EVOLUTIONARY PATTERNS WITHIN FOSSIL LINEAGES: MODEL-BASED ASSESSMENT OF MODES, RATES, PUNCTUATIONS AND PROCESS

GENE HUNT
Department of Paleobiology
National Museum of Natural History, Smithsonian Institution
NHB, MRC 121, P.O. Box 37012
Washington DC 20013-7012

Abstract—Patterns of phenotypic change documented in the fossil record offer the only direct view scientists have of evolutionary transitions arrayed over significant durations of time. What lessons should be drawn from these data, however, has proven to be rather contentious. Although we as paleontologists have made great progress in documenting the geological record of phenotypic evolution with greater thoroughness and sophistication, these successes have been limited by the use of verbal models of how phenotypes change. Descriptive terms such as "gradual" have been understood differently by different authors, and this has led to completely incompatible summary statements about the fossil record of morphological evolution. Here I argue that the solution to this ambiguity lies in insisting that different evolutionary interpretations be represented as explicit, statistical models of evolution. With such an approach, the powerful machinery of likelihood-based inference can be help resolve long-standing paleontological questions.

Here I first review this approach and some aspects of its implementation. Then, I show how this approach leads to new traction on important issues in evolutionary paleobiology, including: understanding modes of evolution and determining their relative importance, separating evolutionary mode from tempo, assessing the evidence for hypotheses of punctuated change, and detecting adaptive evolution in the fossil record.

INTRODUCTION

Of the many lessons that can be learned from the protracted Punctuated Equilibrium debates, two seem particularly lasting. First, paleontology is often data-limited. The debate initially focused much attention on the rather few cases that met the minimum requirements to assess evolutionary mode within fossil lineages: numerous samples, good age control, and carefully measured morphology. Controversy motivated many more studies documenting evolutionary changes in species-level lineages, and we now have a pool of relevant studies that is much larger (Levinton, 2001; Gould, 2002; Hunt, 2007), but still unfortunately thin for many taxa and depositional environments.

A second lesson, obvious even in the midst of the controversy, is that different scientists can draw radically incompatible conclusions from the same set of observations (Gould and Eldredge, 1977; Gingerich, 1985). Indeed, a subtext to much of the disagreement about fossils was the interplay between theory and observation, and how the former frames perception of the latter (Eldredge and Gould, 1972; Fortey, 1988). It was recognized early on that battling verbal descriptions would not resolve these conflicts, and in response a variety of statistical methods were developed (Raup, 1977; Bookstein, 1987; Bookstein, 1988; Gingerich, 1993; Roopnarine et al., 1999; Roopnarine, 2001). These methods employed as a null model a random walk, which is a simple model in which trait increases and decreases are equiprobable. The effect of these developments was to inject much-needed quantitative rigor into discussions that were previously focused mostly on visual impressions.

As useful as these approaches are, there are some limitations inherent to treating the random walk as a null model. Simulation studies showed that these tests can have very low statistical power, and therefore failing to reject the null of a random walk—by far the most common outcome of these analyses—provides little information about the nature of evolutionary changes (Roopnarine et al., 1999; Sheets and Mitch-

In *From Evolution to Geobiology: Research Questions Driving Paleontology at the Start of a New Century, Paleontological Society Short Course, October 4, 2008. Paleontological Society Papers, Volume 14, Patricia H. Kelley and Richard K. Bambach. (Eds.).*

ell, 2001; Hannisdal, 2006). Moreover, what was most urgently needed was the ability to measure statistical support for plausible alternative interpretations, but this is not easily implemented in the null hypothesis-testing framework.

In this paper, I advocate a particular approach for guiding the interpretation of evolutionary change in fossil lineages. The key to this approach is that it insists that all candidate evolutionary explanations be represented as concrete statistical models, which are then evaluated using the standard machinery of likelihood-based inference. A great many practical and theoretical benefits follow from this simple starting point. The structure of this argument will be as follows. First, I start with an overview of this analytical framework and then address a few aspects of its implementation. Then, I will attempt to demonstrate the value of this analytical approach by using it to explore several outstanding issues in evolutionary paleobiology. In the final section, I briefly consider some of the most promising avenues for future work.

FROM VERBAL TO STATISTICAL MODELS

Since Raup's pioneering papers (Raup, 1977; Raup and Crick, 1981), scientists examining evolution within fossil lineages have mostly considered three canonical modes of evolution: directional evolution, random walks and stasis. Of these, however, only the random walk was defined explicitly. Support for the other modes was discerned from long-term evolutionary divergences that were either too great or too limited to be explained by a random walk. These departures from a random walk were attributed to directionality and stasis, respectively (Raup and Crick, 1981; Bookstein, 1987; Gingerich, 1993; Roopnarine, 2001), but these two modes were never fit and compared in their own right.

Over the past few years, I have developed a statistical framework with which these and other kinds of evolutionary dynamics can be represented as statistical models, each of which can be fit to real paleontological data and compared to one another on equal footing (Hunt, 2006). The first step in this procedure is to represent modes of change as fully statistical models. It is most convenient to start with a general version of the random walk (the general random walk of Hunt, 2006). This model occurs in discrete time increments, during each of which an evolutionary change is drawn at random from a distribution of evolutionary transitions or "steps." The long-term dynamics of this model can be shown to depend only on the mean (μ_s) and variance (σ^2_s) of this step distribution. The former determines the directionality of the sequence; on average, positive μ_s values generate increasing trends, and negative values generate decreasing trends. Increasing the step variance results in a greater range in the evolutionary increments and correspondingly more volatile evolutionary changes (Hunt, 2006).

This model provides the basis for understanding both directional evolution and random walks. Directional evolution results whenever the mean of the step distribution is not zero, although subtle trends can be obscured by the variability in evolutionary steps. Random walks are a special case of this more general model for sequences lacking directionality ($\mu_s = 0$). The paleontological literature reflects some inconsistency about what to call these two models. The model referred to here as directional evolution has elsewhere been termed a biased, directional or general random walk. Of these, the term "directional random walk" is perhaps the clearest since it includes this model's most salient feature (directionality) while indicating its underlying similarity to random walks (there are other approaches for modeling directionality that are not random walk-like, e.g., Sheets and Mitchell, 2001). The term random walk is standard, although it is sometimes modified as unbiased or symmetric to indicate its lack of directionality. It is also worth noting that in the phylogenetic methods literature, these models are usually referenced by their continuous time approximations: Brownian motion (or diffusion) for the unbiased random walk, and Brownian motion with a trend for directional evolution. This diversity of terminology has unfortunately obscured the conceptual commonality of these models across and within disciplines.

Several different approaches have been used to model stasis (e.g., Roopnarine, 2001). I follow Sheets and Mitchell (2001) in modeling stasis as normally distributed variation (with variance ω) around a stable phenotypic optimum (θ). This model is simple and analytically tractable, and it produces trait trajectories that conform well to recent qualitative accounts of stasis (Gould, 2002; Eldredge et al., 2005). Stasis

and random walks, while both lacking directionality, differ importantly in that total amount of evolutionary change does not increase with time under stasis. Random walks, by contrast, show increasing evolutionary divergences over time; they drift in morphospace, whereas lineages experiencing stasis fluctuate around a fixed point.

The above information allows the forward simulation of evolutionary sequences, but is not sufficient for the inverse problem of fitting the models, which requires an understanding of what these models predict for real evolutionary divergences. At present, there are two available parameterizations for these models, each of which employs paleontological data somewhat differently. The first considers morphological differences in each ancestor – descendant (AD) pair of populations as separate observations (Hunt, 2006), while the second simultaneously weighs the joint distribution of trait values across all sampled populations (Hunt et al., 2008). The statistical models described above make predictions as to how these different kinds of data should be distributed. Considering separate AD differences (the first parameterization), these are expected to be normally distributed with means and variances that are functions of model parameters and elapsed time (Table 1). Considering all trait means jointly (the second parameterization), an entire sequence of trait means can be considered a single draw from a multivariate normal distribution with a vector of means and covariance matrix that are also functions of the model parameters and the age model (Table 1; note that, in this parameterization, the directional and random walk models require an additional parameter, X_0, equivalent to the intercept or root parameter in comparative methods, that represents the trait mean at the start of the sequence). In either case, the density function of the normal or multivariate normal distribution allows computation of the log-likelihood of observed data, and the best fitting parameter set can be estimated by searching numerically through the space of possible parameter values and choosing those that maximize the likelihood of producing the observed data (Hunt, 2006; Hunt et al., 2008).

While log-likelihoods provide a natural measure of how well data fit models, they are not as well suited for choosing among models because log-likelihoods

	AD parameterization		Joint Parameterization	
Model	Mean AD difference	Variance AD difference	Joint Means	Joint Covariance Matrix
Directional evolution	$E[\Delta X] = \mu_s t_{AD}$	$Var[\Delta X] = \sigma_s^2 t_{AD} + \varepsilon_A + \varepsilon_D$	$E[X_i] = X_0 + \mu_s t_i$	$Var[X_i] = \sigma_s^2 t_i + \varepsilon_i$ $Cov[X_i, X_j] = \sigma_s^2 \cdot \min(t_i, t_j)$
Random walk	$E[\Delta X] = 0$	$Var[\Delta X] = \sigma_s^2 t_{AD} + \varepsilon_A + \varepsilon_D$	$E[X_i] = X_0$	$Var[X_i] = \sigma_s^2 t_i + \varepsilon_i$ $Cov[X_i, X_j] = \sigma_s^2 \cdot \min(t_i, t_j)$
Stasis	$E[\Delta X] = \theta - X_A$	$Var[\Delta X] = \omega + \varepsilon_D$	$E[X_i] = \theta$	$Var[X_i] = \omega + \varepsilon_i$ $Cov[X_i, X_j] = 0$

Table 1—Mathematical basis for fitting three evolutionary models (directional evolution, random walk, and stasis) to empirical paleontological sequences. Two different parameterizations of the problem are possible: one uses the phenotypic differences between ancestor – descendant (AD) pairs of populations, the other considers the distribution of all sample means jointly (Joint). For all models, the expected difference between ancestor and descendant is normally distributed with means and variances that are functions of model parameters and elapsed time. Under the joint approach, each sequence of trait means can be treated as a single draw from a multivariate normal distribution with a vector of means and covariance matrix given below. Model parameters: μ_s, mean of the step distribution; σ^2_s, variance of the step distribution; θ, phenotypic optimum; ω, variance around optimum; X_0, estimated trait mean at the start of the sequence. AD abbreviations: t_{AD}, time elapsed between ancestor and descendant; X_A, phenotypic mean of ancestral populations; ε_A, ε_D, sampling variances of the ancestor and descendant populations. Joint abbreviations: t_i, t_j, time elapsed from the start of the sequence to the i^{th} and j^{th} populations; ε_i, sampling variance of the i^{th} population; min, minimum; $E[x]$, expected (mean) value of x; ΔX, the difference between ancestor and descendant populations ($X_D - X_A$).

generally increase with model complexity. Models with more tunable knobs (parameters) have an unfair advantage over simpler ones, and it is therefore necessary to balance model fit against model complexity. One common metric for doing so is the Akaike Information Criterion (AIC), which is defined as AIC = 2(log-likelihood) + 2(# model parameters) (Akaike, 1974). Lower AIC scores indicate higher model support; this metric reflects the amount of information lost in approximating reality with a model. In practice, it is usually better to use a version of the AIC with a bias correction for small sample sizes called the AIC_C (Anderson et al., 2000). Relative model supports for a set of candidate models are conveniently summarized using Akaike weights, which result from a simple transformation of AIC or AIC_C scores such that total support sums to unity across the models considered (Anderson et al., 2000). The outputs from these calculations are a series of weights that indicate the proportion of total empirical support each model receives (e.g., models A, B and C may receive, respectively, 80%, 15% and 5% of the evidential support).

All the analyses described in this paper can be performed with functions provided in the R package *paleoTS* (Hunt, 2008b), which is publicly available at the Comprehensive R Archive Network (CRAN, see http://www.r-project.org/). This package is most easily downloaded and installed from within R in the usual way for CRAN packages; see software documentation for details. R is a free and cross-platform environment for statistical computing, graphics and programming (R Development Core Team, 2008).

ANALYTICAL ISSUES

Sampling Error is Surprisingly Large

In practice, analyses of trait sequences almost always treat sample means as if they were known without error. Of course they are not—all finite samples entail error in estimating mean values, and this sampling variance has a predictable magnitude equal to within-sample variance divided by the number of measured individuals (Sokal and Rohlf, 1995). In many fossil lineages, measurable individuals are scarce and morphological differences subtle, and thus sampling error may sometimes be quite large. Sampling error can be accommodated quite naturally in the expected evolutionary changes over time (Table 1), and it has been shown that this approach can accurately estimate evolutionary parameters even when data are noisy (Hunt, 2006). To my knowledge, all other methods so far proposed ignore this unavoidable fact that paleontological samples are finite.

It is conceivable that sampling error is generally unimportant although there are indications otherwise (Kinnison and Hendry, 2001; Hunt, 2006). The crucial factor is the magnitude of sampling error relative to true evolutionary differences among fossil populations. This latter quantity—true evolutionary variance—can be estimated as the variance parameter (ω) of the stasis model (this holds regardless of the underlying mode of evolution). One simple way to assess the importance of sampling error is to compare this estimate of true evolutionary variance to the total observed variance of sample means. If sampling error is important, estimates of ω will be much less than the variance among sample means taken at face value.

Judging from the sample of 251 empirical fossil sequences analyzed previously (Hunt, 2007), sampling error is quite often substantial (Fig. 1). On average, about 44% of the variation in trait means is attributable to sampling error, and it is perhaps surprisingly common for this proportion to be very close to 100% (Fig. 1). Ignoring sampling error causes one to overestimate morphological divergence by mistaking noise for true evolutionary differences. This will bias parameter estimates (Hunt, 2006) and has the potential to influence statistical inference (see below). Mundane though it may be, sampling error is nevertheless practically important and should be accounted for in analyzing evolutionary sequences.

Model Selection Performance

Hunt (2006) presented simulations showing that the AD approach can provide good estimates of model parameters, even with noisy and incomplete fossil data. The performance of these methods in terms of model selection has been less explored, and previous results suggest that some models may be easier to detect analytically than others (Sheets and Mitchell, 2001; Hannisdal, 2006). In particular, the stasis model, because it is mathematically similar to sampling error (i.e., both produce uncorrelated Gaussian variation), may be both easier to detect and unduly favored in noisy sequences. In this section, I present some simulations

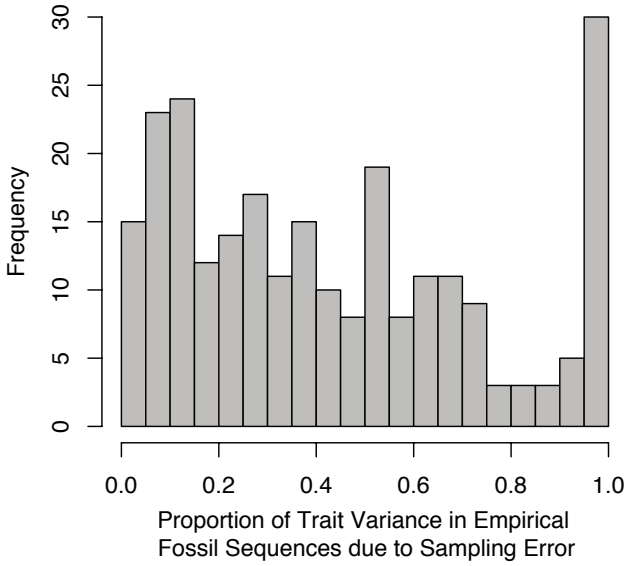

Figure 1—Sampling error in empirical fossil sequences. Histogram shows for 251 fossil time-series, the proportion of observed variance in trait means that is attributable to sampling error. Sampling error is often large—on average, nearly half of all variation in trait means observed in fossil sequences is noise.

that explore the performance of AIC-based model selection criteria under three evolutionary scenarios: a moderate directional trend, an unbiased random walk, and stasis.

These simulations were designed to investigate the relative performance of the two available parameterizations (AD and joint) for a range of sequence lengths and magnitudes of sampling error. There is an enormous range of evolutionary scenarios one could consider and this investigation provides only preliminary guidance for a few empirically reasonable situations. Across all simulations, within-sample phenotypic variance and the elapsed time between each sampled populations were set to unity. Absolute values for these do not matter because statistical inference is insensitive to changing units of both time and morphology (Hunt, 2006). Simulations were run at two different levels of sampling error—relatively low (n = 50 observations per sample), and relatively high (n = 5 observations per sample).

The first scenario investigated is that of a modest directional trend (Fig. 2, left column). The mean step was chosen to produce, over the entire sequence, a net increase of two units of within-population standard deviations. This is an average over many realizations, some of which would show stronger and some weaker trends. The step variance was set so that 99% of simulated sequences would show net increase over time; if the step variance is too large, volatility in steps will swamp directionality. The second scenario is that of an unbiased random walk (Fig. 2, center column), with the step variance chosen such that the standard deviation of trait means at the end of the sequence was equal to twice the within-sample phenotypic standard deviations. The final scenario considered is that of evolutionary stasis, with the evolutionary variance (ω) set equal to the within-sample phenotypic variance (Fig. 2, right column). These particular scenarios were investigated in detail because they produce evolutionary sequences with realistically moderate evolutionary divergence—generally larger than sampling error, but not overwhelmingly so.

Model selection performance under these three scenarios is summarized in Figure 2. I will focus here on three main results. First, while both parameterizations generally perform similarly, the joint parameterization is better able to correctly identify directional trends. This advantage increases with sequence length, and is most pronounced when sampling error is high (Fig. 2). High sampling error obscures the point-to-point differences employed by the AD parameterization, rendering this approach less effective for time-series that are long and noisy. Sampling error does not accumulate, however, and so noise has less effect when considering the distribution of all sample means jointly. Second, sampling noise generally increases the support for the stasis model, and lessens support for the directional change and random walk models (Fig. 2). This is the expected effect because stasis and sampling noise have the same mathematical form. Third, stasis is relatively easy to correctly identify, particularly in sequences that are not short. This finding is consistent with some previous simulation results (Sheets and Mitchell, 2001; Hannisdal, 2006).

While the joint parameterization is better able to detect trends in noisy data, there are several respects in which the AD approach is superior. First, the joint estimation approach occasionally encounters computational difficulties. Calculating log-likelihoods for this approach requires inverting a covariance matrix that may be singular. This does not seem to be very common—it occurred in fewer than ten of the 251 se-

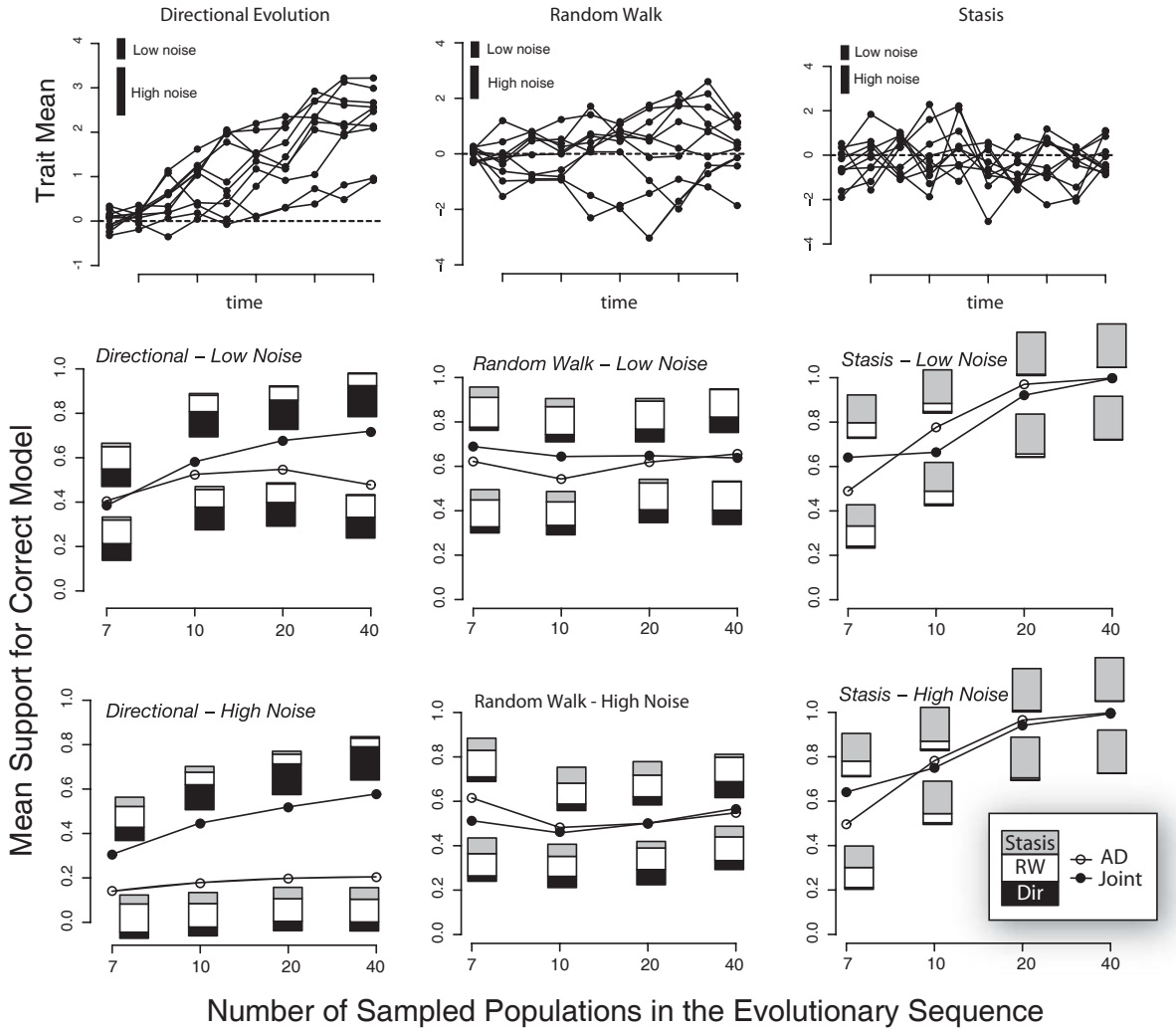

Figure 2—Summary of model selection performance. The top row shows ten realizations for directional evolution (left), random walks (center) and stasis (right). Shown are the true population means for sequences with ten sampled populations; bars in the upper left of each indicate the approximate size of 95% confidence intervals on trait means under low and high noise simulations. These bars do not correspond to specific samples, but rather illustrate typical confidence intervals for the two levels of sampling noise. The bottom two rows summarize the model selection performance for the models with low sampling noise (middle row, $n = 50$ observations per sample), and high sampling noise (bottom row, $n = 5$ observations per sample). Plotted under each scenario is the mean Akaike weight for the correct model, with open circles indicating the ancestor-descendant, and filled circles the joint parameterization. For each set of conditions and parameterization, the rectangles indicate the mean Akaike weight for each of the three models: directional evolution (black bar), random walk (white bar), and stasis (grey bar). Note that the horizontal axis of the bottom two rows is not to scale.

quences analyzed above. Nevertheless, the joint approach may fail for some data sets, or at least require modification to work in some circumstances. Second, these simulations indicate that the parameter estimates for the step variance parameter in the directional evolution model may be biased towards values that are too low, especially for time-series with high sampling error (results not shown). Third and finally, the AD approach is analytically simpler, and more easily extendible to allow for punctuations, covariates and other biologically interesting models (see below). Thus, both approaches have their strengths. Which is most

suitable will depend on the nature of the data and the goals of each particular study.

EVOLUTIONARY INSIGHTS

Relative Importance of Evolutionary Modes

Even the most partisan voices involved in the Punctuated Equilibrium debate acknowledged that no evolutionary mode was universal. Accordingly, the central issue of this debate concerned the relative frequency of stasis versus gradual change. Reviews of the published paleontological literature, however, reached dramatically different conclusions about the dominance of stasis and gradual change (Gingerich, 1985; Erwin and Anstey, 1995; Jackson and Cheetham, 1999; Levinton, 2001; Gould, 2002), mostly because different authors held incompatible interpretations for the same fossil sequences.

One avenue to resolve these disagreements is to apply each of the three canonical modes of evolution—stasis, directional change, and random walks—to all available empirical data sets and summarize support for these three models. Scouring the paleontological literature produced a set of 53 lineages for which the requisite data were published to allow these statistical models to be fit (trait means, variances, and sample sizes, along with an age estimate for each sample in an evolutionary sequence). Some excellent and relevant studies could not be included because the original references did not publish summary statistics at the level of individual samples or stratigraphic levels (e.g., Jackson and Cheetham, 1994; Gingerich and Gunnell, 1995). Of the 53 included lineages, most were measured for multiple morphological traits for a total of 251 evolutionary sequences (Hunt, 2007). Of these, only 13 (5%) were best fit by the directional evolution model, and the remaining split approximately equally between random walks and stasis. Similar levels of support are indicated by the median Akaike weights for each mode (Table 2), suggesting that directional evolution is rarely observed in fossil sequences. This result, which was obtained using the AD parameterization of these models, also holds when the joint parameterization is used instead (Table 2). As might be expected from the simulation results presented above, directional evolution garners slightly more support under the joint approach, but it is still infrequent. Even this relatively low incidence of directional evolution is almost certainly an overestimate because paleontologists have focused greater attention on lineages and traits with prior evidence of gradual change (Gould, 2002). Incidentally, stasis in its strictest sense of no true evolutionary change ($\omega = 0$) is not very common; only about 9% of analyzed sequences (23/251) are consistent with true constancy of form.

While the rarity of directional evolution confirms one key claim of the Punctuated Equilibrium model, it is noteworthy that random walks are at least as common as evolutionary stasis (Hunt, 2007)(Table 2). Qualitatively, fossil sequences seem to meander (random walk) at least as often as they show fluctuations around a relatively stable mean (stasis). In retrospect, it seems that at least some of the disagreements about the relative frequency of stasis and gradual change stemmed from differences in how patterns similar to a random walk were classified. To proponents of Punctuated Equilibrium, their lack of strong directionality rendered these sequences examples of stasis; to its critics, such meandering paths were a kind of gradual change. Neither lumping seems desirable because random walks and stasis are actually distinct, and a fair accounting of the relative importance of different evolutionary modes should reflect this fact.

These three evolutionary modes correspond to identifiable patterns in trait trajectories. How these patterns relate to process is a complex issue because each mode is consistent with a multitude of microevo-

	Median Akaike Weight	
Model	AD parameterization	Joint Parameterization
Directional evolution	0.056	0.069
Random Walk	0.467	0.451
Stasis	0.322	0.409

Table 2—Statistical support, measured as median Akaike weight, for three canonical modes of evolution. Results are shown separately for the ancestor-descendant (AD) and joint parameterizations.

lutionary scenarios (for recent attempts at inferring process from pattern, see Polly, 2004; Estes and Arnold, 2007). For example, populations evolving under neutral genetic drift will show trait trajectories that are random walks (Lande, 1976; Turelli et al., 1988), but so will populations experiencing randomly fluctuating directional selection and those tracking an adaptive optimum that meanders over time. In general, it will be difficult to translate from pattern to microevolutionary process, but I will discuss below a specific example in which the signal of adaptive evolution is apparent in an exceptionally information-rich fossil sequence.

Tempo and Mode Decomposed

The foregoing sections follow current usage in labeling as "modes" qualitatively different kinds of evolutionary patterns. This usage descends from Simpson (1944), who separated how fast evolutionary transitions occurred (tempo) from the nature or mode of the change. Although the three modes commonly considered today differ from Simpson's suite of speciational, phyletic, and quantum modes (Simpson, 1944), the distinction between tempo and mode remains useful.

The fundamental differences between stasis, random walks, and directional change are not in the magnitude or pace of evolutionary divergences, but rather in how constituent evolutionary changes are deployed in a sequence. With directional evolution, changes in the same direction are stacked together, generating persistent trends. For random walks, evolutionary increments are stacked with no preference for one direction over another, and the resulting evolutionary trajectories show increasing but meandering divergence over time. Stasis, by contrast, results when evolutionary increments are stacked antagonistically such that divergences in one direction are preferentially followed by opposing changes so that populations do not wander far from a fixed point.

Within each of these modes, the evolutionary tempo can vary. Random walks, for example, can be faster or slower depending on the magnitude of the underlying step variance. Similarly, evolutionary fluctuations under stasis can be small or large, depending on the value of the variance parameter (ω). Thus, different models correspond to modes of change, with certain parameters of these models governing the tempo of change within that mode. These tempo-controlling parameters are potentially useful as a means to measure evolutionary rates. Particularly promising in this regard is the random walk model, which has just a single parameter—the step variance—that determines the pace of evolutionary change.

Using the estimated step variance of the random walk as a measure of evolutionary rate has a number of advantages over traditionally defined rates such as darwins or haldanes (Haldane, 1949; Gingerich, 1993; Gingerich, 2001). Notably, the step variance: allows for evolutionary reversals; it is estimated in a way that accounts for sampling error; and, at least for true unbiased random walks, its inference is unaffected by the time scale over which it is observed (Hunt, 2006). In addition, Lynch's (1990) rate metric derived for purely neutral evolution is essentially a scaled version of the step variance. When time is measured in organismal generations, Lynch's metric is equal to the estimated step variance standardized by within-sample phenotypic variance. Therefore, the step variance even has a convenient benchmark—the neutral expectation—for judging what constitutes fast or slow evolutionary change. Although parameters related to the step variance are increasingly used to measure phenotypic rates of evolution for phylogenetically related populations (Martins, 1994; O'Meara et al., 2006), they are seldom applied to fossil data.

This recommendation to use the step variance parameter as a rate metric is in conflict with Bookstein's (1987) argument that rates of evolution are undefined for random walks. However, this claim is true only in the technical sense that, as discrete time models, the derivative of a random walk is undefined. However, if differentiability is the sole relevant criterion, evolutionary rates *never* exist because they ultimately change only with the origin or demise of discrete generations or individuals. A broader and more useful definition of evolutionary rate would encompass any model parameter that relates phenotypic divergence to elapsed time (see also Foote, 1991).

Punctuations, and When they are Justified

Some of the most intractable disagreements during the Punctuated Equilibrium battles concerned the distinction between gradual and punctuated change (Gould and Eldredge, 1977; Gingerich, 1985). Punctuations are quite easy to find when looked for, but it is difficult to be sure that hypotheses of pulsed change are truly warranted (Fortey, 1988). Although tests were

developed to detect rate heterogeneity in evolutionary sequences (Charlesworth, 1984; Kitchell et al., 1987), these were seldom applied, and in any case of somewhat limited usefulness because all reasonable models predict some heterogeneity in point-to-point rates of change.

The key to gaining traction on this issue is to recognize that the fundamental claim of punctuational hypotheses is not that rates vary, but rather that evolution is not homogenous. Punctuations are thought to arise when the normal operation of stasis is temporarily suspended, allowing for a period of elevated change that differs qualitatively from stasis. Therefore, this claim should be tested by comparing support for this kind of non-uniform dynamic to that for models in which evolution operates by the same evolutionary rules through the entire sequence, as for example, in a random walk (Hunt, 2008a).

Punctuational explanations posit that evolutionary sequences can be divided into segments, each of which has its own set of evolutionary rules. In practice, each segment can be fit as described above as if it were a complete sequence, and the divisions between segments can be determined by choosing the shift point or points that maximize the log-likelihood of the model (Hunt, 2008a). Punctuations can appear differently depending on their rapidity relative to the temporal resolution of samples; here I will focus only on punctuations that are so fast that intermediate morphologies are not observed. This kind of punctuation can be modeled as two separate intervals of stasis, each with its own evolutionary optimum (θ_1 and θ_2; Fig. 3.1). The magnitude of the pulsed phenotypic change is simply the difference between the two optima. Assuming separate evolutionary variance parameters in each segment, this model has five parameters: two phenotypic optima, each with its own variance parameter, and a parameter that determines the timing of the shift from one segment to the other (Fig. 3.1). AIC_C scores can be used to decide if the log-likelihood advantage of this model more than compensates for its additional complexity, relative to uniform evolutionary models. Note that only the general form of the model is decided a priori (e.g., one punctuation or two punctuations); the actual values of the phenotypic optima and the timing of shifts between stasis regimes are free parameters of the model that are estimated by maximizing the likelihood of the observed sequence.

Although a complete discussion of this class of models is presented elsewhere (Hunt, 2008a), a brief example should suffice to make the general approach clear. Chiba (1996) documented a sequence of evolutionary changes in a suite of shell characters of *Mandarina chichijimana*, a land snail endemic to the Chichijima Islands in the western Pacific Ocean. Radiocarbon dating of shells collected from stratified Quaternary deposits produced a precise and finely resolved age model; the mean elapsed duration between

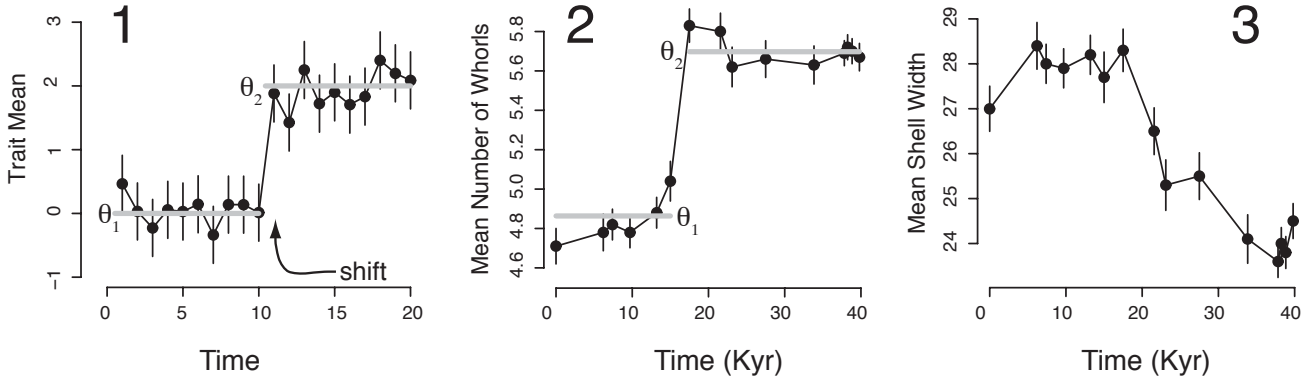

Figure 3—Punctuational models. 3.1, Schematic showing hypothetical punctuated pattern in which the phenotypic optimum shifts suddenly between samples 10 and 11. 3.1 – 3.2, Evolution in shell characters of *Mandarina chichijimana*, a land snail species endemic to the Chichijimana Islands of the western Pacific Ocean (Chiba, 1996). The study interval spans approximately the last 40 Kyr. 3.2, Evolutionary trajectory of the number of shell whorls; a punctuated model decisively outperforms uniform models of change for this character (Table 3). 3.3, Relatively continuous evolution of shell width in the same lineage; a uniform model of change (unbiased random walk) is the best-supported model for this character.

adjacent samples was less than three thousand years. Chiba argued that several of the traits changed in a discontinuous or punctuated manner, including the number of whorls in the shell (Fig. 3.2). Evolution in some other traits, such as overall shell width, seemed more smoothly continuous (Figure 3.3).

To assess these interpretations, I fit to these two traits the standard suite of uniform models (directional change, random walk, stasis), and compared their fit to a model that posited a single, punctuated change. The results corroborate Chiba's interpretations. The model of a single punctuated change accounts for the evolution of whorl number far better than any of the uniform models (Table 3). In contrast, the model of a uniform random walk is best supported for shell width, with the punctuational model accounting for less than 10% of total model support (Table 3).

It may be noted that, although Punctuated Equilibrium posits that punctuations are associated with lineage splitting events, the above example concerns a putative punctuation occurring within an unbranched lineage. This is a fair reflection of the paleontological literature. The claim that morphological jumps coincide with cladogenesis was based on the perceived consequences of allopatric speciation model, not on any direct reading of the fossil record (Eldredge and Gould, 1972). Because there are very few examples for which lineage splitting is thought to be captured in the fossil record (Gingerich, 1985), most discussion about punctuations has focused on examples like Chiba's that document within-lineage evolutionary changes (Gould and Eldredge, 1977; Gould, 2002). If a splitting event is documented directly, however, it would be straightforward to test the Punctuated Equilibrium claim by fitting a model in which the pace of evolutionary change increased during cladogenesis, and testing the fit of this model versus one in which phenotypic evolution during splitting followed the same rules as those operative within lineages.

Natural Selection in Fossil Sequences

Inferring process from pattern is famously difficult, and fossil data offer few exceptions to this rule. Paleontologists have generally attributed patterns of directional evolution to the action of natural selection (e.g., Bell et al., 2006). While it is difficult to imagine trends in fossil lineages arising without the intervention of natural selection, the relationship between patterns of long-term divergence and microevolutionary scenarios can be complex, with multiple processes capable of producing the same macroscopic pattern (Raup and Crick, 1981; Hansen and Martins, 1996). Given this situation, it is worth asking the question: What should a simple bout of adaptive evolution look like in the fossil record?

While this question is a good place to start, it is not precise enough to answer because adaptive evolution can take different forms. If we accept that long-term evolutionary trajectories are best considered in the context of adaptive landscapes (Simpson, 1944;

Trait	Model	Log-likelihood	AIC_C	Akaike weight
Number of whorls	Directional	2.23	0.63	0.000
	Random Walk	1.66	-0.98	0.001
	Stasis	-7.42	19.93	0.000
	Punctuation	16.81	-16.12	**0.999**
Shell width	Directional	-15.55	36.20	**0.239**
	Random Walk	-15.90	34.14	**0.669**
	Stasis	-28.44	61.98	0.000
	Punctuation	-10.30	38.10	0.092

Table 3—Performance of uniform and punctuated evolutionary models for two shell measurements in *Mandarina chichijimana* (Chiba, 1996). Punctuated evolution is strongly supported for whorl counts, but uniform models, especially the random walk, are favored for shell width.

Arnold et al., 2001; Estes and Arnold, 2007), it is possible to focus on one plausible and tractable adaptive scenario: the evolution of populations climbing from suboptimal phenotypes to a nearby peak in the adaptive landscape. This is the expected dynamic for a population that invades an environment with somewhat different selective conditions than its ancestral habitat, or for a population residing in an environment that changes suddenly.

The expected evolutionary dynamic in this scenario is not a simple directional trend, but instead an exponential approach to the new phenotypic optimum (Fig. 4). Change is initially strongly directional, but this directionality tapers rapidly as the new optimum is approached, after which evolutionary stasis ensues. The statistical properties of this model were described by Lande (1976; see also Hansen and Martins, 1996; Hansen, 1997), and it has four key parameters: the initial phenotype, the optimal phenotype, the strength of selection (which determines the rapidity with which the optimum is approached), and the step variance (Lande, 1976; Hansen, 1997; Hunt et al., 2008). This model, which is sometimes referred to as an Ornstein-Uhlenbeck process, can be fit to fossil sequences via maximum likelihood just like the standard modes of change, and its success can be gauged in the normal way using AIC_C scores (Hunt et al., 2008).

This approach was applied to what is probably the most promising example yet described for detecting natural selection in a fossil lineage: Bell et al.'s (2006) study tracking skeletal armor reduction in a stickleback lineage from Miocene lake sediments. In this study, Bell and colleagues documented a tapering decrease in three skeletal traits, including the number of dorsal spines (Fig. 4). These authors applied several methods to test if evolutionary changes had been too rapid or too directional to result from neutral drift, but none of these tests revealed compelling evidence for the action of natural selection. These negative results are noteworthy because (i) the temporal resolution of 250 years—determined by counting yearly varves—is exceptionally fine for fossil studies, and (ii) considerable circumstantial evidence for natural selection exists, including the observation that skeletal reduction is common in modern stickleback that invade lakes with few predatory fishes (as was apparently the case in the paleo-lake studied).

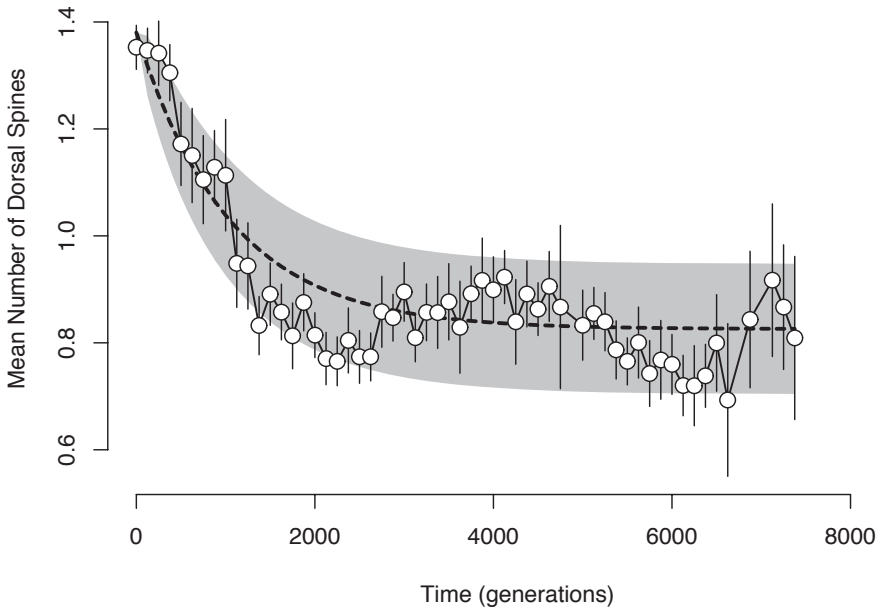

Figure 4—Evolution of dorsal spine counts in a fossil stickleback lineage. Shown is the mean number of dorsal spines (log-transformed as described in Bell et al 2006), with 95% confidence intervals. Time is measured in stickleback generations (= 2 years), with the chronology determined from counting yearly varves. Dotted line shows the best fit adaptive model with its characteristic exponential approach to a new optimal phenotype (see text). Grey region outlines the 95% probability region of the adaptive model.

Rather than attempting to reject a null model of a random walk, a better approach would entail fitting random walk and adaptive models, and then comparing their relative empirical support. When this is done, the adaptive model decisively outperforms the random walk model (the expected form under neutral genetic drift) for all three traits analyzed (Hunt et al., 2008; Fig. 4 shows the model fit for dorsal spine counts). The adaptive model accounts for over 99% of the Akaike weight, and thus there is strong statistical evidence that natural selection has shaped the evolution of these traits, despite the non-significant results from traditional tests (Hunt et al., 2008).

These results demonstrate that, at least under favorable circumstances, it is possible to document natural selection in fossil lineages. It is unclear how often this kind of inference will be possible because this case study benefited from several exceptionally favorable circumstances. Temporal resolution was excellent, and the evolutionary changes were slow enough to occur over several thousand years. While this is a geologically rapid change, much faster instances of adaptive evolution are known from extant populations, including other cases of skeletal reduction in sticklebacks (Bell et al., 2004). Finally, this study benefits from a fortunate window of observation, which happens to include the invasion of a paleo-lake by this particular stickleback lineage (Bell et al., 2006). Adaptive adjustments are expected to occur quickly upon a population's encounter with novel selective conditions, and thus it may be particularly fortuitous to sample the initial stages of an invasion. Thus, while it is possible to infer the action of natural selection in fossil lineages, the requisite depositional and biological conditions may be rather uncommon. Regardless, success in detecting adaptive evolution in the fossil record requires that we are actually looking for its proper signature.

FUTURE DIRECTIONS

The approach described here is not so much a method as it is a general approach that may be applied to evaluate almost any kind of hypothesis about the nature of phenotypic changes within lineages. In this paper, I described several different applications using this approach, but there is still ample room to expand the range of evolutionary models considered beyond those discussed here. One natural extension of this approach would be to incorporate putatively causal factors into evolutionary models. The resulting class of models could be used to assess, for example, the effects of climate or productivity on body size evolution (Schmidt et al., 2004; Hunt and Roy, 2006; Finkel et al., 2007; Novack-Gottshall, 2008), and compare the success of these models to those that lack covariates. Another potentially fruitful class of models to explore are those Alroy (2000) refers to as "structured state space models." These models are characterized by having dynamics that vary with the phenotypic values of the lineage, allowing for the possibility of phenotypic values that attract or repulse evolutionary trajectories. The stasis model is a simple example of this kind of model—the optimum is essentially a very strong attractor—but a whole range of potentially interesting models has yet to be considered.

A second area that could use analytical development is the effect of uncertain chronologies in inferring evolutionary models. These effects are likely to be strongly data and model-dependent, and simulations may help explore the sensitivity of analyses to these kinds of effects (Hunt, 2006; Hunt, 2008a). Some recent work (Hannisdal, 2007) has taken the approach of integrating stratigraphic and evolutionary inferences into a single framework. This approach, though demanding of computational effort and data quantity, offers a very elegant means of accommodating age model uncertainty.

A final area I would point to as profitable for future work is comparing evolutionary patterns within fossil lineages to similar analyses performed with phylogenetically related populations. This area actually does not require much in the way of theory development; for the most part, the models applied in comparative studies are the same as those described in this paper. In particular, the random walk model is dominantly used to explore trait evolution (Felsenstein, 1985; Martins, 1999; Garland and Ives, 2000), although studies that employ directional (Pagel, 2002), adaptive (Hansen, 1997; Butler and King, 2004) and other models (e.g., Pagel, 1999) are increasingly common. Yet at present, two complementary data sets about phenotypic divergence—fossils and phylogenies—employ almost entirely non-overlapping sets of methods. If we are to ever reconcile these two views of evolution, it will

begin with analytical approaches able to evaluate disparate kinds of evidence in comparable terms.

ACKNOWLEDGMENTS

For discussion about many of the ideas presented here, I thank M. Foote, S. Wang, K. R. Thomas, B. Hannisdal, P. Wagner, M. Bell, and C. Marshall. P. Novack-Gottshall and an anonymous reviewer provided extremely helpful and timely comments on the manuscript.

LITERATURE CITED

AKAIKE, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716-723.

ALROY, J. 2000. Understanding the dynamics of trends within evolving lineages. Paleobiology, 26(3):319-329.

ANDERSON, D. R., K. P. BURNHAM, AND W. L. THOMPSON. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management, 64(4):912-923.

ARNOLD, S. J., M. E. PFRENDER, AND A. G. JONES. 2001. The adaptive landscape as a conceptual bridge between micro- and macroevolution. Genetica, 112-113:9-32.

BELL, M. A., W. E. AGUIRRE, AND N. J. BUCK. 2004. Twelve years of contemporary armor evolution in a threespine stickleback population. Evolution, 58(4):814-824.

BELL, M. A., M. P. TRAVIS, AND D. M. BLOUW. 2006. Inferring natural selection in a fossil threespine stickleback. Paleobiology, 32(4):562-577.

BOOKSTEIN, F. L. 1987. Random walk and the existence of evolutionary rates. Paleobiology, 13(4):446-464.

BOOKSTEIN, F. L. 1988. Random walk and the biometrics of morphological characters. Evolutionary Biology, 9:369-398.

BUTLER, M. A., AND A. A. KING. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. American Naturalist, 164(6):683-695.

CHARLESWORTH, B. 1984. Some quantitative methods for studying evolutionary patterns in single characters. Paleobiology, 10(3):308-318.

CHIBA, S. 1996. A 40,000-year record of discontinuous evolution of island snails. Paleobiology, 22(2):177-188.

ELDREDGE, N., AND S. J. GOULD. 1972. Punctuated equilibria: an alternative to phyletic gradualism, p. 82-115. *In* T. J. M. Schopf (ed.), Models in Paleobiology. Freeman, Cooper & Company, 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. Paleobiology, 31(Supplement to 2):133-145.

ERWIN, D. H., AND R. L. ANSTEY. 1995. Speciation in the fossil record, p. 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(2):227-244.

FELSENSTEIN, J. 1985. Phylogenies and the comparative method. American Naturalist, 125(1):1-15.

FINKEL, Z. V., J. SEBBO, S. FEIST-BURKHARDT, A. J. IRWIN, M. E. KATZ, O. M. E. SCHOFIELD, J. R. YOUNG, AND P. G. FALKOWSKI. 2007. A universal driver of macroevolutionary change in the size of marine phytoplankton over the Cenozoic. Proceedings of the National Academy of Sciences USA, 104(51):20416-20420.

FOOTE, M. 1991. Analysis of morphological data, p. 59-86. *In* N. L. Gilinsky and P. W. Signor (eds.), Analytical Paleobiology.Volume 4. The Paleontological Society.

FORTEY, R. A. 1988. Seeing is believing: gradualism and punctuated equilibria in the fossil record. Science Progress, 72:1-19.

GARLAND, T., AND A. R. IVES. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. American Naturalist, 155(3):346-364.

GINGERICH, P. D. 1985. Species in the fossil record: concepts, trends, and transitions. Paleobiology, 11(1):27-41.

GINGERICH, P. D. 1993. Quantification and comparison of evolutionary rates. American Journal of Science, 293-A:453-478.

GINGERICH, P. D. 2001. Rates of evolution on the time scale of evolutionary process. Genetica, 112-113:127-144.

GINGERICH, P. D., AND G. F. GUNNELL. 1995. Rates of evolution in Paleocene-Eocene mammals of the Clarks Fork Basin, Wyoming, and a comparison with Neogene Siwalik lineages of Pakistan. Palaeogeography, Palaeoclimatology, Palaeoecology, 115(1-4):227-247.

GOULD, S. J. 2002. The Structure of Evolutionary Theory. Belknap Press, Cambridge, Massachusetts, 1433 p.

GOULD, S. J., AND N. ELDREDGE. 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology, 3(2):115-151.

HALDANE, J. B. S. 1949. Suggestions as to the quantitative measurement of rates of evolution. Evolution, 3:51-56.

HANNISDAL, B. 2006. Phenotypic evolution in the fossil record: Numerical experiments. Journal of Geology, 114(2):133-153.

HANNISDAL, B. 2007. Inferring phenotypic evolution in the fossil record by Bayesian inversion. Paleobiology, 33(1):98-115.

HANSEN, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution, 51(5):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(4):1404-1417.

HUNT, G. 2006. Fitting and comparing models of phyletic evolution: random walks and beyond. Paleobiology, 32(4):578-601.

HUNT, G. 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(47):18404-18408.

HUNT, G. 2008. Gradual or pulsed evolution: when should punctuational explanations be preferred? Paleobiology, 34(3):360-377.

HUNT, G. 2008b. paleoTS: Modeling evolution in paleontological time-series. 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(3):700-710.

HUNT, G., AND K. ROY. 2006. Climate change, body size evolution, and Cope's Rule in deep-sea ostracodes. Proceedings of the National Academy of Sciences USA, 103(5):1347-1352.

JACKSON, J. B. C., AND A. H. CHEETHAM. 1994. Phylogeny reconstruction and the tempo of speciation in cheilostome Bryozoa. Paleobiology, 20(4):407-423.

JACKSON, J. B. C., AND A. H. CHEETHAM. 1999. Tempo and mode of speciation in the sea. Trends in Ecology and Evolution, 14(2):72-77.

KINNISON, M. T., AND A. P. HENDRY. 2001. The pace of modern life II: from rates of contemporary microevolution to pattern and process. Genetica, 112-113:145-164.

KITCHELL, J. A., G. ESTABROOK, AND N. MACLEOD. 1987. Testing for equality of rates of evolution. Paleobiology, 13(3):272-285.

LANDE, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution, 30:314-334.

LEVINTON, J. S. 2001. Genetics, Paleontology, and Macroevolution. Cambridge University Press, Cambridge.

LYNCH, M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. American Naturalist, 136(6):727-741.

MARTINS, E. P. 1994. Estimating the rate of phenotypic evolution from comparative data. American Naturalist, 144(2):193-209.

MARTINS, E. P. 1999. Estimation of ancestral states of continuous characters: a computer simulation study. Systematic Biology, 48(3):642-650.

NOVACK-GOTTSHALL, P. M. 2008. Ecosystem-wide body-size trends in Cambrian-Devonian marine invertebrate lineages. Paleobiology, 34:210-228.

O'MEARA, B. C., C. C. ANÉ, M. J. SANDERSON, AND P. C. WAINWRIGHT. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution, 60(5):922-933.

PAGEL, M. 1999. Inferring the historical patterns of biological evolution. Nature, 401:877-884.

PAGEL, M. 2002. Modelling the evolution of continuously varying characters on phylogenetic trees: the case of hominid cranial capacity, p. 269-286. *In* N. MacLeod and P. L. Forey (eds.), Morphology, Shape and Phylogeny. Taylor & Francis, London.

POLLY, P. D. 2004. On the simulation of the evolution of morphological shape: multivariate shape under selection and drift. Palaeontologia Electronica, 7(2):1-28.

R DEVELOPMENT CORE TEAM. 2008. R: A language and environment for statistical computing. 2.7.0. R Foundation for Statistical Computing. http://www.R-project.org.

RAUP, D. M. 1977. Stochastic models in evolutionary paleobiology, p. 59-78. *In* A. Hallam (ed.), Patterns of Evolution as Illustrated by the Fossil Record. Volume 5. Elsevier Scientific Publishing Company, Amsterdam.

RAUP, D. M., AND R. E. CRICK. 1981. Evolution of single characters in the Jurassic ammonite *Kosmoceras*. Paleobiology, 7(2):200-215.

ROOPNARINE, P. D. 2001. The description and classification of evolutionary mode: a computational approach. Paleobiology, 27(3):446-465.

ROOPNARINE, P. D., G. BYARS, AND P. FITZGERALD. 1999. Anagenetic evolution, stratophenetic patterns, and random walk models. Paleobiology, 25(1):41-57.

SCHMIDT, D. N., H. R. THIERSTEIN, J. BOLLMAN, AND R. SCHIEBEL. 2004. Abiotic forcing of plankton evolution in the Cenozoic. Science, 303:207-210.

SHEETS, H. D., AND C. E. MITCHELL. 2001. Why the null matters: statistical tests, random walks and evolution. Genetica, 112-113:105-125.

SIMPSON, G. G. 1944. Tempo and Mode in Evolution. Columbia University Press, New York, 237 p.

SOKAL, R. R., AND F. J. ROHLF. 1995. Biometry. W.H. Freeman and Company, New York, 887 p.

TURELLI, M., J. H. GILLESPIE, AND R. LANDE. 1988. Rate tests for selection on quantitative characters during macroevolution and microevolution. Evolution, 42(5):1085-1089.