doi:10.1111/j.1558-5646.2007.00310.x

EVOLUTION TOWARD A NEW ADAPTIVE OPTIMUM: PHENOTYPIC EVOLUTION IN A FOSSIL STICKLEBACK LINEAGE

Gene Hunt,[1,2] Michael A. Bell,[3,4] and Matthew P. Travis[3,5,6]

[1] *Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013*

[2] *E-mail: hunte@si.edu.*

[3] *Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794–5245*

[4] *E-mail: mabell@life.bio.sunysb.edu.*

[6] *E-mail: travism@rowan.edu*

Received November 27, 2007
Accepted November 27, 2007

Natural selection has almost certainly shaped many evolutionary trajectories documented in fossil lineages, but it has proven difficult to demonstrate this claim by analyzing sequences of evolutionary changes. In a recently published and particularly promising test case, an evolutionary time series of populations displaying armor reduction in a fossil stickleback lineage could not be consistently distinguished from a null model of neutral drift, despite excellent temporal resolution and an abundance of indirect evidence implicating natural selection. Here, we revisit this case study, applying analyses that differ from standard approaches in that: (1) we do not treat genetic drift as a null model, and instead assess neutral and adaptive explanations on equal footing using the Akaike Information Criterion; and (2) rather than constant directional selection, the adaptive scenario we consider is that of a population ascending a peak on the adaptive landscape, modeled as an Orstein–Uhlenbeck process. For all three skeletal features measured in the stickleback lineage, the adaptive model decisively outperforms neutral evolution, supporting a role for natural selection in the evolution of these traits. These results demonstrate that, at least under favorable circumstances, it is possible to infer in fossil lineages the relationship between evolutionary change and features of the adaptive landscape.

KEY WORDS: Adaptive evolution, adaptive landscape, neutral evolution, random walk, stickleback, time series, Orstein–Uhlenbeck process.

Paleontologists have traditionally interpreted many morphological changes in the fossil record in terms of adaptive evolution. This viewpoint is particularly evident in the classic works by Simpson (1944, 1953), who modified Wright's genotypic adaptive landscape to apply to the evolution of phenotypic traits. In Simpson's view, the fossil record revealed numerous examples of populations evolving to occupy optima of relatively well-adapted phenotypes (see Lande 1976; 1979; Arnold et al. 2001 for a more modern treatment). More recently, however, many paleontologists have become cautious about attributing change in the fossil record to the action of natural selection. This caution results in part from a greater appreciation of the ability of null models to produce evolutionary patterns similar to those previously interpreted as deterministic (Raup and Gould 1974; Raup 1977; Bookstein 1987). Moreover, paleontologists have become more aware that the temporal resolution of fossil data is usually much coarser than the generational timescales of microevolution (Schindel 1980; Sadler 1981; Kidwell and Behrensmeyer 1993), and that this fact

[5] Current Address: Department of Biological Sciences, Rowan University, Glassboro, New Jersey 08028

complicates attempts to relate paleontological data to population-level processes.

Partly as a result of this skepticism, statistical tests were developed to detect natural selection operating in time series of phenotypic traits through stratigraphic sequences. The null model in these tests is neutral genetic drift (Lande 1976; Turelli et al. 1988; Lynch 1990) or its general form, an unbiased random walk (Raup 1977; Raup and Crick 1981; Bookstein 1987; Gingerich 1993; Roopnarine 2003). These methods test against the neutral expectation by assessing whether trait increases significantly exceeded decreases, or vice versa (Raup 1977; Raup and Crick 1981), if trait increases or decreases were more autocorrelated than expected (runs test, Raup and Crick 1981), if long-term divergence was too great given short-term changes (scaled maximum test, Bookstein 1987), or if the pace of change has been too fast for genetic drift (Lande 1976; Lynch 1990). Application of these tests to detect adaptive evolution in fossil sequences has not been encouraging. The null model of an unbiased random walk frequently cannot be rejected (Raup 1977; Raup and Crick 1981; Malmgren et al. 1983; Bookstein 1987, 1988; Sheets and Mitchell 2001), and evolutionary divergence, rather than being so rapid as to imply directional selection, is usually slower than the neutral expectation (Lynch 1990; Cheetham and Jackson 1995; Clegg et al. 2002; Estes and Arnold 2007; Hunt 2007).

A recent study by Bell et al. (2006) documenting morphological evolution in a stickleback lineage represents an unusually promising test case for detecting natural selection in the fossil record. This lineage showed a steady but tapering reduction in several traits related to the size and robustness of skeletal elements (Fig. 1), and several independent lines of indirect evidence suggest that natural selection influenced these evolutionary trajectories (Bell et al. 2006). In addition, because the lake sediments

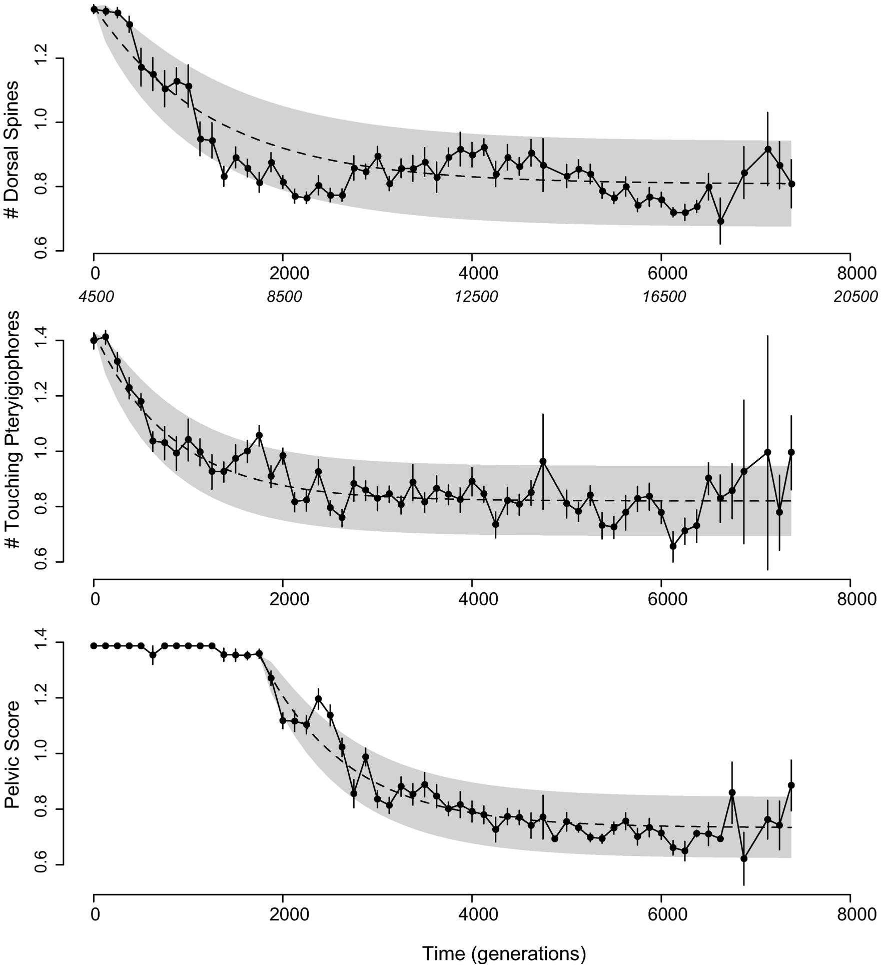

Figure 1. **Plots of mean trait values over time, starting with the first appearance of the highly armored stickleback lineage. Time is measured in organismal generations (= 2 years/generation), with generation zero set at the first sample marking the invasion of the lake by this particular lineage (smaller italicized numbers below the time axis in the top panel show the time scale from Bell et al. [2006], which is in years). Vertical error bars denote one standard error; samples with fewer than five individuals have been omitted. For each trait, the expected evolutionary trajectory of the best-fit adaptive model is shown as a dotted line, with the 95% probability envelope around this solution in gray. There is a delay before the pelvic score character begins to evolve to a lower armored form (see text).**

are preserved in annually resolved layers (diatomaceous varves), temporal resolution in this system is exceptionally fine—about 250 years after strata were lumped to attain useful sample sizes (Bell et al. 2006). Despite these favorable circumstances, scaled maximum and runs tests for all three traits revealed no evidence of directional change. In addition, only one trait showed a marginally significant excess of decreases, and very few of the sample-to-sample differences were larger than expected under drift (Bell et al. 2006). Because of these negative results in the face of independent evidence of selection, Bell et al. (2006) were pessimistic about the likelihood of ever detecting natural selection from the evolutionary trajectories of fossil lineages.

Here we present a reanalysis of these data, which, contrary to the standard tests applied by Bell et al. (2006), strongly favors an adaptive model to account for the evolutionary trajectories of these traits. Our analyses differ from those previously applied in two important respects. First, we do not treat genetic drift as a null model. This approach often has low statistical power (Roopnarine 2001; Sheets and Mitchell 2001), and giving privileged null status for genetic drift is difficult to justify on biological grounds. Instead, we fit two competing models to these data, one neutral and one adaptive, and compare their performance on an equal footing using the Akaike Information Criterion. Second, we do not equate adaptive evolution with relentless directional selection, which is unlikely to persist through the approximately 17,000-year duration of the sequence analyzed. Rather, we model adaptive evolution in terms of a population climbing a peak in the adaptive landscape. When we fit this adaptive model to these stickleback trait sequences, we find that for all analyzed traits it decisively outperforms a model of neutral phenotypic evolution. Accordingly, we conclude that it is probable that natural selection shaped the evolution of these traits, and that at least under favorable circumstances, it is possible to infer the relationship between the evolutionary trajectories of fossil lineages and features of the adaptive landscape.

Materials and Methods

FOSSIL DATA

The fossil stickleback represent *Gasterosteus doryssus* (Jordan 1907), which is a member of the *Gasterosteus aculeatus* species complex (Bell 1994). This complex comprises a large set of extant, phenotypically diverse populations that sometimes represent separate biological species (reviewed by Bell and Foster 1994; McPhail 1994; McKinnon and Rundle 2002). Fossil specimens were obtained in an open pit diatomite mine (39.526°N, 119.094°W, Two Tips, Nevada, 15 min. ser., topographic, U.S.G.S.). Collecting methods, relative dating, and aggregation of data into time intervals were described in detail previously (Bell et al. 1985, 2006), and only methods that affect the present analysis will be detailed here.

The fossil *G. doryssus* sequence analyzed here came from a 21,500-year stratigraphic interval. Specimens were dated approximately to the year of deposition using stratigraphic position and converting it to years using varve counts from a measured, complete stratigraphic section. A low-armored, planktivorous species is initially present, but is replaced after about 4500 years by a more highly armored form that rapidly evolves reduced skeletal elements (Bell et al. 2006). Individuals from this second, initially highly armored species were subject to analysis.

About 5000 fossils were scored for three armor traits: (1) the number of dorsal spines, (2) the number of pterygiophores touching along the dorsal midline, and (3) a score summarizing the size and robustness of the pelvis, based on ordered categories of Bell (1987, fig. 7–10). The pelvic condition ranges from all elements absent to all elements present, with intermediate categories including states with vestiges of various elements present (see Bell 1987; Bell et al. 2006). To obtain analytically useful sample sizes, the morphological data were pooled into consecutive 250-year intervals within each of which trait means and variances were computed (see online Supplementary Table S1; the same data as presented in the appendix of Bell et al. 2006). This time averaging of specimens from multiple generations into the same sample should have minimal effect on the analyses performed here (Appendix).

EVOLUTIONARY MODEL FITTING

Under neutral genetic drift, populations evolve according to an unbiased random walk. This model occurs in discrete time (generations), during each of which an evolutionary transition is drawn at random from a distribution of evolutionary "steps." Long-term divergence from the initial trait mean is determined entirely by the variance of this step distribution. This parameter is sometimes called the step variance in paleontological studies (σ^2_{step}) (Bookstein 1987; Hunt 2006), and its magnitude under drift is determined by population genetic parameters, with two different solutions depending on whether genetic variance is assumed to be constant (Lande 1976) or the result of mutation–drift equilibrium (Turelli et al. 1988). No matter how drift is modeled, however, the expected dynamic is an unbiased random walk, and so fitting this model encompasses both derivations. In the microevolution and comparative methods literature, it is more common to invoke the analytically similar Brownian motion (diffusion) model, which is the limit of a random walk as the step size gets increasingly smaller and the interval of time between steps gets correspondingly shorter. The parameter equivalent to the step variance in this model is commonly called the instantaneous variance (e.g., Hansen and Martins 1996).

Populations evolve toward nearby optima on the adaptive landscape and, under most conditions, eventually occupy a phenotypic position close to the peak (Lande 1976, 1979; Zeng 1988). Because populations are finite, however, genetic drift still operates, and the resulting trajectories can be described as a random walk in the presence of an attracting optimum (Lande 1976; Hansen and Martins 1996). Mathematically, this is an Orstein–Uhlenbeck (OU) process (Lande 1976; Hansen and Martins 1996; Hansen 1997; Butler and King 2004), and its dynamics are governed by four parameters: the initial trait mean of the population, the phenotypic position of the optimum, the strength of the attracting force around this optimum, and the step variance of the random walk component of change (Hansen 1997). Traits subjected to an OU process experience a combination of directional and stabilizing natural selection. Initially, selection is mostly directional, but becomes progressively more stabilizing as the population nears the adaptive optimum. Because stabilizing selection is thought to operate prevalently on morphological traits (Travis 1989; Hansen 1997; Estes and Arnold 2007), this scenario is likely to be a more realistic approximation to adaptation in natural populations than simple directional selection. This general dynamic of a population ascending a nearby adaptive peak is referred to by Estes and Arnold (2007) as the "displaced optimum" model, and they found that, among the six microevolutionary models they considered, it best accounted for the aggregate features of a large sample of phenotypic divergences.

For a sequence of ancestor–descendant populations evolving as a random walk or an OU process, the population means have a joint distribution that is multivariate normal with a mean vector and covariance matrix that are functions of the evolutionary parameters, elapsed time, and the sampling variances of the population means (Appendix). In particular, the random walk model has two parameters that must be estimated: z_0, the initial trait mean at the start of the sequence, and the step variance ($\sigma^2{}_{\text{step}}$). In addition to these, the OU model has two additional parameters: θ, the phenotypic position of the optimum, and α, the strength of the restraining force around this optimum. The parameters of the OU model can all be related to population genetic characteristics of the evolving lineage and to aspects of the shape of the adaptive landscape, including the strength of stabilizing natural selection (Appendix). The probability density function of the multivariate normal distribution can be used to estimate these evolutionary parameters and compute the overall likelihood of each model; for details on how these models were fit to data, see the Appendix.

To account for the differing complexities of the two models compared, we used the bias-corrected version of the Akaike Information Criterion (AIC$_\text{C}$) (Akaike 1974; Hurvich and Tsai 1989; Anderson et al. 2000) as a measure of model fit. AIC$_\text{C}$ scores represent a compromise between goodness of fit, measured as log-likelihood, and model complexity, measured as the number of free model parameters:

$$\text{AIC}_\text{C} = -2\log(L) + 2K + (2K[K+1])/(n-K-1)$$

The AIC$_\text{C}$ measures the amount of information lost in approximating reality with a model; the last term, which includes the number of observations (n) and free parameters of the model (K), is a bias-correction factor that becomes unimportant at high sample sizes (Hurvich and Tsai 1989; Anderson et al. 2000). Using AIC$_\text{C}$ scores, models can be compared equally, without treating one as a null model that is rejected only if conclusively outperformed by more complex models. Akaike weights, which are simple transformations of AIC$_\text{C}$ scores that sum to one, were used to summarize the relative support for the two alternative models (Anderson et al. 2000). This is a similar approach to that used by Butler and King (2004) in evaluating OU and diffusion models in a phylogenetic context.

For two of the three traits studied—the number of dorsal spines and the number of touching pterygiophores—we fit the two models starting with the sample marking the appearance of highly armored stickleback lineage at year 4500 in the time scale of Bell et al. (2006). The third trait, pelvic score, shows a similar decreasing trajectory as the other two traits, but only after a delay of about 3500 years (= 1750 generations, Fig. 1). Bell et al. (2006) interpreted this delay as the result of the strong effect of a single gene for which the derived allele is nearly recessive and thus invisible to selection until its frequency is high enough for substantial numbers of homozygous individuals to occur. Taking this explanation as a working hypothesis, we fit the models for this trait starting from the sample at 8000 years, in which the pronounced evolutionary decrease begins (Fig. 1). Because the microevolutionary processes take place on generational time scales, we converted the ages in years to organismal generations, assuming a generation time of two years for these sticklebacks (Bell et al. 2006).

As indicated by the size of the vertical error bars in Figure 1, sampling error on trait means is rarely small and occasionally very large. As a result, it is important to tease apart true evolutionary variation from that attributable to sampling variance. To reduce sampling noise, we omitted the few samples that were composed of less than five observations and incorporated sampling variance into the likelihood calculations (Appendix). In addition, a few analyzed samples were invariant for two characters (one sample of dorsal spine number, two samples of pelvic score). Taken at face value, these zero variances imply a sampling error of zero because sampling error on a mean is proportional to within-sample variance. Because it is unlikely that these sample means were truly inferred without error (especially because they all have modest sample sizes, $n \leq 10$), we estimated sampling error for these populations assuming that these populations had phenotypic variances equal to the average sample variance pooled across all samples. Results are not affected if variances in these samples are instead assumed to be very low, equal to the smallest nonzero observed sample variance. We did not pool variances in general because of significant heterogeneity among sample variances for all traits

Table 1. Results of the model fits for the three time series of mean trait values. Neutral evolution was modeled as an unbiased random walk; adaptive evolution was modeled as an Orstein–Uhlenbeck process, which is the expected dynamic of a population climbing a peak in the adaptive landscape. For each model fit, the log-likelihood (logL), number of parameters (*K*), and bias-corrected Akaike Information Criterion (AIC_C) are given. The last column gives the result of a likelihood-ratio test (LRT), which tests the significance of the improved fit of the adaptive over the neutral model, with the latter treated as the null. The LRT statistic is distributed as a chi-square, with two degrees of freedom.

Trait	Model	logL	*K*	AIC_C	Akaike weight	LRT
No. of dorsal spines	Neutral	86.48	2	−168.73	0.002	
	Adaptive	94.94	4	−181.11	**0.998**	16.92, $P = 0.0003$
Pterygiophores	Neutral	65.91	2	−127.59	0.001	
	Adaptive	74.80	4	−140.84	**0.999**	17.78, $P = 0.0002$
Pelvic score	Neutral	58.38	2	−112.46	0.001	
	Adaptive	68.33	4	−127.65	**0.999**	19.89, $P = 0.00005$

(Bartlett's test, all $P < 0.0001$), but results do not differ markedly if this is done.

Results

Qualitatively, the observed evolutionary trajectories show the expected pattern for populations climbing an adaptive peak—trait changes are initially large and mostly negative, but then become progressively smaller and nondirectional as the population seems to converge to the new optimum (Fig. 1). This visual impression is confirmed by the model-fitting results: for each trait, the adaptive OU model decisively outperforms the neutral random walk model (Table 1). In fact, the relative support for drift is in each case negligible (all Akaike weights < 0.5%). The match between the observed data and the expected trajectory of each best-fitting OU model is quite good, with none of the observed sample means falling much outside the 95% probability region (Fig. 1). The maximum-likelihood parameter estimates for the adaptive model are listed in Table 2, with approximate standard errors calculated from the curvature of the log-likelihood surface. These model fits indicate a net displacement between the initial phenotype and the adaptive optimum of 2–3 units of within-population standard deviations for each of the three measured traits (Table 3).

The magnitude of short-term trait fluctuations provides an independent consistency check on the plausibility of the adaptive model for these traits. According to this model, these fluctuations around the expected trajectory are caused by genetic drift, which results in a step variance parameter equal to $h^2\sigma^2{}_P/N_e$ where h^2 is trait heritability, $\sigma^2{}_P$ is phenotypic variance, and N_e is the effective population size (Lande 1976). The phenotypic variance can be estimated as the variance pooled across samples, and whereas heritabilities are not known for these traits, realistic values for morphological traits range between 0.1 and 0.7 (Mousseau and Roff 1987; Roff and Mousseau 1987). It is possible to calculated the range of implied N_e separately for the each evolutionary sequence using the estimated step variances, phenotypic variances, and a range of plausible heritability values (Appendix).

The ranges of estimated effective population sizes are similar across all three traits (Table 3). This is obviously expected for traits measured from the same evolving population, but this correspondence need not have been observed because the calculations are independent for each trait. In addition, although little is known about the sizes of these particular fossil populations, N_e estimates of several hundred to several thousand fish are at least compatible with the range inferred for populations of sticklebacks in modern lakes (Hendry et al. 2001).

Table 2. Maximum-likelihood parameter estimates for the adaptive (OU) model. The parameters are the ancestral trait mean at the start of the sequence (z_0), the phenotypic position of the adaptive peak (θ), the step variance ($\sigma^2{}_{step}$) of the random walk component of change attributable to genetic drift, and the strength of the restraining force (α) around the optimum. In parentheses are approximate standard errors calculated from the curvature of the log-likelihood surface. For these model fits, time has been measured in estimated organismal generations. Maximum-likelihood solutions to the unbiased random walk model yielded estimates for z_0 that were nearly the same as in the OU model; step variance estimates for the random walk models were roughly twice as large as those in the OU model.

Trait	z_0	θ	$\sigma^2{}_{step}$	α
No. of dorsal spines	1.36 (0.01)	0.79 (0.05)	7.11×10^{-6} (2.3×10^{-6})	8.12×10^{-4} (2.2×10^{-4})
Pterygiophores	1.43 (0.03)	0.82 (0.04)	9.51×10^{-6} (5.2×10^{-6})	1.19×10^{-3} (9×10^{-4})
Pelvic score	1.36 (0.02)	0.73 (0.04)	6.59×10^{-6} (4.0×10^{-6})	1.09×10^{-3} (2.5×10^{-4})

Table 3. Estimates and ranges for population genetic and selective parameters, derived from the maximum-likelihood parameter estimates for the adaptive model. Displacement is the phenotypic distance between the starting morphology and the optimal morphology, in standard deviation units (computed from the phenotypic variance pooled across samples, $\sigma^2{}_P$). Calculations of the effective population size (N_e) and strength of stabilizing selection (ω^2) require knowledge of trait heritabilities, which were assumed to range between 0.1 and 0.7. The ranges reported for N_e and ω^2 reflect calculation using these two different end-member heritability values. Finally, $t_{1/2}$ is the expected time it takes the population to traverse half the distance to the phenotypic optimum, in generations. For details on the calculation of these quantities, see the Appendix.

Trait	Displacement	$\sigma^2{}_P$	N_e	ω^2	$t_{1/2}$
No. of dorsal spines	−2.80	0.041	575–4023	5.0–35.2	853
Pterygiophores	−2.13	0.081	851–5957	6.7–47.3	580
Pelvic score	−2.57	0.059	889–6222	5.3–37.5	635

Finally, the α parameter from the OU model can be used to calculate the strength of stabilizing selection acting on these traits. As suggested by Hansen (1997), this parameter can be expressed in terms of the amount of time it takes a population to evolve halfway to the optimal phenotype (Appendix), which for the traits measured is about 500 to 1000 generations (Table 3). It is also possible to calculate a parameter, ω^2, which measures the strength of stabilizing selection around the optimal phenotype, again assuming a range of plausible values for trait heritability (Appendix, Table 3). These estimates of ω^2 are much greater than the sample phenotypic variances (Table 3), indicating that these traits experienced weak stabilizing selection (Lande 1976; Estes and Arnold 2007). It is not surprising that selection was weak because convergence to the new optimum is rather slow, taking over several thousand generations. Much faster instances of armor reduction are known for extant stickleback populations (Bell 2001; Bell et al. 2004).

Discussion

When compared on equal footing, the adaptive model decisively outperforms neutral drift for all three of the traits analyzed in this stickleback lineage. In fact, the improvement of fit offered by the adaptive model is large enough that it soundly rejects neutrality, even when the latter is considered the null model in a likelihood ratio test (Table 1). In general, many of the standard tests against neutral evolution look for excessive directionality in evolutionary sequences. However, if adaptive evolution usually involves ascending a stationary peak in adaptive landscape (Estes and Arnold 2007), strongly directional change will only occur early in the evolutionary sequence, with a preponderance of stasis and nondirectional fluctuations once the peak is nearly ascended. Additionally, instability in the location of the adaptive peak (e.g., Grant and Grant 2002) and error due to the limited size of fossil samples (Bell 1994) will tend to prevent rejection of neutral evolution as the null model.

These model-fitting results are corroborated by several lines of independent, circumstantial evidence implicating natural selection in the evolution of the measured traits. In extant threespine sickleback populations, reduction of the number of dorsal spines and of the pelvis to a vestige is infrequent but tends to occur in lakes in which the ionic strength of the water is low and predatory fish are absent (Bell et al. 1993). Although the chemistry of the paleo-lake is unknown, only three specimens of predatory fish have been found in sites that produced many thousands of fossil sticklebacks, indicating that predation by fish was inconsequential as a selection agent on *G. doryssus* (Bell 1994). In addition, the simultaneous reduction of multiple skeletal elements, and the presence of a different very low-armored stickleback lineage in the same paleo-lake also support the claim that the documented skeletal evolution was driven by natural selection (see Bell et al. 2006).

One complication in applying these models is the roughly 3500-year delay observed before pelvic score begins to decrease. Directional asymmetry of the reduced pelvis (Shapiro et al. 2004; Bell et al. 2006) and the trimodal distribution of pelvic scores (M. P. Travis and M. A. Bell, unpubl. data) during the evolution of pelvic reduction both suggest that segregation of the *Pitx1* gene, which has a major effect on pelvic reduction in extant threespine stickleback and causes left-biased pelvic asymmetry (Shapiro et al. 2004), played a major role in expression of reduced pelvic phenotypes. The delay in the response of the character is, according to this explanation, a result of the initially low frequency and recessive nature (Cresko et al. 2004; Shapiro et al. 2004) of the *Pixt1* allele for pelvic reduction. Natural selection on pelvic morphology presumably initiated at year 4500 when the population initially invaded the lake, but an evolutionary response was delayed by the absence of heritable phenotypic variation on which selection could act. Alternatively, it is possible that the delayed response occurred because the position of the optimal pelvic score phenotype shifted position much later than the other two traits. Two observations favor the first of these two explanations:

(1) there is little to no variation for pelvic score in the intervals before the shift in morphology, and (2) although pelvic score did not change immediately, the size of the pelvis started decreasing coincident with the evolutionary shift in the dorsal spine and pterygiophore characters (Bell et al., unpubl. data). If this interpretation is correct, the genetic constraint on pelvic score did not persist for long in a geological sense; in most sedimentary environments, the delay would not even have been resolvable (see Bell and Haglund 1982).

The adaptive model considered here is relatively simple in that the position of the adaptive peak is constant over time. It would be relatively straightforward to extend this model to allow for one or more shifts in the value of the optimal phenotype by adding additional parameters for each additional optimum, and for the timing of the shifts between optima. Minor fluctuations in trait values toward the ends of the sequences may be attributable to subtle shifts in the optimal morphology over time (Fig. 1), although sampling error and genetic drift can also cause such variations. One could also model the movement of the adaptive peak itself, for example positing that the optimal phenotype changes over time according to a random walk or some other mode of change (Estes and Arnold 2007). A simple linear trend may be added in several ways, for example as a phenotypic optimum that trends linearly over time with white noise around the trend (Sheets and Mitchell 2001; Estes and Arnold 2007). Alternatively, one could imagine a distribution of evolutionary steps with a nonzero mean, as if selection coefficients fluctuated over time, but favored decreases more often than increases. This model has been given various names, including Brownian motion with a trend (Hansen 1997, p. 1348), directional constant-variance model (Pagel 2002), and general random walk (Hunt 2006). We have fit this model to these stickleback traits, and in each case it garners only trivial support compared to the OU model (results not shown), suggesting that a simpler selection model with a preferred direction of natural selection is less successful at explaining these data than a model of a population ascending a local adaptive peak.

Our interpretation for this fossil sequence is that it represents a population invading a novel limnetic environment with somewhat different conditions from its more littoral ancestral habitat (Purnell et al. 2007). Presumably because of a change in predation intensity (or perhaps water chemistry) relative to its ancestral environment, the optimal phenotype in the new environment is shifted toward forms with more reduced skeletal armor. The population is thus located initially some distance from the adaptive optimum, which it approaches and eventually surmounts. This kind of dynamic must occur very commonly in natural populations, and in fact it has been suggested to be something like a fundamental unit of phenotypic evolution (Leroi 2000; Estes and Arnold 2007). To our knowledge, however, the dataset analyzed here constitutes the only convincing example for which the combination of excellent stratigraphic resolution, a fortunately placed window of observation, and relatively stable optima actually allow for the direct documentation for fossil lineages of microevolutionary trajectories on the adaptive landscape. If this sedimentary record were more typical in its resolution, this evolutionary shift—if it were detected at all—is likely to appear as a single punctuated burst, even though the continuity of change is manifest at finer temporal scales.

ACKNOWLEDGMENTS

We thank many who broke rocks in the hot sun to obtain the fossil stickleback samples and students who helped score the phenotypes and record the data. The form and content of this article were greatly improved because of comments by T. Hansen, S. Arnold, and two anonymous reviewers. This research was supported by NSF grants BSR8111013, EAR9870337, and DEB0322818, the Center for Field Research (Earthwatch), and the National Geographic Society (2869–84). Contribution 1150 from Ecology and Evolution at Stony Brook University.

LITERATURE CITED

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contrl. 19:716–723.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. J. Wildl. Manage. 64:912–923.

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. 1987. Interacting evolutionary constraints in pelvic reduction of threespine sticklebacks, *Gasterosteus aculeatus* (Pisces: Gasterosteidae). Biol. J. Linn. Soc. 31:347–382.

———. 1994. Paleobiology and evolution of the threespine stickleback. Pp. 438–471 *in* M. A. Bell and S. A. Foster, eds. The evolutionary biology of the threespine stickleback. Oxford Univ. Press, Oxford.

———. 2001. Lateral plate evolution in the threespine stickleback: getting nowhere fast. Genetica 112/113:445–461.

Bell, M. A., and S. A. Foster. 1994. Introduction to the evolutionary biology of the threespine stickleback. Pp. 1–27 *in* M. A. Bell and S. A. Foster, eds. The evolutionary biology of the threespine stickleback. Oxford Univ. Press, Oxford.

Bell, M. A., and T. R. Haglund. 1982. Fine-scale temporal variation of the Miocene stickleback *Gasterosteus doryssus*. Paleobiology 8:282–292.

Bell, M. A., J. V. Baumgartner, and E. C. Olson. 1985. Patterns of temporal change in single morphological characters of a Miocene stickleback fish. Paleobiology 11:258–271.

Bell, M. A., M. S. Sadagursky, and J. V. Baumgartner. 1987. Utility of lacustrine deposits for the study of variation within fossil samples. Palaios 2:455–466.

Bell, M. A., G. Orti, J. A. Walker, and J. P. Koenings. 1993. Evolution of pelvic reduction in threespine stickleback fish: a test of competing hypotheses. Evolution 47:906–914.

Bell, M. A., W. E. Aguirre, and N. J. Buck. 2004. Twelve years of contemporary armor evolution in a threespine stickleback population. Evolution 58:814–824.

Bell, M. A., M. P. Travis, and D. M. Blouw. 2006. Inferring natural selection in a fossil threespine stickleback. Paleobiology 32:562–577.

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

———. 1988. Random walk and the biometrics of morphological characters. Evol. Biol. 9:369–398.

Bush, A., M. G. Powell, W. S. Arnold, T. M. Bert, and G. M. Daley. 2002. Time-averaging, evolution and morphological variation. Paleobiology 28:9–25.

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

Cheetham, A. H., and J. B. C. Jackson. 1995. Process from pattern: tests for selection versus random change in punctuated bryozoan speciation. Pp. 184–207 *in* D. H. Erwin and R. L. Anstey, eds. New approaches to speciation in the fossil record. Columbia Univ. Press, New York.

Clegg, S. M., S. M. Degnan, C. Moritz, A. Estoup, J. Kikkawa, and I. P. F. Owens. 2002. Microevolution in island forms: the roles of drift and directional selection in morphological divergence of a passerine bird. Evolution 56:2090–2099.

Cresko, W. A., A. Amores, C. Wilson, J. Murphy, M. Currey, P. Phillips, M. A. Bell, C. B. Kimmel, and J. H. Postlethwait. 2004. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. Proc. Natl. Acad. Sci. USA 101:6050–6055.

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

Gingerich, P. D. 1993. Quantification and comparison of evolutionary rates. Am. J. Sci. 293-A:453–478.

Grant, P. R., and B. R. Grant. 2002. Unpredictable evolution in a 30-year study of Darwin's finches. Science 296:707–711.

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.

Hendry, A. P., T. Day, and E. B. Taylor. 2001. Population mixing and the adaptive divergence of quantitative traits in discrete populations: a theoretical framework for empirical tests. Evolution 55:459–466.

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

———. 2004b. Phenotypic variance inflation in fossil samples: an empirical assessment. Paleobiology 30:487–506.

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

———. 2007. Evolutionary divergence in directions of high phenotypic variance in the ostracode genus *Poseidonamicus*. Evolution 61:1560–1576.

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

Kidwell, S. M., and A. K. Behrensmeyer. 1993. Taphonomic approaches to time resolution in fossil assemblages. Pp. 302 *in* S. J. Culver, ed. Short courses in paleontology. The Paleontological Society, Knoxville, TN.

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

———. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

Leroi, A. M. 2000. The scale independence of evolution. Evol. Develop. 2:67–77.

Lynch, M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. Am. Nat. 136:727–741.

MacFadden, B. J. 1989. Dental character variation in paleopopulations and morphospecies of fossil horses and extant analogs. Pp. 128–141 *in* D. R. Prothero and R. M. Schoch, eds. The evolution of perissodactyls. Oxford Univ. Press, New York.

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.

McKinnon, J. S., and H. D. Rundle. 2002. Speciation in nature: the threespine stickleback model systems. Trends Ecol. Evol. 17:480–488.

McPhail, J. D. 1994. Speciation and the evolution of reproductive isolation in the sticklebacks (*Gasterosteus*) in south-western British Columbia. Pp. 399–437 *in* M. A. Bell and S. A. Foster, eds. The evolutionary biology of the threespine stickleback. Oxford Univ. Press, Oxford.

Mousseau, T. A., and D. A. Roff. 1987. Natural selection and the heritability of fitness components. Heredity 59:181–197.

Pagel, M. 2002. Modelling the evolution of continuously varying characters on phylogenetic trees: the case of hominid cranial capacity. Pp. 269–286 *in* N. MacLeod and P. L. Forey, eds. Morphology, shape and phylogeny. Taylor & Francis, London.

Purnell, M. A., M. A. Bell, D. C. Baines, P. J. B. Hart, and M. P. Travis. 2007. Correlated evolution and dietary change in fossil stickleback. Science 317:1887.

Raup, D. M. 1977. Stochastic models in evolutionary paleobiology. Pp. 59–78 *in* A. Hallam, ed. Patterns of evolution as illustrated by the fossil record. Elsevier Scientific Publishing Company, Amsterdam.

Raup, D. M., and S. J. Gould. 1974. Stochastic simulation and evolution of morphology- towards a nomothetic paleontology. Syst. Zool. 23:305–322.

Raup, D. M., and R. E. Crick. 1981. Evolution of single characters in the Jurassic ammonite *Kosmoceras*. Paleobiology 7:200–215.

Roff, D. A., and T. A. Mousseau. 1987. Quantitative genetics and fitness: lessons from *Drosophila*. Heredity 58:103–118.

Roopnarine, P. D. 2001. The description and classification of evolutionary mode: a computational approach. Paleobiology 27:446–465.

———. 2003. Analysis of rates of morphologic evolution. Annu. Rev. Ecol. Syst. 34:605–632.

Sadler, P. M. 1981. Sediment accumulation rates and the completeness of stratigraphic sections. J. Geol. 89:569–584.

Schindel, D. E. 1980. Microstratigraphic sampling and the limits of paleontological sampling. Paleobiology 6:408–426.

Shapiro, M. D., M. E. Marks, C. L. Peichel, B. K. Blackman, K. S. Nereng, B. Jonsson, D. Schluter, and D. M. Kingsley. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428:717–723.

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 Univ. Press, New York.

———. 1953. The major features of evolution. Columbia Univ. Press, New York.

Travis, J. 1989. The role of optimizing selection in natural populations. Annu. Rev. Ecol. Syst. 20:279–296.

Turelli, M., J. H. Gillespie, and R. Lande. 1988. Rate tests for selection on quantitative characters during macroevolution and microevolution. Evolution 42:1085–1089.

Zeng, Z.-B. 1988. Long-term correlated response, interpopulation covariation, and interspecific allometry. Evolution 42:363–374.

Associate Editor: T. Hansen

Appendix

FITTING EVOLUTIONARY MODELS

The data to be considered consist of a vector of trait means, **z**, with elements corresponding to the measured samples in the time series

($z_1, z_2, \ldots, z_n$). Under the neutral model of an unbiased random walk, the expected evolutionary divergence between ancestor and descendant is always zero, and so the expected trait mean for each sample is equal to the value of the trait at the start of the sequence (z_0):

$$\mathrm{E}[z_i] = z_0. \tag{A1}$$

The variance around this expectation increases linearly with respect to elapsed time with a slope equal to the step variance ($\sigma^2{}_{\mathrm{step}}$),

$$\mathrm{Var}[z_i] = \sigma^2_{\mathrm{step}} t_i + \varepsilon_i, \tag{A2}$$

where t_i is the elapsed time from the start of the sequence to the ith sample (Hunt 2006). In addition to the variance in evolutionary outcomes ($\sigma^2{}_{\mathrm{step}} t_i$), the observed variance of each sample will be increased by sampling error in estimating the trait means. The magnitude of this sampling variance, ε_i, is equal to the sample variance divided by the number of measured individuals in that sample. The covariance among sample means is computed as

$$\mathrm{Cov}[z_i, z_j] = \sigma^2_{\mathrm{step}} t_{\min}, \tag{A3}$$

where $t_{\min}$ is the amount of elapsed time shared by the two samples, i.e., the duration of time between the start of the sequence and the earliest of the two samples. Because the joint distribution of trait means in a sequence is multivariate normal (Hansen and Martins 1996), their distribution is completely specified by their means, variances, and covariances. Note here multivariate refers to multiple samples within a sequence, not multiple phenotypic traits. In addition, it is the trait means that are normally distributed; no assumption is made about how traits themselves are distributed.

In addition to $\sigma^2{}_{\mathrm{step}}$ and z_0 the adaptive OU model has two additional parameters: the optimum phenotype (θ) and the strength of the restraining force (α) around the optimum. Again considering all samples in a sequence jointly, they are multivariate-normally distributed, with expected values of

$$\mathrm{E}[z_i] = \left[1 - \exp(-\alpha t_i)\right]\theta + \exp(-\alpha t_i) z_0, \tag{A4}$$

and variances and covariances of

$$\mathrm{Var}[z_i] = (\sigma^2_{step}/2\alpha)\left[1 - \exp(-2\alpha t_i)\right] + \varepsilon_i \tag{A5}$$

$$\mathrm{Cov}[z_i, z_j] = (\sigma^2_{step}/2\alpha)\exp(-\alpha t_{ij})[1 - \exp(-2\alpha t_{\min})], \tag{A6}$$

where t_i is the time from the start of the sequence to sample i, t_{ij} is the time separating samples i and j, and $t_{\min}$ is again the time separating the beginning of the sequence and the older of the two samples i and j (Hansen 1997; note typographic error in his eq. 1). Again, note that the variance expression includes a term for sampling error on the mean.

From the probability density function of the multivariate normal distribution, the log-likelihood of a sequence of trait values indicated by the vector **z** is equal to

$$\frac{1}{2}\ln(|\mathbf{V}|) - \frac{1}{2}n\ln(2\pi) - \frac{1}{2}\left([\mathbf{z}-\mathbf{m}]^{\mathbf{T}}\,\mathbf{V}^{-1}\,[\mathbf{z}-\mathbf{m}]\right), \tag{A7}$$

where n is the number of samples in the sequence and **m** and **V** are generic symbols for the multivariate mean and covariance matrix of the normal distribution. The elements of **m** and **V** are determined from the expectations, variances, and covariances given above for the models (eqs. A1–A6). Because **m** and **V** are functions of the evolutionary parameters of the models, the maximum-likelihood parameter estimates can be obtained by numerically searching the space of possible values to find the combination of parameters that maximizes equation (A7). For the unbiased random walk model, two parameters are estimated from the data: $\sigma^2{}_{\mathrm{step}}$ and z_0. For the adaptive model, these same parameters are estimated, along with the position of the optimum (θ) and the strength of the restraining force (α) around the optimum. The phenotypic variances are treated as observed data, not free parameters of the model.

The model optimizations were implemented using R code written by GH and bundled into the R package paleoTS. This package is available from the R website (http://www.r-project.org/), and is most conveniently downloaded and installed from within R using the normal package installation procedures (see software help files for details). These functions take advantage of the built-in optimization routines in R, which can implement a variety of hill-searching algorithms. By default, a quasi-Newton method is used to maximize the log-likelihood while constraining the optimization so that impossible parameter values cannot be obtained (e.g., step variances that are less than zero).

MICROEVOLUTIONARY INTERPRETATION OF THE OU MODEL

The OU model has been applied in both micro- and macroevolutionary situations. For long-term macroevolutionary data, the movement of lineages within and between adaptive zones has been modeled as an OU process (Hansen 1997; Butler and King 2004). It is more relevant to the present study that this model also describes the microevolution of a population in the vicinity of a fixed adaptive peak (Lande 1976; Hansen and Martins 1996). In this scenario, the parameters z_0 and θ are interpreted as the values specifying the initial mean phenotype at the start of the sequence and the position of the adaptive peak, respectively. The two remaining fitted parameters, $\sigma^2{}_{\mathrm{step}}$ and α, determine the nature of the evolutionary trajectory from the initial trait mean as it converges toward the optimal phenotype. These parameters themselves depend on aspects of the evolving population and adaptive landscape. The step variance component of change arises from genetic drift, which produces a step variance equal to $\sigma^2{}_{\text{step}} = h^2\sigma^2{}_{\text{P}}/N_e$, where h^2 is the trait heritability, $\sigma^2{}_{\text{P}}$ is the phenotypic variance of the sample, and N_e is the effective population size (Lande 1976). The strength of the restraining force is $\alpha = h^2\sigma^2{}_{\text{P}}/(\omega^2 + \sigma^2{}_{\text{P}})$, where ω^2 is a parameter that indicates the strength of stabilizing natural selection (Lande 1976). If stabilizing selection is weak and Gaussian in form, the fitness function of individuals can be expressed as (Lande 1976; Estes and Arnold 2007)

$$W(z) = \exp\left[-\frac{(z-\theta)^2}{2\omega^2}\right]. \tag{A8}$$

Under these assumptions, ω^2 is analogous to a variance; as it increases, the fitness function becomes increasingly dispersed and stabilizing selection is progressively weaker. When stabilizing selection is weak, ω^2 is much greater than $\sigma^2{}_{\text{P}}$, and so the parameter α is directly proportional to the additive genetic variance, and approximately inversely proportional to ω^2.

The maximum-likelihood estimates of $\sigma^2{}_{\text{step}}$ and α can be used to compute the underlying population genetic and selective conditions by rearranging the above expressions and solving for other parameters. For example, the expression for $\sigma^2{}_{\text{step}}$ can be rearranged to obtain $N_e = h^2\sigma^2{}_{\text{P}}/\sigma^2{}_{\text{step}}$. Conveniently, each variable on the right-hand side of this expression can be estimated from the model fit ($\sigma^2{}_{\text{step}}$), the samples themselves ($\sigma^2{}_{\text{P}}$), or bracketed by plausible values (h^2, which usually ranges between 0.1 and 0.7 for morphological traits). This calculation offers a useful consistency check on the OU model solution because if the model is correct, N_e estimates should be (1) consistent among different traits measured from the same lineage, and (2) plausible given the biology of the lineage under study. Finally, the above expression for α can be rearranged as $\omega^2 = \sigma^2{}_{\text{P}}(h^2/\alpha - 1)$, which allows the strength of stabilizing selection around the optimal phenotype to be estimated (the derivation of this relationship assumes that the traits and fitness surface are both Gaussian [Lande 1976], and so can only be approximately correct for discrete-valued traits such as those analyzed here). Following Hansen (1997), the strength of the restraining force around the optimum can also be expressed as the expected amount of time it takes the population to evolve halfway to the optimal phenotype. This quantity, the phylogenetic half-life, is equal to $\ln(2)/\alpha$.

PHENOTYPIC MEANS AND VARIANCES IN TIME-AVERAGED SAMPLES

The evolutionary models predict changes in population means with respect to time, and although every population need not be sampled, the calculations do require reliable estimates of trait means and variances over time. However these characteristics of populations are available only indirectly because the analytical samples used in the analyses included specimens time averaged over 250 years (125 generations), and are therefore not population snapshots. The way the analyses were performed essentially treated the mean and variance of each sample as point estimates for the population at the midpoint of each time-averaged interval. For trait means, this is not too problematic because the mean of a time-averaged sample is equal to a weighted average of the means of the constituent populations (with the weights equal to the proportional representation of populations in the time-averaged sample). Under most kinds of evolutionary change, the weighted average of population means should be a reasonable approximation of the mean morphology of the midpoint population, especially when sample-to-sample differences are relatively small, as in the sequences analyzed here.

Sample variances are potentially more of a concern because any evolutionary change that occurs within the time-averaged intervals will be conflated with population phenotypic variance, upwardly biasing sample variance estimates. The key issue is the expected magnitude of this variance inflation. Thus far, empirical studies have found that fossil and modern samples usually have comparable levels of variation (MacFadden 1989; Bush et al. 2002; Hunt 2004b), including fossil sticklebacks time averaged over considerably greater durations than in the present study (Bell et al. 1987). These findings support the assumption that, in general, phenotypic variances can be measured reliably from time-averaged samples.

Moreover, some calculations can show that the specific sample variances measured in the present study are unlikely to be strongly inflated by time averaging. Generally, the most severe variance inflation effects should occur when change is strongly directional because trends within the interval of time averaging will result in a larger range of morphologies all lumped together (Hunt 2004a). If we take as a worst-case situation the initial 12 samples of the evolutionary response of each trait (Fig. 1), we can approximate the evolutionary change during this interval by a general random walk model (Hunt 2006; this model is also called Brownian motion with a trend). When this model is fit, its parameter estimates allow one to predict the degree of variance inflation from time averaging (Hunt 2004a, eq. 4). When these calculations are done, the predicted variance inflation from time averaging is less than 1% for all three traits. Therefore, time averaging is not likely to be an important source of error in estimating trait means and variance in this study system.

Supplementary Material

The following supplementary material is available for this article:

Table S1. Data from Bell et al. (2006) Paleobiology 32:562–577, Yellow shading indicates samples analyzed for the present article.

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1558-5646.2007.00310.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.