Congruence between phylogenetic and stratigraphic data on the history of life

MICHAEL J. BENTON AND REBECCA HITCHIN

Department of Geology, University of Bristol, Bristol BS8 1RJ, UK

SUMMARY

The quality of the fossil record and the accuracy of reconstructed phylogenies have been debated recently, and doubt has been cast on how far current knowledge actually reflects what happened in the past. A survey of 384 published cladograms of a variety of animals (echinoderms, fishes, tetrapods) shows that there is good agreement between phylogenetic (character) data and stratigraphic (age) data, based on a variety of comparative metrics. This congruence of conclusions from two essentially independent sources of data confirms that the majority of cladograms are broadly accurate and that the fossil record, incomplete as it is, gives a reasonably faithful documentation of the sequence of occurrence of organisms through time.

1. INTRODUCTION

The incompleteness of the fossil record has been recognized since before Darwin (1859), and this is a topic of continuing concern (Raup 1972; Paul 1982, 1990; Signor 1982; Allison & Briggs 1991; Benton 1995; Benton & Storrs 1996). There are two kinds of incompleteness. First is the loss of taxa: the majority of individual plants and animals that lived in the past do not become fossils. Indeed, the majority of species have probably never been fossilized, and the same must therefore apply, although presumably in diminishing proportions, to larger constituent clades.

The fossil record is also incomplete in terms of characters: many taxa that have been fossilized are preserved only incompletely, primarily because of the loss of soft tissues. This has led some commentators (e.g. Nelson 1969; Løvtrup 1977; Hennig 1981; Nelson & Platnick 1981; Patterson 1981; Goodman 1989) to suggest that fossils should either be ignored, or accorded a much lower value than extant taxa, in reconstructing phylogeny. This opinion has been refuted by other systematists (Schaeffer *et al.* 1972; Hecht 1976; Gauthier *et al.* 1988; Norell & Novacek 1992*a*, *b*; Smith 1994), who have shown that fossils offer unique character combinations that have been critical in resolving cladogram topologies.

2. TESTING THE CONGRUENCE OF PHYLOGENETIC AND STRATIGRAPHIC DATA

Cladistic analysis of morphological and molecular data may provide the key to testing evolution as a pattern (Patterson 1982). Cladograms are compiled by assessment of character data and without reference to geological age or assumptions about evolutionary process (Platnick 1979; Eldredge & Cracraft 1980; Nelson & Platnick 1981; Forey *et al.* 1992), which makes it possible to compare independently the patterns revealed by cladograms with patterns derived from geological information on the order of occurrence of taxa (Gauthier *et al.* 1988; Norell 1992; Norell & Novacek 1992*a*, *b*; Benton 1994, 1995; Benton & Storrs 1994, 1996; Benton & Hitchin 1996).

A number of metrics have been proposed for assessing the congruence of cladistic and stratigraphic data, and these include: (i) Spearman rank correlation (SRC) (Norell & Novacek 1992*a*, *b*) of the order of group origins from cladistic and stratigraphic data; (ii) the relative completeness index (RCI) (Benton & Storrs 1994), based on a comparison of known and inferred stratigraphic ranges for a cladogram; and (iii) the stratigraphic consistency index (SCI) (Huelsenbeck 1994), based on an assessment of the stratigraphic consistency of each node in a cladogram (figure 1).

It has been suggested (Siddall 1996) that the SCI metric is strongly dependent on both cladogram size (n) and on tree imbalance (Im), the degree to which a cladogram is entirely symmetrical (balanced) or pectinate. By implication, the RCI metric might also suffer the same problems of bias. However, extensive tests of our sample of cladograms (Hitchin & Benton 1997) have not confirmed any evidence for size bias in the SCI or RCI metrics. The only statistically significant correlation between SCI and n was a positive relationship for the SCI measure for fishes ($r^2 = 0.046$, $p = 0.021$), and a negative relationship for the RCI measure for continental tetrapods ($r^2 = 0.109$, $p = 0.002$). Size bias was found for the SRC test, as expected (see below).

Size bias does not seem to be a problem for the RCI and SCI metrics, nor does cladogram imbalance. We found no statistically significant relationship between

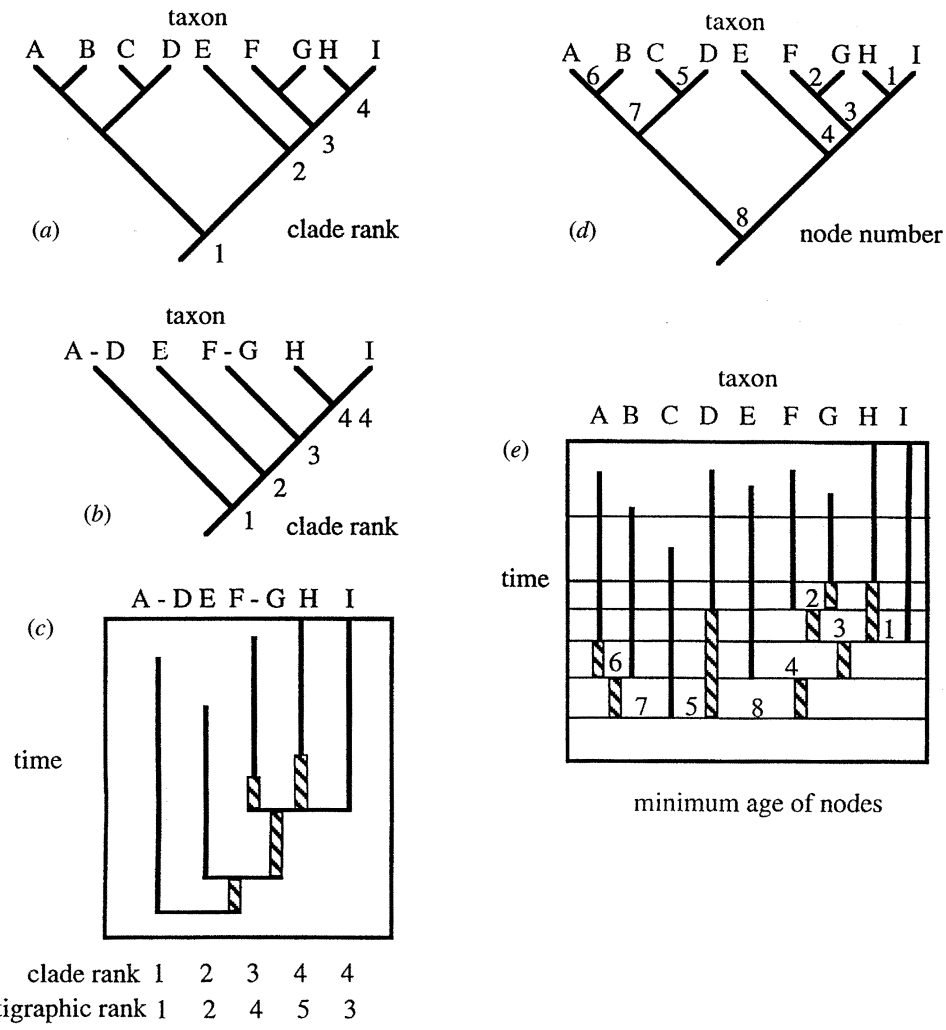

Figure 1. Techniques for assessing the quality of the fossil record. Comparisons are made between branching order in cladograms and stratigraphic data (*a–e*), and between the relative amount of gap and the known record (*e*). The example is a cladogram with nine terminal branches (A–I). For comparisons of clade order and age order, cladistic rank is determined by counting the sequence of primary nodes in a cladogram (*a*); nodes are numbered from one (basal node) upwards to the ultimate node. In cases of non-pectinate cladograms (*a*), the cladogram is reduced to a pectinate form (*b*), and groups of taxa that meet the main axis at the same point are combined and treated as a single unit. The stratigraphic sequence of clade appearance is assessed from the earliest known fossil representative of sister groups, and clade rank and stratigraphic rank may then be compared (*c*). Matching of clade rank and stratigraphic rank may be tested by Spearman rank correlation (SRC). SRC coefficients may range from 1.0 (perfect correlation) through 0 (no correlation) to −1.0 (perfect negative correlation). For assessing the proportion of ghost range, or minimum implied gap (MIG), and known stratigraphic range, the whole cladogram is used (*e*). MIG (hatched bars) is the difference between the age of the first representative of a lineage and that of its sister, as oldest known fossils of sister groups are rarely of the same age. The proportion of MIG to known range is assessed using the relative completeness index (RCI), according to the formula:

$$\mathrm{RCI} = \left(1 - \frac{\Sigma(\mathrm{MIG})}{\Sigma(\mathrm{SRL})}\right) \times 100\%.$$

RCI values may range from 100% (no ghost range) through 0 (ghost range = known range) to high negative values (ghost range ≫ known range). Stratigraphic consistency is assessed (*d*, *e*) as a comparison of the ratio of nodes that are younger than, or of equal age to, the node immediately below (consistent), compared to those that are apparently older (inconsistent). The stratigraphic consistency index (SCI) is assessed on the full cladogram (*d*, *e*). SCI values range from 1.0 (all nodes stratigraphically consistent) to 0 (no nodes stratigraphically consistent).

SCI and *Im* (Hitchin & Benton 1997), whether for the whole dataset; for echinoderms, fishes and continental vertebrates separately; or for culled samples of these datasets, which excluded cladograms with multitomies (multitomies create problems in calculating *Im* values; Siddall 1996). The only significant relationship was a negative correlation of SCI and *Im* ($r^2 = 0.13$; $p = 0.019$) for the full dataset of echinoderms. Plots of RCI and SRC against *Im* for our full dataset, and for various culls of that dataset, also failed to establish any significant relationships, whether negative or positive.

The three metrics of age–clade congruence were applied to a set of 384 published cladograms of echinoderms, fishes and tetrapods (Benton & Storrs

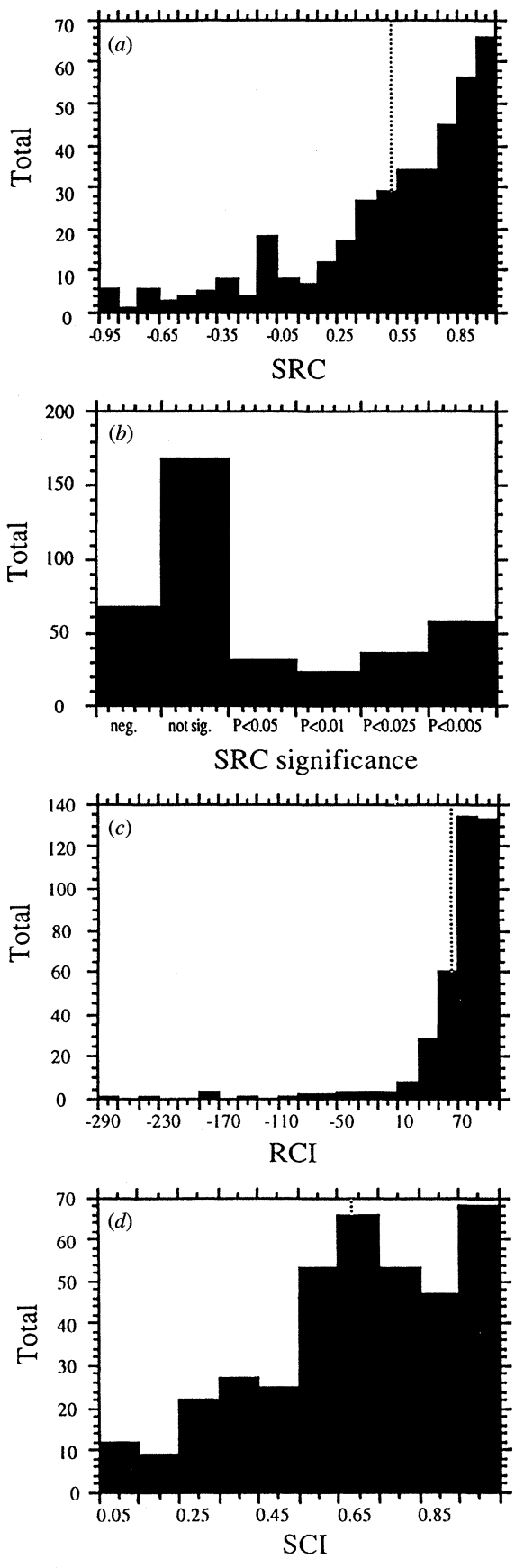

Figure 2. Assessments of congruence between stratigraphic and cladistic data show highly skewed distributions. Values for three metrics calculated on a sample of 384 cladograms of echinoderms, fishes and tetrapods: Spearman rank correlation (SRC) coefficients (*a*); measures of significance of those SRC coefficients, which take account of cladogram size (*b*); relative completeness index (RCI) values (*c*); and stratigraphic consistency index (SCI) values (*d*). Mean values for each sample are indicated by dotted lines.

1994, 1996; Benton & Simms 1995; Benton & Hitchin 1996). This test sample consists of every cladogram of echinoderms we could find published to the end of 1994, and for fishes the sample includes every cladogram we could find published to the end of 1995. For tetrapods the sample is incomplete, but it was randomized by assessing all cladograms presented in five recent multiauthor volumes (Benton 1988; Estes & Pregill 1988; Prothero & Schoch 1989; Weishampel *et al.* 1990; Schultze 1991) and in all issues of the *Journal of Vertebrate Palaeontology* published in 1993, 1994 and 1995.

The question for testing was whether cladistic and stratigraphic data on the history of life are congruent. The null hypothesis was that if the fossil record does not reflect the major patterns of evolution, there would be no evidence for congruence between the two sets of data in our random sample of cladograms. There is no reason why the branching pattern in a cladogram should match the order of occurrence of fossils (SRC), nor why minimum cladistically implied gaps ('ghost ranges'; Norell 1992) should be less than known recorded stratigraphic ranges (RCI), nor indeed why the majority of nodes in cladograms should be stratigraphically consistent (SCI). However, if current methods of cladogram construction actually reveal something like the true phylogeny, and if present knowledge of the fossil record adequately reflects phylogeny, then the null hypothesis should be falsified.

The null model, that there is no congruence between stratigraphic and cladistic data in a random sample of published cladograms, implies a random distribution of scores for the SRC, RCI and SCI metrics. It is assumed here that the null model is represented by a normal distribution of values of each metric, but that assumption requires further testing by modelling. The distributions here are, however, skewed so far from a normal distribution that they provide evidence for strong congruence of the two datasets.

3. RESULTS

Results for all three metrics, SRC, RCI and SCI, falsify the null hypothesis (figure 2). The histograms all differ from normal distributions, the expectation of the null hypothesis (Liliefors test), and they are significantly left-skewed ($p < 0.0005$). Each of the metrics assesses a different aspect of the congruence between cladistic and stratigraphic data, and the results must be assessed carefully.

The left-skewing for SRC values (figure 2*a*) indicates a tendency towards matching of the rank order of nodes based on age and clade data. However, raw SRC coefficients themselves do not indicate correlation, because they depend on the size of the sample, here the number of terminal branches in a cladogram (n) or the number of nodes (n–1). Measures of significance of SRC coefficients, which take account of sample size, indicate that the majority of cladograms in the tested

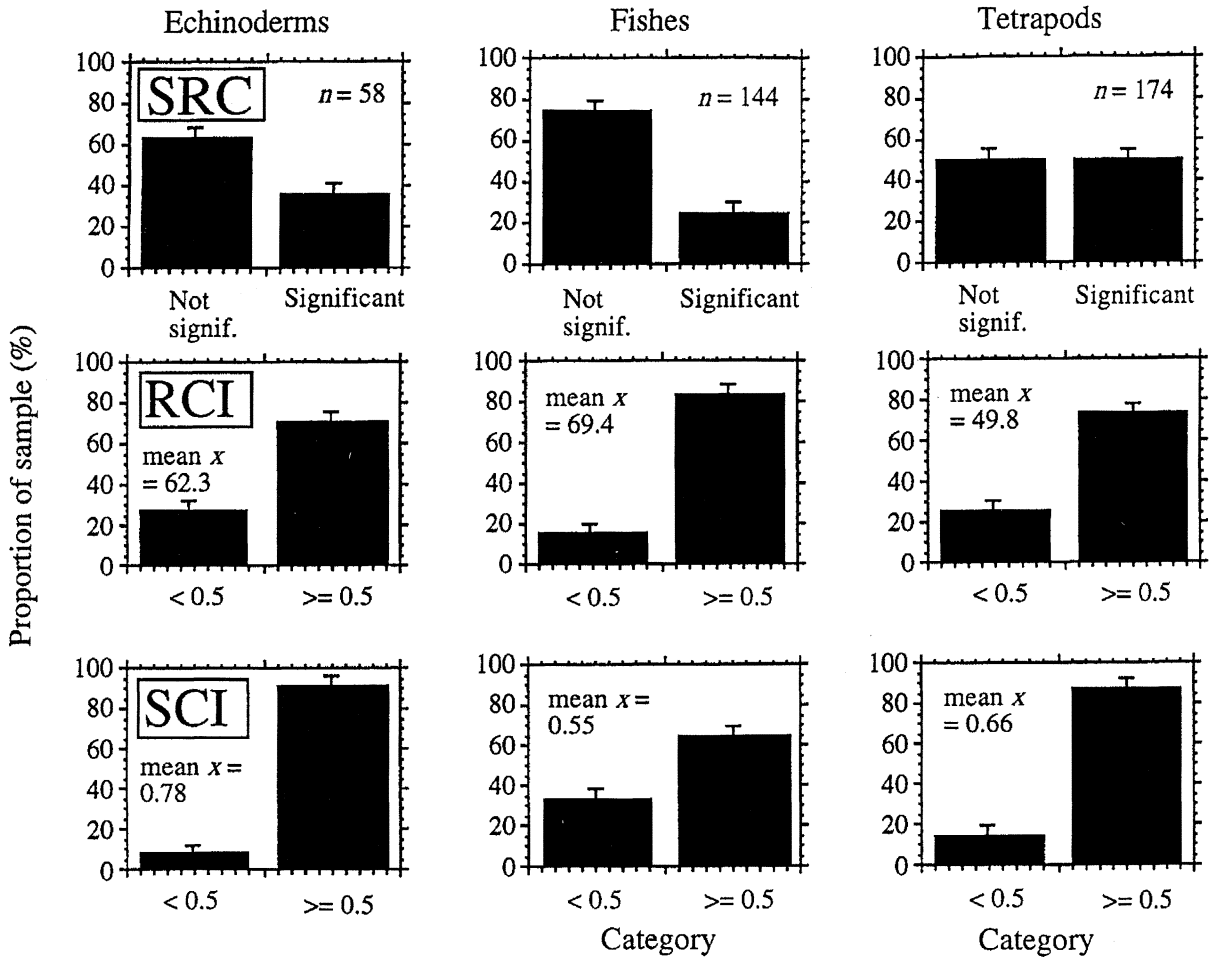

Figure 3. Summary of the metrics for comparison of cladogram data and stratigraphic age data. Metrics indicated are Spearman rank correlation (SRC) of age and clade data, the relative completeness index (RCI), based on comparisons of known and implied stratigraphic ranges, and the stratigraphic consistency index (SCI) of nodes in cladograms. The metrics have been applied to large samples of cladograms (*n*, number of cladograms in sample) for echinoderms, fishes and tetrapods. Comparisons are between significant and non-significant SRC coefficients, between frequencies of values of the RCI above and below 50%, and between frequencies of values of the SCI above and below 0.5. The differences in values among the three groups are significant, based on comparison of the binomial error bars.

sample do not show significant ($p < 0.05$) correlation of age and clade order (figure 2*b*). Out of the sample of 384 cladograms, only 148 (39%) show significant SRC values.

The left-skewing of values for the RCI and SCI metrics (figure 2*c* and 2*d*), however, confirms that our sample of cladograms does show evidence for congruence between stratigraphic and cladistic data. For the RCI metric, 288 of the 384 cladograms (75%) scored better than 50%, which means that they have more than twice as much of their ranges represented by fossils than represented by ghost range (cladistically implied gap). For the SCI metric, 298 of the sample of 384 cladograms (78%) had values greater than or equal to 0.5, which means that half or more of the nodes are stratigraphically consistent (i.e. younger than or equal in age to the node immediately below).

There was some variation in results of the three measures for the three groups of organisms under study. Results from the SRC test varied from group to group (figure 3). For echinoderms, only 24 out of 63 cladograms (36%) showed statistically significant ($p <$ 0.05) matching of clade rank and age rank data. For fishes, the figure was 37 out of 147 cladograms (25%), and for tetrapods, 87 out of 174 cladograms (50%).

The sample of cladograms also showed different results for the measure of completeness, RCI (figure 3). For all three groups, many more cladograms had RCI values equal to or greater than 0.5, than values less than 0.5. An RCI value of more than 0.5 shows that there is more known stratigraphic record than unknown (ghost) range. In all cases the difference between the two categories was statistically significant by a comparison of error bars based on an approximation formula for errors derived from a binomial probability distribution (Raup 1991). For echinoderms, 49 out of 63 cladograms (78%) had RCI values of 0.5 or more, for fishes 124 out of 147 cladograms (84%), and for tetrapods 128 out of 174 cladograms (74%). The difference in mean values of RCI for echinoderms (62.3%) and fishes (69.4%) was modest, but continental tetrapods had a much lower mean value (49.8%).

The SCI measure shows similar high rates for all

three groups (figure 3). In all three sets of cladograms significantly more than half their nodes showed stratigraphic consistency (significance tested by comparison of binomial error bars). The pass rates were 60 out of 63 cladograms (95%) for echinoderms, 102 out of 147 cladograms (69%) for fishes and 152 out of 174 cladograms (87%) for tetrapods. The pass rate for all cladograms is 82%, based on 314 of the 384 cladograms. Mean values of the SCI also differed among the groups, 0.78 for echinoderms, 0.55 for fishes and 0.66 for tetrapods.

4. DISCUSSION

It was disappointing to find from the SRC test that only 39% of the 384 cladograms in the test sample show statistically significant ($p < 0.05$) congruence between cladistic and stratigraphic data. This result contradicts earlier findings on smaller samples of cladograms. For example, Norell & Novacek (1992*a*) found that 18 of their 24 test cladograms of tetrapods (75%) gave statistically significant correlation of cladistic branching order and stratigraphic order. In a slightly expanded study, Norell & Novacek (1992*b*) found that 24 out of 33 cladograms of tetrapods (73%) gave statistically significant correlations. Benton & Storrs (1994, 1996) found that 41 of their 74 tetrapod cladograms (55%) showed statistically significant matching, but Benton & Simms (1995) reported a lower pass rate for echinoderms, only 23 out of 58 cladograms (40%).

There are probably several reasons for the poor matching of the rank order of clades and stratigraphy. Many kinds of cladograms were included in our study that were excluded from previous surveys, such as (i) cladograms that are less parsimonious alternatives, (ii) cladograms based on lower-category taxa (e.g. genera), and (iii) cladograms with fewer terminal taxa. In the first category (i), there is no *a priori* reason why less parsimonious cladograms should match stratigraphy better or worse than the most parsimonious tree(s) (MPTs). Such an assumption lies behind the frequent use of parsimony techniques in reconstructing cladistic phylogenies. It has yet to be demonstrated, however, whether MPTs fit stratigraphic data better or worse than less parsimonious alternatives. Cladograms based on low-category taxa (ii) in our sample were very hard to test. In the case of some of the generic-level and specific-level cladograms of fishes and lizards, there was almost no fossil record, and hence all three metrics showed very low values. Cladograms with few terminal taxa (iii) also tended to give poor values for the three metrics, because these involved comparisons of only two or three nodes in some cases, and age–clade matching has to be perfect to achieve a good result. With more nodes, one or two mismatches may occur, and the SRC and SCI values may still be high.

The three groups of organisms under study, echinoderms, fishes and continental tetrapods, showed a variety of results for the three metrics of congruence between cladistic and stratigraphic data (figure 3). However, these measures did not unequivocally identify one of the groups as having a consistently better or worse fossil record than the other two. Each of the groups performed best with one of the metrics: tetrapods with the SRC, fishes with the RCI and echinoderms with the SCI. Only one group came out worst on two of the tests: fishes had the poorest showing in the SRC and SCI tests. Tetrapods had the worst fossil records according to the RCI test, and echinoderms were not worst of the the three animal groups according to any of the tests. The aggregate results suggest, tentatively, that echinoderms may be said to have a better fossil record than fishes or tetrapods, and tetrapods to have a better fossil record than fishes. Further analyses may change this apparent pecking order.

The three metrics performed differently because they are all measuring different things. The SRC and SCI metrics focus on the order of nodes as compared to the order of fossils, and the RCI measures the relative duration of known record versus the ghost range. Thus, there is no reason why the RCI should give high or low values for particular cladograms that perform well or badly on the rank order metrics (SRC, SCI). Further, the SRC statistic simply tests the order of fossils in the rocks against the order of nodes in a cladogram (i.e. the implied order of origin of clades). The SCI focuses on the order of nodes too, but each node is assessed independently. Hence, one unusually early fossil in a distal part of the cladogram, or an unusually late fossil in a lower part of the cladogram, can lead to a very low SCI metric, even though the majority of nodes match the sequence of fossils. The SRC test assesses rank order for all nodes in an aggregate way, so a single anomalously out-of-sequence fossil or node need not throw the test completely. Finally, the SCI and RCI metrics are based on whole cladograms, as published, whereas the SRC has to be calculated from a pectinate cladogram (figure 1; Norell & Novacek 1992*a*), which means that many of the SRC tests were based on incomplete cladograms.

Even though the results from the SRC test in the present study were not as convincing as earlier analyses, the RCI and SCI metrics showed impressive left skewing; the majority of cladograms tested show good congruence between cladistic and stratigraphic information. Cladists and stratigraphers may breathe easy: the cladistic method appears, on the whole, to be finding phylogenies that may be close to the true phylogeny of life, and the sequence of fossils in the rocks is not misleading.

The approaches explored here are not direct tests of the historical pattern of evolution, merely oblique approaches to such a test. However, it would be hard to explain why the independent evidence of the stratigraphic occurrence of fossils and the patterns of cladograms should show such striking levels of congruence if the fossil record and the cladistic method were hopelessly misleading.

We are grateful to the Leverhulme Trust (Grant F182/AK) for funding, and to two anonymous referees for extremely helpful comments. All data used in this study may be seen in Benton & Hitchin (1996), and at http://www.gly.bris.ac.uk/www/research/palaeo/cladestrat.html
(E-mail: mike.benton@bristol.ac.uk).

REFERENCES

Allison, P. A. & Briggs, D. E. G. 1991 *Taphonomy; releasing the data locked in the fossil record.* New York: Plenum.

Benton, M. J. 1988 *The phylogeny and classification of the tetrapods*, vol. 1 (*Amphibians, reptiles, and birds*), vol. 2 (*Mammals*). Oxford: Clarendon Press.

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

Benton, M. J. 1995 Testing the time axis of phylogenies. *Phil. Trans. R. Soc. Lond.* B **348**, 5–10.

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

Benton, M. J. & Simms, M. J. 1995 Testing the marine and continental fossil records. *Geology* **23**, 601–604.

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. & Storrs, G. W. 1996 Diversity in the past: comparing cladistic phylogenies and stratigraphy. In *Aspects of the genesis and maintenance of biological diversity* (ed. M. E. Hochberg, J. Clobert & R. Barbault), pp. 19–40. Oxford University Press.

Darwin, C. 1859 *On the origin of species by means of natural selection.* London: John Murray.

Eldredge, N. & Cracraft, J. 1980 *Phylogenetic patterns and the evolutionary process.* New York: Columbia University Press.

Estes, R. & Pregill, G. 1988 *Phylogenetic relationships of the lizard families. Essays commemorating Charles L. Camp.* Stanford University Press.

Forey, P. L., Humphries, C. J., Kitching, I. J., Scotland, R. W., Siebert, D. J. & Williams, D. M. 1992 *Cladistics: a practical course in systematics.* Oxford: Clarendon Press.

Gauthier, J., Kluge, A. G. & Rowe, T. 1988 Amniote phylogeny and the importance of fossils. *Cladistics* **4**, 105–209.

Goodman, M. 1989 Emerging alliance of phylogenetic systematics and molecular biology: a new age of exploration. In *The hierarchy of life* (ed. B. Fernholm, K. Bremer & H. Jornvall), pp. 43–61. New York: Elsevier.

Hecht, M. K. 1976 Phylogenetic inference and methodology as applied to the vertebrate record. *Evol. Biol.* **9**, 335–363.

Hennig, W. 1981 *Insect phylogeny.* New York: John Wiley.

Hitchin, R. & Benton, M. J. 1997 Stratigraphic indices and tree balance. *Syst. Biol.* (In the press.)

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

Løvtrup, S. 1977 *The phylogeny of Vertebrata.* London: Wiley.

Nelson, G. J. 1969 Origin and diversification of teleostean fishes. *Ann. N. Y. Acad. Sci.* **167**, 18–30.

Nelson, J. S. & Platnick, N. I. 1981 *Systematics and biogeography. Cladistics and vicariance.* New York: Columbia University Press.

Norell, M. A. 1992 Taxic origin and temporal diversity: the effect of phylogeny. In *Extinction and phylogeny* (ed. M. J. Novacek & Q. D. Qheeler), pp. 89–118. New York: Columbia University Press.

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

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

Patterson, C. 1981 Significance of fossils in determining evolutionary relationships. *A. Rev. Ecol. Syst.* **12**, 195–223.

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.

Paul, C. R. C. 1982 The adequacy of the fossil record. In *Problems of phylogenetic reconstruction* (ed. K. A. Joysey & A. E. Friday), pp. 75–117. Systematics Association Special Volume 21. London: Academic Press

Paul, C. R. C. 1990 Completeness of the fossil record. In *Palaeobiology; a synthesis* (ed. D. E. G. Briggs & P. R. Crowther), pp. 298–303. Oxford: Blackwell Scientific.

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

Prothero, D. R. & Schoch, R. M. 1989 *The evolution of perissodactyls.* New York: Clarendon Press.

Raup, D. M. 1972 Taxonomic diversity during the Phanerozoic. *Science* **177**, 1065–1071.

Raup, D. M. 1991 The future of analytical paleobiology. In *Analytical paleobiology* (ed. N. L. Gilinsky & P. M. Signor), pp. 207–216. Knoxville, TN: The Paleontological Society.

Schaeffer, B., Hecht, M. K. & Eldredge, N. 1972 Phylogeny and paleontology. *Evol. Biol.* **6**, 31–46.

Schultze, H.-P. 1991 A comparison of controversial hypotheses on the origin of tetrapods. In *Origins of the higher groups of tetrapods* (ed. H.-P. Schultze & L. Trueb), pp. 29–67. Ithaca, NY: Cornell University Press.

Siddall, M. E. 1996 Stratigraphic consistency and the shape of things. *Syst. Biol.* **45**, 111–115.

Signor, P. W. 1982 Species richness in the Phanerozoic: compensating for sampling bias. *Geology* **10**, 625–628.

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

Weishampel, D. B., Dodson, P. & Osmólska, H. 1990 *The Dinosauria.* Berkeley: University of California Press.

Received 24 January 1997; accepted 13 February 1997