MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 9, No. 3, June, pp. 398–407, 1998
ARTICLE NO. FY980492

Molecular and Morphological Phylogenies of Mammals: Congruence with Stratigraphic Data

Michael J. Benton

Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, United Kingdom

Received June 13, 1997; revised December 16, 1997

Tests of a sample of 206 cladograms of mammals show that morphological data seem to predict phylogenies that match the known fossil record better than molecular trees. Three metrics that assess the rank order of branching points, the stratigraphic consistency of those nodes, and the ratio of ghost range to known range show a considerable diversity of values. Some published trees show excellent matching with fossil-record data; others show almost no correspondence whatsoever. Morphological trees are nearly twice as good as molecular trees in terms of matching of the rank orders of nodes and oldest fossils, while morphological trees are 10% better than molecular in terms of stratigraphic consistency of the nodes. The ratios of ghost range to known range are lower for molecular trees. Among the molecular trees, those based on gene data are considerably better than those based on protein sequences, at least in terms of the rank order of nodes and the stratigraphic consistency of nodes. Protein trees, however, were best of all in terms of minimizing the proportion of ghost range. These findings probably indicate real phenomena, but the match of molecular trees to the expectations of stratigraphy may improve as the study of molecular phylogeny matures.

INTRODUCTION

Many hundreds of phylogenies of mammals have been published, some based on morphological data and others based on molecular data (proteins, nucleic acids), and yet there is no evidence that these hypotheses of relationship are converging on a single viewpoint. Indeed, for some problems, such as the relationships of the orders of placentals, the reverse seems to be the case: with more work, and the introduction of new data, the variety of postulated phylogenies increases.

All systematists are working toward a single end, which is the reconstruction of some part of the single phylogenetic tree that links all organisms living and extinct. Systematists might wish to determine how good their results are and whether their phylogenetic hypothesis is closer to the truth than an opposing hypothesis. However, "truth" cannot enter into phylogeny reconstruction (Frost and Kluge, 1994), unless the investigator has access to the mind of God or possesses a time machine. The best that can be achieved is to test and retest postulated phylogenies against internal criteria of goodness of fit, such as the consistency index, the retention index, the homoplasy excess ratio, bootstrap values, Bremer support values, and the like. A tree is preferred if these values are optimized.

While it is important to attempt to optimize individual phylogenetic hypotheses, there are also broader questions concerning the entirety of current phylogenetic hypotheses. For example, it would be worth knowing whether the majority of phylogenies are right or wrong, whether phylogenies of one group of organisms are better or worse than another, and whether one method or character set offers better resolution of phylogeny than another. This general approach rests on statistical assessment of congruence between independent evidence about phylogeny.

CONGRUENCE TESTING OF INDEPENDENT DATA SETS ON PHYLOGENY

Independent Data Sets

A number of authors have argued that phylogenetic data and stratigraphic data are independent of each other and that they both offer evidence about the historical shape of phylogenetic trees (Gauthier *et al.*, 1988; Norell, 1992; Norell and Novacek, 1992a; Benton, 1994, 1995; Benton and Hitchin, 1996, 1997; Benton and Storrs, 1996). Morphological characters, as used in cladistic reconstruction of phylogeny, are determined solely by inspection of the organisms, whether living or extinct, and their polarities (primitive → derived) are determined by reference to evidence of distribution, not to stratigraphic age. Indeed, cladistic reconstruction is frequently done without reference to polarity. Trees could be rooted by choosing the oldest fossil in the analysis, but that is generally not done (Smith, 1994).

Further, it could be argued that morphological and molecular data are independent of each other: despite

1055-7903/98 $25.00

the obvious link between genotype and phenotype, it is not clear that genes map directly to phylogenetically informative morphological characters. In addition, each molecule is broadly independent of the others. Thus, it is possible to compare phylogenies based on morphological data, genome data, globins, cytochrome *c*, and other proteins with each other, but more importantly, with independent stratigraphic data.

Benton and Hitchin (1996, 1997) have suggested that this kind of approach offers a unique way of testing history. The congruence-testing approach can be used in specific cases, to assess which phylogenetic hypotheses agree with each other and which fit the known fossil record best. However, when taken in aggregate, as is done here, and tested as a population sample, more general questions may be addressed, without worrying about minor analytical errors and misidentifications. If the phylogenetic information derived from independent data sets agrees, then the methods are probably finding the correct phylogenies; if there are widespread disagreements, then something is wrong with one or more of the data sets. (Agreement could also mean that all data sources are pointing toward uniformly incorrect conclusions, but that view would be hard to sustain unless it could be shown that the supposedly "independent" data sources are in fact linked.)

The Quality of the Fossil Record

Is the fossil record good enough for this role as an arbiter between competing phylogenetic hypotheses? Surely the fossils are so randomly distributed in the rocks, so poorly dated, and so incomplete in themselves, that they are useless as a yardstick? These caveats may be valid for the fossil record of some soft-bodied groups, such as worms or jellyfish, but there is no evidence for such a sceptical view of the fossil record of groups with hard parts, such as vertebrates, echinoderms, molluscs, arthropods, and vascular land plants.

One kind of evidence for a good fossil record is intuitive: if fossils are distributed randomly in the rocks, there should be many dramatic new discoveries. There are not. The broad outlines of, for example, vertebrate evolution were established by about 1860, based largely on the study of European palaeontology. Little has changed since 1860, despite intensive collecting efforts by thousands of palaeontologists professional and amateur, spurred on by the great glory that is associated with dramatic new finds, despite the study of the fossil records of all other continents, despite the invention of new dating techniques, and despite much improved microscopes and other equipment. In 1860, some critics of the apparent progression of fossil forms, and of evolution, expected human fossils in the Silurian and dinosaurs in the Carboniferous; unexpected range extensions of this kind have not happened.

This intuitive observation was confirmed quantitatively by Maxwell and Benton (1990). They compared several stages in the development of knowledge about the history of tetrapods over the past 100 years, using standard compilations of fossil data from 1900, 1933, 1945, 1966, and 1987. There had been huge changes in palaeontological knowledge from 1900 to 1987, but these changes were randomly distributed with respect to time. Global diversities essentially doubled throughout the whole fossil record of tetrapods, but the overall pattern of diversification, and the timing and magnitudes of major extinction events, were unchanged. These findings were confirmed for marine animals in an analogous study by Sepkoski (1993).

Geological approaches have confirmed that the fossil record is adequate for hard-bodied organisms and that it does not necessarily become worse in older rocks. Allison and Briggs (1993) showed that sites of exceptional fossil preservation were more abundant than had been assumed, and they provide controls on the overall diversity of marine settings. More typical marine fossil beds, coquinas, or winnowed accumulations of fossil shells, seem to survive equally well from the early Palaeozoic and the late Cenozoic (Kidwell and Brenchley, 1996). There is no evidence, from geological considerations, that the fossil record gives a misleading representation of the history of life.

Congruence testing, using large samples of cladograms as a standard, has also confirmed that the fossil record tells the correct story, that different parts of the fossil record, assessed by taxa, and by sites of deposition, are comparable, and that fossil collecting fills predicted gaps.

Congruence Testing Methods

There are a variety of metrics for comparing phylogenies and fossil records (Table 1; Fig. 1): Spearman rank correlation (SRC), the stratigraphic consistency index (SCI), and the relative completeness index (RCI). SRC is an established nonparametric statistical test, and it has been used in comparing the order of fossils in the rocks with the implied order of appearance of groups based on the sequence of nodes (branching points) in a cladogram. The first applications of the SRC test for this purpose were by Gauthier *et al.* (1988) and Norell and Novacek (1992a,b).

The SCI was proposed by Huelsenbeck (1994) to assess how well the nodes in cladograms corresponded to the known fossil record. Nodes are dated by the oldest known fossils of either sister group subtended from the node. Each node is compared with the node immediately below it. If the upper node is younger than, or equal in age to, the node below, the node is said to be stratigraphically consistent. If the node below is younger, the upper node is stratigraphically inconsistent. The SCI for a cladogram compares the ratio of the

TABLE 1

Three Metrics for Assessing Congruence between Phylogenetic and Stratigraphic Data

Metric	Spearman rank correlation	Relative completeness index	Stratigraphic consistency index
Abbreviation	SRC	RCI	SCI
Author	Standard technique	Benton (1994)	Huelsenbeck (1994)
Assessment	Compares rank orders of two series of numbers (e.g., numerical orders of nodes and of first fossils)	Compares relative amounts of ghost range and known range	Compares relative numbers of stratigraphically inconsistent and consistent nodes
Statistical test	Yes	No	No
Significance measures	Yes	Yes[a]	Yes[a]
Dependent on tree size?	Yes	No[b]	No[b]
Dependent on tree balance?	No	No[b]	No[b]

[a] Significance tests for the RCI and SCI have not been published, but Matthew Wills (Bristol) has developed a simulation approach to estimating significance for both metrics, available in his program "Ghosts." Further information is available at <http://palaeo.gly.bris.ac.uk/cladestrat/cladestrat.html>.

[b] Siddall (1996) has suggested, on the basis of simulations, that the SCI metric may be affected by tree balance. Empirical studies, however (Hitchin and Benton, 1997b), show that neither the RCI nor the SCI depends on tree balance. In addition, there is no evidence for a strong association of values for these metrics and tree size.

sums of stratigraphically consistent to inconsistent nodes.

The RCI was proposed (Benton, 1994; Benton and Storrs, 1994) to take account of the actual time spans between branching points and of implied gaps before the oldest-known fossils of lineages. Sister groups, by definition, originated from an immediate common ancestor and diverged from that ancestor. Thus, both sister groups should have fossil records that start at essentially the same time. In reality, usually the oldest fossil of one lineage will be older than the oldest fossil of its sister lineage. The time gap between these two oldest fossils is the "ghost range" or minimal cladistically implied gap. The RCI assesses the ratio of ghost range to known range, and high values imply that ghost ranges are short and hence that the fossil record is good.

Results of Congruence Testing

Congruence testing has been applied to a variety of questions in phylogeny reconstruction, but not yet to a comparison of morphological and molecular results. Some background to the results obtained from morphology-only studies will put the present study in context. The first results of congruence testing were encouraging: Norell and Novacek (1992a) found that 18 of 24 test cases of cladograms of vertebrates (75%) gave statistically significant ($P < 0.05$) correlations of clade and age data, using the SRC test, while Benton and Storrs (1994) found significant correlation in 41 of 74 test cases (55%). Subsequent assessments, however, based on a larger sample [384 cladograms, composed of 174 cladograms of tetrapods, 147 cladograms of fishes, and 63 cladograms of echinoderms (Benton and Hitchin, 1996, 1997)], provided more disappointing results. For all cladograms in the test sample, 148 of 384 (38%) showed significant SRC values. These results could indicate poor congruence, or they could simply highlight the fact that the SRC test is rather crude, simply comparing the raw order of points and taking no account of their actual spacing in time nor of the degree of mismatch.

Much better results were obtained with the RCI and the SCI metrics, which measure different aspects of cladogram and fossil record quality. For all three groups assessed, most cladograms have RCI values equal to, or greater than, 50% (Benton and Hitchin, 1996, 1997). The pass rates are 78% for echinoderms, 84% for fishes, 74% for tetrapods and 78% for all cladograms. The pass rates are similarly favorable for the SCI measure. In these cases, all three sets of cladograms have significantly more than half their nodes showing stratigraphic consistency than inconsistency. The pass rates are 95% for echinoderms, 69% for fishes, 87% for tetrapods, and 82% for all cladograms.

These tests show no clear differentiation in quality of the fossil record by major taxonomic groups. The comparisons show that none of them consistently has a better fossil record, or better cladistic resolution, than the others. Each of the animal groups performed best with one of the metrics: tetrapods with SRC, fishes with RCI, and echinoderms with SCI.

There is no strong evidence for differentiation by broad habitats either. Benton and Simms (1995) showed that continental tetrapods have a fossil record that is as good as, or better than, that of echinoderms, based on comparisons of results obtained with the SRC and RCI metrics. This result could not have been predicted from observations of the field occurrence of both groups: tetrapods are found in sporadic and unpredictable sedimentary settings, while echinoderm remains are hugely abundant in many marine shelf deposits. A more detailed comparison yielded mixed results, with

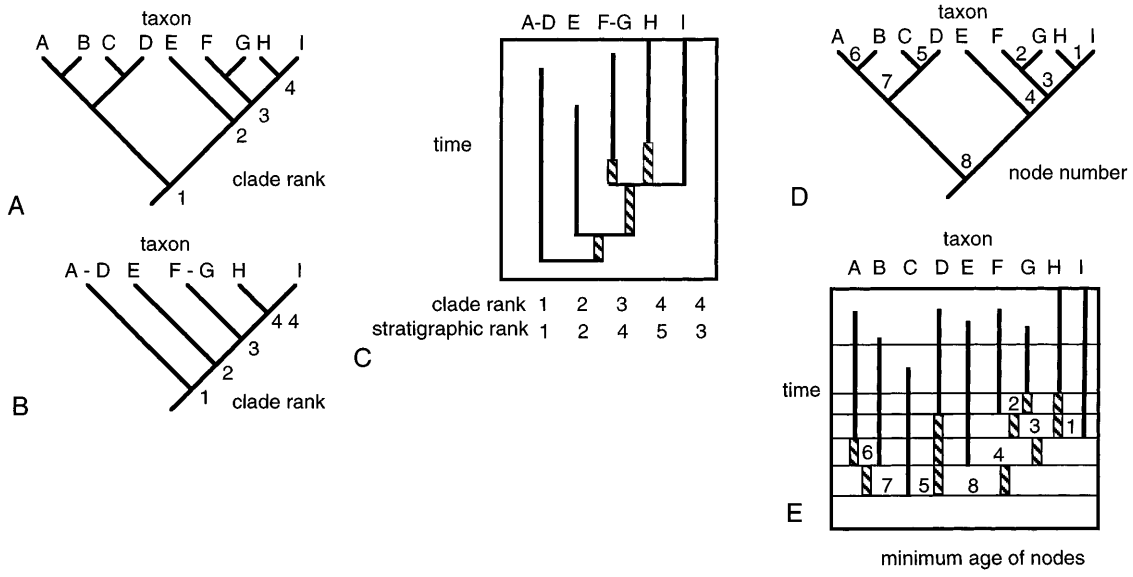

FIG. 1. Techniques for assessing the quality of the fossil record. Comparisons are made between branching order in cladograms and stratigraphic data (A–E) and between the relative amount of gap and the known record (E). The example is a cladogram with nine terminal branches (A–I). For comparisons of 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 nonpectinate cladograms (A), the cladogram is reduced to 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 (SRL), the whole cladogram is used (E). MIG (diagonal rule) 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).

continental cladograms scoring higher values in the SRC and SCI metrics and marine cladograms scoring higher values in the RCI metric (Benton and Hitchin, 1996).

A final conclusion about the fossil record was that new collecting and study tend to fill gaps. Benton and Storrs (1994, 1996) compared standard summaries of fossil record data from 1967 (Harland *et al.*, 1967) and 1993 (Benton, 1993) and found that the mean RCI value for a sample of 87 cladograms of tetrapods improved from 67.9 to 72.3%, a statistically significant difference, according to a Wilcoxon signed ranks test ($P = 0.026$). In other words, comparisons of the relative completeness of cladograms show a significant *improvement*, by about 5%, in knowledge of the fossil record over the past 26 years of research.

MORPHOLOGICAL AND MOLECULAR DATA ON MAMMAL PHYLOGENY

Data and the Testing Methods

A sample of 206 phylogenies of mammals was assessed, 54 of them based on morphological data, 54 based on proteins, and 98 based on genes (RNA, DNA). The sample included trees published from 1910 to 1997, but the majority come from the past 10 years. The morphological cladograms are those used in previous analyses (Benton and Hitchin, 1996, 1997), a random sample of the hundreds available, extracted from a number of multiauthor volumes (Benton, 1988; Prothero and Schoch, 1989) and including every cladogram published in the *Journal of Vertebrate Paleontology* from 1993 to 1995. The molecular phylogenies also

include those from the multiauthor volumes noted, as well as Miyamoto and Cracraft (1991) and Szalay *et al.* (1993). In addition, a large number have been derived from standard journals, *Nature, Science, Journal of Molecular Evolution, Molecular Phylogenetics and Evolution, Molecular Biology and Evolution,* and *Journal of Mammalian Evolution.* None of the data sets is in any way comprehensive. As far as possible, every tree in each publication was assessed, not just the most parsimonious tree (MPT) or the most-favored solution. This allows direct comparison of competing phylogenetic hypotheses.

Each tree was assessed for the SRC, RCI, and SCI metrics. The SRC values were tested for significance (SRC values are dependent on sample size), while the raw RCI and SCI values are given. The results for the morphological trees have been published (Benton and Hitchin, 1996), and for the molecular trees the results may be inspected on the World Wide Web at <http://palaeo.gly.bris.ac.uk/cladestrat/cladestrat.html>. This site includes the full reference for each sampled tree.

For the purposes of comparison and contrast, the data set of 206 cladograms was divided into three subsets, morphological, protein, and genome. These three were compared, and, for some comparisons, the protein and genome trees were united as molecular trees. Comparisons were made using nonparametric statistics. In particular, the Kolmogorov-Smirnov test was found to be a useful way of comparing the shape, mean, and variance of frequency distributions for values of each metric among the samples. This test assesses the probability that one distribution might be derived from another, with a two-sided probability in the range $0.95 < P < 1.00$ if the distributions are closely similar. Nonparametric statistics were used since these make no assumptions about the nature of the samples, permitting, for example, highly skewed distributions.

The comparisons were made for samples containing all cladograms, termed "all," and then for samples containing cladograms with six or more terminals (n), termed "large." Smaller cladograms, with $n = 3$, were excluded altogether, since the SCI metric cannot be applied to such small cladograms, and the significance of SRC values cannot be assessed. Cladograms with $n = 4$ or 5 were retained in the data set, but such small trees were also problematic for some analyses; in particular, trees with $n = 4$ or 5 are too small for the significance of their SRC values to be assessed, and they are also too small for effective assessment of the SCI metric, in which only n-2 nodes may be considered. The sample of all 206 cladograms included 53 with $n = 4$ or 5. Thus the sample of large ($n > 5$) cladograms consisted of 153, 43 morphological, 45 protein, and 65 genome. For the SRC test, some cladograms had to be collapsed before analysis (Fig. 1), so there were rather more cladograms (79) with $n = 4$ or 5 in this case only. Thus the sample of large ($n > 5$) collapsed cladograms consisted of 127, 39 morphological, 35 protein, and 53 genome.

Stratigraphic data were derived from two standard sources, *The Fossil Record 2* (Benton, 1993) for familial data and *Mammalian Paleofaunas of the World* (Savage and Russell, 1983) for generic-level data. These data compilations may be incorrect and out-of-date in places, but the use of standard sources removes a possible subjective element in the testing regime.

Results

The three sets of trees, morphological, protein, and genome, showed different measures of congruence with stratigraphic data, when assessed using the three metrics (Fig. 2). Morphological trees showed best SRC and SCI values and protein trees showed best RCI values. Protein and genome trees showed different levels of congruence with stratigraphic data according to all three metrics, but for the SRC and SCI metrics, genome trees performed better than protein trees.

The SRC test indicated that relatively few of the cladograms showed significant ($P < 0.05$) matching of clade and age order of the nodes (Fig. 3). For all cladograms (Fig. 3A), only 29 of 54 (54%) morphological cladograms yielded significant SRC values, only 8 of 54 (15%) protein trees, and 32 out of 98 (33%) genome. These figures equate to 40 of 152 (26%) molecular trees and 69 out of all 206 cladograms (33%). When only large ($n > 5$) cladograms are considered, the results appear to be better (Fig. 3B). Twenty-eight of 39 (72%) of morphological cladograms yielded significant SRC values, 6 of 35 (17%) protein trees, and 31 of 53 (58%) genome. These figures equate to 37 of 88 (42%) molecular trees and 65 of all 127 cladograms (51%). Kolmogorov-Smirnov tests show that all distributions are different ($P < 0.05$) for all cladograms and for large cladograms.

The RCI metric showed more uniform results (Fig. 4). For all cladograms (Fig. 4A), 42 of 54 (78%) morphological cladograms yielded RCI values greater than 50%, 53 of 54 (98%) protein trees, and 81 of 98 (83%) genome trees. These figures equate to 134 of 152 (88%) molecular trees and 176 of all 206 trees (85%). For large cladograms (Fig. 4B), 33 of 43 (77%) morphological cladograms yielded RCI values greater than 50%, 44 of 45 (98%) protein trees, and 53 of 65 (82%) genome trees. These figures equate to 97 of 110 (88%) of molecular trees and 130 of 153 (85%) of all trees. Mean values for larger cladograms show considerable differences, 56.3 for morphological trees, 71.6 for genome trees, and 82.6 for protein trees. Kolmogorov-Smirnov tests show that, despite the similarities in pass rates, all distributions are different for both the all cladogram and the large cladogram samples (Figs. 4A and 4B), except that the frequency distribution of RCI values for the combined

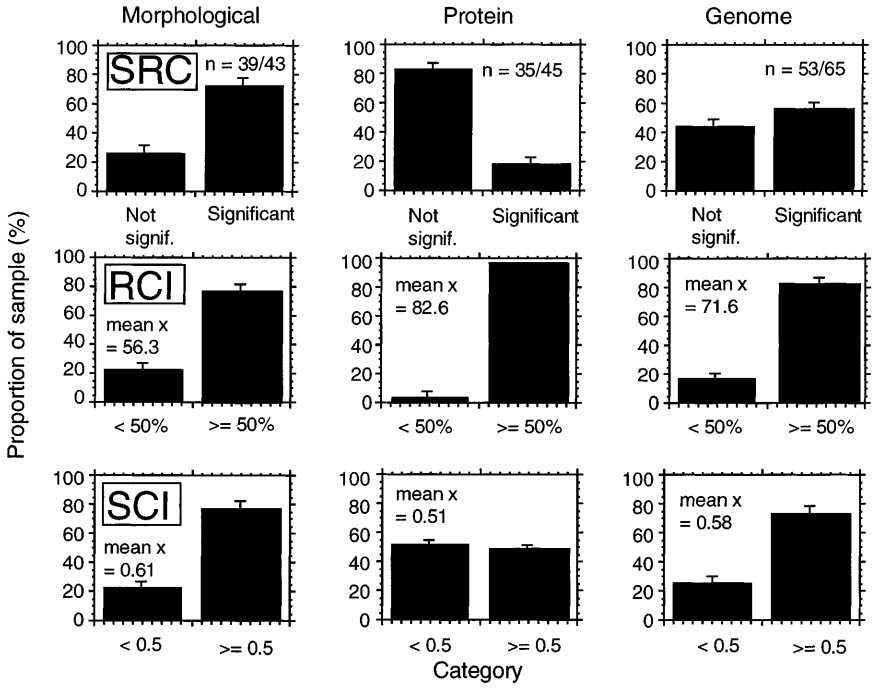

FIG. 2. Summary of the metrics for comparison of published phylogenetic trees of mammals with stratigraphy, based on morphological, protein, and genome data. Metrics indicated are Spearman rank correlation (SRC), relative completeness index (RCI), and stratigraphic consistency index (SCI), as defined in Fig. 1. The metrics have been applied to large samples of cladograms (*n*, number of cladograms in sample). In this case, only cladograms with six or more terminal taxa are included, and this yields lower sample sizes for the SRC test (based on collapsed trees) than the RCI and SCI metrics (based on unmodified trees). Comparisons are between frequencies of significant ($P < 0.05$) and nonsignificant SRC tests, frequencies of values of RCI above and below 50%, and frequencies of values of SCI above and below 0.5. The differences in values among the three groups are significant, based on comparison of the binomial error bars (calculated according to the method of Raup, 1991).

molecular sample could be the same as that for genome trees alone, for all and for large cladograms (in both cases, $P = 0.986$).

The SCI metric shows slightly better results for morphological rather than molecular trees (Fig. 4). For all cladograms (Fig. 5A), 41 of 54 (76%) morphological cladograms yielded RCI values equal to, or greater than, 0.5, compared to 28 of 54 (52%) protein trees and 75 of 98 (77%) genome trees. These figures equate to 103 of 152 (68%) of molecular trees and 144 of 206 (70%) of all trees. For large cladograms (Fig. 5B), 33 of 43 (77%) morphological cladograms yielded RCI values equal to, or greater than, 0.5, 22 of 45 (49%) protein trees, and 48 of 65 (74%) genome trees. These figures equate to 70 of 110 (64%) of molecular trees and 92 of 153 (60%) of all trees. Mean values for larger cladograms show that morphological trees (0.61) and genome trees (0.58) have higher SCI values than protein (0.51). Kolmogorov-Smirnov tests show that most distributions are different for both the all cladogram and the large cladogram samples (Figs. 5A and 5B). The frequency distribution of SCI values for protein trees is similar to that for morphological trees when all cladograms are considered (in both cases, $P = 0.962$), and this distribution becomes identical ($P = 1.000$) for the sample of large cladograms. In addition, for large cladograms, the distribution of frequencies of SCI values is similar ($P = 0.962$) for genome trees and morphological trees.

A broad differentiation was found between morphological and molecular trees in terms of their congruence with stratigraphic data (Fig. 6). Morphological trees appear to show best congruence when assessed with the SRC and SCI metrics, and molecular trees perform better with the RCI. For larger cladograms, the SRC test showed that nearly twice as many morphological trees (72%) as molecular (42%) showed significant correlation. Molecular trees performed considerably better than morphological using the RCI metric (mean values, 75.3 and 56.3% respectively). The SCI metric showed a better matching with stratigraphic data by morphological trees (mean, 0.61) than molecular trees (mean, 0.54).

MOLECULAR OR MORPHOLOGICAL DATA?

Overall Congruence

The present study has confirmed that there is overall good congruence between phylogenetic and strati-

graphic data for all kinds of data, morphological and molecular.

The SRC measure confirmed earlier findings based on morphological data alone (Benton and Hitchin, 1996, 1997) that the rank orders of nodes in large samples of phylogenetic trees do not always match the order of occurrence of the fossils. In this study, pass rates were on the whole better than the global 38% figure found by Benton and Hitchin (1996) for 384 morphological cladograms of tetrapods, fishes, and echinoderms. For mammal trees, 51% of the large trees ($n > 5$) showed statistically significant ($P < 0.05$) matching, although the figure was only 33% for the full sample, including smaller trees where $n = 4$ or 5.

The RCI metric yielded remarkably high levels of congruence for all subsets of the sample, ranging from

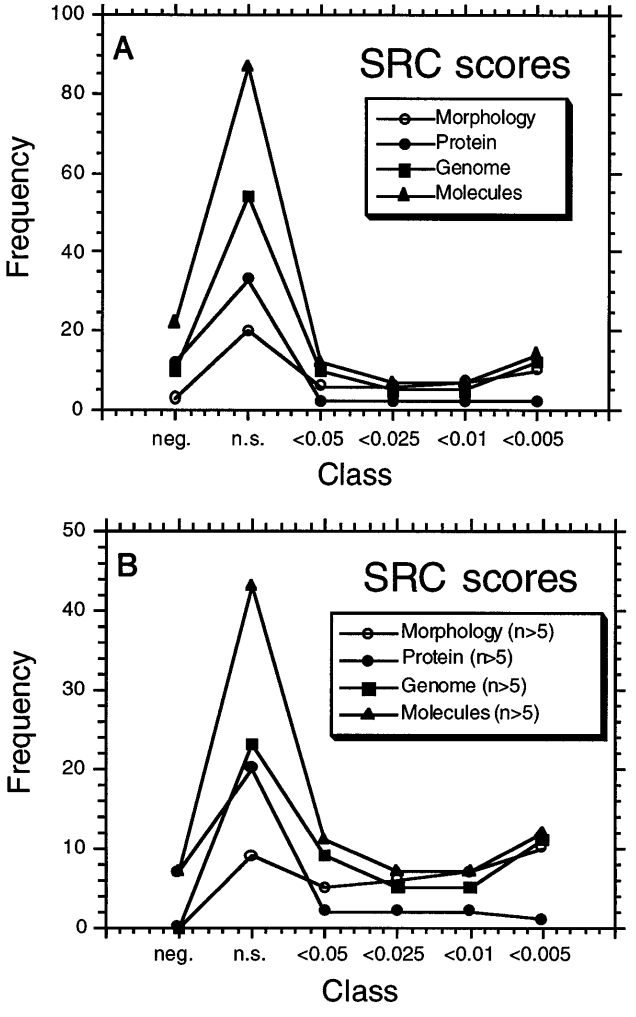

FIG. 3. Frequencies of different levels of significance in SRC tests of published phylogenetic trees of mammals, based on morphological, protein, genome, and molecular (protein + genome) data sets. Frequencies are shown for samples of all cladograms (A) and for cladograms with more than five terminal taxa (B). All frequency distributions in each case are significantly different (Kolmogorov-Smirnov test).

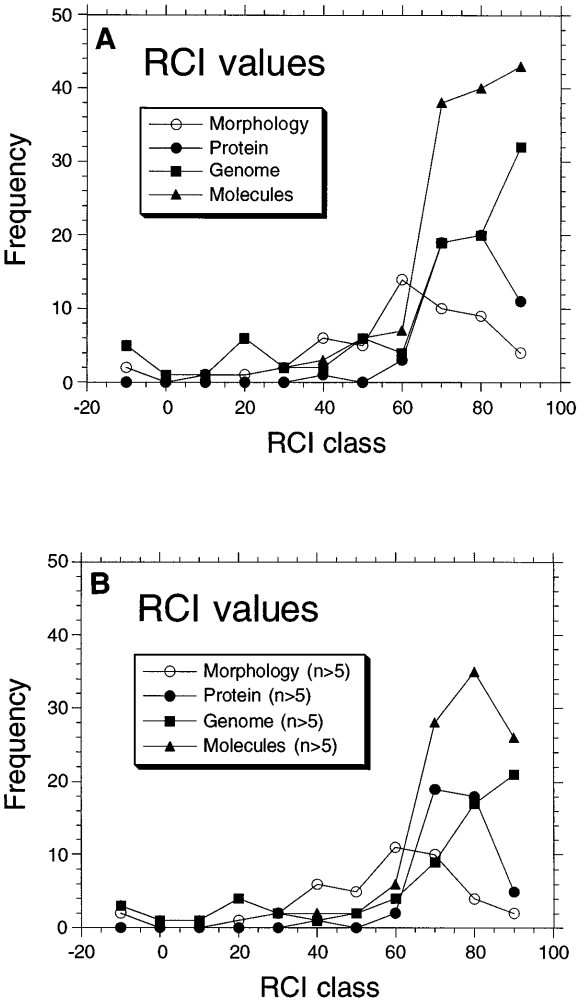

FIG. 4. Frequencies of different values of the RCI metric for published phylogenetic trees of mammals, based on morphological, protein, genome, and molecular (protein + genome) data sets. Frequencies are shown for samples of all cladograms (A) and for cladograms with more than five terminal taxa (B). All frequency distributions in each case are significantly different, except for the genome and molecular curves which could be sampled from the same distribution (Kolmogorov-Smirnov test).

78% of all morphological trees having RCI values in excess of 50 to 98% of all, and large, protein trees. These values compare with generally lower measures obtained by Benton and Hitchin (1996, 1997) for morphological cladograms of tetrapods in general (74%), fishes (83%), and echinoderms (71%).

The SCI metric also yielded high levels of congruence for all subsets of the sample, ranging from 49% of larger protein trees having SCI values equal to, or greater than, 0.5, to 76% of all morphological trees, and 77% of all genome trees and 77% of larger morphological trees. These values compare with SCI results for morphological trees of echinoderms (91%), tetrapods (87%), and fishes (65%) obtained by Benton and Hitchin (1996, 1997).

Molecular or Morphological Data?

Comparisons of congruence measures between the different partitions of the set of 206 cladograms gave mixed results. Most of the evidence suggests that morphological trees match stratigraphic data better than molecular and that genome trees show better congruence with age data than do protein trees, based on the SRC and SCI metrics, although molecular trees performed better, and protein trees performed best, in terms of the RCI metric.

The SRC test of matching of cladistic nodes and stratigraphic ages of first occurrence showed significantly better values for morphologically based cladograms, when compared to molecular trees. For larger trees, 72% of the morphological set, but only 42% of the molecular set, showed significant correlation. Of the molecular trees, those based on protein data (17%) showed poorer matching to stratigraphy than those based on gene data (58%). These results were confirmed for the full data set.

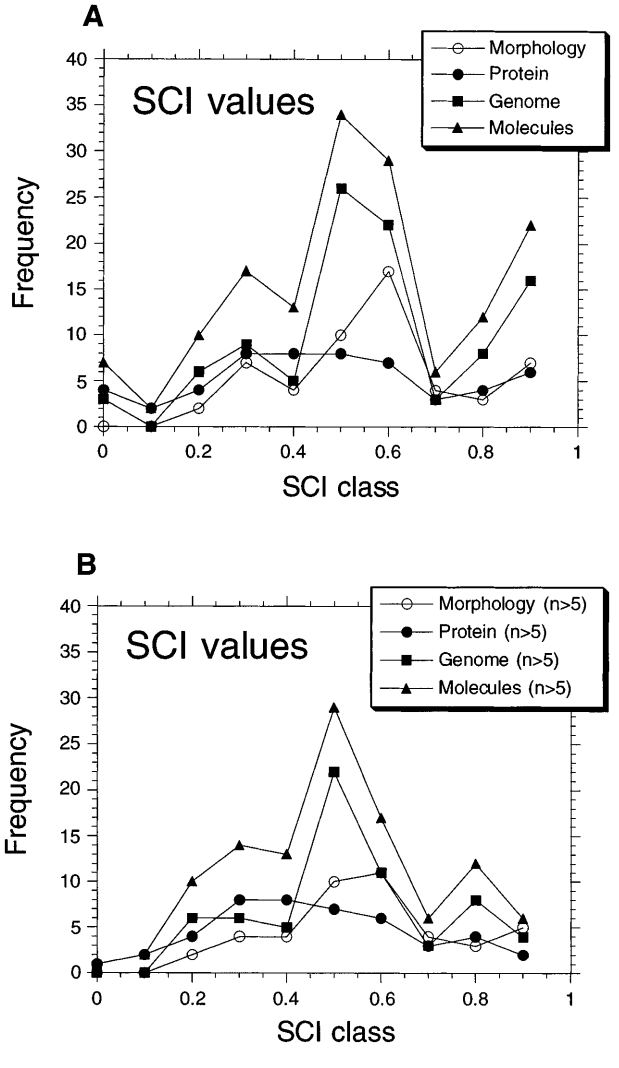

FIG. 5. Frequencies of different values of the SCI metric for published phylogenetic trees of mammals, based on morphological, protein, genome, and molecular (protein + genome) data sets. Frequencies are shown for samples of all cladograms (A) and for cladograms with more than five terminal taxa (B). All frequency distributions in each case are significantly different, except for the morphological and protein curves in both cases (A, B) and the morphological and genome curves for larger cladograms (B), which could be sampled from the same distribution (Kolmogorov-Smirnov test).

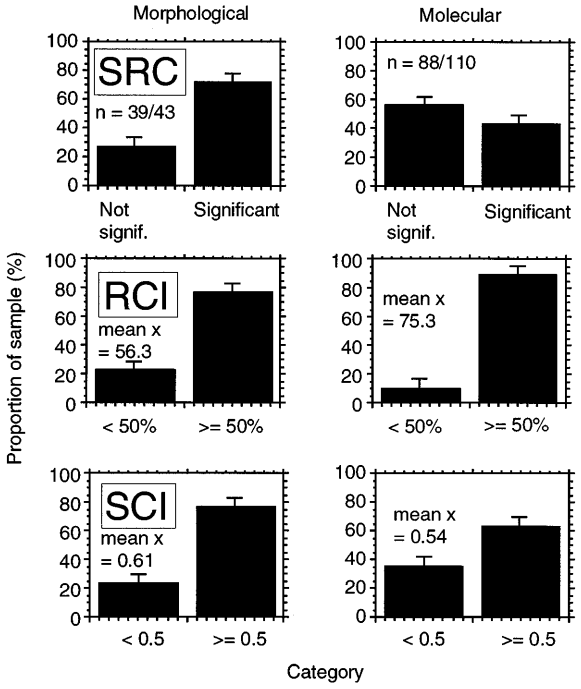

FIG. 6. Summary of the metrics for comparison of published phylogenetic trees of mammals with stratigraphy, based on morphological and molecular (protein + genome) data. The metrics have been applied here only to cladograms with six or more terminal taxa, and this yields lower sample sizes (*n*) for the SRC test (based on collapsed trees) than the RCI and SCI metrics (based on unmodified trees). Comparisons are between frequencies of significant ($P < 0.05$) and nonsignificant SRC tests, frequencies of values of RCI above and below 50%, and frequencies of values of SCI above and below 0.5. Other conventions are explained in Fig. 2.

The RCI metric gave significantly better values for molecular data, with 90% of larger molecular trees showing values greater than 50%, compared to 77% of all morphological trees. This result is confirmed by the higher mean value for molecular trees (75.3%), compared to 56.3% for morphological trees. Protein trees performed best (98%), and morphological trees performed worst (77%). These results were similar for the full data set.

The SCI metric gave results that appear slightly favorable to morphological trees, with 77% of larger trees achieving SCI values equal to, or greater than, 0.5. Protein trees achieved lowest pass rates, at 49% for larger trees, much lower than for genome trees (74%). However, the distributions of SCI values showed less

difference, with the Kolmogorov-Smirnov test finding close similarities between frequency distributions of SCI values for the protein and morphological trees.

CONCLUSIONS

The present study has confirmed for the first time that phylogenetic trees based on molecular data, for mammals at least, show good congruence with expectations from stratigraphic data on the first occurrence of fossils. In addition, it has been shown that trees based on morphological data generally show better congruence with stratigraphic data than do molecular trees. These findings are based on good samples of published trees (54 morphological trees, 152 molecular trees), but larger, or different samples could show different results. Nonetheless, the results reported here show statistically significant differentiation.

The differences between morphological and molecular trees are probably not simply artifacts of sampling: taxic levels and sizes of the trees are equivalent in both data sets. Perhaps the longer span of accumulated research time reflected in the morphological cladograms than in the molecular trees has led to more mature trees that are closer to the truth. Or perhaps morphologists have been more biased in somehow finding MPTs that match the expectations of the order of taxa from the fossil record. More research time, and larger samples of published trees, may shed more light on this dilemma.

An unexpected aspect of the results is that morphological and molecular trees perform differently depending on the congruence metric employed: morphological trees show better congruence with data on stratigraphic data than do molecular trees (SRC, SCI), although molecular trees minimize the ghost range better (RCI). This could be a genuine result, in which case it raises interesting questions about order and gaps in the fossil record, and the ways in which different kinds of character data contribute to the understanding of phylogeny. The result could, however, be an artifact of the sample of protein trees. The great majority of protein trees (43 of 54) are for large parts of the mammalian tree (Mammalia, Theria, Eutheria), rather than for smaller sections of it. Comprehensive cladograms of mammals are dominated by the placental mammal orders, most of which have oldest fossils of very similar age (Paleocene, early Eocene). When the oldest fossils of the majority of branches in a tree are of similar age, the ghost range is much reduced, and the RCI is inevitably high.

Genome trees appear to show better matching with stratigraphic expectations than do protein trees, when assessed by the SRC and SCI metrics, but protein trees show the best matching using the RCI metric. The SRC and SCI metrics assess aspects of the order of nodes in a tree, while the RCI metric focuses on a comparison of known and ghost range. It is not clear why the protein trees show such poor matching of clade and age order, nor why they appear so well to minimize the amount of ghost range in comparison to genome and morphological trees.

These results could be used as part of a debate about whether molecular or morphological cladograms are best; it has been shown here, for example, that morphological trees generally match stratigraphic data better than molecular. However, the molecules vs morphology "controversy," if it ever had any substance, is actually defused by the present study, since *all* kinds of character data, whether morphological, protein, or genome, apparently yield phylogenetic trees of equivalent quality. None of these kinds of character data can be discarded as being consistently uninformative about phylogeny. There are significant differences in congruence of the different data sets, and some of these differences may relate to real aspects of the data and their ability to reconstruct phylogeny, while others may relate to the particular data set used here and of differences in the analytical techniques employed.

The suite of assessment metrics (Fig. 1) offers a simple testing regime that allows systematists to compare the congruence of competing MPTs which represent hypotheses of relationship of specific taxa. Equally, the metrics may be applied in a population-sampling way, as here, to tackle general questions about the relative congruence of different kinds of characters, different taxa, different habitats, different research regimes, and different phases in research. An important future study will to be compare the congruence of trees based on partitioned data sets with "total evidence" cladograms.

ACKNOWLEDGMENTS

I am grateful to Simon Tillier, André Adoutte, and Hervé Philippe for organizing the Paris meeting and for their invitation to contribute this article. I thank the University of Bristol for funding my travel to the conference and the Leverhulme Trust for research funding of this work. I thank Rob DeSalle and an anonymous referee for valuable comments on the manuscript.

REFERENCES

Allison, P. A., and Briggs, D. E. G. (1993). Exceptional fossil record: Distribution of soft-tissue preservation through the Phanerozoic. *Geology* **21:** 527–530.

Benton, M. J. (1988). "The Phylogeny and Classification of the Tetrapods: Amphibians, Reptiles, and Birds," Vol. 1, and "Mammals," Vol. 2, pp. 1–377 and 1–329. Clarendon Press, Oxford.

Benton, M. J. (1993). "The Fossil Record 2," pp. 1–839. Chapman & Hall, London.

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., and Hitchin, R. (1996). Testing the quality of the fossil record by groups and by major habitats. *Hist. Biol.* **12:** 111–157.

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

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

Benton, M. J., and Storrs, G. W. (1996). Diversity in the past: Comparing cladistic phylogenies and stratigraphy. *In* "Aspects of the Genesis and Maintenance of Biological Diversity" (M. E. Hochberg, J. Clobert, and R. Barbault, Ed.), pp. 19–40. Oxford Univ. Press, Oxford.

Frost, D., and Kluge, A. (1994). A consideration of epistemology in systematic biology, with special reference to species. *Cladistics* **10:** 259–294.

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

Harland, W. B., Holland, C. H., House, M. R., Hughes, N. F., Reynolds, A. B., Rudwick, M. J. S., Satterthwaite, G. E., Tarlo, L. B. H., and Willey, E. C. (1967). "The Fossil Record: A Symposium with Documentation," pp. 1–827. Geological Society of London, London.

Hitchin, R., and Benton, M. J. (1997a). Congruence between parsimony and stratigraphy: Comparisons of three indices. *Paleobiology* **23:** 20–32.

Hitchin, R., and Benton, M. J. (1997b). Stratigraphic indices and tree balance. *Syst. Biol.* **46:** 563–569.

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

Kidwell, S. M., and Brenchley, P. J. (1996). Evolution of the fossil record: Thickness trends in marine skeletal accumulations and their implications. *In* "Evolutionary Palaeobiology" (D. Jablonski, D. H. Erwin, and J. H. Lipps, Eds.), pp. 290–336. University of Chicago Press, Chicago.

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

Miyamoto, M. M., and Cracraft, J. (1991). "Phylogenetic Analysis of DNA Sequences," pp. 1–358. Oxford Univ. Press, New York.

Norell, M. A. (1992). Taxic origin and temporal diversity: The effect of phylogeny. *In* "Extinction and Phylogeny" (M. J. Novacek and Q. D. Wheeler, Ed.), pp. 89–118. Columbia Univ. Press, New York.

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

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

Prothero, D. R., and Schoch, R. M. (1989). "The Evolution of Perissodactyls," pp. 1–537. Clarendon Press, New York.

Raup, D. M. (1991). The future of analytical paleobiology. *In* "Analytical Paleobiology" (N. L. Gilinsky and P. M. Signor, Eds.), pp. 207–216. *Short Courses in Paleontology,* The Paleontological Society, Knoxville, TN.

Savage, D. E., and Russell, D. E. (1983). "Mammalian paleofaunas of the world," pp. 1–432, Addison-Wesley, Reading, Mass.

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

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

Smith, A. B. (1994). "Systematics and the Fossil Record," pp. 1–223. Blackwell Sci., Oxford.

Szalay, F. S., Novacek, M. J., and McKenna, M. C. (1993). "Mammal Phylogeny," Vols. 1 and 2. Springer Verlag, New York.