THE ROYAL SOCIETY

Received 17 July 2001
Accepted 19 September 2001
Published online 20 December 2001

Testing the accuracy of methods for reconstructing ancestral states of continuous characters

Andrea J. Webster† and Andy Purvis*

Department of Biological Sciences, Imperial College, Silwood Park, Ascot, Berkshire SL5 7PY, UK

Many methods are available for estimating ancestral values of continuous characteristics, but little is known about how well these methods perform. Here we compare six methods: linear parsimony, squared-change parsimony, one-parameter maximum likelihood (Brownian motion), two-parameter maximum likelihood (Ornstein–Uhlenbeck process), and independent comparisons with and without branch-length information. We apply these methods to data from 20 morphospecies of Pleistocene planktic Foraminifera in order to estimate ancestral size and shape variables, and compare these estimates with measurements on fossils close to the phylogenetic position of 13 ancestors. No method produced accurate estimates for any variable: estimates were consistently less good as predictors of the observed values than were the averages of the observed values. The two-parameter maximum-likelihood model consistently produces the most accurate size estimates overall. Estimation of ancestral sizes is confounded by an evolutionary trend towards increasing size. Shape showed no trend but was still estimated very poorly: we consider possible reasons. We discuss the implications of our results for the use of estimates of ancestral characteristics.

Keywords: accuracy; ancestral states; continuous characters; Foraminifera; maximum likelihood; parsimony

1. INTRODUCTION

Several methods have recently been developed to estimate ancestral states of continuous characters from trait and phylogenetic data on extant descendants (Maddison 1991; Maddison & Maddison 1992; Martins & Hansen 1997; Pagel 1997, 1999; Schluter *et al.* 1997; Garland *et al.* 1999). Such estimates are increasingly used for choosing among alternative evolutionary scenarios (see Pagel 1999; Alroy 2000; Schluter 2000). However, very little is known about the accuracy of the methods. Simulations suggest reasonable performance (Garland *et al.* 1997; Martins 1999*a*), but comparison of actual ancestral trait values with their estimates requires detailed ancestor–descendant relationships in the context of a reasonably sized phylogeny and reliable trait information for ancestors. However, it was the dearth of knowledge about ancestors that led to the development of the methods in the first place. So far only two comparisons have been published. Oakley & Cunningham (2000) studied bacteriophages whose phylogeny was induced experimentally; although the phylogeny had only eight tips it was known with certainty, and ancestral states could be measured directly and compared with the indirect estimates. In that study, a directional trend in character evolution caused estimates to be very inaccurate. An even smaller study of fossil viverrid carnivorans (Polly 2001) found estimates of one character at four nodes to be reasonable, and found no overall trend in character values over time.

Here we present, to our knowledge, the largest test of accuracy of ancestral-trait estimation, and the broadest comparison of methods, to date. We use the Miocene–Pleistocene planktic Foraminifera, which have one of the most comprehensive fossil records of any group (Pearson 1993). The study involved the measurement of foraminiferan body size from 276 scanning electron micrographs of specimens of 20 descendant and 13 ancestral taxa, the use of six different algorithms to estimate the size of the ancestors from the sizes of the descendants, and comparison of these estimates with the measurements.

2. MATERIAL AND METHODS

(a) *Phylogeny*

We have used a clade within the phylogeny presented by Fordham (1986). Although this is not the most recent study (that being the plexigram analysis of Pearson (1993)), we only used the geologically most recent part of the phylogeny, where there is very little disagreement between the two studies. The study reported by Fordham (1986) is more useful for the present purpose, because it is a self-contained analysis containing a phylogeny with many illustrated specimens from precisely identified time periods (biozones). Molecular phylogenies of foraminiferans (e.g. Darling *et al.* 1997, 2000) are currently too incomplete for our purposes; these are discussed further below.

Most of the methods we test are not able to incorporate missing tip data, so those lineages (branches between a cladogenesis event and either another cladogenesis or an extinction) for which specimens of descendants were not available were removed from the analysis. The subset of Fordham's phylogeny used is shown in figure 1.

* Author for correspondence (a.purvis@ic.ac.uk).
† Present address: School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, UK.

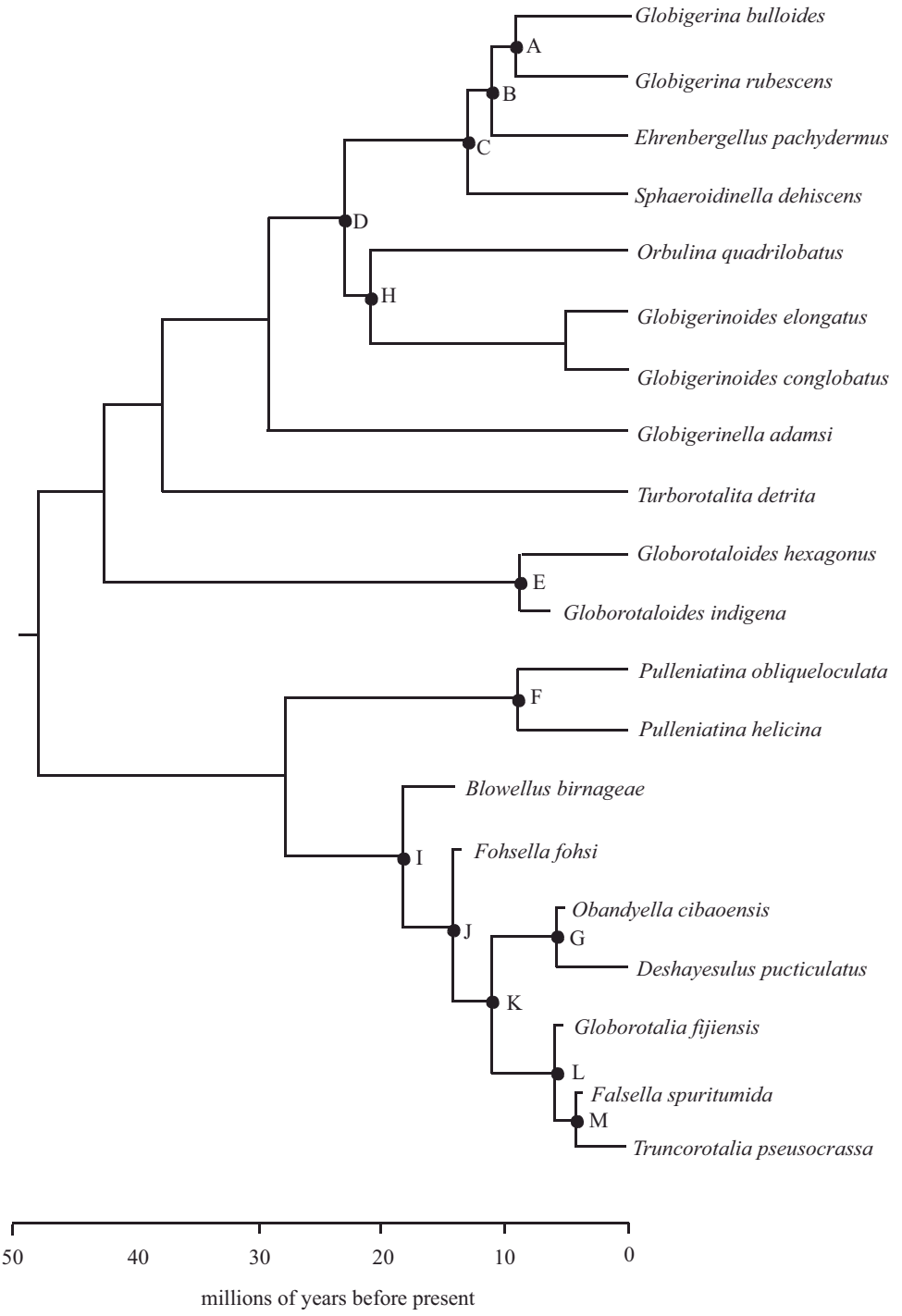

Figure 1. The phylogeny of species used in the current study (after Fordham 1986). Numbers in parentheses and letters at nodes correspond to those in table 1.

(**b**) ***Data***

(i) *Specimens*

Scanning electron micrographs of specimens (Fordham 1986) were scanned into a Macintosh computer and their outlines traced with a pen mouse. The pictures were then analysed using the public domain NIH Image program (developed at the US National Institutes of Health and available at http://rsb.info.nih.gov/nih-image/). The area, ellipse major axis (hereafter, length) and ellipse minor axis (hereafter, width) were calculated with a repeatability error of less than 1% (based on duplicating 20% of the measurements): repeatability was measured as $(|\text{estimate 1} - \text{estimate 2}|/\text{estimate 1}) \times 100$.

Only specimens in biozones adjacent to the age of the node or tip were used, resulting in two possible sets of specimens for each lineage—one from the first biozone of the lineages' duration and one from the last. The size of a morphospecies was estimated as the median value of all measured specimens within it, minimizing the effect of outliers. As above, there were separate size data for the start and end of each lineage. Lineage size was estimated at both the start and end of the lineage as the mean value of the sizes of included morphospecies. The size at a node was estimated as the mean value of the ancestral lineage end size and the two descendant lineage start sizes where available. The reasoning behind this procedure is that the actual

Table 1. The dataset.
(n_a = number of specimens from the end of the ancestral lineage (only ancestral specimens exist for tip species); n_{d1} = number of specimens from the beginning of the daughter lineage nearer the top of figure 1; n_{d2} = number of specimens from the beginning of the other daughter lineage. Numbers for species and letters for nodes correspond to figure 1.)

node or tip	n_a	n_{d1}	n_{d2}	area (mm^2)	length (mm)	width (mm)	width/length
tip species							
1	3	—	—	0.030	0.211	0.167	0.791
2	3	—	—	0.027	0.190	0.173	0.909
3	2	—	—	0.013	0.132	0.125	0.943
4	7	—	—	0.275	0.624	0.554	0.888
5	2	—	—	0.213	0.601	0.434	0.722
6	1	—	—	0.035	0.225	0.198	0.878
7	1	—	—	0.353	0.689	0.652	0.946
8	1	—	—	0.138	0.544	0.323	0.594
9	20	—	—	0.008	0.107	0.092	0.860
10	1	—	—	0.017	0.165	0.132	0.798
11	9	—	—	0.074	0.319	0.295	0.924
12	15	—	—	0.103	0.372	0.344	0.925
13	15	—	—	0.074	0.320	0.282	0.882
14	3	—	—	0.025	0.190	0.163	0.859
15	2	—	—	0.024	0.197	0.154	0.779
16	1	—	—	0.066	0.302	0.277	0.918
17	1	—	—	0.098	0.363	0.342	0.942
18	1	—	—	0.274	0.648	0.538	0.830
19	1	—	—	0.265	0.636	0.530	0.833
20	2	—	—	0.201	0.529	0.466	0.880
ancestral species							
A	28	6	5	0.032	0.213	0.186	0.871
B	10	8	1	0.030	0.196	0.163	0.834
C	16	37	8	0.037	0.222	0.194	0.877
D	0	16	0	0.037	0.224	0.190	0.846
E	6	5	9	0.035	0.201	0.181	0.901
F	4	0	8	0.026	0.188	0.164	0.871
G	0	2	0	0.052	0.270	0.240	0.887
H	0	1	0	0.048	0.262	0.233	0.893
I	0	3	7	0.042	0.243	0.210	0.863
J	7	3	9	0.063	0.298	0.248	0.834
K	9	6	6	0.081	0.329	0.283	0.859
L	2	0	0	0.192	0.502	0.460	0.916
M	2	0	0	0.066	0.313	0.268	0.855

population at the time of cladogenesis presumably had a size somewhere in between the end size of the ancestral lineage and the start sizes of the descendants. Size data were logarithmically transformed prior to analysis, in line with the expectation that the absolute rate of size change will be higher in larger lineages because of the multiplicative nature of growth and hence evolutionary size change. We also calculated a measure of shape—the median width/median length. Shape data were logit transformed ($x' = \log(x/(1 - x))$, where x is the median width/median length). Table 1 shows the data for all four variables and, for each node, the numbers of specimens from which each datum was derived. Because foraminiferans grow continuously, we repeated our analyses using maxima rather than medians (with shape being the ratio of maximum width to maximum length). These analyses gave very similar results, and are not reported further.

Although specimens came from two different sites in the Pacific Ocean (Deep Sea Drilling Project sites 77 (0° 29′ N, 133° 14′ W) and 208 (26° 7′ S, 161° 13′ E); Fordham 1986), data were combined after no significant differences were found in size between sites for comparable lineages, morphospecies or specimens (results not shown).

(c) *Methods of ancestral reconstruction*

Five methods were used to estimate ancestral trait values.

(i) *Linear parsimony*

This was implemented using MacClade (Maddison & Maddison 1992), and calculates the set of ancestral states that minimizes overall change using an algorithm due to Swofford & Maddison (1987). Where an estimate was a range, we used the midrange. Under this method, the amount of change occurring along a branch is independent of its length; this could arise if change occurs only at speciation, or if change is very rare, or if the dynamic of change is very variable through the tree.

(ii) *Unweighted squared-change parsimony*

This was also implemented using MacClade (Maddison & Maddison 1992), and yields the set of ancestral states which minimizes squared change along branches, again with the magnitude of change along branches being independent of their length (Rogers 1984; Felsenstein in Huey & Bennett 1987).

(iii) *One-parameter maximum-likelihood model*

This model (Felsenstein 1981; Schluter *et al.* 1997) yields the most likely ancestral trait values, under the evolutionary model of Brownian motion. An assumption of this model which can be tested with data from descendants alone (the only data available in most studies) is that rates of change are stochastically constant throughout the tree. We tested rate constancy between the clades either side of the root in the phylogeny. Rates were calculated for each pair of sister branches as described by Garland (1992), and compared between the two clades by a *t*-test (16 d.f.) assuming equal variances. This test did not use information from ancestral fossils, because such information is generally not available to workers wishing to test the suitability of the one-parameter maximum-likelihood (ML) model. No rate heterogeneity was found (all $p > 0.5$) indicating that, as far as can be determined from data on descendants alone, use of the model is justified. The computer program ANCML (Schluter *et al.* 1997) was used to estimate ancestral values and the associated standard errors. The standard errors are underestimates in this context, because they do not correct for the joint estimation of trait values for multiple ancestors (Garland *et al.* 1999). Weighted squared-change parsimony (Maddison 1991) yields identical estimates of ancestral values (Schluter *et al.* 1997; Webster & Purvis 2002) and was not considered further.

A drift-based generalized least-squares (GLS) model yields identical estimates of nodal values (Martins & Hansen 1997) but different standard errors associated with them (Martins 1999*b*). Ancestor (Martins 1999*b*) was used to calculate these standard errors too.

(iv) *Two-parameter maximum-likelihood model*

This is an extension of Brownian motion, the Ornstein–Uhlenbeck (O–U) model (Martins 1994). The random walk is constrained, such as would occur when a trait is subjected to a stabilizing selection pressure. We used Ancestor (Martins 1999*b*) to estimate nodal values and standard errors.

(v) *Independent contrasts*

The nodal values estimated as an intermediate stage in the calculation of independent contrasts (Felsenstein 1985) have sometimes been used as estimates of ancestral-trait values (e.g. Owens & Bennett 1995). These values can be estimated with or without branch lengths. We used CAIC (Purvis & Rambaut 1995) to obtain these values, which are weighted means of descendant values. We note that independent contrasts procedures can be modified to yield the same ancestral estimates as the one-parameter ML model and weighted squared-change parsimony, by repeatedly re-rooting the phylogeny (Garland *et al.* 1999).

For a full review of these methods and how they are interconnected, see Webster & Purvis (2002). One further method that has recently been proposed (Pagel 1999) is a directional GLS model. Unlike the drift-based GLS model used here, directional GLS has the strength that it can estimate ancestral values to lie outside the range exhibited by descendants, if descendants differ in their distance from the root. We did not use this method because most of the descendants are from the Upper Pleistocene, so they are roughly equidistant from the root. Molecular phylogenies permit use of directional GLS if the rate of gene evolution varies among lineages (Pagel 1999), as it does in foraminiferans (Darling *et al.* 1997). However, in the discussion we provide evidence that directional GLS may perform badly if used in the future to estimate ancestral sizes of foraminiferans from molecular phylogenies.

(d) *Assessing accuracy*

We used three methods to assess different aspects of the accuracy of ancestral estimates from each method. The first, overall r^2, is the proportion of variance among ancestral values that is explained by the estimates of them, expressed as a percentage. A weakness of this measure is that the correlation between observation and estimates can be high even if the slope relating the two is not unity and/or the intercept of the regression is not zero. We therefore also used matched pairs *t*-tests to test for a significant difference between each set of reconstructions and the observations; the *t*-statistic itself is our accuracy measure (values near to zero indicate high accuracy). Our final measure, which we term overall accuracy, is 1 minus the sum of the squared differences between observed and estimated values, divided by the sum-of-squares of the observed values. An overall accuracy of zero would mean that ancestral estimates were no better than the mean of the true values, while perfect estimation gives an overall accuracy of unity. Unlike the others, this measure of accuracy is comparable not only among methods within this study but also among different studies.

We calculated the confidence limits at each node for the three standard error estimates available (one-parameter ML method (Schluter *et al.* 1997), one- and two-parameter ML method (Martins 1999*b*)), and assessed for each method the percentage of cases where the measured ancestral value lay within the 95% confidence interval associated with the estimate.

3. RESULTS

The comparisons of accuracy are shown in table 2. All methods significantly overestimated all ancestral size variables, with the two-parameter ML method giving the highest overall accuracy for these traits. Shape estimates never differed significantly from the true values, with independent contrasts yielding the smallest *t*-statistics. Alarmingly, overall accuracy is negative for all variables and all methods: overall, the average of the observed data provides a better estimate of the set of ancestral values than do any of the sets of ancestral estimates.

The confidence intervals calculated for the one-parameter ML model using GLS are generally the widest. However, the Ancestor program warned that these particular confidence intervals were unreliable due to an algorithmic failure to converge. All methods gave confidence intervals accurate to 92%.

4. DISCUSSION

The two-parameter ML model performed better than the others, but no method gave good estimates of ancestral values for any trait. For all size variables, the correlation across the 13 nodes between observation and estimate was reasonable, but the estimates were usually significant overestimates, and the overall accuracy was less than zero. Shape is also estimated very poorly: although estimates are unbiased, overall accuracy is again negative, and there is virtually no correlation between estimates and observations.

Three methods provide standard errors. All methods are

Table 2. Comparisons of accuracy of ancestral estimates among six methods for each of four traits.
((*a*) r^2 values (%) for comparison across nodes between estimate and true value. (*b*) *t*-statistics from matched-pairs comparisons between estimates and ancestral values across nodes; $t > \pm 2.189$ indicates significance at $p = 0.05$; positive values indicate estimates higher than true values. (*c*) Overall accuracy. In each case, the best-performing method(s) is highlighted in bold.)

	area	length	width	shape
(*a*)				
LinPars	24.82	20.16	4.14	0.128
SqChPars	31.49	35.15	10.10	1.438
1-par ML	36.46	41.17	38.57	4.503
2-par ML	36.64	41.28	**38.75**	**4.575**
IC BL	**43.45**	**47.57**	16.77	4.239
IC no BL	38.92	43.39	42.88	0.432
(*b*)				
LinPars	2.791	2.978	2.519	−0.287
SqChPars	2.446	2.970	2.578	−0.680
1-par ML	2.263	2.848	2.504	−0.711
2-par ML	**2.236**	2.821	2.476	−0.737
IC BL	2.272	**2.677**	**1.377**	−0.207
IC no BL	2.481	2.866	2.805	**0.095**
(*c*)				
LinPars	−1.759	−2.610	−1.938	−1.127
SqChPars	−0.669	−0.826	−1.244	−1.055
1-par ML	−0.417	−0.524	−0.350	**−0.988**
2-par ML	**−0.407**	**−0.514**	**−0.340**	−0.998
IC BL	−0.937	−1.044	−2.485	−1.332
IC no BL	−1.344	−1.433	−1.158	−1.989

reasonably inaccurate, with 92% of the true ancestral values lying within the confidence intervals of their estimates. However, these confidence intervals are very large, supporting the conclusion of Schluter *et al.* (1997) that ancestral reconstructions are often too variable to be of much use, except to place ancestor sizes within broad limits. Furthermore, real evolutionary dynamics may well be much more complex than any of the models considered here (Alroy 1998, 2000), with many possibilities only distinguishable by using information from observed, rather than estimated, ancestral-trait values. Our results lend weight to the argument that comparative tests should not rely too heavily on precise estimates of ancestral characteristics (Oakley & Cunningham 2000).

The poor accuracy of estimation of ancestral characters is in line with the results reported by Oakley & Cunningham (2000). We computed overall accuracy for the one character (plaque diameter) for which they report sufficient information: it is also negative (range of −1.468 to −0.418) for each of the four methods they employed that did not use the known state of the common ancestor. By contrast, we find Polly's (2001) estimates of the viverrid first lower molar area to have an overall accuracy of 0.567. Performance seems to depend upon whether or not there are evolutionary trends in the characters under study. Like Oakley & Cunningham's (2000) bacteriophages, but unlike the viverrids in Polly's (2001) study, the foraminifera display an evolutionary trend (towards increased size (Arnold *et al.* 1995) that hampers attempts to estimate ancestral character values. This trend is also apparent within our study group (a subset of the species analysed by Arnold *et al.* (1995)): for all size variables, comparisons between measurements of ancestors and those of descendants show that size has tended to increase along lineages (either 20 increases versus 10 decreases, sign test $p = 0.1$; or 21 increases versus 9, sign test $p = 0.04$). Shape does not show a trend over time, however (ancestor–descendant comparisons, sign test, 17 increases and 13 decreases $p = 0.69$). We ascribe the poor estimation of shape to a low signal-to-noise ratio: most species differ only slightly in shape (table 1), and shape is measured as a ratio of two variables and so is likely to contain relatively more error.

The serious effect of trends, also suggested by simulation work (Garland *et al.* 1999; Oakley & Cunningham 2000), leads to two questions. First, are trends sufficiently common to cause problems? Second, can reasonable estimates be obtained when there might be trends, without special knowledge about ancestors?

Trends are not a universal feature of the fossil record (e.g. Jablonski 1997; Roy *et al.* 2000), but there are many well-documented examples of trends (e.g. McNamara 1990; Wagner 1996; Alroy 1998; Saunders *et al.* 1999). The ongoing debate over the nature of evolutionary processes underpinning them (McShea 1994, 1998; Alroy 2000) is only tangential here: any trend will confound attempts to estimate ancestral-trait values, whether it be passive, driven or more complex.

Can trends be accommodated without prior knowledge of their existence? Oakley & Cunningham (2000) showed

that use of an outgroup lineage will improve ancestral estimates only if the outgroup lineage does not show the trend. They also showed that use of a known value at the root greatly improves estimation accuracy; overall accuracy of estimation of plaque diameter then improves from less than 0 to 0.504. Although precise values for the root will seldom be available to those wishing to estimate ancestral values, a broad range of values is more likely to be available; incorporation of such information may prove beneficial. A recently developed method not used here—the directional GLS method (Pagel 1999)—has the potential to identify and accommodate trends, provided that the tips differ in their distance from the root. Put simply, tips nearer to the root are assumed to have diverged less from the root character states: if such tips were all smaller bodied than the fast-evolving tips, the method would infer that the ancestor had been smaller still. Molecular phylogenies of foraminiferans based on small subunit ribosomal DNA show considerable rate heterogeneity among lineages. Although not densely sampled, they permit an informal assessment of whether directional GLS would correctly identify the trend towards larger body size. The phylogeny presented by Darling *et al.* (2000) includes five taxa also present in our study (the same species in two cases; the same genus in the other three). For directional GLS to correctly identify the trend, smaller taxa should tend to have shorter root-to-tip distances in the phylogeny. However, the correlation is if anything negative. In decreasing order of root-to-tip distance, the area measurements for the five taxa (in mm^2) are: *Turborotalita* (0.0080), *Globigerina bulloides* (0.0297), *Globigerinoides conglobatus* (0.3526), *Orbulina* (0.2130) and *Globigerinella* (0.1379), with the first two taxa having very much longer root-to-tip distances than the last three. From this inspection, it would appear that directional GLS might identify a size trend in the wrong direction for this dataset. The likely reason is that smaller foraminiferans have shorter generation times (Arnold *et al.* 1995), which is commonly associated with rapid molecular evolution (Li 1997).

Our study has both weaknesses and strengths when compared with the previous published empirical tests of methods for estimating ancestral states of continuous characters (Oakley & Cunningham 2000; Polly 2001). Ours is, to our knowledge, the largest (20 tips as opposed to 8 and 5), and is able to compare more methods. Oakley & Cunningham (2000) have the currently unique benefits of an independently and perfectly known phylogeny and directly measurable ancestral traits. However, the phylogeny was far from typical, with a totally symmetrical topology and all branches of the same length. Those features limited the range of methods that could be compared (for example, unweighted squared-change parsimony yields the same ancestral estimates as the one-parameter ML model when all branch lengths are equal (Webster and Purvis 2002)). Our study and that reported by Polly (2001) share three difficulties that must be confronted in any study of sexual organisms whose phylogeny and ancestral states cannot be observed directly.

The first issue is the identification of evolutionary lineages. Planktonic forminiferans are easier to work with than most other groups because of the completeness of their fossil record. However, recent molecular genetic evidence (De Vargas *et al.* 1999; Darling *et al.* 2000) has shown that foraminiferan morphospecies can contain several genetically differentiated lineages. This problem is likely to affect any groups in which reproductive isolation commonly arises with little morphological divergence (Knowlton & Weigt 1997). The implication for the present study is that true species' divergences may have occurred earlier than indicated in our phylogeny, leading to overestimation of the rates of change.

The second issue is the precision of measured data for tips and ancestors. The measurement error associated with our data is small (see § 2), but sampling error provides another source of imprecision. Sample sizes are often small; indeed nine of our morphospecies are represented by single specimens (see table 1). The single specimens, though, are those chosen for illustration, and such choices tend to be made on grounds that make them well suited for comparative studies (Arnold *et al.* 1995): specimens are typically well preserved and, importantly when comparing semelparous organisms such as foraminiferans, fully mature. Additionally, most lineages are represented by multiple morphospecies, and ancestors are estimated from the temporally nearest representatives from ancestral and descendant lineages; both these features should help to reduce the effects of sampling error, as on average each descendant or ancestral trait value is derived from over eight specimens. Remaining sampling error will lead to artefactual increases in the rates of change. Whatever the level of sampling error, it is likely to be less than would be found in most other groups with less complete fossil records and so less choice of specimens for illustration.

Finally, the third issue is whether the phylogeny is a correct representation of relationships among lineages. Several estimates of the phylogeny of planktic Foraminifera have recently been constructed using small subunit ribosomal DNA sequences. Although the planktic Foraminifera are not a monophyletic group (Darling *et al.* 1997), the five taxa in our study that have been sequenced are (Darling *et al.* 2000). The molecular and palaeontological estimates of phylogeny within this group are not compatible, but there are reasons for not automatically preferring the molecular estimate: the rate of gene evolution varies by over an order of magnitude among lineages within the group, the genera are separated by long branches with relatively short internal branches, and there is uncertainty over both the intergeneric relationships and the rooting of the group (Darling *et al.* 1997). Further sequencing of more species will clarify the issue.

These problems—lineage identification, data error and phylogenetic uncertainty—are inevitable for any study that aims to compare estimates of ancestral traits based on descendants with estimates based on measurement of putative ancestors themselves. Further studies on groups with adequate fossil records are needed in order to determine the generality of the results found in this study and the robustness of the conclusions they suggest.

This work was supported by the Natural Environment Research Council (UK), through PhD studentship GT4/96/164/T and grant GR3/11526, and by the Leverhulme Trust, through grant F239/AG. We thank three anonymous referees for their comments which helped to improve the manuscript, and in particular the referee who suggested our overall accuracy measure and the logit transformation for our shape data.

REFERENCES

Alroy, J. 1998 Cope's Rule and the dynamics of body mass evolution in North American fossil mammals. *Science* **280**, 731–734.

Alroy, J. 2000 Understanding the dynamics of trends within evolving lineages. *Palaeobiology* **26**, 319–329.

Arnold, A. J., Kelly, D. C. & Parker, W. C. 1995 Causality and Cope's Rule: evidence from the planktonic Foraminifera. *J. Palaeontol.* **69**, 203–210.

Darling, K. F., Wade, C. M., Kroon, D. & Leigh Brown, A. J. 1997 Planktic foraminiferal molecular evolution and their polyphyletic origins from benthic taxa *Mar. Micropalaeontol.* **30**, 251–266.

Darling, K. F., Wade, C. M., Stewart, I. A., Kroon, D., Dingle, R. & Leigh Brown, A. J. 2000 Molecular evidence for genetic mixing of Arctic and Antarctic subpolar populations of planktonic foraminifera. *Nature* **405**, 43–47.

De Vargas, C., Norris, R., Zaninetti, L., Gibbs, S. W. & Pawlowski, J. 1999 Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. *Proc. Natl Acad. Sci. USA* **96**, 2864–2868.

Felsenstein, J. 1981 Evolutionary trees from gene frequencies and quantitative characters—finding maximum likelihood estimates. *Evolution* **35**, 1229–1242.

Felsenstein, J. 1985 Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15.

Fordham, B. G. 1986 Miocene–Pleistocene planktic Foraminifers from D.S.D.P. sites 208 and 77, and phylogeny and classification of Cenozoic species. *Evol. Monogr.* **6**, 1–200.

Garland, T. 1992 Rate tests for phenotypic evolution using phylogenetically independent contrasts. *Am. Nat.* **140**, 509–519.

Garland, T., Martin, K. L. M. & Diaz-Uriarte, R. 1997 Reconstructing ancestral trait values using squared-change parsimony: plasma osmolarity at the origin of amniotes. In *Amniote origins: completing the transition to land* (ed. S. S. Sumida & K. L. M. Martin), pp. 425–501. San Diego: Academic Press.

Garland, T., Midford, P. E. & Ives, A. R. 1999 An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. *Am. Zool.* **39**, 374–388.

Huey, R. B. & Bennett, A. F. 1987 Phylogenetic studies of coadaptation: preferred temperatures versus optimal performance temperatures of lizards. *Evolution* **41**, 1098–1115.

Jablonski, D. 1997 Body-size evolution in Cretaceous molluscs and the status of Cope's rule. *Nature* **385**, 250–252.

Knowlton, N. & Weigt, L. A. 1997 Species of marine invertebrates: a comparison of the biological and phylogenetic species concepts. In *Species: the units of biodiversity* (ed. M. F. Claridge, H. A. Dawah & M. R. Wilson), pp. 199–219. London: Chapman & Hall.

Li, W.-H. 1997 *Molecular evolution*. Sunderland, MA: Sinauer Associates.

McNamara, K. J. (ed.) 1990 *Evolutionary trends*. Tucson: University of Arizona Press.

McShea, D. W. 1994 Mechanisms of large-scale evolutionary trends. *Evolution* **48**, 1747–1763.

McShea, D. W. 1998 Possible largest-scale trends in organismal evolution: eight 'live hypotheses'. *A. Rev. Ecol. Syst.* **29**, 293–318.

Maddison, W. P. 1991 Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. *Syst. Zool.* **40**, 304–314.

Maddison, W. P. & Maddison, D. R. 1992 *MacClade*. Sunderland, MA: Sinauer Associates.

Martins, E. P. 1994 Estimating the rate of phenotypic evolution from comparative data. *Am. Nat.* **144**, 193–209.

Martins, E. P. 1999*a* Estimation of ancestral states of continuous characters: a computer simulation study. *Syst. Biol.* **48**, 642–650.

Martins, E. P. 1999*b* *Compare v. 4.2: computer programs for the statistical analysis of comparative data*. Eugene, OR: Department of Biology, University of Oregon.

Martins, E. P. & Hansen, T. F. 1997 Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* **149**, 646–667.

Oakley, T. H. & Cunningham, C. W. 2000 Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution* **54**, 397–405.

Owens, I. P. F. & Bennett, P. M. 1995 Ancient ecological diversification explains life-history variation among living birds. *Proc. R. Soc. Lond.* B **261**, 227–232.

Pagel, M. 1997 Inferring evolutionary processes from phylogenies. *Zoologica Scripta* **26**, 331–348.

Pagel, M. 1999 Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884.

Pearson, P. N. 1993 A lineage phylogeny for the Paleogene planktonic foraminifera. *Micropalaeontology* **39**, 193–232.

Polly, P. D. 2001 Paleontology and the comparative method: ancestral node reconstructions versus observed values. *Am. Nat.* **157**, 596–609.

Purvis, A. & Rambaut, A. 1995 Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. *Comput. Applicat. Biosci.* **11**, 247–251.

Rogers, J. S. 1984 Deriving phylogenetic trees from allele frequencies. *Syst. Zool.* **33**, 52–63.

Roy, K., Jablonski, D. & Martien, K. K. 2000 Invariant size–frequency distributions along a latitudinal gradient in marine bivalves. *Proc. Natl Acad. Sci. USA* **97**, 13 150–13 155.

Saunders, W. B., Work, D. M. & Nikolaeva, S. V. 1999 Evolution of complexity in paleozoic ammonoid sutures. *Science* **286**, 760–763.

Schluter, D. 2000 *The ecology of adaptive radiation*. Oxford University Press.

Schluter, D., Price, T., Mooers, A. O. & Ludwig, D. 1997 Likelihood of ancestor states in adaptive radiation. *Evolution* **51**, 1699–1711.

Swofford, D. L. & Maddison, W. P. 1987 Reconstructing ancestral character states under Wagner parsimony. *Math. Biosci.* **87**, 199–229.

Wagner, P. J. 1996 Contrasting the underlying patterns of active trends in morphologic evolution. *Evolution* **50**, 990–1007.

Webster, A. J. & Purvis, A. 2002 Ancestral states and evolutionary rates of continuous characters. In *Morphology, shape and phylogenetics* (ed. N. MacLeod, & P. Forey). London: Taylor & Francis. (In the press.)

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.