Points of View

Syst. Biol. 48(3):581–596, 1999

Assessing Congruence Between Cladistic and Stratigraphic Data

MICHAEL J. BENTON,[1] REBECCA HITCHIN, AND MATTHEW A. WILLS[2]

Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, U.K.; E-mail: mike.benton@bristol.ac.uk and R. Hitchin@bristol.ac.uk

Phylogeny offers one of the key pieces of evidence for evolution (Darwin, 1859) and the search for phylogenies draws upon vast stores of data: morphological, molecular, and stratigraphic. Great efforts have been expended in improving methods of phylogeny reconstruction and in trying to assess the quality of the results (e.g., Felsenstein, 1985, 1988; Ax, 1987; Farris, 1989; Forey et al., 1992; Smith, 1994). However, the assessment metrics are internal, based on resampling from within existing data matrices, and hence they do not offer a test against reality. Similarly, many efforts have been made to refine comparisons of stratigraphic data, particularly the completeness of stratigraphic ranges (e.g., Paul, 1982, 1990; Benton, 1987, 1994; Marshall, 1990, 1994). These tests, however, do not address the topologies of phylogenies. One solution to the testing problem is to compare data from different sources, in this context assessing the congruence between phylogenetic and stratigraphic data. This idea was pioneered by Gauthier et al. (1988) and Norell and Novacek (1992a, 1992b) and was further developed by Benton (1994, 1995, 1998a, 1998b), Benton and Storrs (1994, 1996), Huelsenbeck (1994), Siddall (1996, 1997), Benton and Hitchin (1996, 1997), and Hitchin and Benton (1996, 1997).

[1]Address correspondence to Michael J. Benton, Department of Earth Sciences, University of Bristol, Wills Memorial Building, Queen's Road, Bristol, U.K.

[2]Present address: Oxford University Museum of Natural History, Parks Road, Oxford OX1 3PW, U.K.; E-mail: matthew.wills@university-museum.oxford.ac.uk

Assessments of congruence between phylogenetic and stratigraphic data have a variety of uses. First, for a particular phylogenetic problem, it may be of interest to discover which of a number of equally most-parsimonious trees best matches the stratigraphic data. Second, it may also be informative to compare the stratigraphic implications of competing phylogenetic hypotheses for the same taxa published by different authors, perhaps using different suites of characters, such as morphological and molecular (Benton, 1998a). A third use is for larger-scale statistical studies of samples of published cladograms, to determine whether there is overall congruence between stratigraphic and phylogenetic data and to compare different groups of organisms, different parts of the stratigraphic record, different styles of phylogeny reconstruction, and different degrees of completeness of fossil collecting.

Metrics for assessment of congruence between stratigraphic and phylogenetic data have so far been used mainly for the third purpose. Results, based on comparisons of samples of 24 to 376 cladograms, have shown good agreement overall between the stratigraphic order of fossils and the order of cladogram nodes (Norell and Novacek, 1992a, 1992b; Benton and Storrs, 1994, 1996; Benton, 1995, 1998a, 1998b; Benton and Hitchin, 1996, 1997). In addition, echinoderms, fishes, and tetrapods have been shown (Benton and Simms, 1995; Benton and Hitchin, 1996, 1997) to exhibit similar amounts of congruence. Additional

fossil collecting since 1967 has filled appreciable gaps in the fossil record (Benton and Storrs, 1994, 1996), and molecular (protein and genomic) phylogenies show generally a poorer match to stratigraphic data than do morphological phylogenies (Benton, 1998a).

The purpose of this paper is to assess the performance of the three metrics used currently for comparing stratigraphic and phylogenetic data, and to compare their results with a null model of randomized range data distributions across a large sample of published cladograms.

Criticisms

Congruence testing methods have been criticized on three grounds: (1) They give undue prominence to the value of phylogenetic data and downplay stratigraphic data; (2) they make a fundamental, and possibly erroneous, assumption that nodes in cladograms represent real splitting events; and (3) they assume that cladograms are accurate.

The first criticism, that assessments of the quality of the fossil record should not be founded on comparisons with published cladograms because this gives primacy to phylogenetic data and assumes that stratigraphic data are of secondary importance, was made by Clyde and Fisher (1997). Their preference is instead to use a combination of stratigraphic and character data to produce a phylogeny. However, this stratocladistic approach has several fundamental weaknesses: (1) It obscures the real information content of the characters, such that it is impossible to say what the resulting trees mean; (2) it involves mixing of inclusive hierarchical data (from characters) with linear data (from stratigraphy) (Smith, 1994); (3) combining the two kinds of data obscures several specific issues in cladistics, including the meaning of ancestors, initial grouping criteria, and asymmetry in rates of morphological evolution (Norell and Novacek, 1997); and (4) a stratigraphic overlay on a most-parsimonious tree obscures the falsifiability of the phylogenetic hypothesis and confuses two philosophical stances (Rieppel, 1997).

The second criticism was made by Wagner (1998), who questioned the viability of comparisons between cladistic and stratigraphic data, mainly on the grounds that any cladogram could correspond to several phylogenetic trees. In particular, he pointed out that the methods assume that sister groups originated at the same time, whereas in fact one sister might include ancestors of the other. This criticism would be valid for species-level studies of fossils that have excellent fossil records. However, the majority of groups for which cladograms have been produced (e.g., vertebrates, echinoderms, macroplants) have rather patchy distributions of taxa, and there is a strong requirement in cladistics that the terminal taxa be demonstrably monophyletic. In practice, cladograms including fossil taxa have rarely been produced at species level, and the genera or families included in a cladogram usually have an unequivocal earliest member that possesses one or more defining apomorphies of the clade. Hence, the nodes in a cladogram are valid minimum estimates of the timing of branching between two entities. Wills (1999) has developed these points further.

The third caveat, also presented by Wagner (1998), was that the validity of the techniques depends on the accuracy of the cladograms. If we cannot be sure that the cladograms are correct, then we cannot use them to assess completeness of the fossil record. Again, however, this is probably not fundamentally crippling to the technique, since most studies to date have used large data sets of published cladograms, not just a few isolated examples by a small group of authors or those using specific techniques. This allows a statistical approach in which the claim is not that any particular cladogram is correct, but rather that the combined set gives a fair reflection of true phylogenies.

Data

The data set for analysis consisted of 375 cladograms: 58 cladograms of echinoderms, 141 cladograms of fishes, and 176 cladograms of tetrapods (full details may be found at URL http:// palaeo. gly.

bris. ac. uk/cladestrat/cladestrat.html). The samples of echinoderm and fish cladograms are relatively complete, consisting of all the usable cladograms we could locate. The sample of tetrapod cladograms is much less complete, perhaps representing one-tenth of the published total. However, we regard it as comprehensive enough for present purposes, since it was compiled from an unbiased sample of five multiauthor compilations, plus every cladogram from volumes 13 to 15(3) of *Journal of Vertebrate Paleontology* (1993–1995).

We used the *Fossil Record 2* (Benton, 1993) as the sole source of stratigraphic data for the dates of origin of families and suprafamilial taxa. Some of the cladograms include individual genera, and their dates of origin were generally determined from data provided in the paper that presented the cladogram.

In the presentation below, most groups are monophyletic: hence, Dinosauria includes Aves, and Actinopterygii includes teleosts, even though values for Teleostei are also given separately. Some groups, indicated with quotation marks, refer to their traditional paraphyletic usage, namely "fishes," "Agnatha," "Osteichthyes," "Sarcopterygii," "Reptilia," and "Synapsida." However, of course, terminal taxa in the cladograms assessed are claimed to be monophyletic by the authors of the cladograms.

Metrics for Assessing Congruence Between Stratigraphy and Phylogeny

Several metrics for assessing congruence between stratigraphy and phylogeny have been proposed. Three of these have been applied empirically: Spearman rank correlation (SRC; Norell and Novacek, 1992a, 1992b), stratigraphic consistency index (SCI; Huelsenbeck, 1994), relative completeness index (RCI; Benton, 1994), and gap excess ratio (GER; Wills, 1999). Fuller details of the techniques are given in those publications, in Benton and Hitchin (1996, 1997), and in Wagner (1998).

Only the SRC is a statistical test with defined confidence limits, but it has three problems that weaken its use for congruence determination: (1) The test may not be statistically appropriate (Huelsenbeck, 1994); (2) it depends heavily on the spacing of origin points (Benton and Storrs, 1994; Hitchin and Benton, 1996); and (3) the procedure for using the test requires that most cladograms be modified before analysis. The SCI, RCI, and GER offer distinctive approaches to assessing congruence between phylogeny and stratigraphy, comparing the relative order of nodes, the relative completeness of the fossil record, and the relative amount of ghost range. Huelsenbeck (1994) proposed a test to determine the approximate significance of SCI values. A similar approach has been extended to the RCI, the GER, and other potential metrics (Wills, 1999). We think that the SCI, RCI, and GER in conjunction offer a powerful means of comparing cladograms with stratigraphy.

Bias in Stratigraphic Metrics

The SCI and RCI, may suffer from a variety of sources of error, in terms of stratigraphy, taxonomy, cladogram size, and tree balance (Hitchin and Benton, 1996, 1997; Siddall, 1996, 1997; Wills, 1999). Stratigraphic and taxonomic scaling and precision will certainly affect the results, but there is no evidence for a bias beyond what the metrics are designed to assess. The SCI, which is designed to compare the relative order of nodes, will yield low values if the matching is poor. Likewise, a poor fossil record will give low RCI values, but that is precisely the intention.

Cladogram size.—The SCI and RCI would be compromised if they varied with cladogram size (number of taxa, n, or number of internal nodes, nn). Siddall (1996, 1997) argued from simulations and from empirical observations that the SCI is significantly negatively correlated with n; in other words, the nodes become less stratigraphically consistent as cladogram size increases. We could not confirm such a relationship for any of the culls of our data set, and results were similarly inconclusive for the RCI metric (Hitchin and Benton, 1997).

Tree balance.—Siddall (1996, 1997) suggested that the SCI improved as cladograms became less balanced (more pectinate). He found that the SCI was significantly neg-

atively correlated with *Im*, Heard's (1992) tree imbalance index, based on a sample of 14 cladograms. This finding makes sense in light of the evident increase in the proportion of possible balanced (symmetrical) trees (i.e., reduction in *Im* values) as numbers of taxa increase. In addition, perfectly balanced trees must have SCI values in the range 0.50–1.00, whereas unbalanced trees may occupy the whole range of values from 0.00 to 1.00. There is a clear theoretical relationship between the range of possible tree balance values, tree balance, and number of terminal taxa. We have not been able to confirm Siddall's (1996) finding, however, based on our sample of 375 cladograms (Hitchin and Benton, 1997). Plots of RCI against *Im* for our full data set, and for various culls of that data set, also failed to establish any significant relationships, whether negative or positive.

Assessing Statistical Significance for the SCI and RCI

As discussed above, several factors (such as tree size, tree balance, and the temporal spacing and extent of range data) set upper and lower limits on the SCI and RCI values that can be obtained from a given data set (regardless of the actual relationships conveyed by the phylogeny, or the precise distribution of range data on that tree). Benton and Hitchin (1997), in assaying 375 cladograms for echinoderms, fish, and tetrapods, have argued that these biases are probably unimportant in comparing large samples of real data. However, in assessing congruence in detail, a bias-free measure for a particular phylogeny and a particular set of range data is preferable. This permits comparisons of congruence for specific cases. As a partial solution, Wills (1999) forwarded the GER, which expresses the observed minimum cladistically implied gap (MIG) as a fraction of the total range of possible MIG values, given the distribution of range data. This, however, is still biased by differences in tree balance.

Randomization procedures offer an empirical solution. Tree topology is held constant while the range data are randomly reassigned over the tree a large number of times (e.g., 1,000+; see Huelsenbeck, 1994). SCI, RCI, or GER is recalculated for the randomized data sets, and the original value is compared with this distribution. The percentage of randomized values as high as, or higher than, the original provides a measure of the uncertainty that the observed level of congruence is better than a random distribution of range data on the tree (see Wills, 1999, for more discussion). These percentages (rather than raw congruence indices) can then be compared for different data sets. The procedure also provides a test, using a null hypothesis stating that the observed level of congruence is consistent with a random distribution of range data, and the alternative hypothesis stating that congruence is better. Uncertainty of 5% or less leads us to reject H_0. Uncertainty values calculated for the SCI and RCI metrics for all 375 cladograms in our sample are listed in full on the Web.

RESULTS

Significance Tests

The SRC, RCI, SCI, and GER metrics were calculated for all 375 cladograms in the data set. For the RCI and GER, only about one-third of the sampled cladograms gave statistically significant ($P < 0.05$) values (RCI and GER significances are equivalent): for the SCI, < 25% were significant. (For the SRC, 145 cladograms [39%] show statistically significant matching of clade order and age order. We have reservations about the use of this metric, and it is not discussed further.)

Distributions of RCI and SCI significance values are illustrated in Figure 1. The significance test can be couched in one of two ways. For both, the null hypothesis is that the observed level of stratigraphic congruence (observed value for RCI or SCI) is consistent with a random distribution of range data on the tree. As above, an alternative hypothesis states that congruence is better than random. In this form, the test has been used to cull the data set down to include just cladograms with significant fit (those comprising the left bars in Fig. 1). The data set has, therefore, been reduced to 130 cladograms with significant RCI and GER values

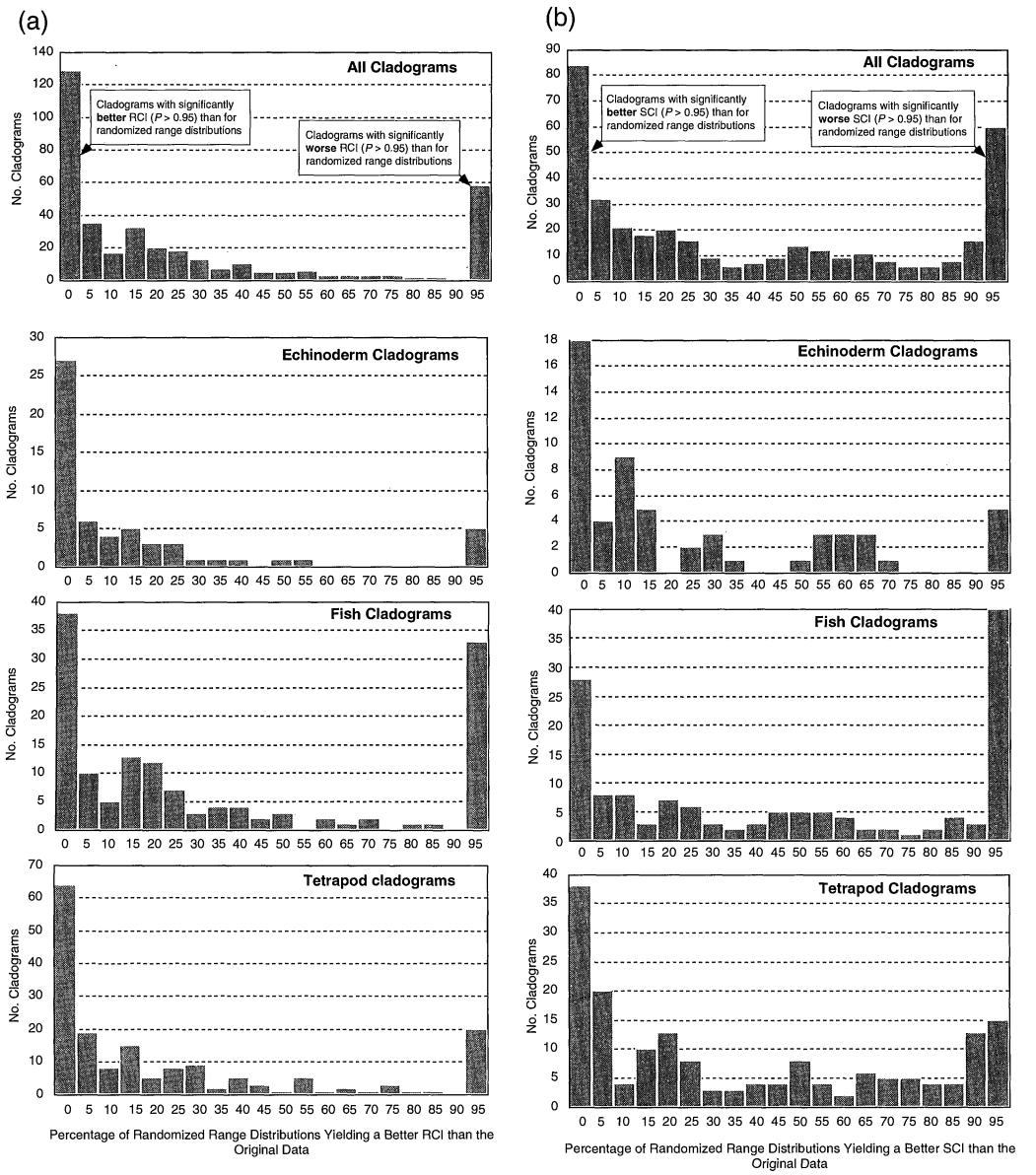

FIGURE 1. Distributions of RCI (a) and SCI (b) significance values for all 375 cladograms and for major groups (echinoderms, fish, and tetrapods).

and 85 with significant SCI values. Only the significant trees have been used to compile the summary statistics in Table 1, which differs from previous publications in which the entire data set was used.

Another alternative hypothesis states that congruence is actually worse than expected from a random model and that the signals from the cladogram and stratigraphy are in direct conflict. Trees for which 95% or more of the randomized distributions yield higher congruence (the right-hand bars in Fig. 1) therefore have significant mismatch.

Figure 1 demonstrates a reasonably even overall distribution of SCI significance values, except at the extremes, where relatively

TABLE 1. Results of tests of congruence between cladograms and stratigraphy. The total number of cladograms is given for each taxonomic group. Other figures refer to significant cladograms only (those whose congruence departs from a random model of range distribution over the tree with $P \geq 0.95$). All taxa are monophyletic, except those in quotation marks, where group names are used in a traditional paraphyletic sense.

		Relative completeness index					Gap excess ratio			Stratigraphic consistency index				
		Significant cladograms[a]			Significant cladograms > 50%[c]			Significant cladograms > 0.5[d]		Significant cladograms[a]			Significant cladograms > 0.5[d]	
Taxon	Total no.	No.	%	Mean RCI[b]	No.	%	Mean GER[b]	No.	%	No.	%	Mean SCI[b]	No.	%
All	375	130	35	56.53	90	69	0.812	125	96	85	23	0.735	78	92
Echinodermata	58	27	47	61.62	20	74	0.877	27	100	18	31	0.773	17	94
Nonechinoids	22	10	45	61.98	7	70	0.907	10	100	6	27	0.849	6	100
Echinoidea	30	14	47	55.48	10	71	0.862	14	100	10	33	0.724	9	90
"Fishes"	141	38	27	65.11	27	71	0.832	35	92	28	20	0.757	26	93
"Agnatha"	24	1	4	40.13	0	0	0.779	1	100	2	8	0.733	2	100
Gnathostomata	117	37	32	65.79	27	73	0.834	34	92	26	22	0.759	24	93
Placodermi	28	6	21	73.26	6	100	0.854	6	100	3	11	0.741	3	100
Acanthodii	2	0	0							0	0			
Chondrichthyes	7	1	14	63.17	1	100	0.990	1	100	1	14	0.750	1	100
"Osteichthyes"	60	24	40	59.43	14	58	0.816	21	88	18	27	0.768	16	89
"Actinopterygii"	35	13	37	50.08	7	54	0.744	10	77	12	34	0.722	12	100
Teleostei	21	6	29	59.11	4	67	0.796	5	83	6	29	0.744	6	100
"Sarcopterygii"	25	11	44	70.48	8	73	0.900	11	100	6	24	0.904	6	100
Tetrapoda	176	65	37	49.20	43	66	0.773	63	97	39	22	0.701	39	100
Amphibia	11	6	55	61.47	4	67	0.753	6	100	1	9	0.625	1	100
Amniota	162	58	36	47.31	39	67	0.773	56	97	38	23	0.703	33	87
"Reptilia"	100	43	43	39.02	23	53	0.768	41	95	28	28	0.668	24	86
Testudines	12	5	42	–29.96	0	0	0.712	5	100	3	25	0.586	3	100
Diapsida	69	33	48	46.13	20	61	0.794	33	100	21	30	0.668	17	81
Lepidosauromorpha	27	8	30	18.87	4	50	0.808	8	100	3	11	0.714	3	100
Archosauromorpha	40	23	58	53.06	14	61	0.783	23	100	18	45	0.66	14	78
Dinosauria	26	15	58	41.99	7	47	0.801	15	100	14	54	0.698	12	86
Synapsida	73	17	23	67.99	15	88	0.756	15	88	12	16	0.798	12	100
"Mammallike reptiles"	19	5	26	61.08	3	60	0.654	3	60	4	21	0.729	4	100
Mammalia	54	12	22	71.12	12	100	0.802	12	100	8	15	0.837	8	100

[a] Number and percentage of cladograms with significant RCI or SCI values. Number and percentage for GER is the same as that for RCI.
[b] Mean values for the respective significant cladograms.
[c] Number and percentage of significant cladograms with RCI values > 50%.
[d] Number and percentage of significant cladograms with GER or SCI values > 0.500.

large numbers of cladograms are significantly better or worse than random. The distribution for RCI values is more skewed towards the left (more congruent) than the SCI distribution; nonetheless, RCI values for relatively large numbers of trees are significantly worse than random. Some cladograms contain taxa with no fossil record. If these are interspersed throughout the tree, congruence is likely to be low and may well appear to be worse than random. Exclusively extant taxa make up only a small proportion ($\sim$ 5%) of those sampled by the 375 cladograms. Moreover, their distribution throughout cladograms in the different SCI and RCI significance bands is remarkably even. According to the SCI test, significantly incongruent cladograms have 3.5% Recent origins on average, whereas significantly congruent cladograms have 2.6% Recent origins. Similarly, for the RCI test, significantly incongruent trees have 3.2% Recent origins, as do 3.0% of the significantly congruent trees. In fact, the proportion of taxa with no fossil record is lower at the extremes of the significance distribution than between them (although none of these differences is statistically significant).

Cladograms with RCI or GER values significantly better than random are very unevenly distributed across the major taxonomic groups: 47% of echinoderm cladograms, 37% of tetrapod cladograms, and only 27% of fish cladograms (Table 1). Proportions are even lower for SCI values: Only 31% of echinoderm cladograms had significant values, 22% of tetrapod cladograms, and 20% of fish cladograms.

The distribution of cladograms with significantly worse congruence than random across the major taxa is even more uneven. For the RCI test, only 9% of echinoderm cladograms and 11% of tetrapod cladograms showed significantly conflicting signals, but 23% of fish cladograms were significantly worse. For the SCI test, the differential is even greater: 9% for echinoderms and tetrapods and 28% for fish.

Cladograms with RCI or GER values significantly better than random were very unevenly distributed among the individual groups of echinoderms, fishes, and tetrapods (Table 1). A few groups, such as amphibians, archosauromorphs, and dinosaurs, yielded significant RCI values in > 50% of their sample of cladograms. Other groups included very low proportions of cladograms with statistically significant RCI or GER values, such as agnathans (4%), chondrichthyans (14%), placoderms (21%), teleosts (29%), mammals (22%), and synapsids (23%).

Far fewer cladograms yielded significant SCI values; only for the dinosaurs were more than half significant. High values (30–50%) were also found for echinoids, tetrapods, archosauromorphs, diapsids, and actinopterygians. Other groups included low proportions of cladograms with statistically significant SCI values, such as agnathans (8%), placoderms (11%), chondrichthyans (14%), amphibians (9%), and lepidosauromorphs (11%).

Relative Completeness Index

The RCI metrics may be represented either as mean RCI values (Fig. 2a) or as proportions of sampled cladograms in which RCI values exceed 50% (an arbitrarily chosen level), a measure of the numbers of cladograms that have a fossil record > 50% complete in terms of the ratio of known to ghost range (Fig. 2b). Fishes have the highest mean RCI value overall (65.11), compared with echinoderms (61.62) and tetrapods (49.20). However, echinoderms show the highest proportion of cladograms with RCI values > 50% (74%), followed by fish (71%) and tetrapods (66%).

Among echinoderms, nonechinoids have a higher mean RCI value (61.98) than echinoids (55.48), although both groups have similar proportions of cladograms with RCI values > 50% (70% and 71%, respectively).

Among fishes, mean RCI values show a relatively narrow range of variation: placoderms (73.26), sarcopterygians (70.48), chondrichthyans (63.17), teleosts (59.11), actinopterygians (50.08), and agnathans (40.13). The order is little different for the proportions of cladograms with RCI values > 50% : placoderms (100%), chondrichthyans (100%), sarcopterygians (73%), teleosts (67%), actinopterygians (54%), and ag-

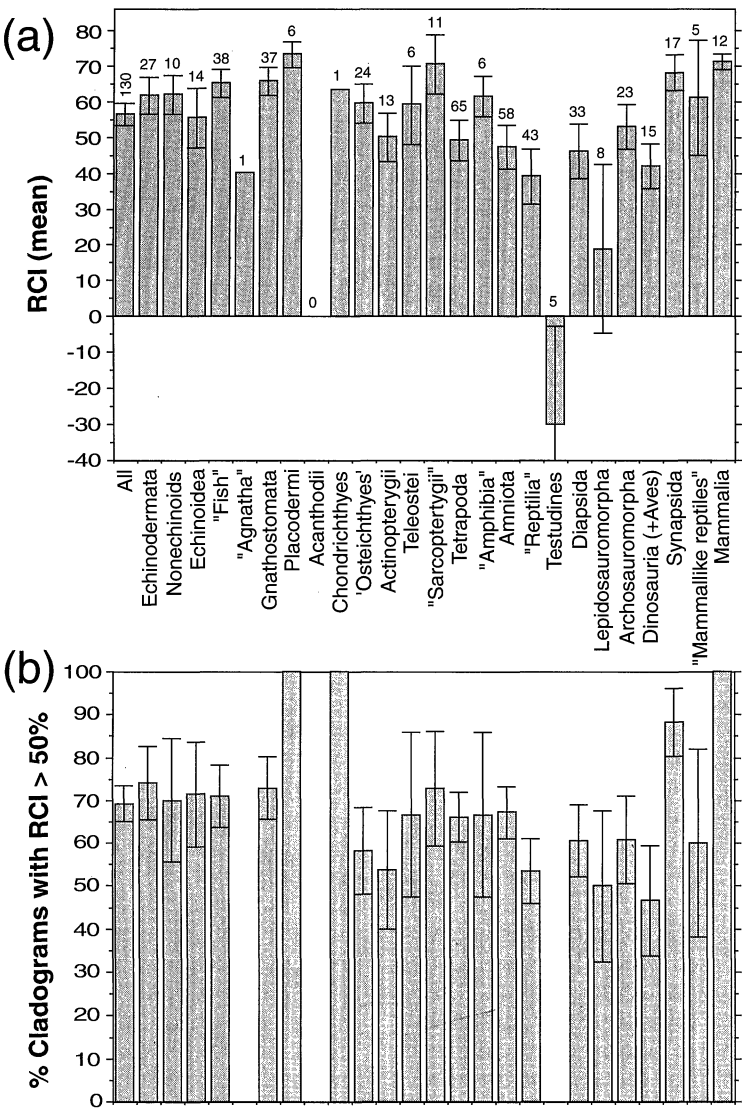

FIGURE 2. Comparisons of the proportions of MIG to known stratigraphic range, as measured by the RCI, for various groups of echinoderms and vertebrates. (a) Mean RCI values. Error bars represent one standard error on either side of the mean. The 95% confidence intervals correspond to bars ~ 1.96 times this length. Numbers of cladograms in each category are indicated at the tops of the columns. All = all echinoderm, fish, and tetrapod cladograms in the sample. (b) Percentage of cladograms with RCI values > 50%. Error bars represent one standard error on either side of the mean. The standard error for a binomial distribution was calculated according to the approximation of Buzas (1990).

nathans (0%). These values are generally not distinguishable according to binomial error bars, although the proportion of high-scoring agnathan trees is significantly lower than for other groups, and the proportions of high-scoring placoderm and chondrichthyan trees are significantly higher than for other groups (Fig. 2b).

The orders of values of the two RCI measures for tetrapods are comparable (Fig. 2). Mean values of RCI show an enormous range for tetrapods: mammals

(71.12), amphibians (61.47), "mammallike reptiles" (61.08), archosauromorphs (53.06), dinosaurs (41.99), turtles (29.96), and lepidosauromorphs (18.87). The proportions that have RCI values > 50% is: mammals (100%), amphibians (66%), "mammallike reptiles" (60%), archosauromorphs (61%), lepidosauromorphs (50%), dinosaurs (47%), and turtles (0%). Binomial error bars indicate that the low mean RCI value for turtles is distinguishable from the high values for synapsids and "mammallike reptiles" (Fig. 2a), and the values for proportions of cladograms of the last two with high RCI values are also distinguishable from the others (Fig. 2b).

Stratigraphic Consistency Index

The SCI metric may also be summarized either as a mean value for each taxonomic group or as a proportion of cladograms that score SCI values of 0.500 or more, an indication that half, or more, of the branches are consistent with stratigraphic evidence. By both measures, fishes and echinoderms score better than tetrapods. Mean SCI values are: echinoderms (0.773), fishes (0.757), and tetrapods (0.701). Proportions of cladograms with SCI values ≥ 0.500 are tetrapods (100%), echinoderms (94%), and fishes (93%). For both measures, values for all three groups are indistinguishable according to binomial error bars (Fig. 3).

Within the sample of echinoderm cladograms, nonechinoids show somewhat better results than echinoids but not significantly so (Fig. 3). The mean SCI value for echinoids is 0.724, and for nonechinoids 0.849; moreover, 90% of echinoid cladograms have SCI values ≥ 0.500, compared with 100% for nonechinoids.

SCI values for fish groups are variable but not significantly different (Fig. 3). For mean SCI values, the order is as follows: sarcopterygians (0.904), teleosts (0.744), placoderms (0.741), agnathans (0.733), and actinopterygians (0.722). In all cases, all sampled cladograms show SCI values > 0.500.

The rankings of tetrapod groups by both aspects of the SCI metric are comparable. Mean SCI values give this sequence: mammals (0.837), "mammallike reptiles" (0.729), lepidosauromorphs (0.714), dinosaurs (0.698), archosauromorphs (0.660), and turtles (0.586). The low value for turtles is significantly lower than the high values for synapsids, mammals, and "mammallike reptiles". Proportions of cladograms with SCI values ≥ 0.500 give this sequence: mammals (100%), "mammallike reptiles" (100%), lepidosauromorphs (100%), turtles (100%), dinosaurs (86%), and archosauromorphs (78%).

Gap Excess Ratio

Values for the significance of the GER are identical (by definition) to those of the RCI and are not discussed here. Echinoderms have the highest mean value (0.877), followed by fishes (0.832) and tetrapods (0.773). All of the significant echinoderm cladograms have a GER > 0.500, and those of nearly all of the tetrapod (97%) and fish (92%) cladograms exceed 0.500 (Fig. 4). Values for all three groups are indistinguishable according to binomial error bars. Mean GER values for significant cladograms in groups show a similar amount of variation to the SCI (and none of the differences is significant).

Within the echinoderms, nonechinoids have a higher mean GER (0.907) than echinoids (0.862), though not significantly so (similar to the performance of the RCI and SCI), and all significant cladograms in both groups have GER values exceeding 0.500.

There is relatively little variation in mean GER values for different fish groups. The highest values are for chondrichthyans (0.990), closely followed by sarcopterygians (0.900), placoderms (0.854), osteichthyans (0.816), teleosts (0.796), agnathans (0.779), and actinopterygians (0.744). Although the majority of cladograms have a GER value > 0.5, the pecentages have a similar order for the different fish groups: chondrichthyans (100%), sarcopterygians (100%), placoderms (100%), agnathans (100%), osteichthyans (88%), teleosts (83%), and actinopterygians (77%).

Groups of tetrapods also show very little variation in mean GER. Better values are for the lepidosauromorphs (0.808), mammals (0.802), dinosaurs (0.801), and archosauro-

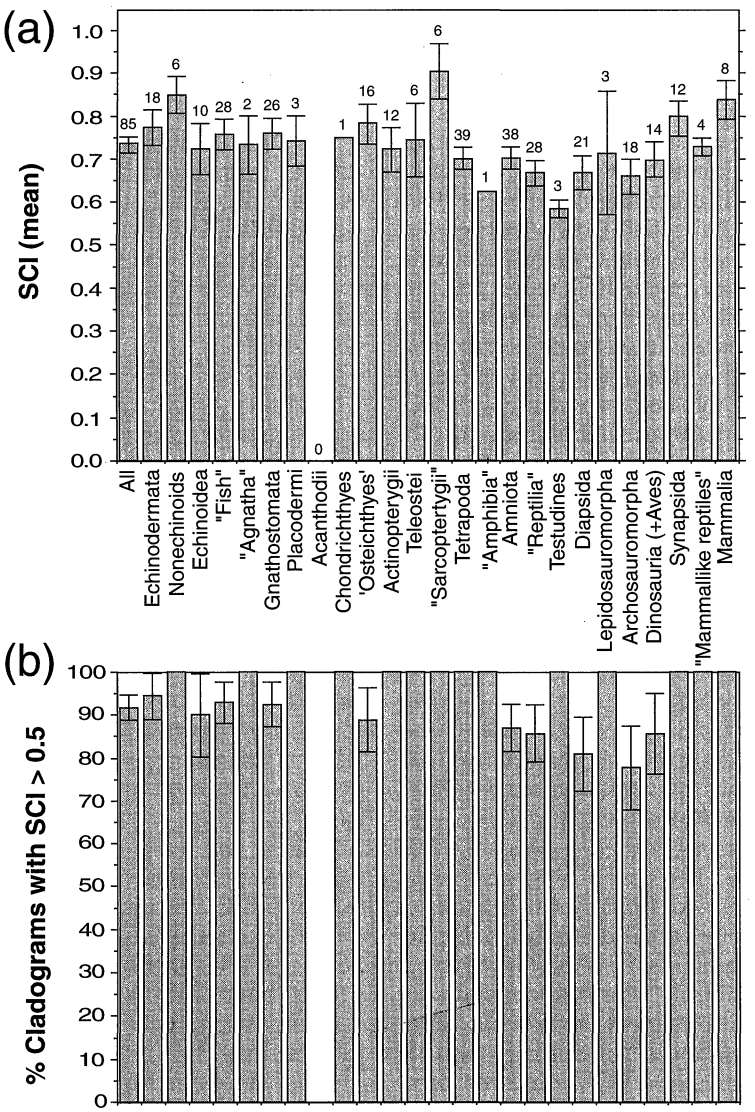

FIGURE 3. Comparisons of the stratigraphic consistency of nodes in cladograms, as measured by the SCI, for various groups of echinoderms and vertebrates. (a) Mean SCI values. Error bars represent one standard error on either side of the mean. The 95% confidence intervals correspond to bars ~ 1.96 times this length. (b) Percentage of cladograms with SCI values > 0.500. Error bars represent one standard error on either side of the mean. The standard error for a binomial distribution was calculated according to the approximation of Buzas (1990).

morphs (0.783), with turtles (0.712) and "mammallike reptiles" (0.654) being the lowest. None of these differences is significant. All significant cladograms in these groups have a GER > 0.500, except the synapsids (88%) and mammallike reptiles (60%) (in the latter case, the sample size of five trees is very small).

Often, 100% of cladograms passing the GER randomization test have GER indices > 0.500 (96% averaged over all 130 trees). Cladograms passing the SCI test also often

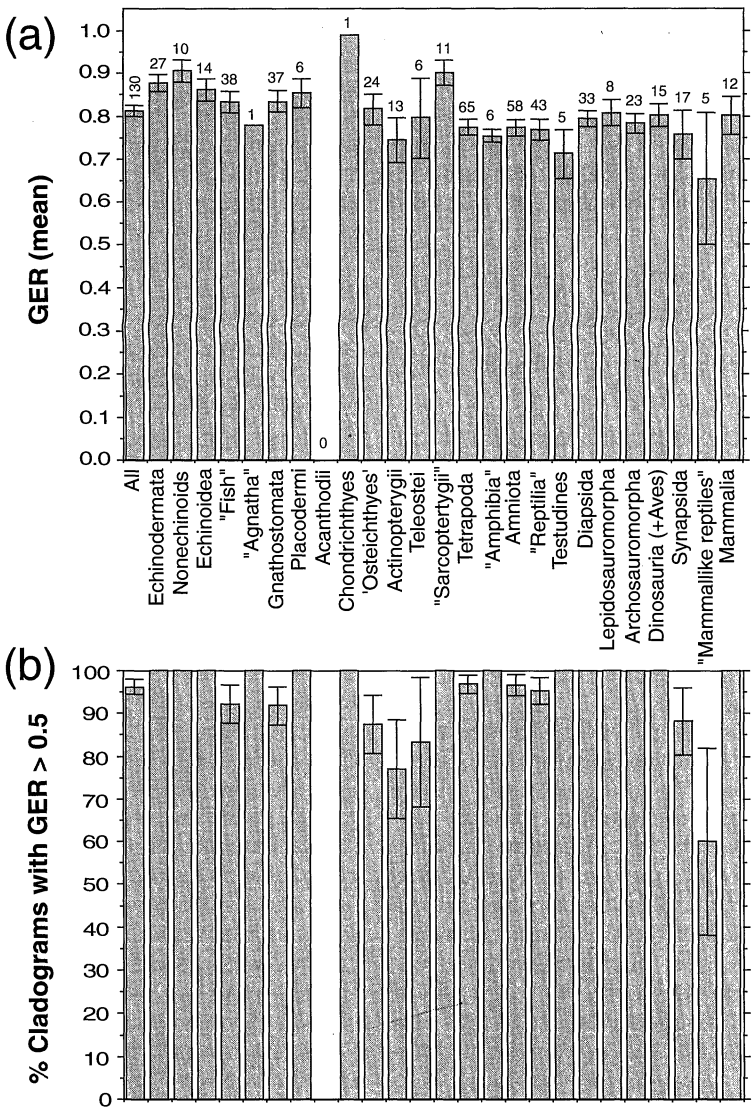

FIGURE 4. Comparisons of the stratigraphic consistency of nodes in cladograms, as measured by the GER, for various groups of echinoderms and vertebrates. (a) Mean GER values. Error bars represent one standard error on either side of the mean. The 95% confidence intervals correspond to bars ~ 1.96 times this length. (b) Percentage of cladograms with SCI values > 0.500. Error bars represent one standard error on either side of the mean. The standard error for a binomial distribution was calculated according to the approximation of Buzas (1990).

have SCI indices > 0.500 (92% over all trees), but the corresponding value for the RCI test is much lower (just 69% of trees passing the test have indices > 50%). The SCI and GER are indices of congruence, whereas the RCI also assesses the extent of observed ranges, the simple range length (SRL). It is therefore not surprising that the SCI and GER have higher values for significant trees. Moreover, since the GER accounts for differences in range distributions, it is also not surprising that the mean GER value for significant trees

(over the whole data set) is higher than the mean SCI value. It can be shown that SCI correlates with its own randomized significance far better than the RCI does, and the GER performs best of all (Wills et al., unpubl.). Moreover, the GER correlates more highly with SCI significance than does the SCI itself.

Discussion

What Causes Poor Matching of Age and Clade Data?

Variations in congruence between cladograms and stratigraphic data result from several factors: (1) differences in the quality of cladograms; (2) differences in the quality of the fossil record; (3) stratigraphic problems; (4) categorical (taxonomic) focus; and (5) sampling density. Only by looking at each case in some detail can we hope to determine the reasons for particularly good or particularly bad matching.

Cladogram quality.—Poorly executed analyses yield artefactually spurious cladograms ("garbage in, garbage out"). In addition, the rapid radiation of groups may mean that a strong phylogenetic signal is absent from the character data (Norell and Novacek, 1992a, 1992b; Benton and Storrs, 1994, 1996). Problems with cladograms (numerous alternative phylogenies and low branch support) often coincide with cases in which some of the metrics are high and others low (e.g., agnathans, some teleosts, amphibians, and basal actinopterygians [i.e., Actinopterygii minus Teleostei]).

Fossil record quality.—The order of fossils in the rocks may not correspond to the true order of appearance of groups for a variety of reasons: (1) Fossils may be genuinely sparse; (2) dating may be wildly inaccurate; and (3) the intensity of collecting and study may be variable, so that important fossils may not have been collected or correctly identified.

A patchy fossil record is likely to be the problem in a case where all metrics are low (e.g., lepidosauromorphs; see Benton, 1987). The MIG can be greatly inflated in cases where most taxa are exclusively extant, and a single taxon (particularly if terminal) has an ancient fossil record. On the other hand, evidence suggests that the record is relatively complete for "mammallike reptiles" and mammals (Benton, 1987; Norell and Novacek, 1992a, 1992b). There is no evidence that variations in the metrics were consistently caused by the quality of dating.

There is an expectation that the Palaeozoic fossil record is poorer than the post-Palaeozoic (Raup, 1972). We have so far found no clear evidence that groups with their main branching points in the Palaeozoic give lower values for the metrics than do those branching in the post-Palaeozoic, but this is an important idea to test further.

Most geologists would also expect that marine groups would have better fossil records, on the whole, than continental groups do. Counterintuitively, echinoderms and fishes show fewer cases than tetrapods in which there is a statistically significant match of age and clade order (Benton and Simms, 1995; Hitchin and Benton, 1996). This could imply that the two aquatic groups have poorer fossil records than the terrestrial tetrapods, a view espoused for echinoderms by Hallam and Wignall (1997:4). However, it could also be a measure of the amount of collecting effort (Benton and Storrs, 1994, 1996). Although echinoderm and fish remains are common, these are usually isolated plates, scales, or teeth, which cannot be reliably classified to genus or family.

The high values for all metrics for "mammallike reptiles" and mammals are probably as much a result of the efforts by palaeontologists to study these groups (Gaston and May, 1992) as the fact that their fossil records are good. Inevitably, vertebrates attract more interest than invertebrates, and mammals more than nonmammals. Comparisons of the state of knowledge at different points in research time (e.g., Maxwell and Benton, 1990; Benton and Storrs, 1994, 1996) may indicate where knowledge is accumulating rapidly (and thus also, perhaps, where data are missing).

Stratigraphic problems.—Three aspects of the geological time scale might affect the congruence metrics: (1) Duration of the

whole cladogram; (2) the mean temporal spacing of nodes; and (3) stratigraphic focus.

Generally, the longer the stratigraphic span of the members of a clade, the better the congruence. Many of the echinoderm and fish cladograms span from the Paleozoic to the present, which might explain their high mean RCI and SCI values. Some of the cladograms of genera of mammals and lizards span in total only 5–30 million years (MY); their congruence values are usually low, because most of these taxa are unknown as fossils.

The mean temporal spacing of nodes is critical for some groups. Stretching of the time span that includes the majority of branching events means that their order is more likely to be maintained, even if the occurrence of fossils is sparse. Many cladograms of higher taxa of echinoderms and fishes have long time intervals between nodes, and hence good SCI values.

The RCI and SCI metrics can give falsely high values if all nodes in a cladogram have the same stratigraphic date, but the SRC test is not compromised in this way. This could be an explanation for the relatively high RCI and SCI values but low SRC values for agnathans and teleosts.

If an attempt is made to date the nodes in a cladogram too precisely, the RCI and SCI values can go down because gaps between known fossil finds are highlighted. In the present study, dating was generally done to the level of the stratigraphic stage (durations range from 2 to 34 MY, mean 7.5 MY), which avoids many of the small-scale uncertainties of disputed dating of particular sites or uncertainties over precise correlation between continents.

Categorical focus.—Most cladograms in our sample have terminal taxa that are traditionally assigned to familial rank. With present knowledge of the fossil record, and current accuracy of dating on a global scale, these represent the best match to stratigraphic divisions at stage level.

Sampling density.—The distribution of taxa within clades can affect the nature of a cladogram and the degree of fit to stratigraphy. Investigators produce cladograms of higher taxa by coding animals in one of three ways: (1) Using representative taxa for established subclades, (2) coding the "average" morphology for established subclades, or (3) coding all taxa.

Many molecular trees are of the first type. Lecointre et al. (1993) have shown how the choice of species to represent families can markedly affect the results obtained. The solution to this problem is to sample more densely within the clade.

Morphological cladograms are often produced by "averaged" sampling. Terminals are higher taxa (genera, families, orders, or phyla) rather than species. Characters may be coded in one of four ways: (1) the most commonly occurring state within the taxon, (2) the most derived state found in one or more species in the taxon, (3) the most plesiomorphic state found in one or more species in the taxon (sometimes, the state seen in the oldest fossil taxon is [often mistakenly] taken as a proxy for this), or (4) all states that occur in any species in the taxon (polymorphic coding). In all cases, the density of taxon sampling cannot be established because it is undefined.

The third sampling type, where all species are included, is rare. Ideally, trees containing all taxa should perhaps be the benchmark against which trees based on representative and averaged morphology are assessed.

The effects of sampling density on cladogram topology and on the stratigraphic metrics have yet to be investigated fully.

Why Are RCI and SCI Values Usually Indistinguishable from Random?

In the sample of 375 cladograms, 130 yielded RCI values and 85 yielded SCI values that were significantly better than random ($P < 0.05$). In addition, 58 yielded RCI values and 60 yielded SCI values significantly worse than random ($P < 0.05$). Why do so many cladograms yield RCI and SCI values that do not differ significantly from random? And why, in this regard, does the SCI yield results that are significantly better than random less often than the RCI metric does?

First, the cladograms that show significantly worse RCI and SCI values than ran-

dom are not explained by particularly poor, or missing, fossil records: Proportions of cladograms with Recent taxa alone are only marginally higher for the significantly incongruent cladograms than for those that were significantly congruent (see above). Significant incongruence might relate to cladogram size but, again, there is no clear relation. The remainder of the discussion focuses on the significantly congruent cladograms.

Cladogram size is important. The nature of the RCI and SCI metrics, and of the permutation procedure, means that more small ($n < 7$) cladograms will yield values that are not significantly better than random than will large ($n \geq 7$) cladograms (discussed by Wills, 1999). Indeed, this is the case. In the sample of 375 cladograms, 112 are small, and only 18 of these (16%) yielded significant RCI values, compared with 35% of the whole sample. The same is true for the SCI measure: Of the 112 small cladograms, only nine (8%) yielded significant SCI values, compared with 23% of the whole sample.

The samples of cladograms of echinoderms, fishes, and tetrapods showed very different proportions of cladograms having RCI and SCI values significantly better than random, but the proportions of small cladograms in each group are different. For example, the low passing rate for fishes (28%) is not explained by an unusually high number of small fish cladograms. On the other hand, nearly half the echinoderm cladograms were small, and yet more RCI and SCI values for echinoderm cladograms were significantly better than random than for fishes and tetrapods.

Congruence of significance values is not obviously related to the total duration of a group: High values are found in echinoderm and fish groups measured over time spans of 400 MY and also in many mammal phylogenies based on time spans of just 30 MY or so. Tree balance does not appear to be the explanation either (Hitchin and Benton, 1997), nor is taxic level a clear determinant.

Particular subgroupings of taxa tend to have significant RCI and SCI values. For example, significant RCI and SCI values are obtained for most cladograms of asteroids, certain echinoid subgroups, arthrodire placoderms, actinopterygians (higher-level cladograms), teleosts (higher-level cladograms), dipnoans, amphibians, turtles (higher-level cladograms), lepidosauromorphs (higher-level cladograms), archosaurs, crocodiles, dinosaurs, "mammallike reptiles," and perissodactyls. By contrast, relatively few cladograms with significant RCI or SCI values are observed for crinoids, agnathans, placoderms (higher-level cladograms), acanthodians, chondrichthyans, teleosts (lower-level cladograms), sarcopterygians (higher-level cladograms), turtles (lower-level cladograms), lizards, mammals (higher-level cladograms), eutherians (higher-level cladograms), artiodactyls, carnivores, primates, proboscideans, and rodents.

Taxa with comparable preservability are found in both the "good" and "bad" lists above, suggesting that the quality of the fossil record is not an overriding factor. Equally, the apparent quality of cladograms cannot be related to a single author or team; in most cases, the independent work of different authors yields cladograms with consistently significant or nonsignificant RCI and SCI values. More probably, some fundamental aspect of the cladogram, its resolvability against the true pattern in nature, contributes to the constancy of significance of RCI and SCI values.

Cladograms with significant RCI values tend to have significant SCI values, although there are more of the former than of the latter. SCI calculations involve smaller numbers of nodes ($n - 2$) than do RCI calculations (n), and small cladograms (more often than large cladograms) tend to fail to achieve significant differences from random. In addition, as a ratio of numbers of nodes, there are far fewer possible SCI values for any cladogram than possible RCI values.

Conclusions

The RCI, SCI, and GER metrics are valid techniques for comparing large samples of cladograms to try to establish variations in congruence between the fossil record and cladograms for different groups of organ-

isms and for different habitats. We do not recommend further use of the SRC test in this context because of the number of practical and theoretical problems with its use.

The three congruence metrics indicate variations in the quality of the fossil record, the intensity of study and collection effort, and the quality of cladogram reconstruction for the different groups of echinoderms and vertebrates. These results suggest areas of improvement in cladistic and fossil-collecting practice. However, studies of cladograms of other groups of organisms, and of their molecular phylogenetic trees, may provide more meaningful generalizations.

Testing the significance of the RCI and SCI metrics for each cladogram provides a measure of confidence that the values depart significantly from a random distribution of range data across the tree. Significance levels do not appear to depend on cladogram size, tree balance, total stratigraphic range, or taxic hierarchical level. Rather, significance depends on the taxic group being studied. Some groups consistently yield cladograms with significant RCI and SCI measures, and other groups do not.

Acknowledgments

We thank the Leverhulme Trust (Grant F/182/AK) for funding and Paul Pearson, Charles Marshall, David Cannatella, Richard Olmstead, and two anonymous reviewers for useful comments on the manuscript.

References

Ax, P. 1987. The phylogenetic system. Wiley, New York.

Benton, M. J. 1987. Mass extinctions among families of non-marine tetrapods: The data. Mém. Soc. Géol. Fr. 150:21–32.

Benton, M. J. 1993. The fossil record 2. 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 349:5–10.

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

Benton, M. J. 1998b. The quality of the fossil record of the vertebrates. Pages 269–303 *in* The adequacy of the fossil record (S. K. Donovan and C. R. C. Paul, eds.). Wiley, Chichester, England.

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

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

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

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

Benton, M. J., and G. W. Storrs. 1996. Diversity in the past: Comparing cladistic phylogenies and stratigraphy. Pages 20–40 *in* Phylogeny and biodiversity (M. Hochberg, J. Clobert, and R. Barbault, eds.). Oxford Univ. Press, Oxford, England.

Buzas, M. A. 1990. Another look at confidence limits for species proportions. J. Paleontol. 64:842–843.

Clyde, W. C., and D. C. Fisher. 1997. Comparing the fit of stratigraphic and morphologic data in phylogenetic analysis. Paleobiology 23:1–19.

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

Farris, J. S. 1989. The retention index and the rescaled consistency index. Cladistics 7:81–91.

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

Felsenstein, J. 1988. Phylogenies and quantitative characters. Ann. Rev. Ecol. Syst. 19:445–471.

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

Gaston, K. J., and R. M. May. 1992. Taxonomy of taxonomists. Nature 356:281–282.

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

Hallam, A., and P. Wignall. 1997. Mass extinctions and their aftermath. Oxford Univ. Press, Oxford, England.

Heard, S. P. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution 46:1818–1826.

Hitchin, R., and M. J. Benton. 1996. Congruence between parsimony and stratigraphy: Comparisons of three indices. Paleobiology 21:20–32.

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

Huelsenbeck, J. P. 1994. Comparing the stratigraphic record to estimates of phylogeny. Paleobiology 40:470–483.

Lecointre, G., H. Philippe, H. L. V. Lé, and H. Le Guyader. 1993. Species sampling has a major impact on phylogenetic inference. Mol. Phylogenet. Evol. 2:205–224.

Marshall, C. R. 1990. Confidence intervals on stratigraphic ranges. Paleobiology 16:1–10.

Marshall, C. R. 1994. Confidence intervals on stratigraphic ranges: Partial relaxation of the assumption

of randomly distributed fossil horizons. Paleobiology 20:459–469.

MAXWELL, W. D., AND M. J. BENTON. 1990. Historical tests of the absolute completeness of the fossil record of tetrapods. Paleobiology 16:322–335.

NORELL, M. A., AND M. J. NOVACEK. 1992a. Congruence between superpositional and phylogenetic patterns: Comparing cladistic patterns with fossil records. Cladistics 8:319–337.

NORELL, M. A., AND M. J. NOVACEK. 1992b. The fossil record and evolution: Comparing cladistic and paleontologic evidence for vertebrate history. Science 255:1690–1693.

NORELL, M. A., AND M. J. NOVACEK. 1997. The ghost dance: A cladistic critique of stratigraphic approaches to paleobiology and phylogeny. J. Vertebr. Paleontol. Suppl. 17:67A.

PAUL, C. R. C. 1982. The adequacy of the fossil record. Pages 75–117 *in* Problems of phylogenetic reconstruction (K. A. Joysey and A. E. Friday, eds.). Academic Press, London.

PAUL, C. R. C. 1990. Completeness of the fossil record. Pages 293–303 *in* Palaeobiology, a synthesis (D. E. G. Briggs and P. R. Crowther, eds.). Blackwell, Oxford, England.

RAUP, D. M. 1972. Taxonomic diversity during the Phanerozoic. Science 177:1065–1071.

RIEPPEL, O. 1997. Falsificationist versus verificationist approaches to history. J. Vertebr. Paleontol. Suppl. 17:71A.

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

SIDDALL, M. E. 1997. Stratigraphic indices and tree balance: A reply to Hitchin and Benton. Syst. Biol. 46:569–573.

SMITH, A. B. 1994. Systematics and the fossil record. Blackwell, Oxford, U.K.

WAGNER, P. J. 1998. Phylogenetic analyses and the quality of the fossil record. Pages 165–187 *in* The adequacy of the fossil record (S. K. Donovan and C. R. C. Paul, eds.). Wiley, Chichester, England

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

Received 18 February 1998; accepted 15 November 1998
Associate Editor: R. Olmstead

Syst. Biol. 48(3):596–603, 1999

Sampling Confidence Envelopes of Phylogenetic Trees for Combinability Testing: A Reply to Rodrigo

FRANÇOIS LUTZONI[1,4] AND F. KEITH BARKER[2,3]

[1]*Department of Botany and* [2] *Department of Zoology, Field Museum of Natural History, Roosevelt Road at Lake Shore Drive, Chicago, Illinois, 60605, USA* [3] *Department of Ecology and Evolution, Committee on Evolutionary Biology, University of Chicago, Chicago, Illinois, 60637-1573, USA*

In 1997, Lutzoni pointed out two main caveats of a method (referred to here as RKB3) proposed by Rodrigo et al. (1993) to determine whether trees derived from different data sets are sample estimates of a parametric phylogenetic tree. The first problem was the high and undetermined number of bootstrap replicates necessary to implement correctly the second part of the RKB3 method. The second problem was the difficulty of handling the huge bootstrap tree files for the tree-to-tree comparisons needed. This second problem was most acute when analyzing data sets with differential (low versus high) resolving power. The term "resolving power" refers here to the relative number of equally most-parsimonious trees associated with a given data set.

Rodrigo (1998) acknowledged part of the first problem and proposed, using an adapted mark–capture–recapture approach, to estimate the number of bootstrap replicates necessary to adequately sample 95% of the unique trees in the "confidence envelope" surrounding the optimal tree(s). To avoid the prohibitive amount of time needed to phylogenetically analyze the extremely high number of bootstrapped data sets (e.g., > 1,000,000) that would be needed in many cases, Rodrigo (1998) proposed using distance-based methods (e.g., neighbor-joining) instead of maximum parsimony. In this reply to Rodrigo (1998), we demonstrate that estimates of the appropriate number of bootstrap replicates (*b*),

[4]Address correspondence to this author. E-mail: flutzoni@fmnh.org.