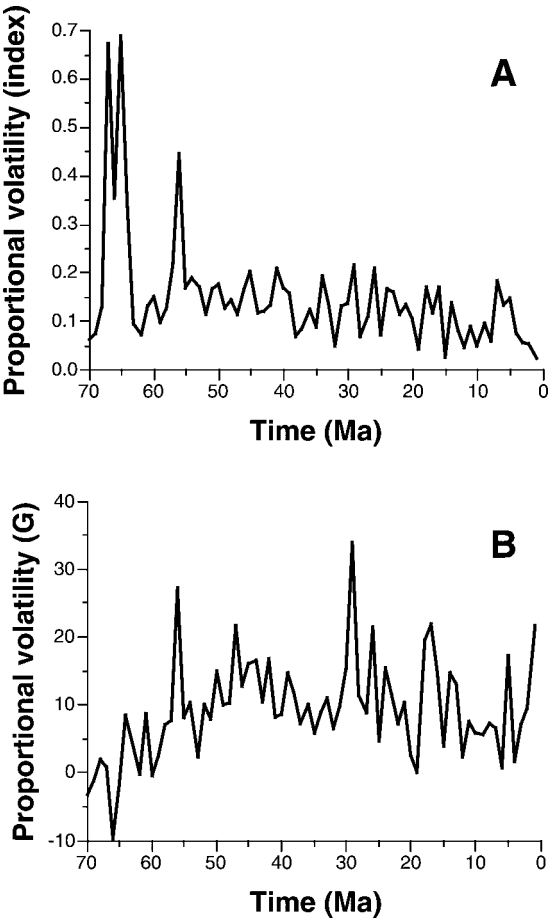

FIGURE 9. Two measures of the proportional volatility of ordinal diversity. A, Proportional volatility index: sum of changes in proportional diversity of major groups. B, Proportional volatility *G*-statistic: summary of goodness-of-fit between observed and expected counts of orginations and extinction within each time interval and major group.

groups prospering at the expense of others. So virtually every peak in this time series can be matched with one in the origination and/or extinction curves (Fig. 7A,B): its rank-order cross-correlation with either variable over the last 65 m.y. is significant and fairly strong (λ: $r = +0.425$; $t = 3.729$; $p < 0.001$; μ: $r = +0.351$; $t = 2.975$; $p < 0.01$).

A second way to summarize the data removes any built-in correlation with turnover rates. This involves computing a ''proportional volatility *G* statistic'' that summarizes departures of observed extinction and origination rates within groups from expected values given random turnover. Expected values are

based on observed average turnover rates within groups:

$$\bar{e_j} = \left(\sum_{i=1}^{T} e_{i,j}\right) \Big/ T \quad \text{and} \tag{15}$$

$$\bar{o_j} = \left(\sum_{i=1}^{T} o_{i,j}\right) \Big/ T, \tag{16}$$

where e = the per-species extinction rate, o = the per-species origination rate, i = the index variable for the T time intervals, and j = the index for the N orders. Assuming as a null hypothesis that underlying turnover rates are relatively constant among groups but vary among time intervals, the expected values are

$$\hat{E}_{i,j} = \frac{\sum_{k=1}^{N} e_{i,k} R_{i,k}}{\sum_{k=1}^{N} \bar{e}_k R_{i,k}} \bar{e}_j R_{i,j} \tag{17}$$

$$\hat{O}_{i,j} = \frac{\sum_{k=1}^{N} o_{i,k} R_{i,k}}{\sum_{k=1}^{N} \bar{o}_k R_{i,k}} \bar{o}_j R_{i,j}, \tag{18}$$

where R = species richness of the jth order at the beginning of the ith time interval (set to one instead of zero if the order first appears during this interval). The next step is to compute the G statistic for the goodness-of-fit between observed and expected absolute frequencies within each time interval:

$$G_i = 2\left(\sum_{j=1}^{N} e_{i,j} R_{i,j} \ln \frac{e_{i,j} R_{i,j}}{\hat{E}_{i,j}} + \sum_{j=1}^{N} o_{i,j} R_{i,j} \ln \frac{o_{i,j} R_{i,j}}{\hat{O}_{i,j}}\right). \tag{19}$$

The statistic is modified by Yates's correction for continuity (Sokal and Rohlf 1995: p. 730), both for the sake of conservativeness and in order to increase the likelihood that each computed value of G will have a similar number of degrees of freedom. The correction involves adjusting the expected values upwards or downwards by 0.5 depending on whether they fall short of or exceed the observed values. The resulting G values are shown in Figure 9B and in the supplementary material, Table 2. High values indicate that the proportional richness of the orders has shifted more quickly than expected at random; values close to zero indicate that random turnover can explain these proportional shifts. Unlike the simpler proportional volatility index, the rank-order correlation between the G statistic and extinction is near zero ($n = 65$; $r = -0.191$; $t = 1.083$; n.s.), although there is still a residual correlation with origination ($r = +0.272$; $t = 2.241$; $p < 0.05$). Although the asymmetry here is not overwhelming, it suggests that major replacements among orders may be mediated by bursts of speciation in ''winning'' groups, not bursts of extinction in ''losing'' groups.

As discussed by Alroy et al. (2000), the dominant feature of the G value time series is a major biotic transition at the Paleocene/Eocene boundary (55.5-Ma bin), although three additional peaks have biological significance (earliest Paleocene, 64.5 Ma; mid-Eocene, 46.5 Ma; mid-Oligocene, 28.5 Ma). Although the earliest Paleocene and Paleocene/Eocene events are associated with major spikes in origination rates, the other two are not, as one might expect from the essential independence of the new statistic from underlying turnover rates. In fact, some dramatic turnover events including an extinction pulse in the 57.5-Ma bin are barely reflected by the G statistic.

Body Mass Distributions

General Patterns.—Among-species body mass distributions have been the subject of considerable study in both the paleobiological and macroecological literature (e.g., Alroy 1998b; Clyde and Gingerich 1998; Brown and Nicoletto 1991). Here I argue for summarizing these distributions using four simple univariate statistics: the mean, standard deviation, skewness, and kurtosis (Fig. 10; supplementary material, Table 3). Earlier studies employing these data focused just on the mean and standard deviation (Alroy 1999a,b). Values are presented here for individual distributions within each 1.0-m.y. sampling bin (Fig. 10). All species ranging into each bin were included in the calculations, so unlike the boundary-crossing diversity counts (see above) these statistics describe a time-averaged population and do not control for sampling intensity. Time-averaging may introduce perceptible

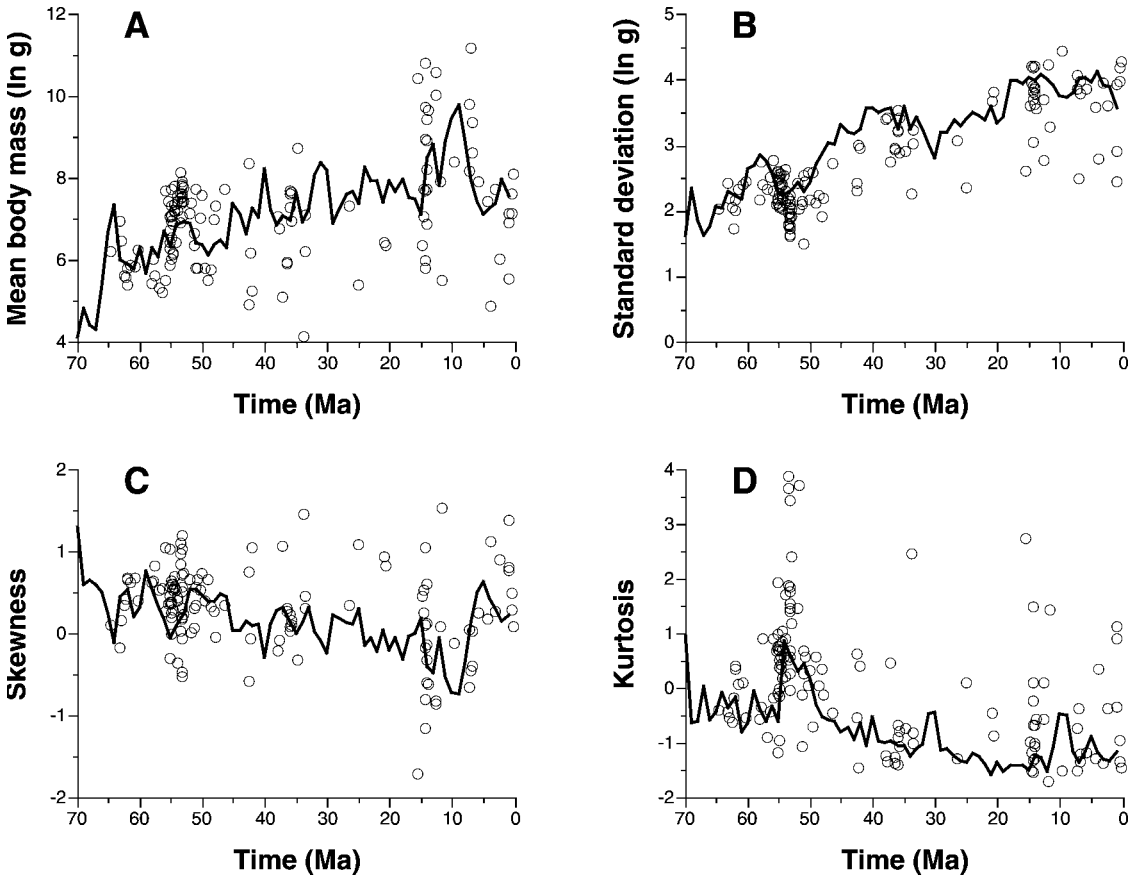

FIGURE 10. Univariate statistics describing body mass distributions. Line shows statistics describing the distribution of body mass estimates across all species ranging into each 1.0-m.y.-long sampling bin; points show same statistics for 133 local faunal assemblages from a narrow geographic area within the Western Interior. Each assemblage includes at least 25 species. A, Mean. B, Standard deviation. C, Skewness. D, Kurtosis.

distortions, but this step was made unavoidable by the need to guarantee reasonable sample sizes. The median Cenozoic bin includes 83 species, and the range is 30 (31–30 Ma) to 174 (15–14 Ma). Also, sampling intensity is unlikely to be important because statistics like the mean and standard deviation are designed to show no systematic relationship to sample size (even though the random error in these statistics does decrease as more data are acquired).

The generally monotonic trend throughout the Tertiary toward a higher mean and a larger standard deviation has been discussed elsewhere (Alroy 1998b, 1999a,b). The new data confirm the earlier results and do not need to be discussed in detail. However, the data for skewness and kurtosis (Fig. 10C,D) are both of note. The skewness data suggest a long-term drift toward negative values, meaning that the distribution's mode has shifted from being relatively low to being relatively high. The pattern is consistent with the nonlinear within-lineage evolutionary dynamic documented by Alroy (1998b), which predicts that lineages will actively evolve upwards from the middle size range and therefore shift the overall mode upwards. However, the trend reverses abruptly toward the very end of the Miocene (i.e., around 6 Ma), which may be explained by a shift in proportional diversity away from ungulates and toward rodents (Fig. 8) (see also Alroy 1996; Fig. 8). Notably, no extraordinary patterns are seen either in turnover rates or in proportional volatility statistics at this time.

Meanwhile, the kurtosis data suggest a sudden shift toward negative (platykurtic) values during the mid-Eocene. This pattern involves

the opening up of the middle size range, which creates a bimodal distribution (also visible in the raw data of Alroy 1998b). The shift may be attributable to environmental perturbations such as increased seasonality; further details including a possible connection to global climate changes are discussed by Alroy et al. (2000).

Local-Scale Data.—An issue of concern is the fact that the per-bin values used to construct these curves are based on lumped, time-averaged data for the entire continent. Therefore, the sets of species that are examined in each bin do not constitute actual ecological assemblages pertaining to individual habitats. This is worrisome because very different body mass distributions have been observed at local, regional, and continental scales in Recent North American mammals (Brown and Nicoletto 1991).

Can similar differences be demonstrated using paleontological data? Unfortunately, the answer appears to be no at this time. In addition to overall continental values, Figure 10 also presents univariate statistics for 133 diverse fossil assemblages that span the Cenozoic. Each assemblage includes at least 25 species. For this analysis, body masses of species lacking direct estimates and masses of specifically indeterminate but generically determinate taxa were estimated on the basis of among-species generic means. Figure 10 illustrates only assemblages that fall within a geographic rectangle spanning 35–46°N and 98–112°W (essentially Kansas, Nebraska, South Dakota, Colorado, Wyoming, and parts of adjacent states to the south and west); another 33 equally diverse assemblages were excluded. The region spans less than 6% of the area of North America, includes 2966 (60%) of the 4978 lists, and is roughly comparable to a single biome (see Brown and Nicoletto 1991).

The scatter of the points representing these local assemblages suggests three general conclusions: (1) Data are not completely adequate to establish a reliable time series of local-scale data points for the entire Cenozoic, with significant gaps in the mid-Tertiary. (2) Even within a relatively narrow time slice, the scatter of points for any variable may be extreme (e.g., data falling at about 15 Ma). (3) Relative to continental data, local data consistently underestimate the standard deviation and overestimate kurtosis, for which a minority of assemblages exhibit extraordinarily high values. Mean and skewness values show less systematic departures, although the amount of scatter seems to increase through time.

The highly variable local-scale mean, standard deviation, and skewness values all suggest taphonomic effects. Faunal assemblages that are largely restricted to large mammals exhibit high means and low skewness; small-mammal assemblages exhibit opposite patterns; and both types of size-biased assemblages exhibit low standard deviations. The number of size-biased assemblages seems to increase going into the Neogene. As for kurtosis, this statistic seems vulnerable to clumps of tied or nearly tied values in distributions with relatively small numbers of species like those seen in the local-scale data. Thus, the differences between local- and continental-scale data for these variables have no necessary biological basis. In general, the scatter is not attributable to biogeographic or climatic signals because the localities are restricted to a geographic area that is simply too small to pick up such signals.

Given the present evidence, local-scale data are simply not complete or reliable enough to be interpreted biologically. Thus, it seems that no meaningful paleoecological conclusions can be drawn on the basis of local assemblages including just 20 or 30 species. Some of the small-scale variation in the continental data also may relate to taphonomic effects. However, given the wide variation among samples that is suggested by Figure 10, the relatively steady, long-term trends seen in all four continental-scale curves are likely to reflect biological signals. Indeed, the continental data appear to benefit by sampling rare species and canceling out opposed biases in local assemblages. The asymptotic approach of continental data to underlying values is suggested by the very fact that departures of individual samples from the overall trend are so easily attributed to taphonomic effects.

Use of Cenograms.—At least for continental or regional data sets, univariate moment statistics may capture meaningful biological sig-

nals. Nonetheless, these statistics are not normally used by paleobiologists to describe mammalian body mass distributions (but see Clyde and Gingerich 1998). Instead, a large body of literature has emphasized mostly qualitative interpretations of log mass versus rank of mass plots called ''cenograms'' (Legendre 1989). In addition to the fact that local body mass distributions appear to be extraordinarily variable for reasons of sampling, there are at least five strong reasons to reject this method in favor of moment statistics.

First, most quantitative methods related to cenogram analysis (e.g., computation of slopes across parts of cenogram plots, or counts of species in particular body mass categories) require defining arbitrary splits of the body mass spectrum (e.g., Gingerich 1989; Wilf et al. 1998). Moment statistics require no arbitrary assumptions: the data are not binned into size categories. Furthermore, these statistics are universally employed in scientific analyses of frequency distributions.

Second, much of the key information represented by cenograms is captured by moment statistics: skewness indicates higher species richness in either small or large body mass categories, and negative kurtosis indicates the presence of a gap in the medium size range.

Third, the key information not captured by moment statistics is species richness, which partially determines the slopes of cenogram plots. Richness is notoriously hard to estimate in mammalian fossil assemblages. Of the 4978 faunal lists currently in the database, only 690 (13.9%) include at least 10 identified species, 212 (4.3%) at least 20, and just 54 (1.1%) at least 30—but Recent mammalian assemblages almost always include more than 20 terrestrial species. Temperate North American assemblages have been severely disturbed by the anthropogenic extinction of dozens of geographically wide-ranging large mammals (Alroy 1999b), and yet Brown and Nicoletto (1991) still found an average of 28 species in 24 North American habitats, with a minimum of 18. Tropical richness is still higher: Patton et al. (2000) found an average of 42 species (range 37–49) in seven habitats scattered across the western Amazon.

The possibility that many cenograms have been misinterpreted because of failure to recognize undersampling has been raised by other authors (e.g., Morgan et al. 1995; Wilf et al. 1998). In any event, the cenogram method conflates richness and the range of body mass in the form of a single statistic (the slope), which is a needless waste of information.

Fourth, empirically speaking there is no evidence that cenograms bear a strong and systematic relationship to independent measures of climate and habitat. In an exhaustive study, Rodriguez (1999) found few strong rank-order correlations between 16 different quantitative body mass measures based on cenogram analysis and eight important measures of climate and vegetation. Furthermore, relationships that were specifically predicted by earlier authors were not found.

Fifth and last, it is clear that the mammalian body mass spectrum was not nearly filled during the early, and arguably even middle, Tertiary. The main reason for this fundamental difference was an evolutionary lag following the opening of the high end of the body mass spectrum in the wake of the Cretaceous/Tertiary mass extinction (Wing and Tiffney 1987; Alroy 1998b, 1999a). Therefore, strictly ecological interpretations that hinge on observing low species richness among medium- or large-sized mammals are meaningless in the Paleogene. This issue has been raised but not explored in detail by earlier authors (e.g., Morgan et al. 1995).

Conclusions

Over the last few years, much ink has been spilt over methodological issues related to the preparation of macroevolutionary time series—and I am among the culprits. Indeed, this paper argues for modifying almost every step of the analysis that I outlined in an even more lengthy paper just two years ago (Alroy 1998d). These seemingly endless methodological revisions beg the question of whether the additional effort really matters.

The answer is both yes and no, depending on the biological signal at hand. The complex new methods for maximimum likelihood appearance event ordination, calibration of the appearance event sequence, and randomized

subsampling of faunal lists all seem to generate much the same patterns. Likewise, turnover rates computed with either the new equations of Foote (1999, 2000) or the older equations of Alroy (1996, 1998d) yield almost identical values, at least for this particular data set.

Each of these methods, however, has a very firm conceptual justification. For example, the maximum likelihood approach makes it possible to explicitly test for the importance of differences among taxa in sampling probabilities; the new calibration method makes use of as much information as possible; variations on subsampling methods bring assumptions about alpha diversity to the foreground; and Foote's equations directly capture the instantaneous turnover rates that are of fundamental concern in macroevolution.

Much more importantly, though, this study again highlights the importance of not preparing diversity data using traditional approaches. The level of precision in the underlying age ranges could not have been obtained using the conventional mammalian timescale (Alroy 1998c), correcting for variation in sampling intensity has an enormous impact (Alroy 1998d), and both Foote's equations and the older ones used by this author have robust properties (Foote 2000).

Of possibly greater interest are this paper's new, simple approaches for quantifying body mass distributions and changes in the proportional diversity of major groups. These methods could be applied easily to many different data sets, and paleobiologists should consider adding them to the roster of fundamental macroevolutionary variables. Examining these new variables is important because they capture distinct patterns of undoubted biological interest. Body mass and proportional diversity data not only emphasize the importance of the well-known Cretaceous/Tertiary and Paleocene/Eocene transitions, but point to additional shifts that otherwise might have been overlooked because they are not marked by extraordinarily high origination and extinction rates (e.g., mid-Eocene, mid-Oligocene).

The fact that major biotic transitions sometimes register in such different ways suggests that simple models of biotic change may be too limiting—no model of diversity dynamics could have predicted a second major ecomorphological transition just a few million years after the Paleocene/Eocene event. This study's discovery of such unexpected patterns highlights the importance of exploring new macroevolutionary methods.

Acknowledgments

I thank D. Fox, M. Foote, J. Hunter, J. Zachos, and especially P. Koch for comments on the manuscript, and S. Wing for his editorial solution to a Gordian-knot-like problem. This work was conducted while I was a postdoctoral associate at the National Center for Ecological Analysis and Synthesis, a center funded by the National Science Foundation (grant DEB-94-21535); the University of California, Santa Barbara; the California Resources Agency; and the California Environmental Protection Agency.

Literature Cited

Agterberg, F. P., and F. M. Gradstein. 1999. The RASC method for ranking and scaling of biostratigraphic events. Earth Science Reviews 46:1–25.

Alroy, J. 1992. Conjunction among taxonomic distributions and the Miocene mammalian biochronology of the Great Plains. Paleobiology 18:326–343.

———. 1994. Appearance event ordination: a new biochronologic method. Paleobiology 20:191–207.

———. 1996. Constant extinction, constrained diversification, and uncoordinated stasis in North American mammals. Palaeogeography Palaeoclimatology Palaeoecology 127:285–311.

———. 1998a. Diachrony of mammalian appearance events: implications for biochronology. Geology 26:23–27.

———. 1998b. Cope's rule and the dynamics of body mass evolution in North American mammals. Science 280:731–734.

———. 1998c. Diachrony of mammalian appearance events: implications for biochronology—Reply. Geology 26:956–958.

———. 1998d. Equilibrial diversity dynamics in North American mammals. Pp. 232–287 *in* M. L. McKinney and J. Drake, eds. Biodiversity dynamics: turnover of populations, taxa and communities. Columbia University Press, New York.

———. 1999a. The fossil record of North American mammals: evidence for a Paleocene evolutionary radiation. Systematic Biology 48:107–118.

———. 1999b. Putting North America's end-Pleistocene megafaunal extinction in context: large scale analyses of spatial patterns, extinction rates, and size distributions. Pp. 105–143 *in* R. D. E. MacPhee, ed. Extinctions in near time: causes, contexts, and consequences. Plenum, New York.

Alroy, J., P. L. Koch, and J. C. Zachos. 2000. Global climate change and North American mammalian evolution. *In* D. H. Erwin and S. L. Wing, eds. Deep time: *Paleobiology*'s perspective. Paleobiology 26(Suppl. to No. 4):259–288.

Bloch, J. I., K. D. Rose, and P. D. Gingerich. 1998. New species of *Batodonoides* (Lipotyphla, Geolabididae) from the early Eocene of Wyoming: smallest known mammal? Journal of Mammalogy 79:804–827.

Brown, J. H. 1995. Macroecology. University of Chicago Press, Chicago.
Brown, J. H., and P. F. Nicoletto. 1991. Spatial scaling of species composition: body masses of North American land mammals. American Naturalist 138:1478–1512.
Carr, T. R., and J. A. Kitchell. 1980. Dynamics of taxonomic diversity. Paleobiology 6:427–443.
Clyde, W. C., and P. D. Gingerich. 1998. Mammalian community response to the latest Paleocene thermal maximum: an isotaphonomic study in the northern Bighorn Basin, Wyoming. Geology 26:1011–1014.
Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B 345:101–118.
Damuth, J., and B. J. MacFadden. 1990. Body size in mammalian paleobiology: estimation and biological implications. Cambridge University Press, Cambridge.
Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368–376.
Flessa, K. W., and J. Imbrie. 1973. Evolutionary pulsations: evidence from Phanerozoic diversity patterns. Pp. 247–285 *in* D. H. Tarling and S. K. Runcorn, eds. Implications of continental drift to the earth sciences. Academic Press, London.
Foote, M. 1991. Morphological patterns of diversification—examples from trilobites. Palaeontology 34:461–485.
———. 1994. Temporal variation in extinction risk and temporal scaling of extinction metrics. Paleobiology 20:424–444.
———. 1999. Morphological diversity in the evolutionary radiation of Paleozoic and post-Paleozoic crinoids. Paleobiology 25:1–115.
———. 2000. Origination and extinction components of taxonomic diversity: general problems. *In* D. H. Erwin and S. L. Wing, eds. Deep time: *Paleobiology*'s perspective. Paleobiology 26(Suppl. to No. 4):578–605.
Foote, M., and D. M. Raup. 1996. Fossil preservation and the stratigraphic ranges of taxa. Paleobiology 22:121–140.
Gilinksy, N. L., and I. J. Good. 1991. Probabilities of origination, persistence, and extinction of families of marine invertebrate life. Paleobiology 17:145–166.
Gingerich, P. D. 1974. Size variability of teeth in living mammals and diagnoses of closely related sympatric fossil species. Journal of Paleontology 48:895–903.
———. 1989. New earliest Wasatchian mammalian fauna from the Eocene of northwestern Wyoming: composition and diversity in a rarely sampled high-floodplain assemblage. University of Michigan Museum of Paleontology Papers on Paleontology 28:1–97.
Gunnell, G. F. 1998. Mammalian faunal composition and the Paleocene/Eocene Epoch/Series boundary: evidence from the northern Bighorn Basin, Wyoming. Pp. 409–427 *in* M.-P. Aubry, S. G. Lucas, and W. A. Berggren, eds. Late Paleocene-early Eocene climatic and biotic events in the marine and terrestrial records. Columbia University Press, New York.
Gunnell, G. F., M. E. Morgan, M. C. Maas, and P. D. Gingerich. 1995. Comparative paleoecology of Paleogene and Neogene mammalian faunas: trophic structure and composition. Palaeogeography Palaeoclimatology Palaeoecology 115:265–286.
Harper, C. W. 1996. Patterns of diversity, extinction and origination in the Ordovician-Devonian Stropheodontacea. Historical Biology 11:267–288.
Hilborn, R., and M. Mangel. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, N.J.
Huelsenbeck, J. P., and K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annual Review of Ecology and Systematics 28:437–466.
Hunter, J. P., and J. Jernvall. 1995. The hypocone as a key innovation in mammalian evolution. Proceedings of the National Academy of Sciences USA 92:10718–10722.
Janis, C. M. 1993. Tertiary mammal evolution in the context of changing climates, vegetation, and tectonic events. Annual Review of Ecology and Systematics 24:467–500.
———. 1997. Ungulate teeth, diets, and climatic changes at the Eocene/Oligocene boundary. Zoology: Analysis of Complex Systems 100:203–220.
Janis, C. M., and P. B. Wilhelm. 1993. Were there mammalian pursuit predators in the Tertiary? Dances with wolf avatars. Journal of Mammalian Evolution 1:103–125.
Jernvall, J., J. P. Hunter, and M. Fortelius. 1996. Molar tooth diversity, disparity, and ecology in Cenozoic ungulate radiations. Science 274:1489–1492.
Legendre, S. 1989. Les communautés de mammifères du Paléogène (Éocène supérieur et Oligocène) d'Europe occidentale: structures, milieux et évolution. Münchner Geowissenschaftliche Abhandlungen, Reihe A, Geologie und Paläontologie 16:1–110.
Macdonald, L. J. 1972. Monroe Creek (early Miocene) microfossils from the Wounded Knee area, South Dakota. South Dakota Geological Survey Report of Investigations No. 105.
Markwick, P. J. 1998. Crocodilian diversity in space and time: the role of climate in paleoecology and its implication for understanding K/T extinctions. Paleobiology 24:470–497.
Maurer, B. A. 1989. Diversity-dependent species dynamics: incorporating the effects of population-level processes on species dynamics. Paleobiology 15:133–146.
May, R. M. 1975. Patterns of species abundance and diversity. Pp. 81–120 *in* M. L. Cody and J. M. Diamond, eds. Ecology and evolution of communities. Harvard University Press, Cambridge.
McKenna, M. C., and S. K. Bell. 1997. Classification of mammals above the species level. Columbia University Press, New York.
Miller, A. I., and M. Foote. 1996. Calibrating the Ordovician Radiation of marine life: implications for Phanerozoic diversity trends. Paleobiology 22:304–309.
Miller, A. I., and J. J. Sepkoski Jr. 1988. Modelling bivalve diversification: the effect of interaction on a macroevolutionary system. Paleobiology 14:364–369.
Morgan, M. E., C. Badgley, G. F. Gunnell, P. D. Gingerich, J. W. Kappelman, and M. C. Maas. 1995. Comparative paleoecology of Paleogene and Neogene mammalian faunas: body-size structure. Palaeogeography Palaeoclimatology Palaeoecology 115:287–317.
Murray, B. G., Jr. 1997. On calculating birth and death rates. Oikos 78:384–387.
Nee, S., A. O. Mooers, and P. H. Harvey. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proceedings of the National Academy of Sciences USA 89:8322–8326.
Patton, J. L., M. N. F. da Silva, and J. R. Malcolm. 2000. Mammals of the Rio Juruá and the evolutionary and ecological diversification of Amazonia. Bulletin of the American Museum of Natural History 244:1–306.
Raup, D. M. 1966. Geometric analysis of shell coiling: general problems. Journal of Paleontology 40:1178–1190.
———. 1975. Taxonomic diversity estimation using rarefaction. Paleobiology 1:333–342.
———. 1976. Species diversity in the Phanerozoic: an interpretation. Paleobiology 2:289–297.
———. 1978. Cohort analysis of generic survivorship. Paleobiology 4:1–15.
———. 1985. Mathematical models of cladogenesis. Paleobiology 11:42–52.
Raup, D. M., and J. J. Sepkoski Jr. 1982. Mass extinctions in the marine fossil record. Science 215:1501–1503.
———. 1984. Periodicity of extinctions in the geologic past. Pro-

ceedings of the National Academy of Sciences USA 81:801–805.
Rodriguez, J. 1999. Use of cenograms in mammalian palaeoecology: a critical review. Lethaia 32:331–347.
Rosenzweig, M. L. 1975. On continental steady states of species diversity. Pp. 121–140 *in* M. L. Cody and J. M. Diamond, eds. Ecology and evolution of communities. Harvard University Press, Cambridge.
Sanders, H. L. 1968. Marine benthic diversity: a comparative study. American Naturalist 102:243–282.
Sepkoski, J. J., Jr. 1978. A kinetic model of Phanerozoic taxonomic diversity. I. Analysis of marine orders. Paleobiology 4:223–251.
———. 1981. A factor analytic description of the Phanerozoic marine fossil record. Paleobiology 7:36–53.
———. 1984. A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. Paleobiology 10:246–267.
———. 1988. Alpha, beta, or gamma: where does all the diversity go? Paleobiology 14:221–234.
Shinozaki, K. 1963. Notes on the species-area curve. P. 5 *in* Proceedings of the tenth annual meeting of the Ecological Society of Japan.
Smith, E. P., P. M. Stewart, and J. Cairns Jr. 1985. Similarities between rarefaction methods. Hydrobiologia 120:167–170.
Sokal, R. R., and F. J. Rohlf. 1995. Biometry, 3d ed. W. H. Freeman, New York.
Storer, J. E. 1984. Mammals of the Swift Current Creek local fauna (Eocene: Uintan, Saskatchewan). Natural History Contributions 7:1–158.
Stucky, R. K. 1990. Evolution of land mammal diversity in North America during the Cenozoic. Pp. 375–430 *in* H. H. Genoways, ed. Current mammalogy, Vol. 2. Plenum, New York.
Tipper, J. C. 1979. Rarefaction and rarefiction—the use and abuse of a method in paleoecology. Paleobiology 5:423–434.
Van Valkenburgh, B. 1985. Locomotor diversity within past and present guilds of large predatory mammals. Paleobiology 11: 406–428.
———. 1988. Trophic diversity in past and present guilds of large predatory mammals. Paleobiology 14:155–173.
Voorhies, M. R. 1990. Vertebrate paleontology of the proposed Norden Reservoir Area, Brown, Cherry, and Keya Paha counties, Nebraska. Division of Archeological Research, Department of Anthropology, University of Nebraska, Technical Report 82-09.
Wagner, P. J. 1998. A likelihood approach for evaluating estimates of phylogenetic relationships among fossil taxa. Paleobiology 24:430–449.
Walker, T. D., and J. W. Valentine. 1984. Equilibrium models of evolutionary species diversity and the number of empty niches. American Naturalist 124:887–899.
Walsh, S. L. 1998. Diachrony of mammalian appearance events: implications for biochronology—Comment. Geology 26:955–956.
Webb, S. D. 1977. A history of savannah vertebrates in the New World, Part I: North America. Annual Review of Ecology and Systematics 8:355–380.
Wilf, P., K. C. Beard, K. S. Davies-Vollum, and J. W. Norejko. 1998. Portrait of a late Paleocene (early Clarkforkian) terrestrial ecosystem: Big Multi Quarry and associated strata, Washakie Basin, Southwestern Wyoming. Palaios 13:514–532.
Wing, S. L., J. Alroy, and L. J. Hickey. 1995. Plant and mammal diversity in the Paleocene to early Eocene of the Bighorn Basin. Palaeogeography Palaeoclimatology Palaeoecology 115: 117–155.
Wing, S. L., and B. H. Tiffney. 1987. The reciprocal interaction of angiosperm evolution and tetrapod herbivory. Review of Palaeobotany and Palynology 50:179–210.
Woodburne, M. O., and C. C. Swisher III. 1995. Land mammal high-resolution geochronology, intercontinental overland dispersals, sea level, climate, and vicariance. SEPM (Society for Sedimentary Geology) Special Publication 54:335–365.