THE ROYAL SOCIETY

doi 10.1098/rspb.2001.1769

Finding the tree of life: matching phylogenetic trees to the fossil record through the 20th century

Michael J. Benton

Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK (*mike.benton@bristol.ac.uk*)

Phylogenies, or evolutionary trees, are fundamental to biology. Systematists have laboured since the time of Darwin to discover the tree of life. Recent developments in systematics, such as cladistics and molecular sequencing, have led practitioners to believe that their phylogenies are more testable now than equivalent efforts from the 1960s or earlier. Whole trees, and nodes within trees, may be assessed for their robustness. However, these quantitative approaches cannot be used to demonstrate that one tree is more likely to be correct than another. Congruence assessments may help. Comparison of a sample of 1000 published trees with an essentially independent standard (dates of origin of groups in geological time) shows that the order of branching has improved slightly, but the disparity between estimated times of origination from phylogeny and stratigraphy has, if anything, become worse. Controlled comparisons of phylogenies of four major groups (Agnatha, Sarcopterygii, Sauria and Mammalia) do not show uniform improvement, or decline, of fit to stratigraphy through the twentieth century. Nor do morphological or molecular trees differ uniformly in their performance.

Keywords: cladistics; phylogeny; fossil record; stratigraphy; tree

1. INTRODUCTION

Systematics, the discovery of the pattern of the evolution of life, has never been more important. There has been an explosion in the number of new phylogenies published each year (Pagel 1999), and they are being used as a fundamental tool in a wide range of investigations in comparative and evolutionary biology, ecology, behaviour, and biodiversity studies (Harvey *et al.* 1996; Hillis 1997; Huelsenbeck & Rannala 1997; Pagel 1999; Purvis & Hector 2000).

A variety of approaches have been adopted in the past decades to improve the quality of trees. Cladistics (Hennig 1966; Kitching *et al.* 1998) offers techniques for extracting best-fitting trees from matrices of characters, and computer algorithms now ensure that the shortest (i.e. the most parsimonious) trees are extracted (Farris 1988; Felsenstein 1991; Swofford 1999). Other tree-making techniques, commonly used for molecular data, include maximum likelihood, where trees are reconstructed according to their fit to models of evolution: unweighted pair-group method with arithmetic mean, where a tree is reconstructed from pairwise distances among molecular sequences, and neighbour-joining, a distance method in which rate constancy of molecular change is not assumed (Hillis *et al.* 1996; Huelsenbeck & Rannala 1997).

Various approaches have been developed to assess the robustness of trees, their closeness of fit to the data. The consistency index (CI; Kluge & Farris 1969) measures the closeness of fit of all characters in a data matrix to the most parsimonious tree. However, the CI cannot be used to compare trees based on different datasets since it is influenced by the size of the data matrix (number of taxa and number of characters) and by the relative frequencies of character states (Sanderson & Donoghue 1989; Archie & Felsenstein 1993). Techniques have been developed to take account of relative tree length. For example, the overall retention index (RI; Farris 1989) and the homoplasy excess ratio maximum (HERM; Archie 1989) express the actual tree length as a fraction of the range of possible lengths for the data. Two further measures, the homoplasy excess ratio (HER; Archie 1989) and the permutation tail probability (PTP; Faith 1991), compare the favoured tree with trees reconstructed from datasets that have been permuted in various ways, and so give an assessment of whether the structure of a tree is better or worse than random. The robustness of individual branching points (nodes) within trees is commonly quantified by non-parametric bootstrapping (Felsenstein 1985), where the character support for each node is compared with the distribution of character data derived from randomized datasets, and by the decay index, where parsimony is relaxed step-by-step until the node fails (Bremer 1988).

These techniques give quantified measures of robustness, but they cannot offer a statistical measure of confidence since the tree structure they describe may actually reflect something other than a phylogenetic signal (Kitching *et al.* 1998). Tempting as it might be, therefore, the various tree metrics, the CI, RI, HERM, HER, PTP, bootstrapping, and the decay index, cannot be used as comparators among different trees. It would be ideal if systematists simply had to maximize these measures to know that they were approaching the most reliable tree. However, the fact that trees derived from widely different kinds of data, for example morphological and molecular, are often congruent, suggests that the structure is founded on phylogeny (Patterson 1987; Miyamoto & Fitch 1995; Donoghue *et al.* 1995; Hillis 1997). Does this mean, then, that systematists must labour in a vacuum, sequencing and computing into the wee small hours, never knowing whether their trees, produced with so much effort, are any more accurate than a quick intuition scribbled on the back of a beer mat?

The aim of this paper is to explore a possible solution to the dilemma. Correct phylogenies and relatively well-sampled fossil records should tell the same story, if all is well. A large-scale statistically based study of 1000 phylogenies,

Received 14 May 2001 *Accepted* 4 June 2001

dating from 1910 to 1998, is used to explore whether more recently published phylogenies match stratigraphy any better or worse than earlier efforts. Then, in order to control for some of the many potential sources of error in the large-scale study, alternative published phylogenies of four major clades (Agnatha, Sarcopterygii, Sauria, Mammalia) are compared through the 20th century.

2. STRATIGRAPHY VERSUS CLADOGRAM SHAPE

Stratigraphy, in particular the order and spacing through time of fossils in the rocks, is generally accepted as independent of cladistic and molecular phylogeny techniques since there is no consideration of geological age in the choice of characters, nor in the processing techniques (Platnick 1979; Norell & Novacek 1992*a*,*b*; Patterson 1982; Smith 1994; Benton & Hitchin 1996, 1997). Complete independence probably cannot be claimed, despite the evident distance between the molecular biologist's laboratory and the fossil quarry, because (i) unwitting bias may affect the choice of taxa in an analysis (Smith 1994); (ii) the geometry of trees may impose some time-related trends in the acquisition of characters and homoplasies (confounding convergences and reversals; Wagner 2000*a*); and (iii) the mix of monophyletic and paraphyletic groups as terminals may affect the statistics (Wagner 2000*b*; Wagner & Sidor 2000).

A major criticism of the approach adopted here is that it can only work when the fossil record is relatively complete: if there are many gaps, then there is no evidence that the correct tree will fit stratigraphy any better or worse than any of the numerous incorrect trees (Fortey & Jefferies 1982; Wagner 2000*b*). With significant missing fossils, the recorded order of origination of groups could easily be the exact opposite to that implied by the 'correct' tree. So, hanging over any comparisons between stratigraphy and phylogeny is the concern that modest changes in the degree of congruence may say nothing about improvements in the quality of the cladogram. This will be considered further below.

The quality of the fossil record is a contentious issue. There are two key matters: is the fossil record as it exists well-sampled or not, and does the fossil record actually represent the history of life?

There is growing evidence that the fossil record is well-sampled. For example, in a study of the quality of the fossil record of tetrapods from 1967 to 1993, there was a statistically significant improvement in the quality of fit to a fixed set of cladograms (Benton 1994; Benton & Storrs 1994): new discoveries were filling predicted gaps, not creating new gaps. If, on the other hand, new fossil finds created gaps more often than they filled them, palaeontologists would have to retire from the discussion until they could demonstrate some stability in their knowledge of the fossil record. Indeed, Weishampel (1996) found, in comparing the addition of new taxa to trees of dinosaurs, horses, and hominids through the past 120 years, that many new finds did extend trees, and add gaps. However, this effect was seen only for dinosaurs and horses, and the pattern is one of fluctuation, with highest mean ghost range measures in 1900 and 1916, respectively, and lower values since. For hominids, mean ghost ranges have declined from 1920 onwards.

The discovery of constant, or declining mean ghost range (i.e. gaps remaining constant, or being filled by new finds) is borne out by further studies of the effects of new fossil finds on knowledge of the fossil record: Maxwell & Benton (1990) and Sepkoski (1993) found essentially no change in the broad shape of diversification curves computed from 1900 to 1987 for tetrapods and from 1982 to 1992 for marine animals, respectively.

Third, simulations have shown that a very incomplete fossil record can still record the correct pattern of taxon durations, providing the gaps are randomly distributed (Foote 1997). The claim is not that the fossil record is perfect, merely that the distribution of gaps probably does not bias the results. At the scale of the analyses presented here, local biases, such as gaps in successions or particular sites of exceptional fossil preservation, probably do not act as global biases.

There is, then, strong evidence that fossils are well sampled. But how well do those fossils represent the history of life? The assumption implicit in palaeontology textbooks is that the fossil record says enough, even though certain clades of soft-bodied organisms are unknown and unknowable. The relatively comprehensive fossil records of organisms with preservable skeletons are assumed to show what happened in the past, and the worms and jellyfish can be imagined. Indeed, sites of exceptional fossil preservation do preserve worms and jellyfish, but do not hint at vast arrays of unpredicted, major, soft-bodied clades.

An opposing view is that important sectors of the tree of life are not preserved as fossils. Evidence comes from two sources: (i) the discovery that the distribution of sedimentary rocks may control the preservation of fossils; and (ii) molecular evidence that certain major clades originated much earlier than had been determined from fossils. Indeed, this assumption, that global biodiversity curves are strongly dependent on the vagaries of sampling, is a key element in Alroy *et al.*'s (2001) re-analysis of the marine fossil record.

In a study of the marine fossil record of the post-Palaeozoic, Smith (2001) found that outcrop area and sea-level changes correlated with some aspects of diversity change. The large-scale global rise in diversity through the past 250 million years (Myr) was unrelated to outcrop area, but smaller-scale changes in diversity and in origination rate were related to the surface area of outcrop. Most peaks in extinction did not correspond to changes in outcrop area, but two occurred towards the culmination of stacked transgressive system tracts and close to system bases. The question unanswered by this study was whether the sea-level changes were driving evolution or driving preservation. If the former, then sampling is not indicted; if the latter, then the fossil record may not adequately sample certain aspects of the true patterns of diversification. If Smith (2001) is right, and the shape of the rock record limits our ability to sample (and his study did not *prove* that), an interesting further test would be to compare the results with the deep sea, where sedimentation is essentially continuous, and with continental settings, where sedimentation is often supposed to be even more sporadic than in shallow seas. If it really is possible to prove that there are long global-scale hiatuses in the rock record, these should be minimal for open-ocean organisms, and maximal for terrestrial organisms.

Two further observations may weaken Smith's (2001) conclusions. The first is scaling of observation. Advances and retreats of the sea occurring on a scale of 5–20 Myr could clearly affect the accuracy of species-level or generic-level counts, but perhaps longer-spanning higher taxa such as families and orders would be unaffected. The second is the geographical limitation of Smith's (2001) measure of rock outcrop area, which came from geological maps of Britain and France only. Certainly there are global patterns of sea-level change, but it is unclear whether these alone could explain wholesale increases and decreases in the preservation of fossils. Some times of minimal marine deposition in Western Europe, such as the Early Cretaceous, are well represented by marine successions elsewhere (e.g. Russia). Global-scale compilations on fossil diversity can follow the evolution of groups from basin to basin.

Perhaps Smith (2001) has only found evidence for regional bias in current data compilations, reflecting poor sampling of the existing fossil record, and not poor sampling of the tree of life. Datasets on the fossil record are still dominated by knowledge of well-studied parts of the world, such as Europe and North America, where substantial collecting and monographic work began over 200 years ago. Addition of modern data from relatively less well-known parts of the world, such as the southern continents and Asia, may remove or diminish the linkage Smith (2001) found between the 'global' fossil record compilations and areas of outcrop in Western Europe.

The proposal that metazoans originated and diversified some 1000 million years ago, 400 million years before the first fossils (Wray *et al.* 1996; Gu 1998; Wang *et al.* 1999; Cutler 2000), raises the possibility that the first half of the history of most animal phyla was not represented by fossils. Similar claims for the early origins of modern orders of birds and mammals (Hedges *et al.* 1996; Cooper & Penny 1997) have also suggested phylogenetically important gaps. These proposals have been debated (e.g. Ayala *et al.* 1998; Benton 1999; Lee 1999; Morris 2000), but, if correct, a major bias in the fossil record has been identified. Whether this bias applies only in the identified cases, or is indicative of a wider problem in the fossil record, is debatable.

Contrary to these findings is that a number of studies have shown congruence between trees and stratigraphy for a wide range of organisms (Norell & Novacek 1992*a*,*b*; Benton & Storrs 1994; Smith & Littlewood 1994; Benton & Hitchin 1996; Benton *et al.* 1999, 2000). Admittedly, many of the assessed trees cover groups with accepted 'good' fossil records, but others are molecular trees that include clades with poor fossil records. In either case, the finding of congruence is looking beyond the preserved fossil record to the larger pattern of what actually existed. Sedimentary control of the fossil record, as indicated by Smith (2001) would prevent the preservation of whole swathes of hard-bodied and soft-bodied organisms alike. This suggests that, although neither cladograms nor stratigraphy are perfect, they are telling the same story, since it is unlikely that geological biases, which affect the known distributions of fossils, would also affect the ways in which systematists choose and use morphological and molecular characters.

3. MATERIAL AND METHODS

The dataset used here consists of 1000 published cladograms, including one cladogram of 'all life', 33 of plants, 9 of coelenterates, 1 of molluscs, 179 of arthropods, 14 of brachiopods, 1 each of bryozoans and graptolites, 60 of echinoderms, 34 of deuterostomes including calcichordates, 157 of fishes, and 510 of tetrapods, including 26 of amphibians, 203 of reptiles, 8 of birds, and 269 of mammals (Benton *et al.* 2000). Some parts of this compilation are relatively 'complete', such as arthropods, brachiopods, echinoderms, and fishes, for which every accessible cladogram published to the end of 1998 was assessed. The fossil plant cladograms came largely from a single recent compilation (Kenrick & Crane 1997), and the fossil tetrapod cladograms came from five multi-author compilations, plus every cladogram from volumes 1 to 18 of the *Journal of Vertebrate Paleontology* (1981–1998). In addition, every cladogram published in volumes 31–41 of *Palaeontology* (1988–1998) was included. The tetrapod sample also includes 144 molecular trees of mammal phylogeny extracted from a thorough search of recent journals and multi-author volumes (Benton 1998). The full dataset of 1000 cladograms may be found at http://www.palaeo.gly.bris.ac.uk/cladestrat/cladestrat.html.

The fossil record 2 (Benton 1993) was used as the primary source of stratigraphic data for the dates of origin of families and suprafamilial taxa. Some of the cladograms include individual genera or species, and their dates of origin were determined as far as possible from *The fossil record 2*, but also from data within the papers containing the trees (this applied to less than 2% of the 10 388 terminal taxa in the dataset). Origins and extinctions of clades were assessed to the level of the stratigraphic stage or series (mean duration of the 79 time-units used for the Phanerozoic is 6.8 Myr). Geological dates for these stages in Myr were taken from one source (Harland *et al.* 1990), and in every case origins were taken to the start of the stage, extinctions to the end.

The congruence between trees and stratigraphy is assessed by a number of metrics, the stratigraphic consistency index, SCI (Huelsenbeck 1994), the relative completeness index (RCI; Benton & Storrs 1994), and the gap excess ratio (GER; Wills 1999), the relation of the actual summed ghost range in a cladogram to the minimum and maximum possible ghost range. The SCI is the proportion of stratigraphically consistent nodes (those younger than, or equal in age to, the node immediately below) to total nodes in a cladogram. The RCI is:

$$\mathrm{RCI} = 1 - \left[\frac{\sum \mathrm{MIG}}{\sum \mathrm{SRL}}\right] \times 100\%, \tag{3.1}$$

where MIG is the minimum implied gap, or ghost range, and SRL is the standard range length, the known fossil record. The GER is:

$$\mathrm{GER} = 1 - \frac{(\mathrm{MIG} - G_{\min})}{(G_{\max} - G_{\min})}, \tag{3.2}$$

where $G_{\min}$ is the minimum possible sum of ghost ranges and $G_{\max}$ the maximum, for any given distribution of origination dates. These techniques are described fully in the papers cited, together with comments on biases and significance tests. The three metrics were calculated for each of the 1000 trees using

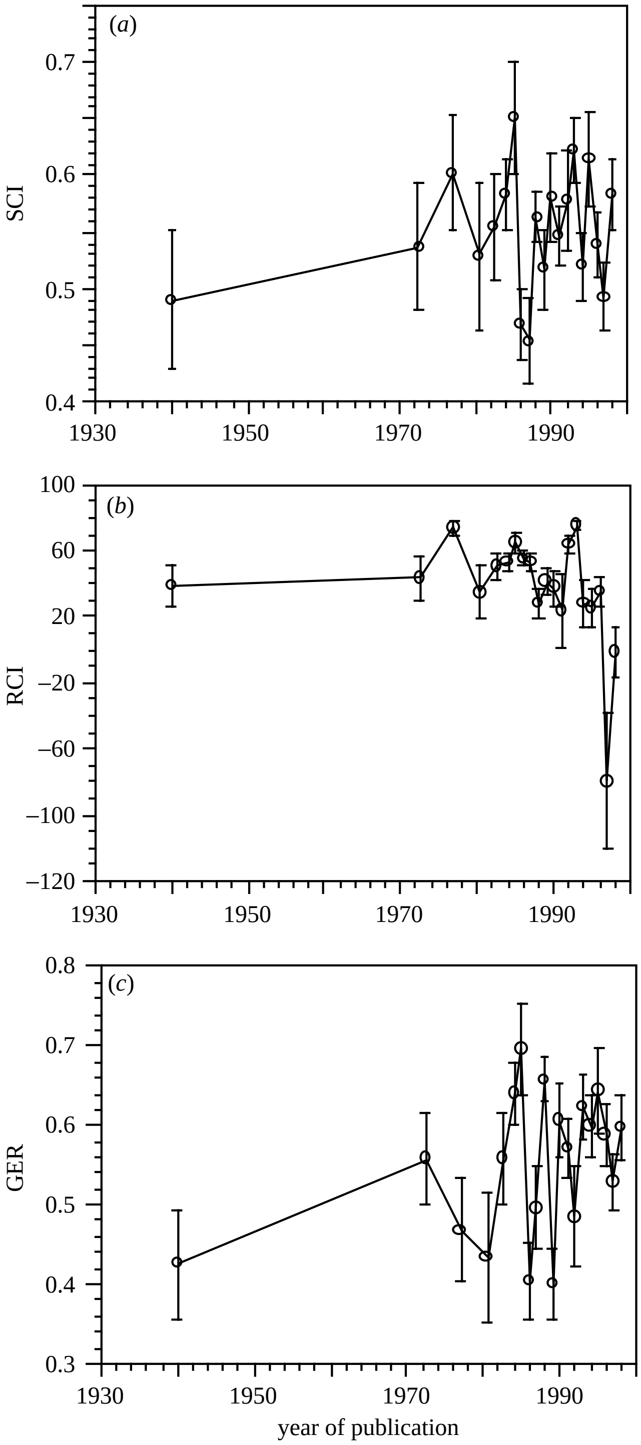

Figure 1. The congruence of cladograms and stratigraphy has perhaps improved through the 20th century, but the evidence is limited. Plots of mean values of the mean stratigraphic consistency index, SCI (*a*), relative completeness index, RCI (*b*), and gap excess ratio, GER (*c*) for a sample of 1000 published cladograms, divided into years of publication. Since relatively few older trees were included, midpoint dates represent a range of years: 1940 (1910–1969), 1972 (1970–75), 1977 (1976–79), 1980 (1980–81), 1982 (1982–83). After 1984, partitions represent a single year of publication. The SCI apparently increases through research time, but not significantly ($y = -1.370 + 0.001$; $r = 0.236$, n.s.), while the RCI apparently declines, but again not significantly ($y = 1420–0.696$; $r = 0.266$, n.s.). The GER improves significantly ($y = -5.097+0.003$; $r = 0.405$, $p < 0.05$).

the software GHOSTS 2.4 (written by Matt Wills), available from http://www.palaeo.gly.bris.ac.uk/cladestrat/cladestrat.html.

Each of these metrics looks at a different aspect of age versus clade congruence and in combination they give a detailed picture of interactions of tree topology and the stratigraphic distribution of known fossils. In each case, an increase in the value of the metric indicates an improvement in congruence (high RCI values can indicate high congruence, good fossil-record completeness, or both).

4. RESULTS OF THE LARGE-SCALE SURVEY

The three age versus clade metrics were assessed for all 1000 cladograms in the sample, and two analyses were carried out: one in which all 1000 points were included, and one in which the cladograms were partitioned by year of publication. Sample sizes in each partition ranged from 16 to 102 (mean, 50). The all-data-point survey (not shown) indicated no statistically significant relationship for the SCI, a marked decline in the RCI ($r = 0.107$, $p < 0.001$), and a slight improvement in the GER ($r = 0.053$, $p < 0.05$). The year-by-year survey (figure 1) also shows considerable fluctuations in all three metrics through the 20th century. There is no significant relationship for the SCI, an apparent, but not statistically significant, decline in the RCI, and a significant ($p < 0.05$) improvement in the GER. Exclusion of the point for 1940 does not affect these statistical measures.

The very low negative value for RCI in 1997 (figure 1*b*) is something of an anomaly, the result of including 10 cladograms of basal land plants (Kenrick & Crane 1997) that cover long spans of time, but have a poor fossil record (hence giving relatively large amounts of ghost range). Without those 10 cladograms, the 1997 RCI value recovers to 27.15 ± 8.04, close to the values for neighbouring publication years. The SCI and GER are not affected by the anomaly (figure 1*a*,*c*), which supports the value of using all three metrics for comparisons.

Perhaps more informative would be a simple division of the dataset into a pre-cladistic and post-cladistic partition. Two partitions were attempted, the first for pre- and post-1985 publications, and the second for pre- and post-1989 publications. The years 1985 and 1989 were chosen, not to represent the introduction of cladistics, but to represent the introduction of computer programs in published cladograms (1985) and the more widespread application of sophisticated computing techniques and the rise of molecular trees (1989). The results (figure 2) show little apparent change in the SCI or GER metrics, but a marked reduction in the RCI metric. The change is not related to changes in tree size. Modern systematists are not, on the whole, publishing larger trees than their forebears: trees in all four time partitions are in the range 9.36 ± 0.81 to 10.60 ± 0.44 terminal taxa, which represents minor variation about the all-tree mean of 10.40 ± 0.40. Equally, there are no other apparent changes in tree geometry, groups tackled, or other factors, among the major time partitions (Benton *et al.* 2000).

5. CASE STUDIES

These comparisons of partitions of the dataset of 1000 trees may mask many variable factors other than tree

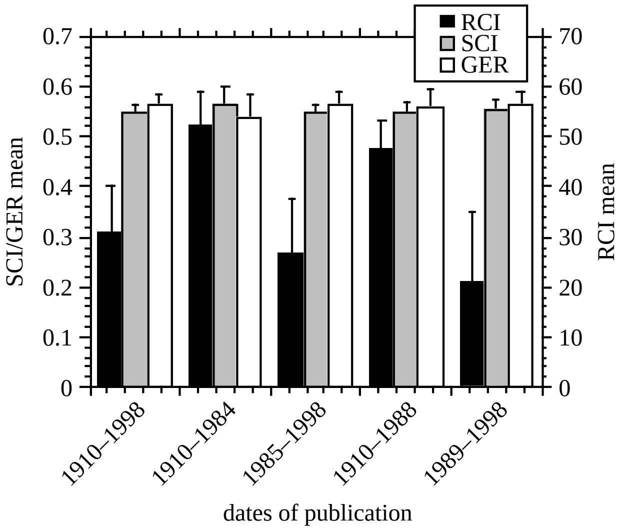

Figure 2. The 'cladistic revolution' (1985 onwards and 1989 onwards) is not detectable by an improvement in age versus clade congruence. Mean values of the SCI and GER are indistinguishable for pre-1985 and post-1985, and the pre-1989 and post-1989 phylogenetic trees, and these are indistinguishable from those of the whole sample (1910–1998). The RCI plummets after the 'cladistic revolution', whether the post-1985 or the post-1989 trees define it. Error bars represent two standard errors. Sample sizes: 1910–1998 ($n = 1000$), 1910–1984 ($n = 167$), 1985–1998 ($n = 833$), 1910–1988 ($n = 385$), 1989–1998 ($n = 615$).

quality. Statistical approaches, including calculation of error bars (figures 1 and 2), should take account of random variations in tree geometry and analytical approaches. A more controlled approach is to compare standardized trees in which different resolutions of the phylogeny of a fixed group of taxa are compared. In these cases, not only is the stratigraphy and distribution of fossils held constant, as it was in the comprehensive survey, but the exact taxa under consideration are also fixed. In combination, these controls should exclude suggestions that all we are seeing are variations in the quality of the fossil record, since the taxa and the stratigraphy are identical throughout each case study.

Four case studies were selected, Agnatha (jawless fishes), Sarcopterygii (lobe-finned fishes), Sauria (lizards) and Eutheria (placental mammals). Each was the subject of a number of major systematic revisions during the 20th century, and it should be possible to track changes in the quality of matching of the trees to stratigraphy through research time.

In each case, a core set of the major taxa within each clade was selected (Agnatha, 6 orders; Sarcopterygii, 8 orders; Sauria, 17 families; Eutheria, 18 orders), and the trees presented by each author were standardized to those taxa. The trees show considerable variation in the three age versus clade metrics (figure 3), sometimes varying seemingly chaotically (figure 3*a*,*d*), and in other cases seemingly varying in concert (figure 3*b*,*c*). None of the four case studies, however, shows any hint of a significant improvement or decline in the fit of cladograms to stratigraphy through research time. In the case of the Eutheria, a mix of morphological and molecular trees was sampled. The protein (p) and gene (g) trees are distinguished (figure 3*d*), but neither of these types of tree seems to be consistently better or worse than the other, or than the morphological trees (unlabelled). The best values for all three metrics are for two protein trees, that of Shoshani (1986), an immunological distance tree based on mammalian sera and albumins, and that of De Jongh *et al.* 1993), based on alpha-lens crystallin sequences, and a gene tree, that of Emerson *et al.* (1999), based on complete 12S rRNA sequences from mitochondria.

6. DISCUSSION AND CONCLUSIONS

The most striking finding of the comprehensive study, and of the case studies, is that there was relatively little change in congruence between stratigraphy and phylogeny through the 20th century, and especially in the past 30 years, a time of major revolution in methods and data sources. Bearing in mind that the stratigraphic distribution of the fossils was held constant by using a single primary source of data from 1993, the results might suggest that the advent of cladistics and of molecular phylogeny reconstruction techniques has had a minimal effect on the results obtained by systematists.

Why did the RCI drop through the 20th century (figure 1*b* and figure 2), indicating an increase in the relative proportion of ghost ranges to known ranges? There are five possible suggestions: perhaps the more recent trees differ from the earlier ones in their size or shape (i), taxic level (ii), influence of stratigraphy in their construction (iii), proportions of morphological to molecular trees (iv), or ease of resolution (v).

The first three suggestions may be rejected. The RCI measure may indeed reflect variation in tree size and geometry (Benton *et al.* 1999, 2000; Wills 1999), but there is no evidence for regular change in these factors during the 20th century (table 1). Nor does the mean taxic level of trees (e.g. species, family, order) change through the sampled time-span. Third, it cannot be claimed that being founded on stratigraphy biases all the older trees: of the pre-1985 sample of 167 trees, only 25 are non-cladistic or non-molecular. In any case, the SCI and GER metrics are unaffected.

The drop in the RCI may reflect a combination of the fourth and fifth suggestions: that after 1985 there are more molecular trees in the sample, and systematists are now tackling harder-to-resolve trees. Perhaps molecular trees simply match stratigraphy less well than morphological trees? This was indeed found in a study of mammal trees (Benton 1998), in which RCI values were on average 10% better for morphological than for molecular trees. Harder-to-resolve phylogenies may also play a part. The post-1985 sample includes large-scale trees of 'all life', 'all land plants', and deuterostomes, as well as trees of the major clades of brachiopods, bryozoans, molluscs and other groups. Before 1985, it was rare for systematists to attempt such broad-scale trees based on detailed character analysis.

The apparent improvement in the GER through the 20th century (figure 1*c*) indicates that more recently published trees imply sums of ghost ranges that are closer to the theoretical minimum than earlier trees. The distribution of taxa in newer trees evidently matches more closely their order of appearance in the fossil record than in older trees. Since the current fossil record is the control here, the ordering of branches in newer cladograms has evidently improved. Since the SCI has not changed significantly through research time (figure 1*a*), this improvement in ordering of taxa has not been accompanied by an improvement in the proportion of stratigraphically consistent nodes.

The generally modest changes in the congruence of stratigraphy and phylogeny through the 20th century, and in particular the seemingly unchanged SCI metric for the large-scale sample (figure 1*a*), and the mixed results from the case studies (figure 2), could suggest that stratigraphy is inadequate as a yardstick for assessing cladograms (Wagner 2000*b*; Smith 2001). According to this argument, trees *have* improved through the 20th century, but the stratigraphic record is so poorly sampled that it is incapable of detecting that improvement. As estimates of phylogeny improved, and branches in phylogenies moved around, there has been no reduction in ghost ranges (and hence increases in the RCI and GER), nor have the sequences of branches come to match the sequences of fossils any better (leading to higher SCI values).

Neither viewpoint can be demonstrated conclusively. The supporters and the opponents of the fossil record as a control against which to compare trees would both argue

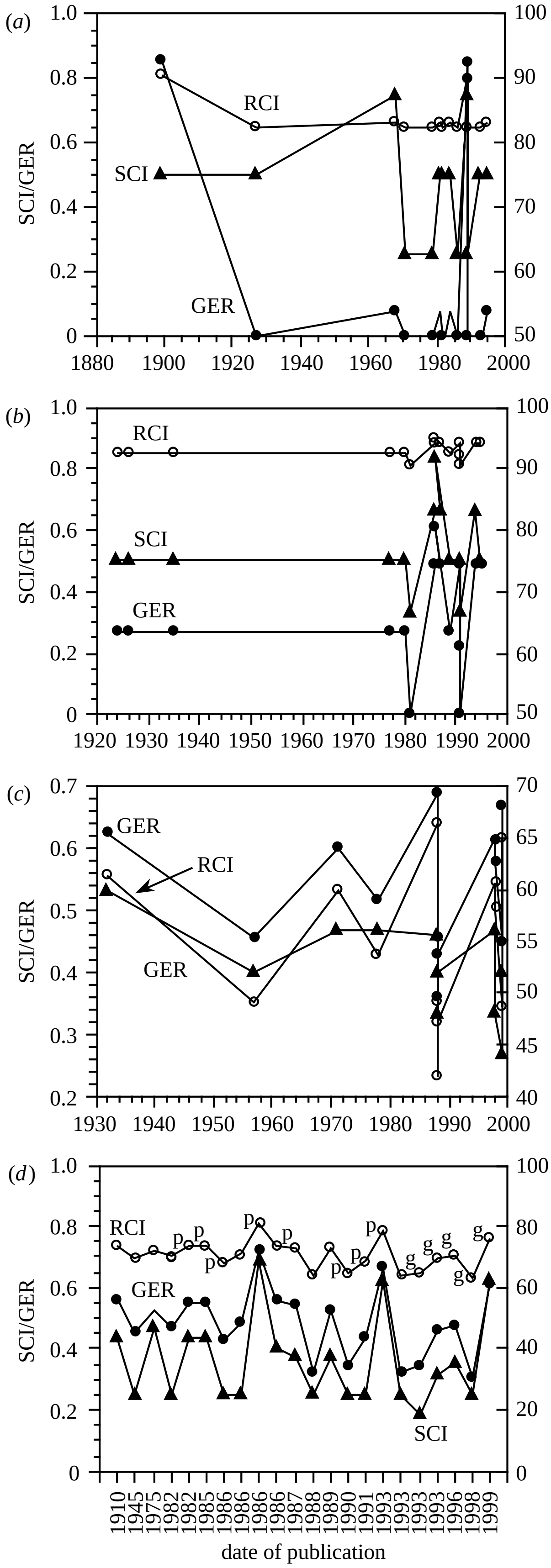

Figure 3. Individual published trees for a variety of vertebrate groups show no convincing trends, either of improvement or worsening, through research time. Trees were analysed for (*a*) Agnatha (jawless fishes; 13 trees, standardized for 6 orders and supraordinal taxa), (*b*) Sarcopterygii (lobe-finned fishes; 15 trees, standardized for 8 orders and supraordinal taxa), (*c*) Sauria (lizards; 13 trees, standardized for 17 families), and (*d*) Eutheria (placental mammals; 22 trees, standardized for 18 orders). Published trees were standardized for taxa so that age versus clades metrics may be compared directly, the only variable being the arrangement of the clades in the trees. In all cases, the values of the metrics are plotted against a time-scale of publication dates, arranged as a time axis in (*a*)–(*c*), and as unit dates in (*d*). Trees were all based on morphological data for Agnatha, Sarcopterygii, and Sauria (except the last, 1999, tree), and on a mix of morphological (unlabelled), protein (p), and gene (g) trees for Eutheria. Pre-cladistic (i.e largely pre-1970) trees are no better or worse in their fit to stratigraphy than cladistic and molecular trees. The data points are derived from analyses of standardized trees of (*a*) Agnatha from Cope (1899), Stensiö (1927), Jarvik (1968), Moy-Thomas and Miles (1971), Janvier & Blieck (1979), Janvier (1981), Halstead (1982), Forey (1984), Maisey (1986), Novitskaya & Talimaa (1989), Gagnier (1989), Forey & Janvier (1993), and Forey (1995); (*b*) Sarcopterygii from Goodrich (1924), Watson (1926), Säve-Söderbergh (1935), Miles (1977), Gardiner (1980), Rosen *et al.* (1981), Maisey (1986), Panchen & Smithson (1987), Schultze (1987), Ahlberg (1989), Ahlberg (1991), Chang (1991), Forey *et al.* (1991), Schultze (1994), Cloutier & Ahlberg (1995); (*c*) Sauria from Camp (1923), Underwood (1957), Underwood (1971), Northcutt (1978), Estes *et al.* (1988, two trees), Presch (1988), Rieppel (1988), and Schwenk (1988); and (*d*) Eutheria from Gregory (1910), Simpson (1945), McKenna (1975), Novacek (1982), De Jongh (1982), Goodman (1985), Miyamoto and Goodman (1986), Novacek and Wyss (1986), Shoshani (1986, two trees), Wyss *et al.* (1987), Novacek (1988), Novacek (1989), Czelusniak *et al.* (1990), Pettigrew (1991), De Jongh *et al.* (1993), Honeycutt & Adkins (1993, two trees), Springer & Kirsch (1993), Porter *et al.* (1996), Stanhope *et al.* (1998), and Emerson *et al.* (1999).

Table 1. Mean tree statistics and age versus clade metrics for a sample of 1000 cladograms, divided according to date of publication.

(The number of trees in each sample is noted, together with mean values of the ghost range (minimum implied gap, MIG); the duration of origins of all taxa within each cladogram (G_{min}), the relative completeness index (RCI), the stratigraphic consistency index (SCI), and the gap excess ratio (GER) for each year sample are also given. The standard error (s.e.) for each mean measure is also noted.)

dates	no. trees	tree size (no. of taxa)	s.e.	MIG	s.e.	G_{min}	s.e.	RCI	s.e.	SCI	s.e.	GER	s.e.
1910–1969	16	12.31	1.164	499.25	64.100	139.25	18.731	38.394	6.5044	0.490	0.031	0.426	0.034
1970–1975	17	11.00	2.033	505.37	64.396	182.18	15.304	43.465	6.8220	0.537	0.028	0.557	0.029
1976–1979	32	8.41	0.798	267.06	21.704	120.53	8.195	73.781	2.0867	0.602	0.025	0.468	0.032
1980–1981	21	6.33	0.558	277.71	32.280	155.71	17.971	34.863	8.1570	0.529	0.033	0.434	0.041
1982–1983	30	10.00	1.020	358.07	32.350	156.20	12.704	50.740	4.0205	0.554	0.023	0.558	0.028
1984	50	9.36	0.617	393.63	28.981	161.62	8.681	53.341	2.3179	0.584	0.016	0.641	0.020
1985	29	7.90	0.683	243.68	23.794	128.31	10.553	65.351	3.0031	0.651	0.025	0.696	0.029
1986	49	10.80	0.858	452.02	53.058	134.71	11.075	55.327	2.4420	0.468	0.015	0.405	0.024
1987	37	10.73	0.639	301.70	26.859	99.95	8.567	53.096	3.1457	0.454	0.019	0.496	0.026
1988	102	12.22	0.656	227.62	10.240	88.79	3.408	28.451	4.6091	0.563	0.011	0.657	0.014
1989	65	10.17	0.689	227.28	17.770	86.54	6.610	41.193	3.8851	0.517	0.018	0.400	0.023
1990	51	9.43	0.875	264.06	26.305	103.16	7.546	37.570	5.4588	0.581	0.020	0.606	0.023
1991	76	10.05	0.797	343.07	42.498	101.76	5.526	23.814	10.896	0.546	0.013	0.571	0.018
1992	42	8.26	0.649	267.15	30.485	118.31	9.960	64.032	2.4732	0.578	0.022	0.485	0.031
1993	77	8.82	0.497	201.42	19.894	89.18	7.883	75.006	1.3807	0.622	0.014	0.623	0.021
1994	62	10.50	0.668	245.18	22.385	84.76	5.919	27.920	6.7724	0.520	0.015	0.599	0.019
1995	40	9.12	0.543	335.52	28.446	164.25	13.708	25.893	5.7821	0.614	0.021	0.643	0.026
1996	64	10.66	0.555	543.13	30.539	205.77	11.313	35.045	4.8428	0.539	0.015	0.588	0.019
1997	88	12.76	1.073	468.75	23.881	190.97	8.164	−79.818	20.554	0.492	0.015	0.529	0.017
1998	48	13.44	1.155	534.13	62.816	120.42	10.316	−1.497	7.4865	0.583	0.015	0.597	0.021

that trees have surely improved through the 20th century as a result of the introduction of cladistics, tree-finding algorithms, and molecular sequencing methods. However, these techniques have been widely used only since the 1980s, and methods are still very much under discussion and development. As cladistic and molecular approaches are refined, it will be interesting to revisit this question in 10 or 20 years' time. Further, the ability of the fossil record to document the history of life requires further assessment.

This study was funded by the Leverhulme Trust (grant F182/AK) and a University of Bristol Research Fellowship. Thanks are due to Mike Foote, Paul Pearson, Andrew Smith, and an anonymous referee for helpful comments on earlier versions of the manuscript. All data used in this study, including all references cited in the caption to figure 4, are given at http://www.palaeo.bris. ac.uk/cladestrat/cladestrat.html.

REFERENCES

Alroy, J. (and 25 others) 2001 Effects of sampling standardization on estimates of phanerozoic marine diversification. *Proc. Natl. Acad. Sci. USA* **98**, 6261–6266.

Archie, J. W. 1989 Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. *Syst. Zool.* **38**, 253–269.

Archie, J. W. & Felsenstein, J. 1993 The number of evolutionary steps on random and minimum length trees for random evolutionary data. *Theor. Popul. Biol.* **43**, 52–79.

Ayala, F. J., Rzhetsky, A. & Ayala, F. J. 1998 Origin of metazoan phyla: molecular clocks confirm paleontological estimates. *Proc. Natl Acad. Sci. USA* **95**, 606–611.

Benton, M. J. (ed.) 1993 *The fossil record 2*. London: Chapman & Hall.

Benton, M. J. 1994 Palaeontological data, and identifying mass extinctions. *Trends Ecol. Evol.* **9**, 181–185.

Benton, M. J. 1998 Molecular and morphological phylogenies of mammals: congruence with stratigraphic data. *Mol. Phylogenet. Evol.* **9**, 398–407.

Benton, M. J. 1999 Early origins of modern birds and mammals: molecules vs. morphology. *BioEssays* **21**, 1043–1051.

Benton, M. J. & Hitchin, R. 1996 Testing the quality of the fossil record by groups and by major habitats. *Historical Biol.* **12**, 111–157.

Benton, M. J. & Hitchin, R. 1997 Congruence between phylogenetic and stratigraphic data on the history of life. *Proc. R. Soc. Lond.* B **264**, 885–890.

Benton, M. J. & Storrs, G. W. 1994. Testing the quality of the fossil record: paleontological knowledge is improving. *Geology* **22**, 111–114.

Benton, M. J., Hitchin, R. & Wills, M. 1999 Assessing congruence between cladistic and stratigraphic data. *Systematic Biology* **48**, 581–596.

Benton, M. J., Wills, M. & Hitchin, R. 2000 Quality of the fossil record through time. *Nature* **403**, 534–537.

Bremer, K. 1988 The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42**, 795–803.

Cooper, A. & Penny, D. 1997. Mass survival of birds across the Cretaceous–Tertiary boundary: molecular evidence. *Science* **275**, 1109–1113.

Cutler, D. J. 2000 Estimating divergence times in the presence of an overdispersed molecular clock. *Mol. Biol. Evol.* **17**, 1647–1660.

De Jongh, W. W., Leunissen, J. A. M. & Wistow, G. J. 1993 Eye lens crystallins and the phylogeny of placental orders:

evidence for a macroscelid–paenungulate clade? In *Mammal phylogeny*, vol. 1 (ed. F. S. Szalay, M. J. Novacek & M. C. McKenna), pp. 5–12. New York: Springer.

Donoghue, M. J., Kim, J. & de Queiroz, A. 1995 Separate versus combined analysis of phylogenetic evidence. *A. Rev. Ecol. Syst.* **26**, 657–681.

Emerson, G. L., Kilpatrick, C. W., McNiff, B. E., Ottenwalder, J. & Allard, M. W. 1999 Phylogenetic relationships of the Order Insectivora based on complete 12S rRNA sequences from mitochondria. *Cladistics* **15**, 221–230.

Faith, D. P. 1991 Cladistic permutation tests for monophyly and nonmonophyly. *Syst. Zool.* **40**, 366–375.

Farris, J. S. 1988 *Hennig86, v. 1.5.* Port Jefferson, NY: J. S. Farris.

Farris, J. S. 1989 The retention index and homoplasy excess. *Syst. Zool.* **34**, 406–407.

Felsenstein, J. 1985 Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.

Felsenstein, J. 1991 *Phylip: phylogenetic inference package, v. 3.4.* Seattle, WA: University of Washington.

Foote, M. 1997 Estimating taxonomic durations and preservation probability. *Paleobiology* **23**, 278–300.

Fortey, R. A. & Jefferies, R. P. S. 1982 Fossils and phylogeny—a compromise approach. In *Problems of phylogenetic reconstruction* (ed. K. A. Joysey & A. E. Friday), pp. 197–234. Systematics Association Special Volume 21. London: Academic Press.

Gu, X. 1998 Early metazoan divergence was about 830 million years ago. *J. Mol. Evol.* **47**, 369–371.

Harland, W. B., Armstrong, R. L., Cox, A. V., Craig, L. E., Smith, A. G. & Smith, D. G. 1990 *A geologic time scale 1989.* Cambridge University Press.

Harvey, P. H., Leigh Brown, A. J., Smith, J. M. & Nee, S. (eds) 1996 *New uses for new phylogenies*. Oxford University Press.

Hedges, S. B., Parker, P. H., Sibley, C. G. & Kumar, S. 1996 Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**, 226–229.

Hennig, W. 1966 *Phylogenetic systematics*. Urbana, IL: University of Illinois Press.

Hillis, D. M. 1997 Biology recapitulates phylogeny. *Science* **276**, 218–219.

Hillis, D. M., Moritz, C. & Mable, B. K. 1996 *Molecular systematics*, 2nd edn. Sunderland, MA: Sinauer Associates.

Huelsenbeck, J. P. 1994 Comparing the stratigraphic record to estimates of phylogeny. *Paleobiology* **20**, 470–483.

Huelsenbeck, J. P. & Rannala, B. 1997 Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**, 227–232.

Kenrick, P. & Crane, P. 1997 *The origin and early diversification of land plants: a cladistic study.* Washington, DC: Smithsonian Institution Press.

Kitching, I. J., Forey, P. L., Humphries, C. J. & Williams, D. M. 1998 *Cladistics; the theory and practice of parsimony analysis*, 2nd edn. Oxford University Press.

Kluge, A. G. & Farris, J. S. 1969 Quantitative phyletics and the evolution of anurans. *Syst. Zool.* **18**, 1–32.

Lee, M. 1999 Molecular clock calibrations and metazoan divergence dates. *J. Mol. Evol.* **49**, 385–391.

Maxwell, W. D. & Benton, M. J. 1990 Historical tests of the absolute completeness of the fossil record of tetrapods. *Paleobiology* **16**, 322–335.

Miyamoto, M. M. & Fitch, W. M. 1995 Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* **44**, 64–76.

Morris, S. C. 2000 The Cambrian 'explosion': slow-fuse or megatonnage? *Proc. Natl Acad. Sci. USA* **97**, 4426–4429.

Norell, M. A. & Novacek, M. J. 1992*a* Congruence between superpositional and phylogenetic patterns: comparing cladistic patterns with fossil records. *Cladistics* **8**, 319–337.

Norell, M. A. & Novacek, M. J. 1992*b* The fossil record and evolution: comparing cladistic and paleontologic evidence for vertebrate history. *Science* **255**, 1690–1693.

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

Patterson, C. 1982 Morphological characters and homology. In *Problems of phylogenetic reconstruction* (ed. K. A. Joysey & A. E. Friday), pp. 21–74. Systematics Association Special Volume 21. London: Academic Press.

Patterson, C. (ed.) 1987 *Molecules and morphology in evolution: conflict or compromise?* Cambridge University Press.

Platnick, N. I. 1979 Philosophy and the transformation of cladistics. *Syst. Zool.* **28**, 537–546.

Purvis, A. & Hector, A. 2000 Getting the measure of biodiversity. *Nature* **405**, 212–219.

Sanderson, M. J. & Donoghue, M. J. 1989 Patterns of variation in levels of homoplasy. *Evolution* **43**, 1781–1795.

Sepkoski Jr, J. J. 1993 Ten years in the library: how changes in taxonomic data bases affect perception of macroevolutionary pattern. *Paleobiology* **19**, 43–51.

Shoshani, J. 1986 Mammalian phylogeny: comparison of morphological and molecular results. *Mol. Biol. Evol.* **3**, 222–242.

Smith, A. B. 1994 *Systematics and the fossil record.* Oxford: Blackwell Scientific.

Smith, A. B. 2001 Large-scale heterogeneity of the fossil record: implications for Phanerozoic biodiversity studies *Phil. Trans. R. Soc. Lond.* B **356**, 351–367.

Smith, A. B. & Littlewood, D. T. J. 1994 Paleontological data and molecular phylogenetic analysis. *Paleobiology* **20**, 259–273.

Swofford, D. L. 1999 *Paup: phylogenetic analysis using parsimony, Paup* 4.0 beta 4a.* New York: Sinauer Associates.

Wagner, P. J. 2000*a* Exhaustion of morphologic character states among fossil taxa. *Evolution* **54**, 365–386.

Wagner, P. J. 2000*b* The quality of the fossil record and the accuracy of phylogenetic inferences about sampling and diversity. *Syst. Biol.* **49**, 65–86.

Wagner, P. J. & Sidor, C. A. 2000 Age rank/clade rank metrics—sampling, taxonomy, and the meaning of 'stratigraphic consistency'. *Syst. Biol.* **49**, 463–479.

Wang, D. Y. C., Kumar, S. & Hedges, S. B. 1999 Divergence time estimates for the early history of animal phyla and the origins of plants, animals and fungi. *Proc. R. Soc. Lond.* B **266**, 163–171.

Weishampel, D. B. 1996 Fossils, phylogeny, and discovery: a cladistic study of the history of tree topologies and ghost lineage durations. *J. Vertebr. Paleontol.* **16**, 191–197.

Wills, M. A. 1999 The gap excess ratio, randomization tests, and the goodness of fit of trees to stratigraphy. *Syst. Biol.* **48**, 559–580.

Wray, G. A., Levinton, J. S. & Shapiro, L. H. 1996 Molecular evidence for deep precambrian divergences among metazoan phyla. *Science* **274**, 568–573.

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