MISSING ENTRY REPLACEMENT DATA ANALYSIS: A REPLACEMENT APPROACH TO DEALING WITH MISSING DATA IN PALEONTOLOGICAL AND TOTAL EVIDENCE DATA SETS

MARK A. NORELL[1] and WARD C. WHEELER[2]

[1]Division of Paleontology, American Museum of Natural History, 79th Street at Central Park West, New York,
New York 10024-5192, norell@amnh.org;

[2]Division of Invertebrate Zoology, American Museum of Natural History, 79th Street at Central Park West, New York,
New York 10024-5192, wheeler@amnh.org

ABSTRACT—Almost all phylogenetic studies utilizing fossils are faced with the problem of missing data. With the development of methods allowing total evidence phylogenies where fossils are combined with extant taxa, and molecular evidence is available, this problem is compounded. Standard methods ignore missing data and often result in poorly resolved trees, and procedures such as culling of taxa or partitioning data have been proposed to improve resolution. Here, we take a different tack, one grounded in the fact that because we do not know what the behavior of missing data would be, it is worthwhile to examine the universe of possible outcomes. The MERDA value is the frequency with which a particular clade is recovered in replicated analyses where missing observations are replaced randomly with observable states. A technique to de-resolve missing data-dependent clades is also proposed. We also show that, in published data sets, there is little obvious relationship between MERDA performance and standard measures of clade support such as Bremer and Jacknife indices.

INTRODUCTION

Missing data can influence phylogenetic analyses in nefarious ways. The most apparent is that they can lead to the discovery of multiple parsimonious trees (Gauthier 1986, Novacek 1992, Nixon and Wheeler, 1992), which when consensed leave little in the way of resolution. Various approaches have been proposed to deal with such data sets. The most common is to cull taxa, either through their empirical identification as wildcard, or redundant taxa (Wilkinson, 1995) or simply eliminate them from the analysis at some arbitrary threshold level of missing data (Gao and Norell, 1998, see Bemis and Grande 1998). Other proposals divide data into partitions and use a type of taxonomic congruence among these partitions to reflect relationships (O'Leary, 2000) or utilize Adams consensus methods (Adams, 1972). A second problem with missing data is that they can inflate measures of tree and node support (Makovicky, 2000).

Apart from these methodological issues, contemporary approaches to incorporating missing data suffer from other operational problems. The most obvious is that real data do not behave in the same way as missing data. Platnick et al. (1991) showed in small data sets that if all possible character combinations were substituted for missing data in no case could their original tree be recovered as the shortest. The only way that the missing value could prefer the same tree is if they were polymorphic—hence returning to the original problem. It is this behavior that we attempt to examine and quantify here.

The issues regarding missing data come from two different aspects of homoplasy. The first is that missing character observations cannot be homoplastic. In this sense, the observable (non-missing data) can support alternative topologies, while the missing data are ostensibly treated as congruent with the observed. A second related issue is the multiple placements of taxa due to high levels of missing data. This results in almost entirely uninformative strict consensus cladograms.

Here, we propose a method (Missing Entry Replacement Data Analysis—MERDA) that takes a novel approach to identify problematic areas. Instead of identifying which taxa should be eliminated from the analysis a priori, we instead look at a sample of the universe of possible character distributions and determine the frequency of clade or node recovery. In this way, we can identify those clades which are most dependent on the unique properties of missing data (maximal conformance to all topologies), as separate from those of any particular set of real observations.

METHODS

Data were treated as Hennig86/Nona/CLADOS files. Script files were run for each of the input files to automate the missing element replacement. There were three steps to the missing element replacement. First, the input data file was supplied to the program "MERDA" (available via anon ftp ftp.amnh.org). This program scanned the data file for missing values ("?") and created a new file with replaced values drawn equiprobably from the state set defined by the non-missing characters. Hence, if the state were binary, then 50% of the time the missing values were replaced by 0 and 50% by 1; if there were four states 25% for 0, 1, 2, and 3 and so on. Character ordering can be maintained, but we ran all example data as unordered so the effects of this procedure could be more clearly evaluated. Inapplicable values, "-", were unaffected. This differential treatment of inapplicable and truly missing values comes from their separate origins. Missing data have not been observed, inapplicable data have, but the investigator for whatever reason has not made a state assignment (=putative homology) for that taxon for that character. This modified data set was then subjected to parsimony analysis using POY vers. 2.7 (Gladstein and Wheeler, 1997; also available via anon ftp ftp.amnh.org) with the options "-seed-1 maxtrees 10-fuselimit 100-multibuild 12-buildmaxtrees 2-buildspr-treefusespr-approxquickspr-treefuse-norandomizeoutgroup-fitchtrees-approxbuild." This meant that each analysis was subject to 12 random addition sequences, SPR and TBR branchswapping and treefusing with SPR branch swapping (Goloboff, 2000). Thirdly, the strict consensus of the resultant cladograms was calculated and appended to the ensemble output file. After 1,000 replicates of substitution, analysis, and consensus were performed, the fre-
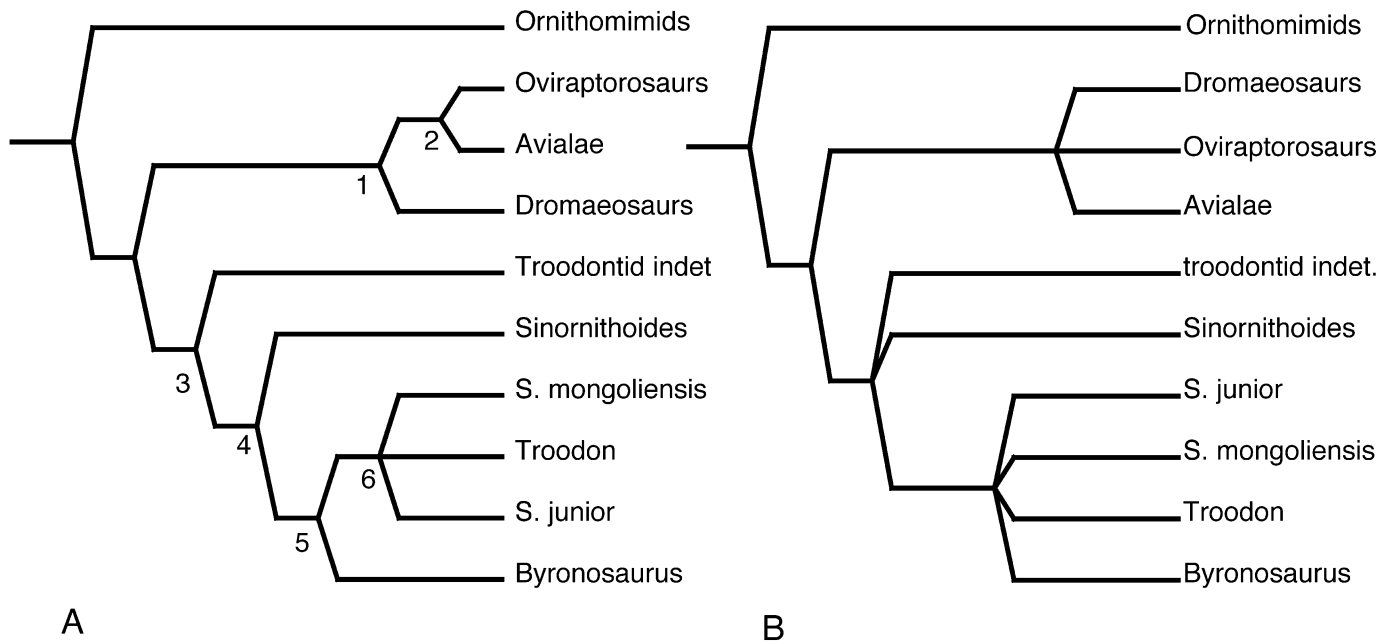
FIGURE 1. **A**, shortest calculated tree (from Norell et al., 2000). **B**, MERDA consensus of these data.

quency of each group present in the strict consensus cladograms was tabulated. Jacknife (1,000 replicates) (Farris et al., 1996) and approximate Bremer values (Bremer, 1998) were also calculated via POY.

The frequency of recovery of each clade is the MERDA value. These values identify clades that are dependant on the idiosyncratic behavior of missing observations (they have low MERDA values). Furthermore, comparison of the contents of groups ranked according to MERDA values with groups found in a tree calculated on the original data matrix that includes missing data often show discrepancies. These discrepancies are groups that are not present in the original analysis, yet have higher MERDA values than groups in the original tree. This observation leads us to propose a second protocol—MERDA collapse: (1) A phylogenetic tree based on all of the taxa and all of the characters is calculated. (2) Groups in the original tree are identified in the ranked MERDA analysis above. (3) If groups not found in the analysis of the original data are found to have higher MERDA values than groups recovered in the original data analysis, this region of the tree is collapsed. Implementing this procedure as a simple script is straightforward (Appendix 1). This sort of collapse criterion is both non-arbitrary and data set specific.

TABLE 1. MERDA, Jacknife and Bremer values from the data of Norell et al. (2000).

Clade	MERDA	Bremer	Jack-nife	MERDA rank	Bremer rank	Jack-nife rank
3	0.632	2	0.878	6	2	6
1	0.522	3	0.716	5	3	5
5	0.482	2	0.573	4	2	1
4	0.468	1	0.574	3	1	2
2	0.295	1	0.685	2	1	4
6	0.056	1	0.611	1	1	3
*	0.32					

SAMPLE CASES

We examined several sample cases that display a wide variety of data matrices, types of data and patterns of missing data. We consider each of these cases below. From these a few general patterns or tendencies emerge. First, there are no general rules concerning support. Bremer indices and Jacknife support are not necessarily correlated with MERDA indices. Furthermore, amount of missing data and overall CIs and RIs are not predictive as to the stability of a particular group or topology. This means that such analyses are data set specific and what is being assayed is the interaction between homoplasy
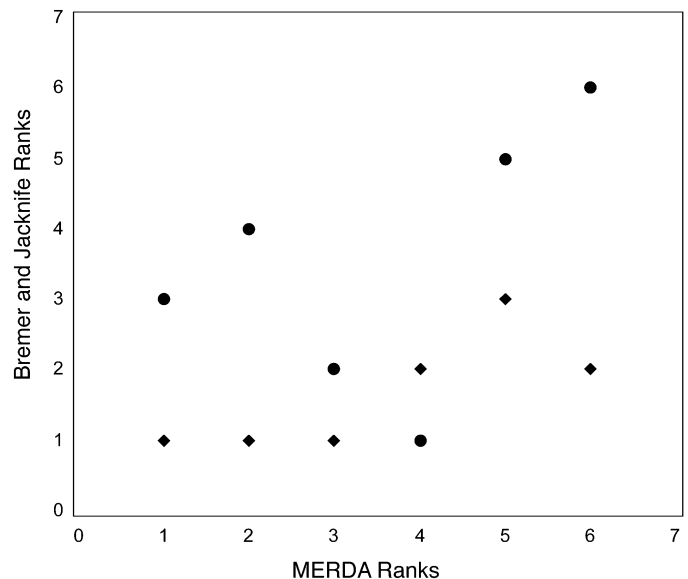

FIGURE 2. Graphical representation between both Bremer support (diamonds) and Jacknife value (dots) and MERDA rank for results calculated from data in Norell et al. (2000).
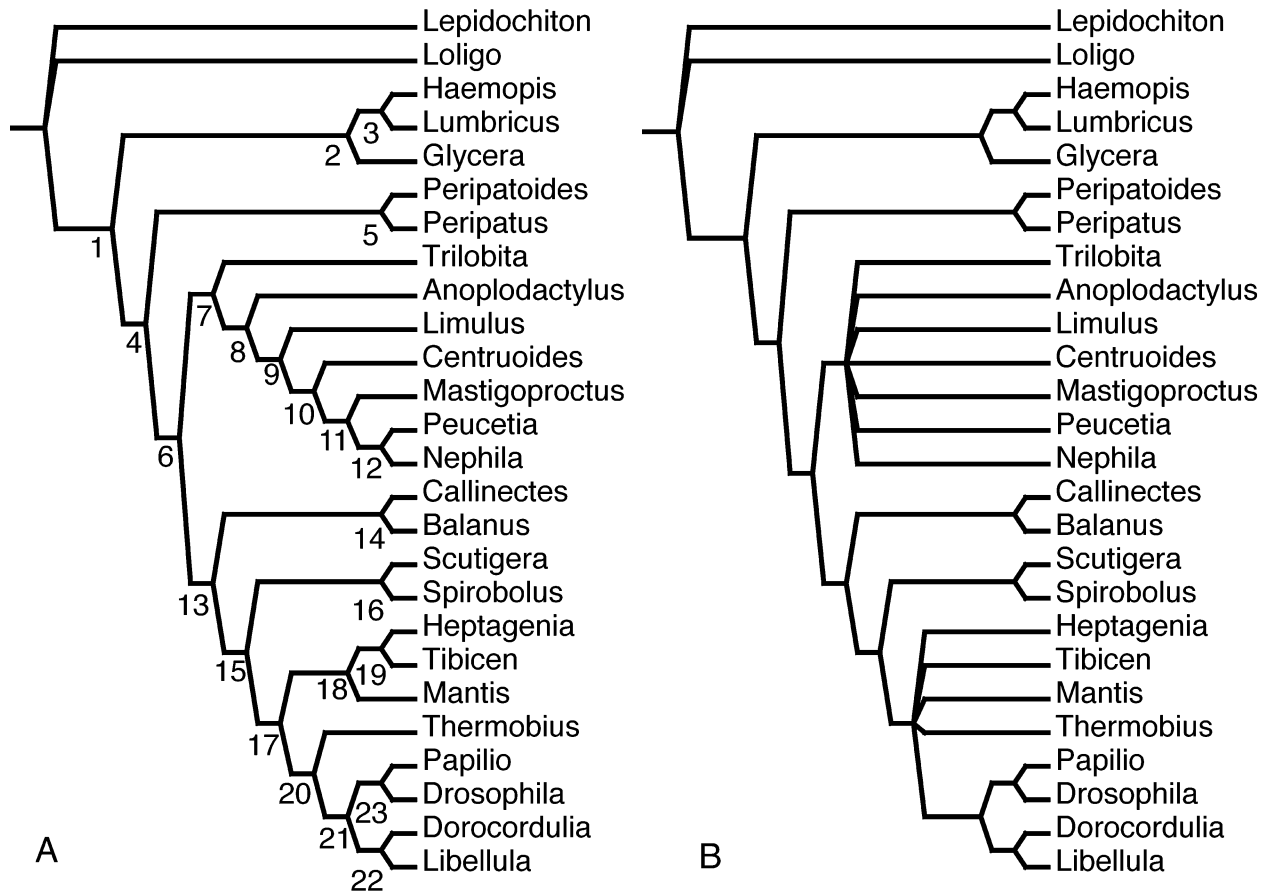
FIGURE 3. **A**, shortest calculated tree from data in Wheeler et al. (1993). **B**, MERDA consensus of these data.

TABLE 2. MERDA, Jacknife and Bremer values from the data of Wheeler et al. (1993).

Clade	MERDA	Bremer	Jack-nife	MERDA rank	Bremer rank	Jack-nife rank
16	1.00	2	0.80	12	2	9
1	1.00	6	0.88	12	5	11
2	1.00	11	0.97	12	8	14
23	1.00	9	0.98	12	7	14
5	1.00	15	0.98	12	9	14
22	0.99	11	0.94	11	8	12
17	0.99	9	0.96	11	7	13
6	0.99	8	0.98	11	6	14
4	0.99	11	0.98	11	8	14
3	0.99	15	0.98	11	9	14
15	0.97	5	0.97	10	4	13
14	0.93	5	0.98	9	4	14
7	0.91	1	0.51	8	1	5
13	0.89	1	0.57	7	1	6
18	0.84	2	0.61	6	2	7
19	0.60	1	0.48	5	1	4
21	0.57	1	0.00	4	1	1
20	0.56	1	0.12	3	1	2
8	0.08	2	0.70	2	2	8
9	0.08	2	0.86	2	2	0
10	0.08	9	0.97	2	7	13
12	0.08	4	0.86	2	3	10
11	0.07	2	0.26	1	2	3
*		0.91				

and missing data. Lastly, these examples are not commentary on the "quality" of any particular data set. Instead they are simply illustrative exercises about the use and behavior of the MERDA metric and the relative importance of missing data for the support of clades. For each of the cases, voluminous output data of these analyses are available at http://research.amnh.org/vertpaleo/norell.html.
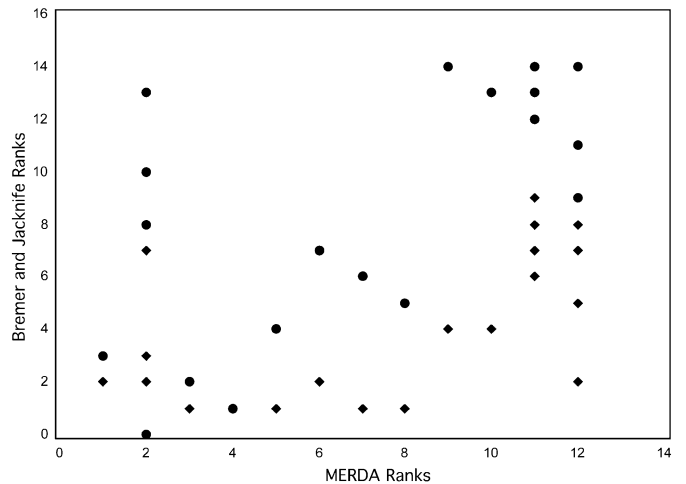

FIGURE 4. Graphical representation between both Bremer support (diamonds) and Jacknife value (dots) and MERDA rank for results calculated from data in Wheeler et al. (1993).
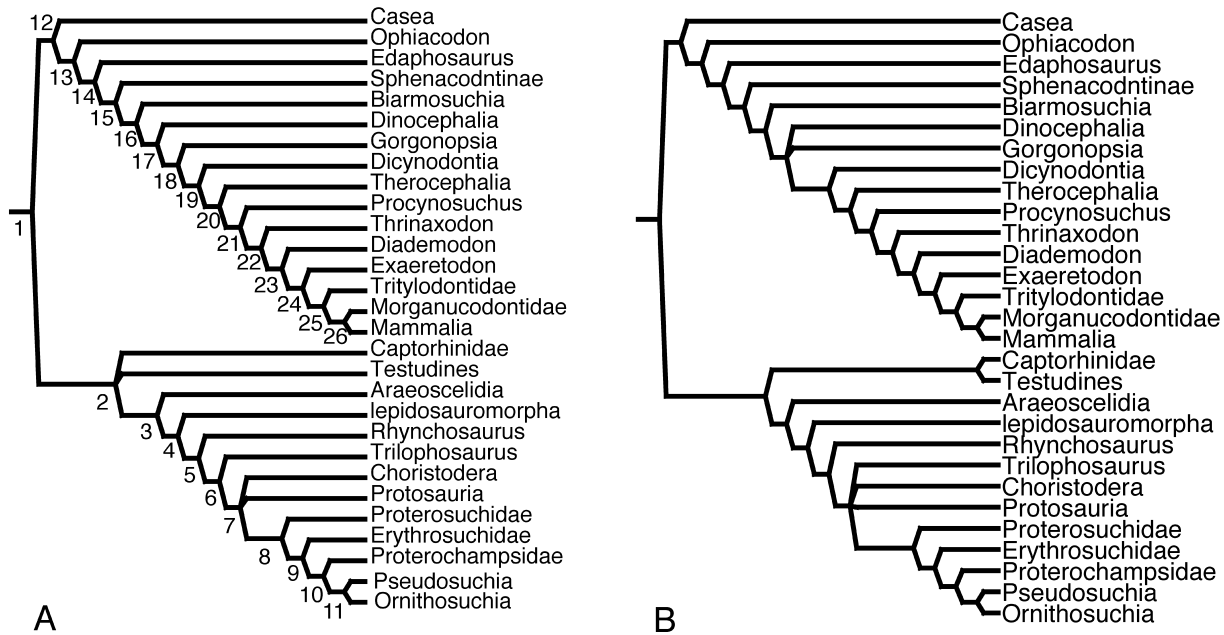
FIGURE 5. **A**, shortest calculated tree (from Gauthier et al., 1988). **B**, MERDA consensus of these data.

Sample Case 1: Troodontids—Small Fossil Data Set with a High Level of Missing Data

Norell et al. (2000) analyzed the relationships among troodontid theropod dinosaurs. This data set has 10 taxa, 38 characters and 24% missing data. The best tree (Fig. 1) indicated that *Saurornithoides* (both *mongoliensis* and *junior*) formed a group with the North American *Troodon* and that *Byronosaurus*

is the sister taxon to this clade. Other Central Asian troodontids and outgroup taxa are more distantly related.

These are extremely rare animals. Of the six ingroup troodontids, only two *Byronosaurus* and *Troodon* are known from more than one specimen. *Byronosaurus* is known from two specimens, but most of the postcrania are unknown. Troodontid indet. is fragmentary, and no articulated specimen of the North American *Troodon* has been recovered. This tree is somewhat at odds with trees presented by Xu et al. (2002) and Norell et al. (2001), in regards to taxa outside the Troodontidae (dromaeosaurs etc.) based on much larger data sets with both more characters and more taxa. In light of this, this is a heuristic example using a published dataset as opposed to an up-to-date evaluation of the relationships of these taxa.

This clade was subjected to 1,000 replications of a MERDA

TABLE 3. MERDA, Jacknife and Bremer and values from the data of Gauthier et al., 1989.

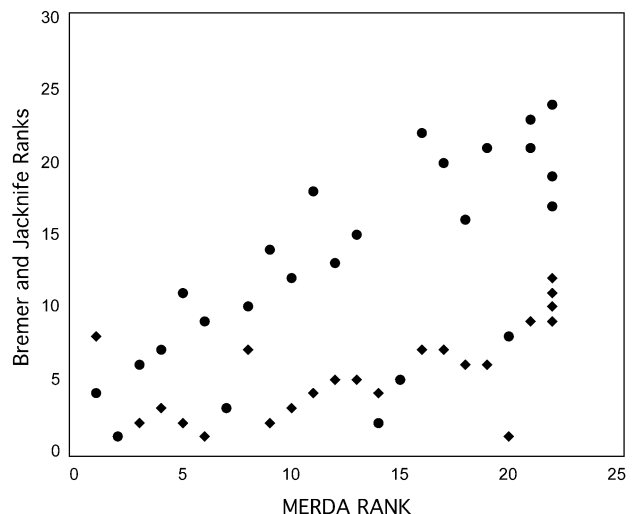| Clade	MERDA	Bremer	Jack-nife	MERDA rank	Bremer rank	Jack-nife rank
4	1.000	17	0.951	22	10	17
5	1.000	12	0.951	22	9	17
16	1.000	29	0.973	22	12	19
25	1.000	18	0.983	22	11	24
19	0.997	12	0.976	21	9	21
21	0.997	12	0.982	21	9	23
11	0.985	1	0.671	20	1	8
9	0.957	6	0.976	19	6	21
20	0.930	6	0.950	18	6	16
10	0.927	7	0.975	17	7	20
24	0.923	7	0.979	16	7	22
15	0.867	5	0.288	15	5	5
12	0.865	4	0.098	14	4	2
8	0.839	5	0.928	13	5	15
17	0.835	5	0.890	12	5	13
22	0.712	4	0.971	11	4	18
18	0.636	3	0.810	10	3	12
26	0.536	2	0.919	9	2	14
3	0.489	7	0.770	8	7	10
14	0.442	3	0.143	7	3	3
23	0.428	1	0.760	6	1	9
6	0.411	2	0.779	5	2	11
2	0.313	3	0.485	4	3	7
7	0.275	2	0.394	3	2	6
13	0.247	1	0.059	2	1	1
1	0.149	8	0.161	1	8	4
*	0.364					

FIGURE 6. Graphical representation between both Bremer support (diamonds) and Jacknife value (dots) and MERDA rank for results calculated from data in Gauthier et al. (1986).
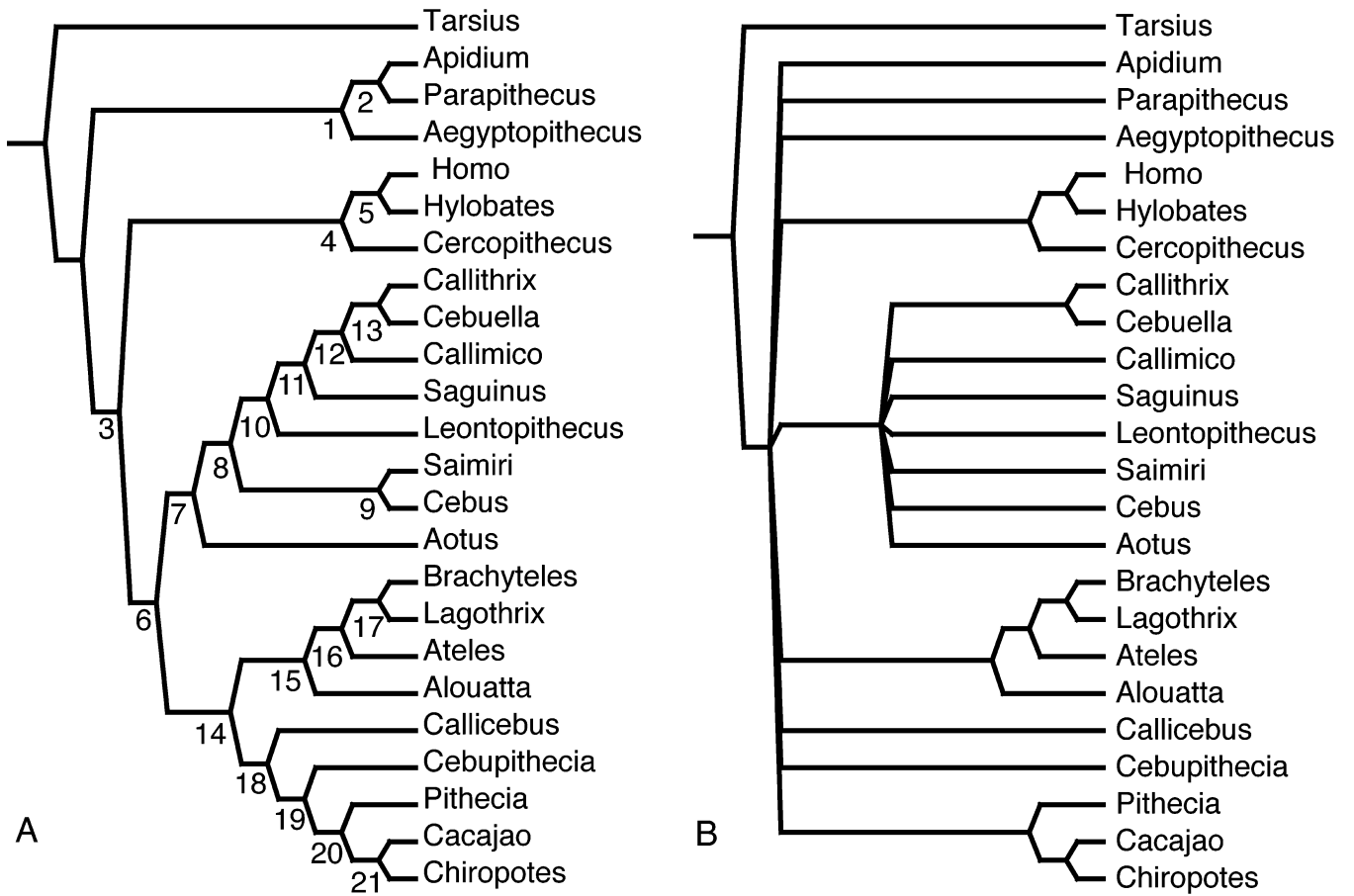
FIGURE 7. **A**, shortest calculated tree from data in Horovitz et al. (1998). **B**, MERDA consensus of these data.
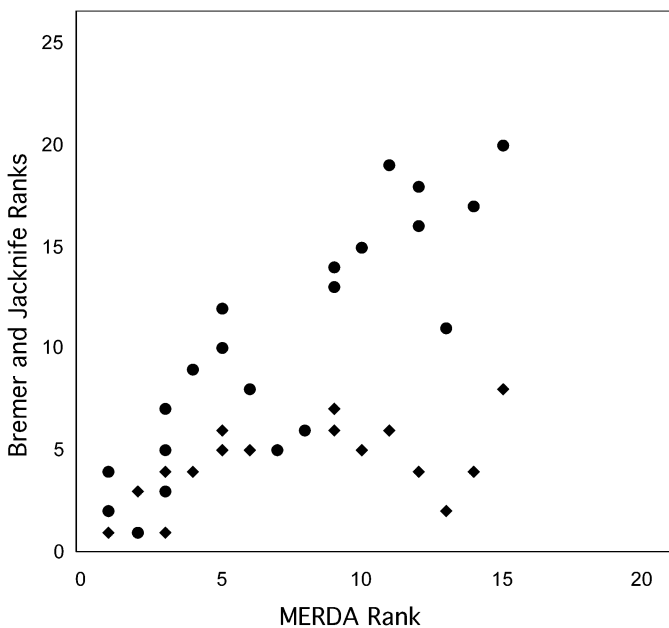

FIGURE 8. Graphical representation between both Bremer support (diamonds) and Jacknife value (dots) and MERDA rank for results calculated from data in Horovitz et al. (1998).

TABLE 4. MERDA, Jacknife and Bremer values from the data of Horovitz et al., 1998.

| Clade	MERDA	Bremer	Jack-nife	MERDA rank	Bremer rank	Jack-nife rank
13	0.708	12	0.624	15	8	20
5	0.646	4	0.615	14	4	17
20	0.583	2	0.484	13	2	11
4	0.542	4	0.595	12	4	16
21	0.542	4	0.618	12	4	18
16	0.438	6	0.621	11	6	19
15	0.417	5	0.590	10	5	15
9	0.375	7	0.563	9	7	13
11	0.375	6	0.584	9	6	14
10	0.292	6	0.419	8	6	6
12	0.208	5	0.401	7	5	5
7	0.188	5	0.443	6	5	8
6	0.167	5	0.478	5	5	10
17	0.167	6	0.549	5	6	12
19	0.125	4	0.464	4	4	9
14	0.042	4	0.263	3	4	3
2	0.042	1	0.401	3	1	5
18	0.042	4	0.425	3	4	7
8	0.021	3	0.071	2	3	1
1	0.000	1	0.174	1	1	2
3	0.000	2	0.304	1	2	4
*	0.479					
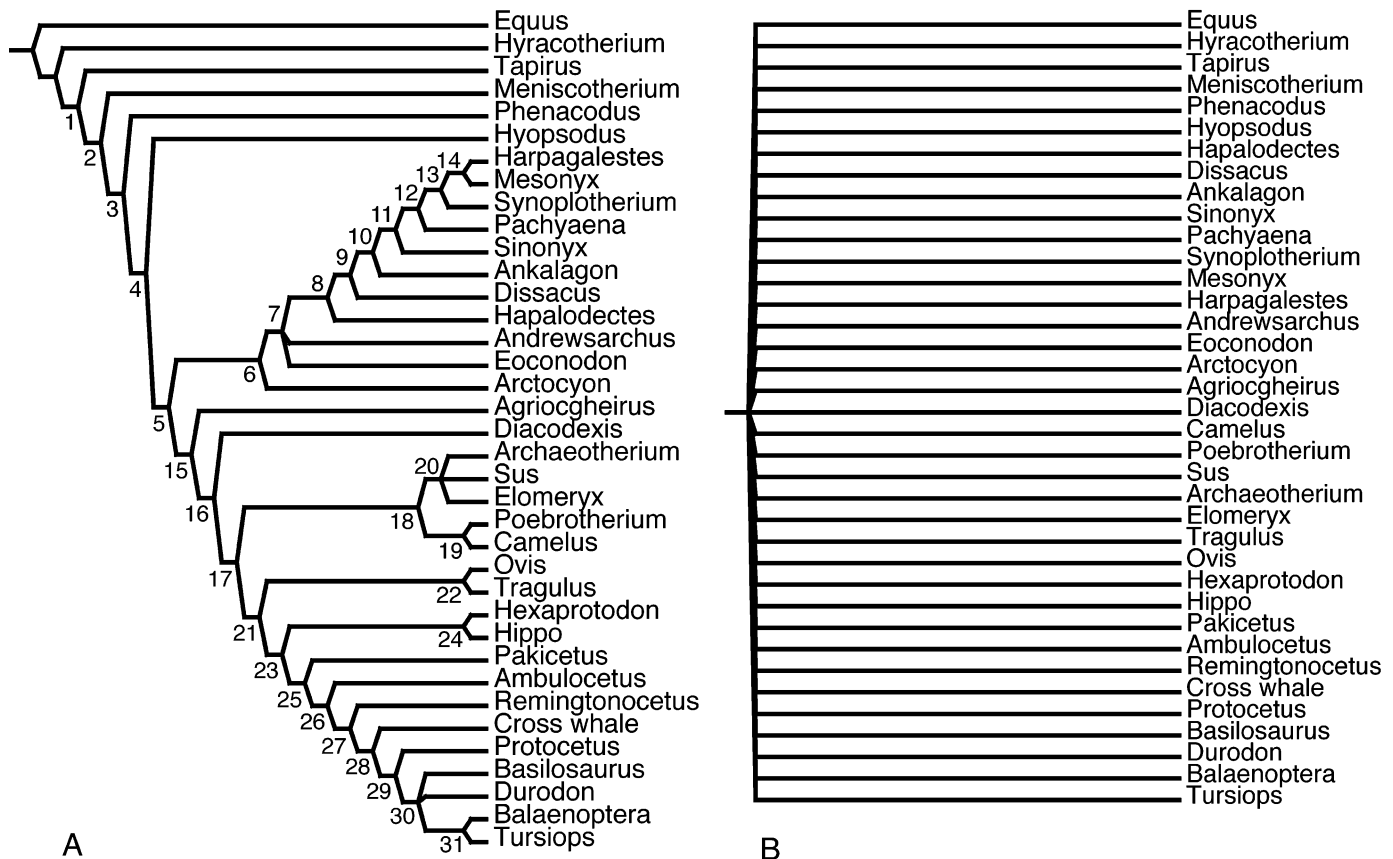
FIGURE 9. **A**, shortest calculated tree calculated from O'Leary (1999). Difference in topology between this tree and thee tree presented in O'Leary (1999) are due to run conditions where we did not order characters. **B**, MERDA consensus of these data.

analysis. Sixty-two different groups were found. Table 1 shows MERDA values, Bremer supports and parsimony Jacknife values for clades in Figure 1. To investigate the relationship between these indices, values were ranked and compared in Figure 2. As is evident from the graph, there is no real relationship between high Bremer, Jackknife, or MERDA values in this analysis.

In Table 1, the taxon with the asterisk is the 'clade' with the highest MERDA value that did not occur in the original tree as it has a higher MERDA value than clade 2 of the original analysis. This convention will be followed in all analyses. The tree in Figure 1 (right) is the MERDA consensus tree that accounts for this grouping. This tree differs from the original tree in two places where branches have been collapsed. Dromaeosaurs, oviraptors and avialans are collapsed as well as crown group troodontids. Both of these indicate that there were arrangements found at a higher frequency in the MERDA analysis than the groups of crown group troodontids or oviraptorosaurs + avialans, etc. found in the original data analysis.

Sample Case 2: Arthropods

The data set of Wheeler at al. (1993) combined morphological character data (100 characters) of 26 arthropod taxa with molecular data (897 aligned positions) on the 25 extant lineages. The single extinct lineage, Trilobita, was well characterized by the morphological data, but molecular data were (obviously) unavailable. Of the 25,922 potential observations, there were 1,311 missing (5.1%), 932 of which were found among the trilobite data.

When analyzed, this data set produced a single tree at length

1,037 steps (CI = 0.43; RI = 0.61) supporting a monophyletic Arthropoda, Arachnata, Mandibulata, Tracheata, Insecta, and Myriapoda. Additionally, the chelicerate portion of the tree was completely resolved, with support for Chelicerata, Euchelicerata, Arachnida, and Araneae (Fig. 3).

Upon subjection to MERDA analysis, a clade consisting of Trilobita + *Nephila* (a spider) appeared in 91% of the replicates. This sister taxon relationship would overturn the entire chelicerate clade. Although many of the major lineages (6 of 24) were found in every replicate (e.g., Arthropoda), the chelicerate lineages were not. Hence, several groups were collapsed including all the chelicerate and several insect groupings (Fig. 3, Table 2).

The rank correlation among Jacknife, Bremer, and MERDA values shows little relationship (Table 2, Fig. 4).

Sample Case 3: Amniotes—Large Fossil Data Set with Little Missing Data

One of the classic data sets that has been used to evaluate problems of missing data is the Gauthier et al. (1988) data set of amniote relationships. The Gauthier et al. (1988) phylogeny is reproduced in Figure 5, and displays a fairly conventional arrangement of a division between Synapsida and Reptilia, with subdivisions including Therapsida, Archosauria, Diapsida and Anapsida. This data set consists of 30 taxa and 316 characters. It has a relatively low amount of missing data (28%) and a relatively high RI (0.85).

The MERDA groupings are shown in Table 3. In this case the RI is predictive in that the MERDA consensus of this data set results in only two collapses—*Casea* + other synapsids and

TABLE 5. MERDA, Jacknife and Bremer values from the data of O'Leary 1999.

Clade	MERDA	Bremer	Jack-nife	MERDA rank	Bremer rank	Jack rank
1	0.000	1	0.000	1	1	1
2	0.000	6	0.000	1	6	1
3	0.000	3	0.000	1	3	1
4	0.000	6	0.000	1	6	1
5	0.000	4	0.000	1	4	1
6	0.000	2	0.000	1	2	1
7	0.000	1	0.000	1	1	1
8	0.000	5	0.000	1	5	1
9	0.000	3	0.000	1	3	1
10	0.000	2	0.000	1	2	1
11	0.000	1	0.000	1	1	1
12	0.000	1	0.000	1	1	1
13	0.000	1	0.000	1	1	1
14	0.032	1	0.000	2	1	1
15	0.000	2	0.000	1	2	1
16	0.000	1	0.000	1	1	1
17	0.000	1	0.000	1	1	1
18	0.000	1	0.000	1	1	1
19	0.000	5	0.534	1	5	9
20	0.000	1	0.000	1	1	1
21	0.000	1	0.000	1	1	1
22	1.000	7	0.913	3	7	12
23	0.000	1	0.060	1	1	2
24	0.000	3	0.895	1	3	11
25	0.000	5	0.533	1	5	8
26	0.000	3	0.425	1	3	6
27	0.000	2	0.401	1	2	4
28	0.000	3	0.417	1	3	5
29	0.000	1	0.284	1	1	3
30	0.000	3	0.494	1	3	7
31	1.000	3	0.836	3	3	10
*		1				

FIGURE 10. Graphical representation between both Bremer support (diamonds) and Jacknife value (clear dots) and MERDA rank for results calculated from data in O'Leary (1999).

turtles and captorhinids relative to other reptiles (Fig. 5). Both *Casea* and captorhinids are stem taxa. Several groups (Therapsida, Mammaliamorpha, Sauria and Archosauromorpha) have extremely high support and are found in all of the replicates. In this case obvious correlation exists between Bremer, Jacknife and Merda values (Fig. 6).

Sample Case 4: Platyrrhines—Total Evidence Analysis of Fossils and Extant Taxa

Horovitz et al. (1998) presented an analysis of living and fossil New World monkeys. This analysis was based on 76 morphological characters, 142 16S rDNA, 951 12S rDNA, 261 ε-globin, and 332 IRBP informative sites for a total of 1,762 informative characters coded for 24 taxa. Of the 127,656 original observations (morphological characters + aligned sequence positions) 29,371 were missing (23%). The most parsimonious reconstruction of these data yields a single tree at length 4473 (CI = 0.51; RI = 0.58; Fig. 7).

When subjected to MERDA replacement, all but five groups have lower recovery frequencies than a novel grouping of (*Cacajao* + *Chiropotes* + *Pithecia* + *Callicebus*) at 48%. The subsequent collapse of lower recovery branches yields a dramatically less resolved depiction of relationships (Figs. 7, 8, Table 4).

Sample Case 5: Artiodactyls—A Large Total Evidence Analysis with Extensive Missing Data

Recently, O'Leary (1999) published a total evidence data set for several artiodactyl and cetacean taxa (37 taxa by 2,780 characters). In addition to the morphological characters that could be assayed in all taxa where relevant aspects of anatomy are preserved, molecular characters from a variety of genes and
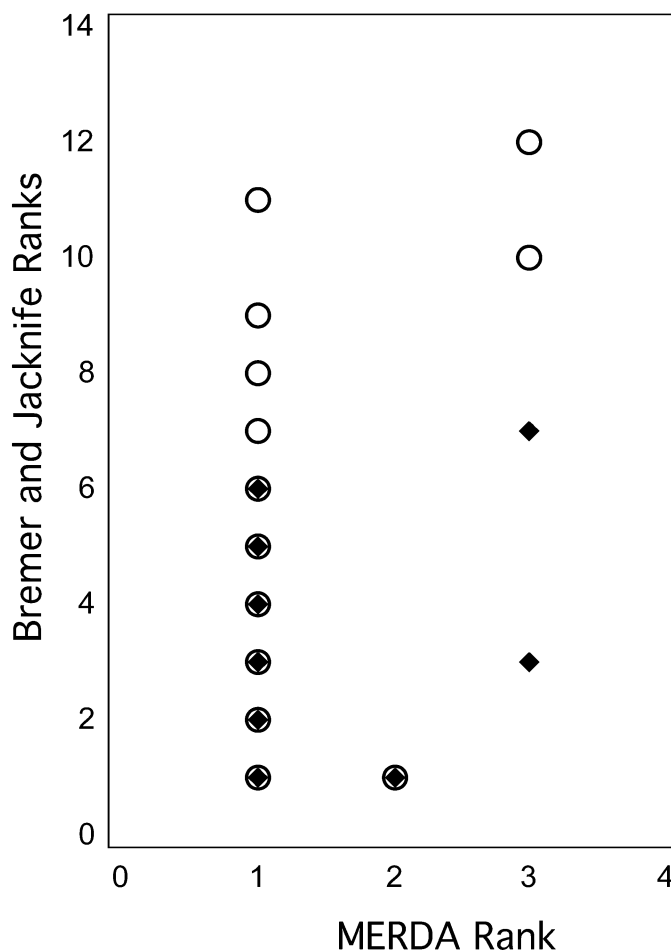
transposons and a few soft tissue characters were also considered. Of the 102,860 potential observations, 74,862 were missing (72.8%). This high number comes from the fact that 27 of the taxa were present only as fossils. We reanalyzed these data provided by O'Leary. This resulted in 2 topologies of 2,793 steps (Fig. 9). Presumably because we did not order any multistate characters our topology differs somewhat from that presented by O'Leary (2001:fig. 6A).

When subjected to MERDA analysis, this topology is highly unstable. Novel groupings are found immediately and many original groupings have extremely low MERDA values (Table 5). Only eight of the 5,565 groups found occurred in all of the replicates, and only two of these were in the original cladogram (Table 1). This results in a completely unresolved tree (Fig. 9). The collapse of branches is due to the high frequency of recovery of groups that did not contain *Camelus*, *Sus*, and *Ovis*. Notably all these are extant taxa with dense character sampling. Consequently, this collapse effect is due to large amounts of randomly replaced missing variables interacting with homoplasy to create new novel groupings that exclude the relatively stable densely sampled taxa. Figure 10 shows the relationship between MERDA values and Bremer and Jacknife indices, again showing the typical condition of no relationship among these values.

DISCUSSION AND CONCLUSIONS

Problems associated with missing data are fundamental aspects of any total evidence analysis that incorporates fossils. Previously, this problem has been dealt with either by excluding taxa with copious amounts of missing data and/or using procedures that basically assume that the missing data are congruent with actual observations. But what if they are not? It is this question that we have attempted to address here.

In creating the MERDA metric, we have not attempted to create the final word on cladogram stability, but to define a tool to help assay the interaction effects of homoplasy and missing data on clade support. A descriptor to note the contribution of the unique properties of missing observations on cladogram resolution. Since the replacement values are random, and random noise has been shown to have almost no affect on cladistic results (Whiting and Siddall, 1999), the disrupting influence of missing data must come from their compliant nature. When the missing data are replaced with fixed values, they can no longer conform to every cladogram upon which they are optimized. It is this property we are attempting to measure. Any group that disappears with replaced values must to some degree be dependent on that one behavior of missing data no observation can have. Given the lack of necessary correlation with other measures of nodal support (e.g., Bremer values, Jacknife levels), this approach does measure another dimension of cladogram resolution.

Our approach has been to see the effect of random observation addition and to assay how well the assumption of non-incongruence performs relative to a sample of random results. Although more cases need to be examined before any general properties can be determined, it is apparent that randomly incongruent patterns of missing data can often lead to phylogenetic conclusions which are at variance with the original (missing) data. Since the randomly replaced data have no phylogenetic signal, this suggests that these resolutions are dependent on the specific behavior of missing data. These multifarious entities behave unlike any actual observation, and clades dependent on this property are unlikely to remain long.

ACKNOWLEDGMENTS

LITERATURE CITED

Adams, E. N. 1972. Consensus techniques and the comparison of taxonomic trees. Systematic Zoology 34:46–58.

Grande, L., and W. E. Bemis. 1998. A comprehensive phylogenetic study of amiid fishes (Amiidae) based on comparative skeletal anatomy. An empirical search for interconnected patterns of natural history. Society of Vertebrate Paleontology Memoir 4:1–690.

Farris, J. S., V. A. Albert, M. Lallersjo, D. Lipscomb, and A. Kluge. 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12:99–124.

Gao, K., and M. A. Norell. 1998. Taxonomic revision of *Carusia* (Reptilia: Squamata) from the Late Cretaceous of the Gobi Desert and phylogenetic relationships of anguimorphan lizards. American Museum Novitates 3230:1–51.

Gauthier, J. A. 1986. Saurischian monophyly and the origin of birds. California Academy of Sciences Memoir 8:1–55.

———, A. G. Kluge, and T. Rowe. 1988. Amniotes and the importance of fossils. Cladistics 4:105–209.

Horovitz, I., R. Zardoya, and A. Meyer. 1998. Platyrrhine systematics: a simultaneous analysis of molecular and morphological data. American Journal of Physical Anthropology 106:261–281.

Makovicky, P. J. 2000. Effects of missing data on support measures and weighted analyses. Journal of Vertebrate Paleontology 20(3, suppl.):56A.

Nixon, K., and Q. Wheeler. 1992. Extinction and the origin of species; pp. 119–143 *in* M. J. Novacek and Q. D. Wheeler (eds.), Extinction and Phylogeny. Columbia University Press, New York.

Norell, M. A., P. Makovicky, and J. M. Clark. 2000. A new troodontid from Ukhaa Tolgod, Late Cretaceous, Mongolia. Journal of Vertebrate Paleontology 20:7–11.

———, J. M. Clark, and P. J. Makovicky. 2001. Relationships among Maniraptora: problems and prospects; pp. 49–67 *in* J. Gauthier and L. F. Gall, (eds.), New Perspectives on the Origin and Early Evolution of Birds: Proceedings of the International Symposium in Honor of John H. Ostrom. Peabody Museum of Natural History, Yale University, New Haven, Connecticut.

Novacek, M. J. 1992. Fossils as critical data or phylogeny; pp. 46–88 *in* M. J. Novacek and Q. D. Wheeler (eds.), Extinction and Phylogeny. Columbia University Press, New York.

O'Leary, M. A. 2000. Operational obstacles to total evidence analyses considering that 99% of life is extinct. Journal of Vertebrate Paleontology 20(3, suppl.):61A.

O'Leary, M. A. 1999. Parsimony analysis of total evidence from extinct and extant taxa and the cetacean-artiodactyl question (Mammalia, Ungulata). Cladistics 15:315–330.

Platnick, N. I., C. E. Griswold, and J. A. Coddington. 1991. On missing entries in cladistic analysis. Cladistics 7:337–343.

Siddall, M. E., and M. F. Whiting. 1999. Long branch abstractions. Cladistics 15:9–24.

Wheeler, W. C., P. Cartwright, and C. Y. Hayashi. 1993. Arthropod phylogeny: a combined approach. Cladistics 9:1–39.

Wilkinson, M. 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Systematic Biology 44:501–514.

Xu, X., M. A. Norell, X.-L. Wang, P. J. Makovicky, and X.-C. Wu. 2002. A basal troodontid from the early Cretaceous of China. Nature 415:780–784.

APPENDIX 1

Script for running MERDA in POY.

```
for DATA in x
do
for MORE in 1 2 3 4 5 6 7 8 9 10
do
for OTHER in 1 2 3 4
do
for REP in 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
do
norell < $DATA.hen > $DATA.tmp
poy -parallel $DATA.tmp -maxtrees 50 -multibuild 4 -norandomizeoutgroup
-fitchtrees -multiratchet -ratchettbr 20 -ratchetseverity 3
-ratchetpercent 33 > $DATA.out
jack2top 100 < $DATA.out >> $DATA'1000'.out
echo $MORE $OTHER $REP
done
done
done
cat semicolon.file ≫ $DATA'1000'.out
groupnumsf < $DATA'1000'.out > $DATA'1000'.gps
done
```