Paleobiology, 34(3), 2008, pp. 360–377

Gradual or pulsed evolution: when should punctuational explanations be preferred?

Gene Hunt

Abstract.—The problem of gradual versus punctuated change within phyletic lineages can be understood in terms of the homogeneity of evolutionary dynamics. Hypotheses of punctuated change imply that the rules governing evolutionary change shift over time such that the normal dynamics of stasis are temporarily suspended, permitting a period of net evolutionary change. Such explanations are members of a larger class of models in which evolutionary dynamics are in some way heterogeneous over time. In this paper, I develop a likelihood-based statistical framework to evaluate the support for this kind of evolutionary model. This approach divides evolutionary sequences into nonoverlapping segments, each of which is fit to a separate evolutionary model. Models with heterogeneous dynamics are generally more complex—they require more parameters to specify—than uniform evolutionary models such as random walks and stasis. The Akaike Information Criterion can be used to judge whether the greater complexity of punctuational models is offset by a sufficient gain in log-likelihood for these models to be preferred.

I use this approach to analyze three case studies for which punctuational explanations have been proposed. In the first, a model of punctuated evolution best accounted for changes in pygidial morphology within a lineage of the trilobite *Flexicalymene*, but the uniform model of an unbiased random walk remains a plausible alternative. Body size evolution in the radiolarian *Pseudocubus vema* was neither purely gradual nor completely pulsed. Instead, the best-supported explanation posited a single, pulsed increase, followed later by a shift to an unbiased random walk. Finally, for the much-analyzed claim of ''punctuated gradualism'' in the foraminifera *Globorotalia*, the best-supported model implied two periods of stasis separated by a period of elevated but not inherently directional evolution. Although the conclusions supported by these analyses generally refined rather than overturned previous views, the present approach differs from those prior in that all competing interpretations were formalized into explicit statistical models, allowing their relative support to be unambiguously compared.

Gene Hunt. Department of Paleobiology, National Museum of Natural History Smithsonian Institution, NHB, MRC 121, Post Office Box 37012, Washington D.C. 20013-7012. E-mail: hunte@si.edu

Accepted: 7 April 2008

Introduction

The punctuated equilibria model makes two central claims: stasis is prevalent within lineages and that phenotypic evolution is concentrated in punctuations associated with speciation (Eldredge and Gould 1972). These two claims are not on equal evidentiary footing. Stasis is straightforward to document, and the commonness of stasis in the fossil record is fairly uncontroversial (Hoffman 1989; Jackson and Cheetham 1999; Eldredge et al. 2005). In contrast, universally convincing examples of punctuations are far fewer. Moreover, because fossil species are identified on the basis of morphology, the link between cladogenesis and morphological evolution is very difficult to ascertain, at least directly (Hoffman 1989; Levinton 2001).

Whereas determining the relationship between lineage splitting and phenotypic evolution is understandably difficult, it should be much easier to assess gradual versus punctuated change within unbranched lineages. However, even attempts at this simpler task have been contentious (Gould and Eldredge 1977; Malmgren et al. 1983; Gingerich 1985; Levinton 2001; Gould 2002). With few exceptions (Bookstein et al. 1978; Roopnarine 2001), these modes of change have been identified qualitatively by visually inspecting plots of trait values over time, and it has been widely acknowledged that preconceived expectations can influence the conclusions drawn (Gould and Eldredge 1977; Fortey 1985; Erwin and Anstey 1995).

Taken to exaggerated extremes, the differ-

 0094-8373/08/3403-0004/$1.00

ences between gradual and punctuational patterns are clear: for the former, all rates of change are equal; for the latter, all rates are either zero (stasis) or high (punctuations), with instances of stasis outnumbering punctuations. However, even if these simplified versions of gradualism and punctuation were fair (which they are not), their greater defect is that they are not directly useful because real evolutionary sequences never show such clear-cut patterns. Instead, fossil lineages always show a mixture of slower and faster rates of change. How rapid is punctuation, and how sluggish is stasis? There seems to be no non-arbitrary way to answer this question, and therefore it may not be surprising that the same data sets could yield conflicting evolutionary interpretations.

I propose that this question of gradual versus pulsed evolution within lineages can be framed such that the preference for different evolutionary models is not arbitrary, or simply a restatement of prior beliefs. The key to this reformulation is the recognition that although punctuations have implications for the distribution of evolutionary rates, more fundamentally they are claims about the homogeneity of evolution. Essentially, punctuational explanations posit that intervals of accelerated evolution differ qualitatively from normal evolutionary dynamics. Numerous mechanisms have been proposed to account for these two regimes of punctuation and stasis, but what matters is that punctuations are postulated to represent evolutionary change that is governed by a different set of rules than those that operate during stasis. Testing claims of punctuation therefore ought to assess the fit of models incorporating this kind of heterogeneity versus reasonable models for which evolutionary dynamics are uniform, as in, for example, a simple random walk. It is very important to note that uniform dynamics generally do not imply absolutely constant evolutionary rates; all useful homogeneous models produce evolutionary trajectories with varying rates. Consequently, models of pulsed evolutionary change will outperform homogenous models when the amount and temporal distribution of rate variation could not plausibly obtain unless the rules governing evolutionary divergence shifted within an evolving lineage.

In this paper, I present a framework to evaluate statistically the general class of models with heterogeneous evolutionary dynamics, with a focus on those that imply punctuated change. This approach extends a previous treatment of modes of evolution as statistical models (Hunt 2006), and the first section of this paper briefly reviews how the traditionally recognized modes of evolutionary change—directional change, random walk, and stasis—can be analyzed as statistical models. Next, I describe how this approach can be generalized to incorporate models in which evolutionary dynamics shift at one or more points in time, as in notions of punctuated evolutionary change. The last section of this paper applies this general approach to analyze three case studies for which the claim of punctuated evolution has been made, and then assesses the robustness of the results with respect to age model error.

In general, these analyses find substantial support for shifting dynamics models in real paleontological sequences. The interpretations that result are generally refinements of previous views, rather than radical revisions. In each example considered, however, the differences between competing explanations are clarified by their conversion to explicit statistical models, allowing the relative merit of differing interpretations to be judged.

Homogenous Evolutionary Dynamics

Models with shifting evolutionary dynamics are elaborations of simpler models in which evolutionary rules do not change over time. Thus, dealing with shifting dynamics requires first an overview of homogeneous evolutionary models. In this section, I review three kinds of homogeneous evolutionary modes: directional change, unbiased random walk, and stasis. The first two modes have their basis in the general random walk model (Hunt 2006), and the last follows the formulation of stasis by Sheets and Mitchell (2001). What follows is a brief summary of statistical inference using these models; further details and derivations are given by Hunt (2006). Although I focus on these standard modes of

evolutionary change (Gingerich 1993; Roopnarine et al. 1999), other specified models can be substituted readily.

Directional Evolution and Random Walks

The general random walk (see Hunt 2004, 2006) is useful for modeling both directional evolutionary change and unbiased random walks. This model imagines time in discrete increments, during each of which an evolutionary transition is drawn at random from a distribution of evolutionary ''steps.'' The absolute timing between steps does not matter, so long as they are much more finely spaced than the temporal resolution of the fossil samples. Under these conditions, the long-term dynamics of an evolving lineage only depend on the mean (μ_{step}) and variance (σ^2_{step}) of the step distribution. Specifically, the expected change in traits after t time steps is normally distributed with mean $t\mu_{step}$ and variance $t\sigma^2_{step}$. For real data, the variance is increased by the sampling error in estimating trait means (Hunt 2006).

For non-zero values of μ_{step}, the general random walk is inherently directional with the sign and strength of the evolutionary trend determined by μ_{step}. Trait values will tend to increase over time if μ_{step} is positive and decrease if μ_{step} is negative. When μ_{step} is exactly zero, the general random walk reduces to the special case of an unbiased random walk that, on average, produces no net evolutionary trends. Thus, this relatively simple model can, depending on the value of the μ_{step} parameter, produce either directional change or random walks.

Stasis

The term stasis is used to describe a pattern of limited net evolutionary change over time (Eldredge et al. 2005), and this qualitative notion has been translated into several different quantitative models (Gingerich 1993; Roopnarine 2001). Here I follow the parameterization of Sheets and Mitchell (2001) in imagining an evolutionary optimum (θ), around which evolutionary fluctuations occur with a variance of ω. Trait values are distributed normally around a central value and evolutionary divergence does not accrue over time. Regardless of the time separating ancestral and descendant populations, the expected trait value of the descendant population is normally distributed with mean θ and variance ω (Hunt 2006). As with the general random walk, sampling error also contributes variance to the expected evolutionary divergence for real data (Hunt 2006). Although this stasis model and the general random walk both have mean and variance parameters, these parameter pairs refer to different distributions. The stasis parameters (θ and ω) characterize the distribution of actual trait values, whereas μ_{step} and σ^2_{step} of the general random walk specify the mean and variance of evolutionary increments.

Of those proposed, this formulation of stasis is probably the simplest and most analytically tractable. Despite its simplicity, this model often provides a good fit to real paleontological time series (Hunt 2006, 2007). In fact, this model may be unduly favored when evolutionary changes are small because it has the same mathematical form as sampling noise (Sheets and Mitchell 2001; Hannisdal 2006). Stasis modeled in this way also resembles a special case of an Ornstein-Uhlenbeck process (Hansen 1997) with a very strong restraining force. Such models are sometimes adopted in studies of stabilizing selection and macroevolution (Hansen and Martins 1996; Hansen 1997; Martins et al. 2002; Butler and King 2004; Hunt et al. 2008).

Fitting and Comparing Models

For all three homogenous modes of evolution, expected evolutionary divergences are normally distributed with means and variances that depend on evolutionary parameters and magnitudes of sampling error. This predicted probability density is sufficient to compute log-likelihoods of evolutionary transitions. The best estimates for model parameters are those that maximize the likelihood of observed evolutionary transitions. The log-likelihood also serves as a convenient measure of the goodness-of-fit of any model to real data. This fitting process—determining parameter estimates and log-likelihoods—is described in detail by Hunt (2006).

Fitting models and comparing models are

usually separate steps of statistical inference. Once several candidate models are fit to the same data, it is desirable to compare their performances as explanations for the phenomena under study. Because models with more parameters generally fit data better than those with fewer parameters, these comparisons must account for model complexity. At present, one of the most useful means of balancing goodness-of-fit and model complexity is achieved by using the Akaike Information Criterion (AIC; Akaike 1974), which is computed as $-2\log(L) + 2K$, where $\log(L)$ is log-likelihood and K is the number of free parameters in the model. In practice, it is preferable to use a bias-corrected form of the AIC to prevent over-fitting (Hurvich and Tsai 1989; Anderson et al. 2000):

$$\mathrm{AIC_C} = \mathrm{AIC} + (2K[K+1])/(n-K-1),$$

where n is sample size (the number of evolutionary transitions). The AIC represents the amount of information lost in approximating reality with a particular model (Anderson et al. 2000), and therefore the model with the lowest AIC score is preferred. The relative support for different models can be quantified as Akaike weights, which are computed from the differences between each AIC score and the score for the best model: $\Delta_i = \mathrm{AIC}_i - \min(\mathrm{AIC})$; either AIC or $\mathrm{AIC_C}$ values can be used in this calculation. The weight for each of G models is computed as

$$w_i = \exp\left(-\frac{1}{2}\Delta_i\right) \Big/ \left[\sum_{j=1}^{G} \exp\left(-\frac{1}{2}\Delta_j\right)\right].$$

Akaike weights sum to one and are a convenient means of summarizing the proportion of total evidential support each model receives (Anderson et al. 2000). It is a great advantage of likelihood-based methods that there are well grounded statistical means for judging the success of different models.

Shifting Evolutionary Dynamics

Adding a shift in dynamics to an evolving lineage represents a straightforward elaboration of homogenous dynamics. This class of models divides an observed evolutionary sequence into two or more segments, each of which evolves according to its own evolutionary dynamic. Segments may evolve according to different evolutionary modes, as in a lineage that evolves directionally at first but then experiences stasis. Or, the same general mode of evolution may apply but with distinct parameter values in each segment, for example a directional random walk that shows first increasing ($\mu_{step} > 0$) and then decreasing ($\mu_{step} < 0$) evolutionary trends.

Statistical Inference

Clearly, determining exactly when in a sequence evolutionary dynamics change is a crucial part of any shifting dynamics model. If we set aside this question temporarily and assume the shift point or points are given, statistical inference of shifting dynamics models is not different from inferring dynamics in a set of different sequences (Hunt 2006). The models are optimized separately by segment, yielding parameter estimates and log-likelihoods for each. Because the dynamics within each segment are assumed to be independent of the others, the total log-likelihood of a sequence is simply the sum of the log-likelihoods of its constituent segments (see Wagner 2000 for an analogous analytical approach for assessing heterogeneity in fossil occurrences).

In fitting these models, it is necessary to specify what kind of shifting dynamics are hypothesized, including the number of independently evolving segments and the general form of morphological evolution in each segment. Different kinds of shifting dynamics are likely to vary substantially in performance, and appropriate models should be chosen carefully on the basis of the empirical and theoretical context of the lineage under study.

The important question of how to determine the timing of evolutionary shifts has so far been ignored. In rare circumstances, there may be a standing hypothesis about when shifts might occur. If so, these hypothesized shifts can be taken as part of the model. Much more often, however, there are no such a priori expectations, and it is more appropriate to treat shift points as free parameters of the model and estimate them from the data. As a result, the total number of parameters in a shifting dynamics model is equal to the sum of the parameters in each segment, plus an ad-

ditional parameter for each shift that specifies its timing.

Optimal shift point parameters can be determined in a variety of ways. The approach I use here is simply to try all possible shift points and retain the one (or ones) most consistent with the data (i.e., yielding the highest log-likelihood). It is reasonable to set a minimum segment length as a constraint on the possible shift points to be evaluated. Very short segments do not permit confident estimation of models, and this constraint also reduces the number of shift points that need be considered, speeding up analysis of very long sequences. I have found that restricting segments to those with five samples or more works well in practice; allowing shorter segments sometimes produces spurious solutions for complex models (unpublished simulation results).

In addition to estimating when dynamics change, it is important to know how precisely shift points have been inferred. There are several means of computing confidence intervals within a likelihood framework. Here I use what are sometimes called profile confidence intervals (Kalinowski and Taper 2005). An approximate confidence region (at level $1 - \alpha$) includes all model solutions with a likelihood ratio statistic within $C/2$ units of the maximum-likelihood solution, where C is the $1 - \alpha$ quantile of the chi-square distribution with degrees of freedom equal to the number of parameters of interest (Meeker and Escobar 1995). If the goal is to produce confidence intervals only for the shift points, the number of parameters is taken as the number of shift points, not the full number of parameters of the model (see Meeker and Escobar 1995). Ninety-five percent confidence intervals include all solutions within 1.92 log-likelihood units of the maximum-likelihood solution when there is one shift point, and within 3.00 units of log-likelihood when there are two shift points.

Two Kinds of Punctuations

There are numerous conceivable models in which evolutionary dynamics shift within an evolving lineage. Many of these may be useful for paleobiological inference, but models that imply pulsed phenotypic change have generated particular interest. Punctuation is a three-phase model: lineages initially experience stasis, then undergo a period of rapid change after which stasis is resumed. In terms of analytic strategy, the best way to model punctuated change depends on the rapidity of the punctuation relative to the resolution of paleontological sampling.

Unsampled Punctuations.—If the pulsed change is very rapid compared to the temporal spacing of samples, no samples of intermediate morphology are observed and the punctuation is inferred from a shift in morphology between two adjacent samples. This scenario can be modeled as a two-segment stasis model in which the trait optimum shifts abruptly between the two segments (Fig. 1A). In this scenario, a lineage initially experiencing stasis around a phenotype of θ_1 suddenly experiences stasis around a new optimum (θ_2), to which the phenotype immediately converges. Because the shift to the new optimum is instantaneous at the level of sampling resolution, this model is well suited for approximating rapid punctuations, but will poorly fit sequences with well-sampled transitional intervals.

Under this approach, only the beginning and ending stasis segments are modeled explicitly; the punctuation registers as the difference between two estimated optima. This is a reasonable strategy because it matches the preserved information; there is little point modeling in any detail a pulsed change that is not observed. In terms of microevolutionary scenarios, this model corresponds well to a situation in which a population follows an adaptive optimum that shifts rapidly relative to the temporal resolution of samples (e.g., the displaced optimum model of Estes and Arnold 2007).

Sampled Punctuations.—When there are more than a few intermediate populations sampled within a putative punctuation interval, it becomes desirable to explicitly model the evolutionary change occurring within the transition zone. Punctuations are generally envisioned as periods of directional change, as opposed to non-directional oscillation or fluctuation. Therefore, a reasonable way to model

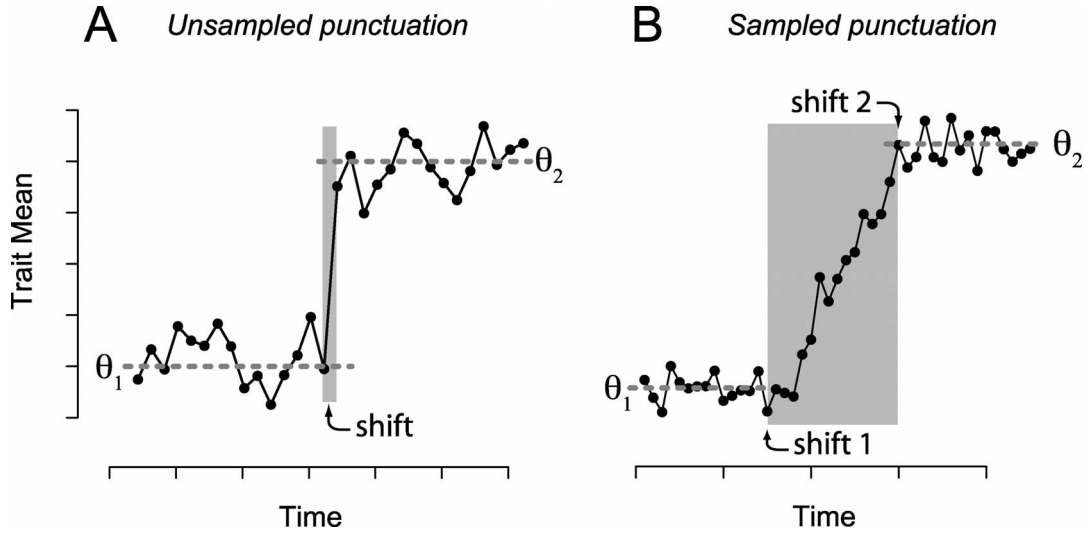

FIGURE 1. Illustrations of two kinds of punctuated change: unsampled punctuations (A), in which few or no intermediate populations are captured; and sampled punctuations (B), in which multiple populations from the transition interval are observed. Unsampled punctuations can be modeled effectively as two intervals of stasis with different optima (θ_1 and θ_2); the magnitude of punctuated change is determined by the difference in the two optimal values. For sampled punctuations, the interval of directional change is modeled explicitly as a general random walk inserted between two periods of stasis.

sampled punctuations is to insert a period of directional change (using a general random walk) in between two episodes of stasis (Fig. 1B). One could also relax the assumption of directionality and instead fit an unbiased random walk between the two intervals of stasis. Unbiased random walks produce accumulating but not inherently directional evolution, and this kind of punctuation might be thought of as a temporary relaxation of whatever processes limit evolutionary divergence during stasis. In either case, the transition interval is now sampled, and it is therefore reasonable to model its evolutionary dynamics.

As modeled here, all punctuations start and end in episodes of stasis. These two periods of stasis may be similar in their magnitudes of evolutionary fluctuations (ω), or these fluctuations may be greater before or after the evolutionary pulse. Because neither option seems strongly preferable a priori, I fit both versions in the examples that follow and retain whichever is better supported by the data.

Clearly, which kind of punctuation is most appropriate will depend on both the underlying evolutionary trajectory and the fineness with which it has been sampled. This dependence on resolution is entirely appropriate; few paleontological patterns are likely to appear the same at all scales of observation, and it is important that models and data operate at commensurate scales.

Analyses reported here were implemented in the R programming language (R Development Core Team 2007), using functions from the package *paleoTS* (Hunt 2008). This package can be downloaded and installed from within the R program in the usual way (see the documentation for the R software). For a given set of shift points, the model optimizations are quite fast—a few seconds or less, depending on the length of the sequence and the speed of the computer. However, with long sequences, there are many possible shift points to explore, and these analyses can take some time. The most time-consuming analysis involved a sequence of 95 samples and models with three segments, which took about 15 minutes to analyze on a relatively speedy desktop computer.

Case Studies

Punctuation in *Flexicalymene* Pygidia

Cisne et al. (1980) documented spatial and temporal variation in pygidial ring counts

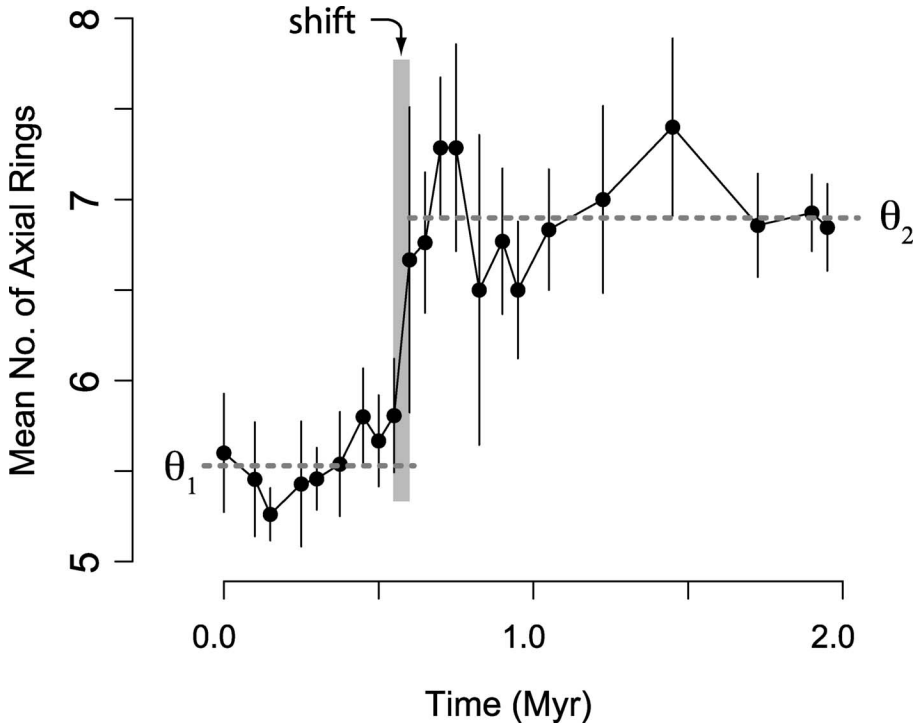

FIGURE 2. Mean number of pygidial axial rings in a stratigraphic sequence of the trilobite *Flexicalymene* (Cisne et al. 1980). The best-supported model for these data implies an unsampled punctuation event between the ninth and tenth samples (vertical gray rectangle); dashed horizontal lines indicate the estimated stasis optima (θ_1 and θ_2) for this model. Time is measured in millions of years from the first population. Vertical bars show 95% confidence intervals around the sample means.

within a single lineage of the trilobite genus *Flexicalymene*. This trait showed a rapid increase within a short stratigraphic interval, with more or less stable values before and after this transition zone (Fig. 2). These authors interpret the period of rapid increase as a punctuation lasting about 200 Kyr, preceded and followed by intervals of stasis. A seemingly different interpretation is held by Levinton (2001, Table 6.1), who includes this example in a list of gradual evolutionary transitions.

From the reported frequencies of each axial ring count morph, I computed the mean and variance of this trait through the reported composite section. For stratigraphic levels with few specimens, I combined adjacent samples so that all had five or more pygidia. The age model for these samples was determined from the reported heights in a composite section, each meter of which was estimated to span 50 Kyr on average (Cisne et al. 1980).

I fit four evolutionary models to the resulting sequence of 22 samples. These included the three standard homogeneous models and a fourth model that postulated a single unsampled punctuation event, corresponding to the interpretation of Cisne et al. (1980). Note that the timing of this punctuation is a free parameter of the model that is estimated from the data.

The performance of each of these models is summarized in Table 1. Of the homogenous dynamics models, the unbiased random walk is best supported (it has the lowest AIC_C). The general random walk offers only a slight improvement in log-likelihood over the unbiased random walk in exchange for its extra parameter, and its AIC_C score is substantially worse (Table 1). As might be expected, the stasis model offers a very poor fit because early axial ring counts are systematically lower, and later counts are systematically higher, than the putative optimum of intermediate value.

The model of a single, unsampled punctuation event fits these data better than any of

TABLE 1. Model fits to *Flexicalymene* pygidial ring number data. Three homogenous evolutionary dynamics models (models 1–3) were fit, along with a model postulating a single, unsampled punctuation event (model 4). Segments refer to the sections of the evolutionary sequence with independent evolutionary dynamics. Abbreviations: log(L), log-likelihood; K, the number of free model parameters; AIC_C, bias-corrected Akaike Information Criterion; w, Akaike weight. The punctuated model fit for these data assumes the same evolutionary variance before and after the punctuation event. Parameter estimates for the best-supported model (4): $\theta_1 = 5.52$ rings, $\theta_2 = 6.90$ rings, $\omega = 0.017$ rings^2; dynamics shift after the ninth sample. Akaike weights for models receiving at least moderate support ($w > 0.2$) are in bold.

No.	Model	No. of segments	K	log(L)	AIC_C	w
1	Unbiased random walk	1	1	−2.90	8.01	**0.375**
2	Directional evolution	1	2	−2.82	10.31	0.119
3	Stasis	1	2	−22.43	49.48	0.000
4	Punctuation-1	2	4	1.54	7.42	**0.505**

the homogeneous dynamics models (Table 1). By far the best-supported position for the shift in dynamics is after the ninth sample (Fig. 2), which corresponds to the interpretation of the original authors. The gain in log-likelihood (>4 units) is more than enough to offset the greater number of parameters of this model ($K = 4$ relative to $K = 1$ of the unbiased random walk; this punctuational model assumes the same stasis variance [ω] before and after the punctuation). However, the performance of the punctuational model does not unambiguously rule out the possibility of homogeneous dynamics. In particular, the unbiased random walk model is only moderately less successful than the punctuation model; its Akaike weight is large enough so that it should be retained as a plausible explanation of this trait's evolutionary dynamics ($w = 0.375$, compared to $w = 0.505$ for the punctuation model). There is less, but non-negligible, support for the general random walk as well ($w = 0.119$), but this is expected because it is not possible for its fit to be too much worse than that of an unbiased random walk (see Hunt 2006: p. 596).

In summary, the punctuational interpretation of this case study is supported, but it is not possible to rule out a uniform unbiased random walk through the whole sequence. Collecting from additional stratigraphic levels, or of more pygidia within the present intervals, could help to discriminate between these two models better. Some of the levels have means with rather wide confidence intervals (Fig. 2), and this broad uncertainty allows for these data to be consistent with a range of evolutionary dynamics.

Gradual versus Stepped Evolution in *Pseudocubus vema*

Kellogg (1975) documented several million years of body size evolution in the radiolarian lineage *Pseudocubus*. Body size (measured as thoracic width) increased by approximately 50% during this interval, with the increasing trend described as ''stepped'' rather than constant (Fig. 3). Gould and Eldredge (1977), while acknowledging the nuance of Kellogg's description, argued against a gradualistic interpretation of this sequence. Instead, they suggested the existence of three intervals of stasis separated by two punctuations toward larger size. This punctuational reinterpretation was viewed somewhat skeptically by Bookstein et al. (1978), who seemed to suggest that evolution in this lineage was actually more complicated than either a uniform directional trend or two neat punctuation events.

Using the software WinDig (Lovy 1996), I digitized sample means and confidence intervals with respect to depth in core (Kellogg 1975: Fig. 4). I used the confidence intervals and sample sizes to calculate sample variances and the sedimentation rates of Kellogg (1975) to translate core depths into ages.

In addition to the three standard homogenous models (Table 2), I fit to this sequence a model with two unsampled punctuations as per the explanation invoked by Gould and Eldredge (1977). However, in making the case for their three plateaus of stasis, Gould and Eldredge omitted several data points, including the last two samples in the sequence. With

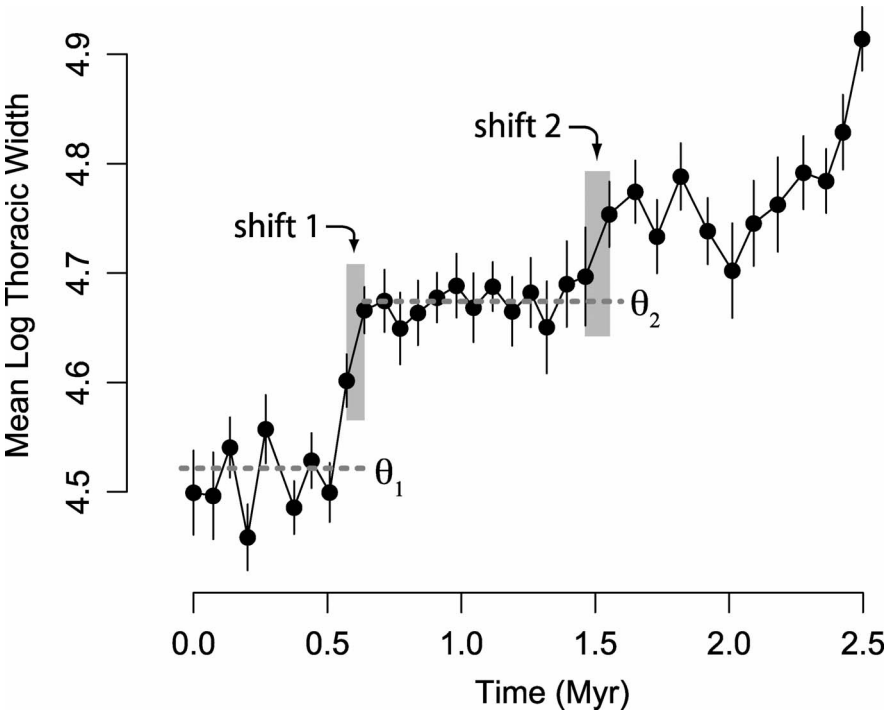

FIGURE 3. Mean thoracic width in the radiolarian lineage *Pseudocubus vema* (Kellogg 1975). Error bars indicate 95% confidence intervals on sample means, and time is measured in millions of years after the first sample. Shown is the best-supported model, which implies an unsampled punctuation event (shift 1, from optimum θ_1 to θ_2) and a subsequent shift from stasis to an unbiased random walk (shift 2).

these samples included, the last section of the sequence appears decidedly less like the fluctuating evolution of stasis and more like a meandering random walk. Reflecting this possibility, I fit a sixth model to this sequence, which replaces the last putative interval of stasis with an unbiased random walk (Table 2).

Of the homogenous evolutionary dynamics, the general and unbiased random walk models perform about equally well (Table 2). Thus, if only homogenous dynamics are considered, a directional trend is tenable but not conclusive. Postulating a single unsampled punctuation event (model 4, Table 2) worsens model fit, but allowing for two unsampled punctuations produces large improvements in both log-likelihood and AIC_C scores (model 5, Table 2). This model has different evolutionary variances for each segment of stasis; fluctuations are very small during the middle interval, and much larger in the first and third segments. This difference in evolutionary variance explains why the ninth population, while closer in thoracic width to the second optimum (θ_2) than the first (θ_1), is nevertheless inferred to be part of the first segment. The spread around θ_2 is consistently very small (the maximum-likelihood estimate of ω_2 is in fact zero) and so it is more probable that this population is a relatively outlying deviation from the less tightly clustered first optimum (Fig. 3).

Although postulating two unsampled punctuations is markedly better than all homogenous dynamics models, the case for stasis in the third segment is rather weak. Positing instead a shift to an unbiased random walk rather than stasis after the second segment produces a model that is by far the best supported, and accounts for nearly all of the Akaike weight (model 6, Table 2). This favored model shares important features with the two-punctuation model. Both have two intervals of stasis separated by an unsampled punctuation between the ninth and tenth samples, and both result in exactly the same parameter estimates for these segments. In addition, both also imply a shift in evolutionary

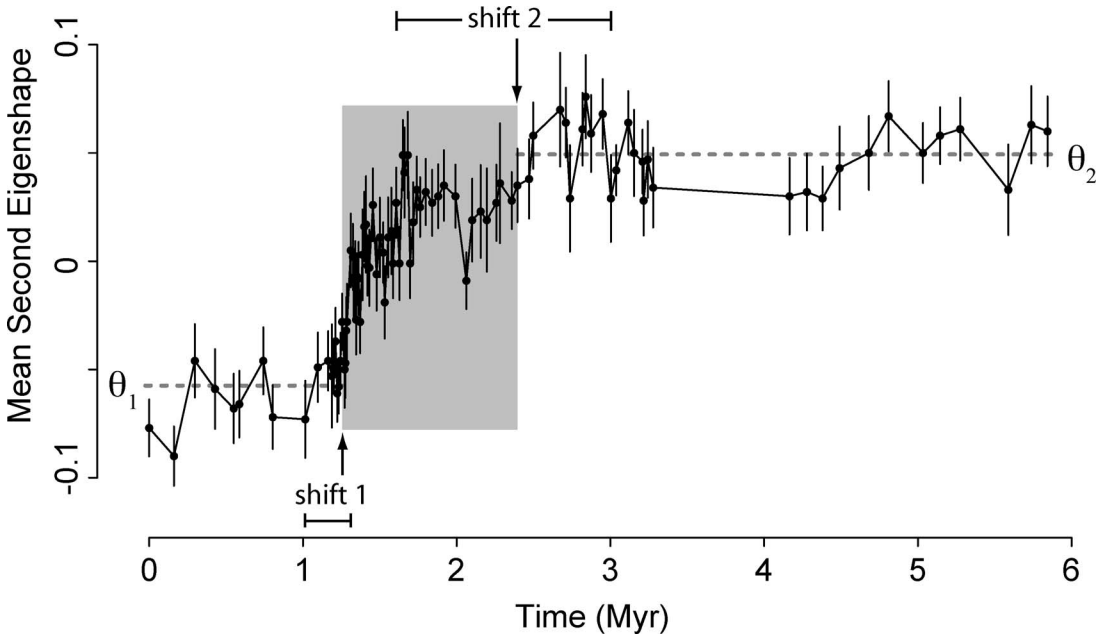

FIGURE 4. Test shape (second eigenshape score) for the foraminifera lineage *Globorotalia* (Malmgren et al. 1983). Vertical error bars indicate 95% confidence intervals on sample means, and time is measured in millions of years after the first sample. Shown is the best-supported model, which implies a single sampled punctuation event (from optimum θ_1 to θ_2), with the transition zone exhibiting an unbiased random walk (gray rectangle). The indicated shift points (arrows) delimit the maximum-likelihood estimates of the transition interval, and the 95% confidence interval of the shift points are indicated as horizontal error bars around the shift. The onset of the punctuated interval (shift 1) is estimated fairly precisely, but the subsequent return to stasis (shift 2) is much more poorly constrained.

dynamics after the 22nd sample. These models differ only as to whether the third segment fluctuates (stasis) or drifts (random walk).

For the best-supported model, the timing of the first shift in dynamics is well constrained; all solutions in the 95% confidence set shift after the ninth sample. The second transition, from the second stasis optimum to random walk dynamics, is somewhat less tightly inferred. The best solution has the evolutionary dynamics changing after the 22nd sample, but the 95% confidence region includes solutions in which the shift occurs up to four samples earlier.

Thus, body size evolution in *Pseudocubus vema* appears to be neither uniformly gradual nor completely pulsed. Instead, the net evolutionary increase in thoracic width is best ac-

TABLE 2. Model fits to *Pseudocubus vema* thoracic width data. All punctuations are of the unsampled variety (see text); models 4 and 5 posit one and two punctuations, respectively. The punctuated models fit here assume the separate evolutionary variances for each interval of stasis. Model 6 postulates a single punctuation, and then a later transition to an unbiased random walk. Parameter estimates for the best-supported model (6): $\theta_1 = 4.52$ log μm, $\theta_2 = 4.67$ log μm, $\omega_1 = 0.0017$ log μm², $\omega_2 = 0$ log μm², $\sigma^2_{step} = 0.018$ log μm/Myr; shift points after the ninth and 22nd samples. Akaike weights for models receiving at least moderate support ($w > 0.2$) are in bold. Abbreviations as for Table 1.

No.	Model	No. of segments	K	log(L)	AIC_C	w
1	Unbiased random walk	1	1	53.71	−105.29	0.000
2	Directional evolution	1	2	54.81	−105.21	0.000
3	Stasis	1	2	27.56	−50.73	0.000
4	Punctuation-1	2	5	47.67	−83.11	0.000
5	Punctuation-2	3	8	69.87	−117.74	0.040
6	Punctuation-1, then unbiased random walk	3	7	71.28	−124.08	**0.960**

counted for by a single punctuated increase, followed later by a shift to an unbiased random walk.

Punctuated Gradualism in *Globorotalia*

Malmgren and colleagues (1983) measured test size and shape in the planktonic foraminifera lineage *Globorotalia tumida* over roughly the last 10 Myr. Over this interval, they documented a pattern of character change that they called ''punctuated gradualism.'' According to their interpretation, this lineage initially experienced stasis, which then shifted to a fluctuating but generally directional mode of evolution for about 0.6 Myr, after which the lineage experienced stasis for the remainder of the study interval (Fig. 4). This general pattern of change matches the three-stage description of a sampled punctuation given above (stasis–directional change–stasis; Fig. 1B), although the authors are somewhat equivocal as to whether the transitional interval features truly directional evolution. They present several tests that show that an unbiased random walk cannot be rejected for this transition zone, but nevertheless they seem to prefer directional over non-directional change for this interval (Malmgren et al. 1983: p. 382).

This data set has been reanalyzed several times since Malmgren et al.'s original publication. Different studies have focused on the size (test area) or shape (second eigenshape score) traits separately, but these two variables are highly correlated (log area and second eigenshape score, $r = 0.93$, $p < 0.00001$) and show very similar patterns of overall change. Both of the detailed reanalyses of this example—Bookstein 1987 (test size) and Roopnarine 2001 (test shape)—agree with Malmgren et al. (1983) that there are three stages in the evolution of these traits and that net rates of change are highest during the middle, transitional interval (see also Charlesworth 1984; Kitchell et al. 1987; MacLeod 1991). Roopnarine's interpretation paralleled Malmgren et al.'s closely except that he rejected directionality in favor of an unbiased random walk for the punctuation interval. Bookstein's (1987) interpretation differed more fundamentally from Malmgren et al. in that he denied any evidence for stasis in this lineage. In his view, this sequence was better explained as a series of three unbiased random walks with differing step variances. In summary, there are two points of disagreement among these authors: (1) whether the beginning and ending segments represent stasis or unbiased random walks, and (2) whether evolution during the transition phase is inherently directional.

I reanalyzed foraminifera test shape (second eigenshape score) using MacLeod's (1991) ''best case'' chronology for this core. Because of the stratigraphic uncertainty surrounding the ten oldest samples (Malmgren et al. 1983), these samples were not included in the analysis. Omitting these data points shortens the duration of the first segment of the sequence but does little to alter the overall character trajectory. To the remaining 95 samples (Fig. 4), I fit three uniform and four heterogeneous dynamics models, the latter including one for each of the proposed explanations for this pattern (Table 3).

Of the homogeneous models, the unbiased random walk is the best supported (Table 3). Elaborating on this model to allow for a random walk with two and then three different step variances (models 4 and 5) substantially improves model support with each additional parameter. Consequently, Bookstein's three-stage random walk is markedly better than any of the simpler models of evolutionary change (Table 3). However, modeling the beginning and ending segments as stasis rather than random walks (models 6 and 7) results in a rather large increase in model support, and in fact, these two models account for nearly all the Akaike weight among all models. The superior performance of these models supports the conclusion of Malmgren et al. (1983) and Roopnarine (2001) that this sequence of trait values starts and ends in stasis; these segments achieve too little net divergence to be adequately accounted for by a random walk. As for the directionality of the transition phase, modeling the transition zone as a general random walk rather than an unbiased random walk does not improve the log-likelihood enough to make up for its additional parameter (Table 3). Thus, the overall interpretation of this sequence matches that of Roopnarine: test shape experiences stasis, in-

TABLE 3. Model fits to putative punctuated anagenesis of test shape in the *Globorotalia tumida* lineage. ''Unbiased random walk-2'' and ''Unbiased random walk-3'' have two and three segments, respectively, each with separate step variance parameter values. The punctuation models (6, 7) differ as to whether the transition interval exhibits inherently directional change (7) or not (6); both assume equal evolutionary variances before and after punctuation event. Parameter estimates for the best-supported model (6): $\theta_1 = -0.057$ units, $\theta_2 = 0.049$ units, $\omega = 0.00013$ units^2, $\sigma^2_{step} = 0.013$ units^2/Myr, shift points after the 17th and 65th samples (see Fig. 4). Akaike weights for models receiving at least moderate support ($w > 0.2$) are in bold. Abbreviations as for Table 1.

No.	Model	No. of segments	K	log(L)	AIC_C	w
1	Unbiased random walk	1	1	232.18	−462.32	0.000
2	Directional evolution	1	2	232.40	−460.67	0.000
3	Stasis	1	2	167.51	−331.02	0.000
4	Unbiased random walk-2	2	3	238.57	−470.86	0.000
5	Unbiased random walk-3	3	5	243.12	−475.55	0.002
6	Sampled punctuation (unbiased random walk transition)	3	6	250.37	−487.76	**0.763**
7	Sampled punctuation (directional evolution transition)	3	7	250.50	−485.70	**0.263**

terrupted by an interval of accumulating but not inherently directional change before stasis is again resumed.

Unlike the previous two cases, this example exhibits considerable uncertainty in the timing of shifts in dynamics. The solution to the best-fit model (6) starts the transition phase after the 17th sample and resumes stasis after the 65th sample (Fig. 4). The onset of this transition phase has a moderately narrow 95% confidence interval that ranges from about 250 Kyr before to about 50 Kyr after the estimated shift point (see Fig. 4). The confidence region for the end of the punctuated interval is much broader, including samples 750 Kyr before and 600 Kyr after the estimated shift point. Thus, although the estimated duration of the transition interval according to model 6 is approximately 1.1 Myr, the much shorter duration estimated by previous authors (~600 Kyr) is also consistent with these results. It is worth noting that the only other model with any substantial support (model 7) has identical estimated shift points to model 6, and nearly the same confidence intervals on those shift points.

Bookstein's three-stage unbiased random walk model does not perform as well as models that invoke stasis for these data, but he analyzed test size rather than test shape. Although size and shape are highly correlated, they do differ in the degree of constancy in their beginning and ending segments (Malmgren et al. 1983; compare their figures 4 and 5). Test size meanders more than shape in these portions of the sequence, and in fact Bookstein's three-stage unbiased random walk does outperform all other models for the size data (results not shown). This difference is in line with the general finding that shape characters tend to exhibit stasis more so than size-related features (Hunt 2007).

Age Model Error

General Considerations.—Age models are approximations that are sometimes systematically biased and always subject to random error. As a result, it is important to consider how errors in age determination may influence the results of any evolutionary analysis.

It is useful to begin by noting that some kinds of age model error do not influence these analyses at all. Uniform additive error has no effect whatsoever on any of these models; all parameter estimates and log-likelihoods are unchanged if each sample's age is underestimated or overestimated by a fixed amount. Uniform multiplicative errors—for example, if sedimentation is correctly inferred to be constant but at a incorrect rate—affect the absolute value of parameter estimates but do not change log-likelihoods. Constant multiplicative errors are equivalent to changing the units with which time is measured, which has no effect on model support (Hunt 2006). The stasis model and its elaborations (such as unsampled punctuations) are furthermore insensitive to all errors that do not alter the order in which samples occur, because evolution is not time-dependent under stasis (i.e.,

TABLE 4. Sensitivity of the empirical examples to age model error. For each of the three empirical examples, the second column lists the best-supported model and the third column its degree of support (Akaike weight) assuming the original age model. Next is listed the Akaike weight for that same model using a revised or alternative age model (if available). The last column lists the range of Akaike weights observed in the middle 90% of randomized age models. These simulated age models assume that the order of samples is correct but that the relative spacing given by the age model is completely unreliable (see text). These simulations included 1000 replicates for the *Flexicalymene* data, 200 replicates for *Pseudocubus*, and 100 replicates for *Globorotalia*. Larger data sets and more complex models produced longer run times, and the number of replicates was scaled accordingly.

		Akaike weights		
Lineage	Best model	Original ages	Alternative ages	Randomized ages
Flexicalymene	Punctuation-1	0.505	—	0.380–0.510
Pseudocubus	Punctuation-1, then unbiased random walk	0.960	0.958	0.548–0.982
Globorotalia	Sampled punctuation (unbiased random walk transition)	0.763	0.731	0.672–0.762

elapsed time does not enter into the likelihood calculations, Hunt 2006: eq. 9). Thus, unless the actual succession of samples is incorrectly inferred, parameter estimates and log-likelihoods are not affected. Directional evolution and random walks are time dependent, however, and so their parameter estimates and log-likelihoods are influenced by age errors. Moreover, as long as at least one of the evolutionary models under consideration is time dependent, the relative support for different candidate models will change whenever the temporal spacing of samples is altered. The sensitivity of an analysis can range from trivial to substantial, depending on the particular example.

Updating Age Models.—None of the examples analyzed in this paper were published recently, and their age models rely on correspondingly dated chronostratigraphy. For example, the calculated sedimentation rates for Kellogg's *Pseudocubus* data are tied to absolute age estimates for magnetic reversals that are over 40 years old (Kellogg 1975). Using the more modern values in Gradstein et al. (2004) changes the inferred sedimentation rates in different parts of the core but has almost no effect on evolutionary inference. Even with more current age determinations, the model of an unsampled punctuation followed by a transition to an unbiased random walk is still overwhelmingly supported (Table 4).

MacLeod (1991) explored the role of age model uncertainty in the case of *Globorotalia*, and the preceding analysis used his ''best-case'' chronology. To explore a maximally different but still plausible age model, I reanalyzed these data using MacLeod's ''worst-case'' chronology (''best'' and ''worst'' refer to models that minimize and maximize, respectively, variation in sediment accumulation rates through the core). As in the *Pseudocubus* example, the relative support for the candidate evolutionary models is essentially unaffected by using an alternative age model (Table 4).

The *Flexicalymene* sequence has perhaps the greatest vulnerability to age model error because it includes specimens collected at multiple stratigraphic sections. In addition to the normal sources of age error within a site, this age model also depends crucially on the regional correlation of Cisne and Rabe (1978). Although the volcanic ash marker beds used to correlate different sections are still deemed reliable, some of the stratigraphic concepts from Cisne and Rabe (1978) have not been supported by more recent work (Brett and Baird 2002). Unfortunately, it is not possible to update this sequence with revised ages because Cisne et al. (1980) did not report detailed information for their numerous samples, instead summarizing evolutionary patterns using a composite standard section. As a result, it is difficult to reevaluate the robustness of evolutionary inferences in the light of improved chronostratigraphic understanding. However, this example's best-supported model—a single unsampled punctuation—is robust to age model error, provided that samples are at least correctly assigned to pre- and post-punctuation intervals. Reexamination of the stratigraphic context of the original sam-

ples or new collecting from the region should be able to resolve this uncertainty.

Simulating Lousy Age Control.—Although uncertainty remains for the *Flexicalymene* example, evolutionary inferences for the two other case studies appear to be robust to age model revision. Age models like these routinely assume constant sedimentation rates at some scale of stratigraphic resolution, yet most depositional environments are likely to exhibit short-term variation in sediment accumulation rates, which will result in random, point-to-point errors in determining sample ages. One could explore this kind of age model error in many ways. In this section, I assume a near worst-case scenario of extremely poor age control. If conclusions are robust to this unrealistically high age model error, it is likely that more plausible errors will also be of no consequence.

The age model error for a particular sequence of n samples was simulated as follows. The age of the oldest sample was set to zero and time was counting forward to the time elapsed until the last sample, t. Next, n ages were drawn at random from a uniform distribution between zero and t. Once sorted, these were assumed to be the ages of the n samples. After this sorting, a minimum spacing between samples was enforced to avoid unrealistically short durations between samples. The value of this minimum spacing was set to be much smaller than the minimum observed time between adjacent samples in the age model—10 Kyr for *Flexicalymene*, 5 Kyr for *Pseudocubus*, and 1 Kyr for *Globorotalia*. In essence, this procedure models a situation in which the order of samples is reliable but the relative temporal spacing of samples is very poorly constrained, with an abundance of undetected short-term variation in sediment accumulation rates.

In general, this rather extreme age error has only modest effect on the analyses reported here (Table 4). The *Flexicalymene* data almost always fit the punctuational model best (Table 4), with most of the remaining Akaike weight supporting an unbiased random walk. Conclusions about the *Pseudocubus* lineage are also robust, although in some cases the two-punctuation model (model 5 in Table 2) accounts for a much larger portion of the available Akaike weight than when the actual or updated age models are used (Table 4). Similarly, the *Globorotalia* results under these degraded age models differ only modestly from their values under the actual age models (Table 4). These results do not necessarily imply that all conceivable age errors do not matter, but at least the conclusions supported here are not overly dependent on the details of the age models used.

Discussion

Fortey (1985) dissects some of the interpretive difficulties inherent to the gradualism/punctuated equilibria dichotomy. He notes an asymmetry between the two paradigms in that, even when morphological and stratigraphic intermediates indicate gradual change, a pulsed interpretation can be shielded from rejection by postulating additional punctuation events between sampled populations. Fortey suggests that the principle of parsimony can be used to prefer gradualistic models because they subsume all the observations with fewer ad hoc events (punctuations), but he does not indicate exactly how simplicity should be weighed against the improved fit of more complicated, punctuational models.

The approach outlined in this paper can be viewed as a statistical implementation of Fortey's reasoning. In essence, AIC and related metrics achieve a compromise between goodness-of-fit (measured as log-likelihood) and parsimony (measured as the number of model parameters). Relative to uniform evolutionary change, explanations invoking multiple punctuations are penalized because they are more complicated—they require more parameters to specify. Therefore, punctuational models should be preferred only when their log-likelihood advantage more than offsets their increased complexity.

In this paper, I have used AIC_C to balance model fit and complexity. This metric is widely used and has a firm epistemic foundation (Forster and Sober 1994). There are some concerns that this metric may unduly favor models of high complexity (e.g., Link and Barker 2006), and other metrics, such as the Bayes In-

formation Criterion (BIC) are sometimes used instead to avoid this over-fitting. Most alternatives, however, ultimately rely on log-likelihood as a measure of model fit and so can be implemented rather easily in the framework described here. In practice, whether the AIC_C over-fits may depend on the complexity of the phenomena being modeled. The BIC may outperform the AIC when the generative processes have only a few parameters, but the opposite can be true when reality is highly dimensional (Forster 2001; Burnham and Anderson 2004). Given the interplay of ecology and population genetics that ultimately governs the evolutionary trajectory of lineages, it seems likely that the evolutionary reality to be modeled is usually rather complex, and thus the AIC_C may be particularly justified in this context.

There have been several previous attempts to formulate gradual and punctuational models in statistical terms. A few statistical tests for rate heterogeneity have been developed (Charlesworth 1984; Kitchell et al. 1987), but evidence for rate heterogeneity is less useful than explicit comparisons of pulsed and gradual explanations. An early attempt at the latter was made by Bookstein et al. (1978), and it shares with the present approach the strategy of fitting sets of models and choosing the best performing model or models. However, that approach fits evolutionary models via linear regression assuming independence among different stratigraphic levels, an assumption that is violated for most kinds of evolutionary change.

More recently, Roopnarine (2001) developed a procedure to examine evolutionary series for evidence of changing evolutionary dynamics. This procedure distinguished unbiased random walks from other evolutionary dynamics by using the Hurst exponent, a metric that relates the trait range and temporal span of a sequence (Roopnarine et al. 1999; Roopnarine 2001). A randomization procedure is used to test observed Hurst exponents against a null model of an unbiased random walk, and the analysis is repeated in a moving window through the sequence to explore if the dynamics change over time. In the case of the *Globorotalia* lineage, this approach leads to an account of morphological evolution that is very similar to the best-supported model here. However, because it employs randomization tests of a null model, it cannot easily be used to compare the statistical support among competing explanations.

Most of the models fit to data in this paper derive from qualitative accounts of previous workers. However, these prior accounts were, at least in part, inspired by the actual pattern of trait changes over time. This poses a potential problem for statistical inference because models should not be derived from the data to which they are applied (Bookstein 1987; Hunt 2006). One possible resolution to this difficulty is to decide on a set of canonical models that should be applied routinely to all data sets. Several classes of models could be included, and within a class (e.g., unsampled punctuations), one would fit a suite of models (e.g., one, two, three, etc. punctuations) rather than just the specific model that looks most promising. In practice, it will not be fruitful to fit very complex models unless a sequence has a very large number of samples—the parsimony component to AIC_C scores will strongly penalize complex models when there are few samples. According to this strategy, candidate models are decided a priori and there is no danger of unfairly focusing on those tailored to a particular data set. However, the class of potential heterogeneous models is very large, and it may not be easy to decide ahead of time which models should be included in a standard analysis. Absent an established set of canonical models, it is probably prudent to demand a high level of support from models that are potentially inspired by the data to be analyzed.

Thus far, I have not much considered possible artifactual causes of punctuated patterns. Ecophenotypic response to environmental change, immigration of related taxa into the study area, and condensed stratigraphic intervals can all produce patterns that mimic punctuated evolution within a lineage. Ecophenotypy and lineage replacement are always difficult to evaluate in fossils, although some understanding of how the studied traits vary with environmental conditions can help. Regardless, to the extent that these two factors

operate, they will confound any attempts to analyze evolutionary changes in fossil lineages, and so the problems they pose are not specific to the methods described here. The possibility that a punctuation is an artifact of normal rates of change observed in a stratigraphic section with missing or condensed time can be evaluated from sedimentological and stratigraphic evidence, and the effects of these and other age model errors can be evaluated by simulation, as was done for the examples presented here. Alternatively, methods are now being developed that attempt to infer parameters of morphological evolution that explicitly employ uncertainty in sample ages and other geological parameters (Hannisdal 2007).

In Eldredge and Gould's (1972) original formulation, the pace of phyletic gradualism is said to be ''even and slow.'' Here, I operationalize gradual change as that occurring under a uniform evolutionary dynamic. In practice, when punctuational interpretations are tenable, the best-supported uniform model is likely to be some kind of random walk. I believe that this formulation of gradualism corresponds well to most historical uses of the term (e.g., Gingerich 1985), although some authors might add a requirement of directionality (e.g., Levinton 2001: p. 320). Regardless, my main concern is not historical fidelity but rather scientific utility. After all, many unresolved arguments from this debate involved defining exactly what is meant by terms such as ''gradual'' and ''sudden.'' Evaluating punctuational hypotheses in terms of homogeneous versus heterogeneous dynamics has the virtues of being reasonable, biologically interesting, and above all, unambiguously testable.

The methodological framework described in this paper allows hypotheses of punctuated evolution to be evaluated, but it is actually much more general, encompassing any evolutionary scenario in which dynamics change over time. A subset of these heterogeneous dynamics models imply pulsed change, but many other biologically interesting models do not (e.g., Bookstein's multi-stage random walk). Theory necessarily influences observation, and one unintended consequence of the gradualism—punctuated equilibria dichotomy may have been that diverse evolutionary patterns have been coerced into two exclusive categories (Bookstein et al. 1978; Malmgren et al. 1983), perhaps to the detriment of our understanding of evolutionary patterns as preserved in the fossil record.

Summary

1. The question of gradual versus pulsed evolution can be made tractable by focusing on the homogeneity of evolutionary dynamics. Gradual explanations invoke uniform dynamics within sequences and punctuational models imply heterogeneity in the rules governing evolutionary divergence.

2. The class of heterogeneous dynamics models includes some that imply pulsed evolutionary change. These models can be assessed statistically by extending the likelihood framework used to evaluate the traditional homogeneous modes of evolutionary change (directional change, unbiased random walk, stasis).

3. Two kinds of punctuated change are particularly useful to consider in this framework: (a) unsampled punctuations, in which the evolutionary change is too rapid relative to the temporal spacing of samples to be captured, and (b) sampled punctuations, in which the transitional period of rapid change is represented by intermediate populations. Unsampled punctuations are conveniently modeled as two periods of stasis with different optima; the punctuation results from an instantaneous shift from one optimum to the other. When more than a few intermediate populations are known, evolutionary change will usually be better modeled as a sampled punctuation. In this kind of punctuation, the transitional period is modeled explicitly as a period of directional (or at least accumulating) change sandwiched between intervals of stasis.

4. This framework was applied to test several empirical claims of punctuated change. Results support the finding of punctuated change in the pygidia of *Flexicalymene*, although a uniform random walk is also plausible. The evolutionary trend of increasing size in the radiolarian *Pseudocubus vema* was neither purely gradual nor completely pulsed. Instead, the best-supported model entails a sin-

gle punctuated increase, followed later by a shift to an unbiased random walk. Finally, test shape in the foraminifera *Globorotalia* was best accounted for as a sampled punctuation, with a transitional period of elevated (but not directional) change intercalated between intervals of stasis. Simulations suggest that these interpretations are robust to plausible errors in the age models, although some uncertainty remains for the *Flexicalymene* time series.

Acknowledgments

This manuscript benefited from discussions with P. Wagner and P. Novack-Gottshall, and from a careful reading by K. R. Thomas. I thank S. Wang and B. Hannisdal for their thoughtful and constructive reviews. I am particularly grateful to B. Hannisdal for insisting that I not cop out about age model error.

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912–923.

Bookstein, F. L. 1987. Random walk and the existence of evolutionary rates. Paleobiology 13:446–464.

Bookstein, F. L., P. D. Gingerich, and A. G. Kluge. 1978. Hierarchical linear modeling of the tempo and mode of evolution. Paleobiology 4:120–134.

Brett, C. E., and G. C. Baird. 2002. Revised stratigraphy of the Trenton Group in its type area, central New York State: sedimentology and tectonics of a Middle Ordovician shelf-to-basin succession. Physics and Chemistry of the Earth 27:231–263.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference. Understanding AIC and BIC in model selection. Sociological Methods and Research 33:261–304.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. American Naturalist 164:683–695.

Charlesworth, B. 1984. Some quantitative methods for studying evolutionary patterns in single characters. Paleobiology 10: 308–318.

Cisne, J. L., and B. D. Rabe. 1978. Coenocorrelation: gradient analysis of fossil communities and its applications to stratigraphy. Lethaia 11:341–364.

Cisne, J. L., G. O. Chandlee, B. D. Rabe, and J. A. Cohen. 1980. Geographic variation and episodic evolution in an Ordovician trilobite. Science 209:925–927.

Eldredge, N., and S. J. Gould. 1972. Punctuated equilibria: an alternative to phyletic gradualism. Pp. 82–115 *in* T. J. M. Schopf, ed. Models in paleobiology. Freeman, Cooper, San Francisco.

Eldredge, N., J. N. Thompson, P. M. Brakefield, S. Gavrilets, D. Jablonski, J. B. C. Jackson, R. E. Lenski, B. S. Lieberman, M. A. McPeek, and W. I. Miller. 2005. The dynamics of evolutionary stasis. *In* E. S. Vrba and N. Eldredge, eds. Macroevolution: diversity, disparity, contingency. Paleobiology 31(Suppl. to No. 2):133–145.

Erwin, D. H., and R. L. Anstey. 1995. Speciation in the fossil record. Pp. 11–38 *in* D. H. Erwin and R. L. Anstey, eds. New approaches to speciation in the fossil record. Columbia University Press, New York.

Estes, S., and S. J. Arnold. 2007. Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales. American Naturalist 169:227–244.

Forster, M. 2001. The new science of simplicity. Pp. 83–119 *in* A. Zellner, A. Keuzenkamp, and M. McAleer, eds. Simplicity, inference and modelling. Cambridge University Press, Cambridge.

Forster, M., and E. Sober. 1994. How to tell when simpler, more unified or less *ad hoc* theories will provide more accurate predictions. British Journal of the Philosophy of Science 45:1–35.

Fortey, R. A. 1985. Gradualism and punctuated equilibria as competing and complementary theories. Special Papers in Palaeontology 33:17–28.

Gingerich, P. D. 1985. Species in the fossil record: concepts, trends, and transitions. Paleobiology 11:27–41.

———. 1993. Quantification and comparison of evolutionary rates. American Journal of Science 293-A:453–478.

Gould, S. J. 2002. The structure of evolutionary theory. Belknap Press of Harvard University Press, Cambridge.

Gould, S. J., and N. Eldredge. 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology 3: 115–151.

Gradstein, F. M., J. G. Ogg, and A. G. Smith, eds. 2004. A geological time scale 2004. Cambridge University Press, Cambridge.

Hannisdal, B. 2006. Phenotypic evolution in the fossil record: numerical experiments. Journal of Geology 114:133–153.

———. 2007. Inferring phenotypic evolution in the fossil record by Bayesian inversion. Paleobiology 33:98–115.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution 50:1404–1417.

Hoffman, A. 1989. Arguments on evolution. Oxford University Press, New York.

Hunt, G. 2004. Phenotypic variation in fossil samples: modeling the consequences of time-averaging. Paleobiology 30:426–443.

———. 2006. Fitting and comparing models of phyletic evolution: random walks and beyond. Paleobiology 32:578–601.

———. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. Proceedings of the National Academy of Sciences USA 104: 18404–18408.

———. 2008. PaleoTS: modeling evolution in paleontological time-series, Version 0.3–1.

Hunt, G., M. A. Bell, and M. P. Travis. 2008. Evolution toward a new adaptive optimum: phenotypic evolution in a fossil stickleback lineage. Evolution 62:700–710.

Hurvich, C. M., and C.-L. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.

Jackson, J. B. C., and A. H. Cheetham. 1999. Tempo and mode of speciation in the sea. Trends in Ecology and Evolution 14: 72–77.

Kalinowski, S. T., and M. L. Taper. 2005. Likelihood-based confidence intervals of relative fitness for a common experimental design. Canadian Journal of Fisheries and Aquatic Science 62:693–699.

Kellogg, D. E. 1975. The role of phyletic change in the evolution of *Pseudocubus vema*. Paleobiology 1:359–370.

Kitchell, J. A., G. Estabrook, and N. MacLeod. 1987. Testing for equality of rates of evolution. Paleobiology 13:272–285.
Levinton, J. S. 2001. Genetics, paleontology, and macroevolution. Cambridge University Press, Cambridge.
Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel inference. Ecology 87:2626–2635.
Lovy, D. 1996. WinDig, Version 2. 5.
MacLeod, N. 1991. Punctuated anagenesis and the importance of stratigraphy to paleobiology. Paleobiology 17:167–188.
Malmgren, B. A., W. A. Berggren, and G. P. Lohmann. 1983. Evidence for punctuated gradualism in the Late Neogene *Globorotalia tumida* lineage of planktonic foraminifera. Paleobiology 9:377–389.
Martins, E. P., J. A. F. Diniz-Filho, and E. A. Housworth. 2002. Adaptive constraints and the phylogenetic comparative method: a computer simulation test. Evolution 56:1–13.
Meeker, W. Q., and L. A. Escobar. 1995. Teaching about approximate confidence regions based on maximum likelihood estimation. American Statistician 49:48–53.
R Development Core Team. 2007. R: a language and environment for statistical computing, Version 2.6.1. R Foundation for Statistical Computing, Vienna.
Roopnarine, P. D. 2001. The description and classification of evolutionary mode: a computational approach. Paleobiology 27: 446–465.
Roopnarine, P. D., G. Byars, and P. Fitzgerald. 1999. Anagenetic evolution, stratophenetic patterns, and random walk models. Paleobiology 25:41–57.
Sheets, H. D., and C. E. Mitchell. 2001. Why the null matters: statistical tests, random walks and evolution. Genetica 112–113:105–125.
Wagner, P. J. 2000. Likelihood tests of hypothesized durations: determining and accommodating biasing factors. Paleobiology 26:431–449.