An internet publication of the Sequence Monkey Anti-Dogma League, Ann Arbor, Michigan

The Naïve Verificationist

An international outlet for controversial papers on the subject of molecular phylogenetics

Sometimes Too Much is Not Enough: A Comment on the Use of mtDNA to Recover Ancient Divergences

by

Daniel L. Graf & John S. Sparks

University of Michigan, Museum of Zoology, Ann Arbor, Michigan 48109;
dgraf@umich.edu, jsparks@umich.edu

Volume I **17 July 2000** **Number 1**

Visit the Sequence Monkey Home Page — Got Monkey? http://www.members.tripod.com/sequence_monkey

D.L. Graf & J.S. Sparks. *Naïve Verificationist* 1(1): 1-13, July 2000.

Sometimes Too Much is Not Enough: A Comment on the Use of mtDNA to Recover Ancient Divergences

BY DANIEL L. GRAF AND JOHN S. SPARKS
Museum of Zoology, University of Michigan, Ann Arbor, Michigan 48109;
dgraf@umich.edu, jsparks@umich.edu

Authors' Note: This paper was intended as a 'Point of View' article in response to a specific symposium published in Systematic Biology (1999: volume 48, number 1): Symposium on the Origin of Mammalian Orders. *It, however, has wider implications and is actually directed towards a whole school of phylogenetics. The original manuscript was submitted to and rejected from that journal.*

The nature of mitochondrial genes— *i.e.*, single copies, lack of introns, maternal inheritance (but see Hoeh *et al.*, 1996), availability of 'universal' PCR primers (Palumbi, 1996)— makes mtDNA ideal for recovering phylogeny among Eukaryota. However, there are also drawbacks to the use of mtDNA sequences in molecular phylogenetic studies. Specifically, the high nucleotide substitution rate (Naylor and Brown, 1998) is such that the phylogenetic utility of mitochondrial genes decreases rapidly as divergence time among taxa increases. For more ancient splits, this can lead to 'leafy' trees (*i.e.*, those where the terminal branches are much longer than the internal branches; Salisbury, 1999) and long-branch attraction (Felsenstein, 1978,1981).

Alternatives that may be implemented to improve phylogenies based upon such 'leafy' reconstructions boil down to (1) making the terminal branches shorter or (2) making the internal branches longer. Although the merits of these alternatives are obvious to most molecular systematists, we feel that many researchers that advocate the use of mtDNA sequences for recovering ancient divergences have made less-than-optimal choices when confronted with data sets prone to long-branch problems. In this paper, we briefly review and comment on the performance of mitochondrial and nuclear genes for recovering late Mesozoic divergences in light of several recent analyses using only mtDNA for similarly deep splits. It is our hope that the examples presented here (and the numerous references pointed to) will both emphasize the limitations of adopting a mitochondria-only approach and encourage the exploration of nuclear protein-coding and ribosomal DNA when appropriate.

The mitochondria-only approach is highlighted in the results of the "Symposium on the Origin of Mammalian Orders" published in a recent issue of *Systematic Biology* (1999: volume 48, number 1). While some effort was made to include data derived from nuclear genes (*e.g.*, Springer *et al.*, 1999), the apparent *modus operandi* of other molecular systematists of 'higher' vertebrates is to address the interordinal relationships of mammals and birds using entire, or major portions of, mitochondrial genomes (*e.g.*, Waddell *et al.*, 1999b, 1999c; Penny *et al.*, 1999; Mindell *et al.*, 1999; see Curole and Kocher, 1999: Fig. 1). Previous studies of this type applying the same general methodology, far too numerous to list, are cited in these references and can also be found in the bibliography to Hillis *et al.* (1996) and in Curole and Kocher (1999).

We advocate alternative #1 for dealing with 'leafy' phylogenies: shortening the terminal branches, and thereby reducing the number of misinformative characters (see below). As a focal point for discussion, we present our reanalyses of two previously published phylogenies. Firstly, the phylogeny of freshwater pearly mussel families (Mollusca: Bivalvia: Unionoida) recovered by Graf and Ó Foighil (2000, in review), and second, that of the placental mammal orders, specifically the relationship of Primates to Glires, presented by Waddell *et al.* (1999a: cover illustration). In contrast to previous opinions (Simpson, 1944), mammalian orders and freshwater mussel families represent analogous systems for the purpose of this study—modern taxa have proliferated throughout the Cenozoic but had their origins in the Cretaceous (Haas, 1969; Alroy, 1999). Each taxon set was initially analyzed using the Folmer *et al.* (1994) fragment of cytochrome *c* oxidase subunit I. For the mussels, we attempted to improve the 'leafy' phylogeny by substituting a slower, nuclear rDNA fragment to shorten terminal

branches, whereas for the mammals we tried to lengthen the internal branches by adding more mtDNA characters to the data matrix (alternative #2: lengthening the internal branches and hopefully character support). Our results indicate that: (1) a single mitochondrial gene fragment is insufficient to reliably recover either the interordinal relationships of mammals or the interfamilial relationships of mussels as a result of weak, 'leafy' topologies; and (2) the 'stemminess' of these topologies can be improved by employing a more evolutionarily-conservative data set but not by simply adding more similarly-homoplastic mitochondrial sequence data.

METHODS AND MATERIALS

Characters and Taxon Sets.— Two, nine-taxon sets were analyzed for this study: one comprised of freshwater mussels of the bivalve order Unionoida, the other made up of placental mammals. The unionoid taxa were chosen as representatives of the major clades studied by Graf and Ó Foighil (2000, in review). Mammalian taxa were chosen to represent Rodentia, Lagomorpha, Primates, Eulipotyphyta, and Cetartiodactyla (the latter as an outgroup) and from among those species for which the entire mtDNA sequence is available. The taxon sets utilized for this study are listed in Tables 1 and 2.

A total of four character sets were analyzed, two for each of the mammal and the freshwater mussel taxon sets. For each taxon set, a matrix comprised of the 650 nt Folmer *et al.* (1994) fragment of cytochrome *c* oxidase subunit I (COI) was assembled. It was presumed— based on previous studies, *i.e.*, Graf and Ó Foighil (2000), Waddell *et al.* (1999b), Sparks (unpubl. data)— that the shapes of the phylogenies, with respect to the clades of interest, derived from these character sets would be highly 'leafy.' That is, the internal branches would be short relative to the terminals. In order to correct for this, a second character set was compiled for each group of taxa in an attempt to recover a more 'stemmy' phylogeny. The 'corrected' data sets differ for the two taxon sets and correspond to the two different alternatives given above for dealing with 'leafy' topologies.

For the second mammal data set, the objective was to lengthen (*i.e.*, add more characters) to the internal branches. To accomplish this, the complete mtDNA sequences of cytochrome *c* oxidase subunit III (COIII) and cytochrome *b* (cyt b) were added to the partial COI sequence for a total of 2580 nt. For the mussels, the objective was to shorten the terminal branches. This was done by abandoning COI in favor of a 450 nt fragment of a more slowly evolving (*i.e.*, more conservative) nuclear rDNA gene: domain 2 of 28S (Park and Ó Foighil, 2000). All sequences were obtained through GenBank (National Center for Biotechnology Information, National Institutes of Health; http://www.ncbi.nlm.nih.gov) (Tables 1 and 2). Multisequence alignments were compiled and manipulated using Sequence Monkey 2.8.0 (available from http://www.members.tripod.com/sequence_monkey) and Clustal_X (Thompson *et al.*, 1994, 1997; available from http://ncbi.nlm.nih.gov) and refined manually where necessary.

Ideally, a single taxon set would be used to illustrate these different methodologies. That was not possible due, we think, to the differences in philosophies of the groups collecting the data— those molecular systematists using nuclear rDNA would see no value in sequencing several Kb of mtDNA, and vice versa. The result has been very little overlap among the taxon and character sets available in GenBank. Since the purpose of these data are purely as examples of appropriate vs. inappropriate phylogenetic methodologies, our results are still useful for evaluating the performance of these data sets with these taxa. These analyses provide a focal point for discussion and are intended as representative of many more comprehensive studies recently undertaken (numerous references cited throughout this paper).

Phylogenetic Analyses.— Phylogenetic analyses for all four data matrices (*i.e.*, mussel COI, mussel 28S, mammal COI, mammal COI + COIII + cyt b) were identical. Parsimony analyses were conducted using PAUP* 4.0b3 (Swofford, 1998). All parsimony tree-searches were exhaustive. Maximum likelihood (ML) was also applied using PAUP* [heuristic searches, 10

random sequence additions, HKY (Hasegawa *et al.*, 1985) with rate heterogeneity (8 rate categories)], and model parameters (ti/tv ratio, proportion of invariant sites, Γ-curve shape for rate heterogeneity) were estimated from the 'preferred' tree for each taxon set (Table 3). For the mussels, the preferred topology was the one recovered by analysis of the COI and 28S matrices; for the mammals, the preferred topology was that of Waddell *et al.* (1999a: cover illustration) (choice of the preferred trees is explained below). In all analyses, alignment gaps were defined as missing data.

Table 1. — GenBank Accession Numbers of the Freshwater Mussel Taxa Used. COI and 28S sequences were published previously in Graf and Ó Foighil (2000, in review).

TAXON	COI	28S
UNIONIDAE		
UNIONINAE		
UNIONINI		
Unio (Unio) pictorum	AF156499	N/A
Unio (Cafferia) caffer	AF156500	N/A
ANODONTINI		
Pyganodon grandis	AF156504	N/A
Lasmigona compressa	AF156503	N/A
AMBLEMINAE		
Lampsilis cardium	AF156518	N/A
Amblema plicata	AF156512	N/A
MARGARITIFERIDAE		
Cumberlandia monodonta	AF156497	N/A
HYRIIDAE		
Hyridella depressa	AF156496	N/A
Velesunio ambigua	N/A	N/A

Table 2. — Mammal classification and GenBank Accession Numbers. Classification follows that of Waddell *et al.* (1999a). GenBank accessions refer to whole mitochondrial sequences from which cytochrome c oxidase subunits I (partial) and III and cytochrome *b* were excised.

TAXON	GenBank #	Reference
LAURASIATHERIA		
EULIPOTYPHLA		
Erinaceus europaeus	X88898	Krettek *et al.* (1995)
CETARRTIODACTYLA		
Bos taurus	J01394	Anderson *et al.* (1982)
Balaenoptera musculus	X72204	Arnason and Gullberg (1993)
PRIMATES		
Hylobates lar	NC002082	Arnason *et al.* (1996)
Homo sapiens	J01415	Anderson *et al.* (1981)
GLIRES		
LAGOMORPHA		
Oryctolagus cuniculus	AJ001588	Gissi *et al.* (1998)
RODENTIA		
Cavia porcellus	NC 000884	D'Erchia *et al.* (1996)
MURIDAE		
Mus musculus	J01420	Bibb *et al.* (1981)
Rattus norvegicus	X14848	Gadaleta *et al.* (1989

Table 3. — Model Parameters Used for Maximum Likelihood Searches. For the mussel COI and mussel 28S data sets, the estimated parameters — ti/tv ratio, proportion of invariant sites, and gamma distribution shape (Γ) for variable rates— were estimated from the cladogram recovered from phylogenetic analysis of those data sets (mussel topology I, Table 4). For the mammal COI and COI + COIII + Cyt b matrices, the estimated parameters were based on the tree of Waddell *et al.* (1999a) (mammal topology VI, Table 4). In all cases, the Hasegawa *et al.* (1985) model (HKY) with rate heterogeneity was applied, nucleotide frequencies were based on empirical values, the number of Γ rate categories = 8, and all other parameters matched the PAUP* factory settings.

DATA SET	ti/tv	invariant sites	Γ
Unionoidea COI	5.75686	0.582549	0.870669
Unionoidea 28S	1.18695	0.419530	3.596530
Mammalia COI	5.320006	0.000000	0.135862
Mammalia COI + COIII + Cyt b	2.763990	0.384557	0.651337

In order to evaluate the 'robustness' of the hypotheses being addressed, jackknife resampling (%JK) (1000 replicates, heuristic searches of 10 random sequence additions each) was conducted using PAUP* and Bremer-Decay Index (BDI) values (Bremer, 1995) were calculated using TreeRot v.2 (available from http://mightyduck.bu.edu/TreeRot). A third metric that we will refer to as 'percent perfect characters' (%PC) was also calculated; it denotes the percentage of unambiguous, unreversed characters (CI = 1.00) of all characters supporting a node. The number of misinformative characters (upper bound) was estimated as the number of characters tracing to terminal branches with a CI < 1.00. Unambiguous and misinformative characters were traced using the "Describe Trees: List of Apomorphies" option of PAUP*. The 'stemminess' of a particular clade was quantified as the mean of the internal branch to terminal branch ratio of all the taxa in a clade (after Fiala and Sokal, 1985; as modified by Salisbury, 1999), where the terminal branch length is the sum of all branches between the node in question and the tip. Theoretically, 'stemminess' can vary from 0 (internal branch length = 0) to infinity (terminal branch length = 0). In all cases, branch lengths were assigned by PAUP* under the accelerated-transformation model (ACCTRAN).

RESULTS

The trees recovered by the parsimony analyses of the mussel COI and mussel 28S data sets are shown in Figure 1. The phylogenetic relationships suggested by these two cladograms are identical and were the same as those recovered by ML (Table 3-4). This topology (mussel topology I, Table 4) supports both monophyly of Unioninae (= (Unionini, Anodontini)), and Unionidae (= (Unioninae, Ambleminae)). This hypothesis differs from the topology of Graf and Ó Foighil (2000) (mussel topology II, Table 4) which placed Unionini as sister to an (Anodontini, Ambleminae) clade. The mussel COI phylogeny is 'leafy' (Figure 1A), with generally low branch support for Unioninae and Unionidae; this is evident in the low 'stemminess' values for those clades (Table 5). The mussel 28S phylogeny, however, is 'stemmy,' with 'stemminess' improved by an order of magnitude for both clades compared to the COI phylogeny. This increase in 'stemminess' is reflected in the improved branch support for the respective nodes (Table 5).

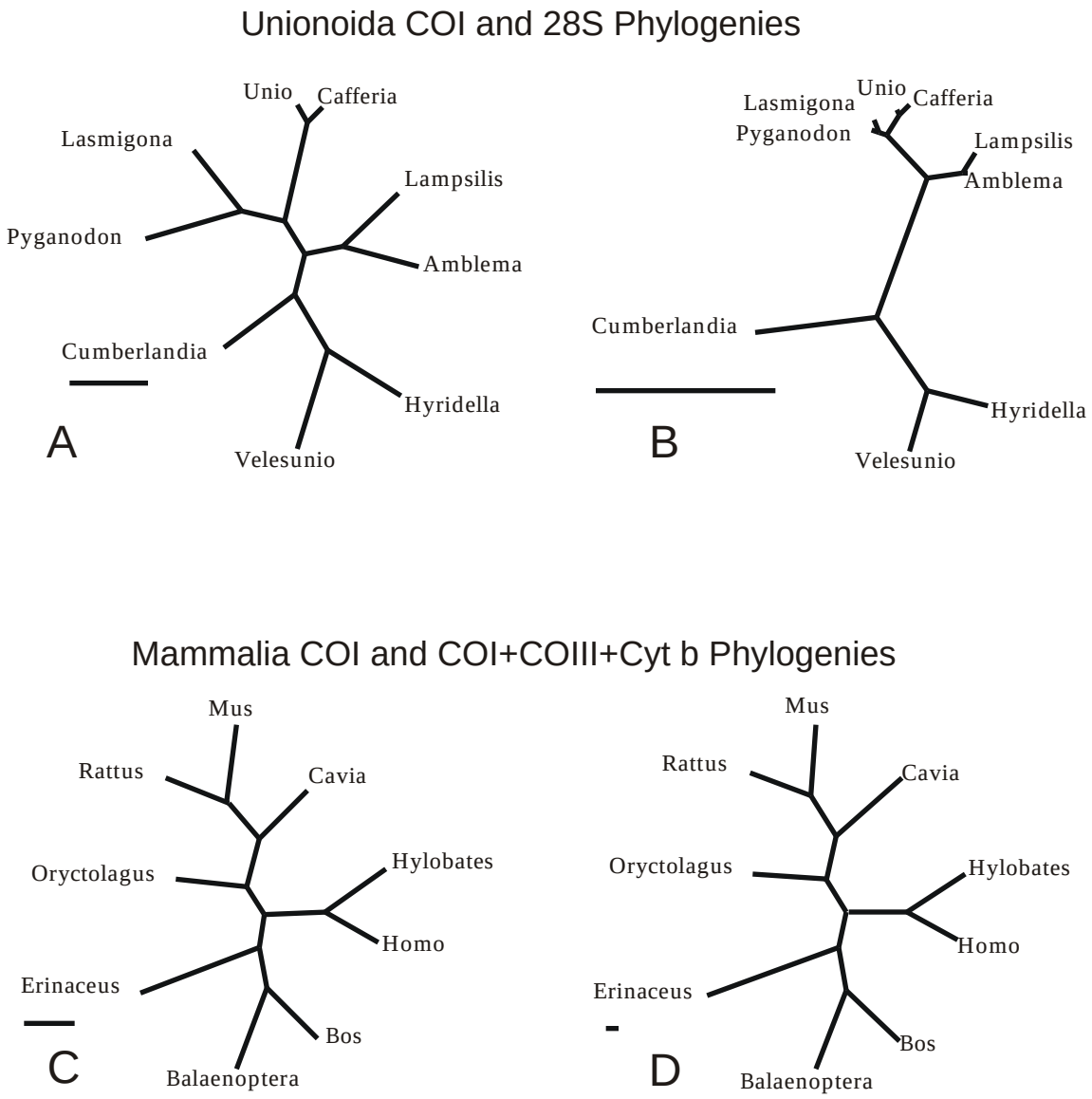

Figure 1. — Unrooted phylograms of mussel and mammal relationships. Phylograms A and B depict the parsimonious trees for the mussel COI and 28S data sets, respectively. C and D show the COI and combined mammal phylogenies. Branch lengths are proportional to the number of steps traced along them; scale bars show the length of 40 steps. See Table 6 for tree statistics.

Figure 1 also shows the mammal COI and mammal combined character sets traced on the "best estimate" tree of Waddell *et al.* (1999a) (mammal topology V of Table 4). Four different mammal topologies were recovered by the four analyses (COI and combined, each with parsimony and ML), and these correspond to mammal topologies I-IV in Table 4. None of the recovered topologies support monophyly of Rodentia. To recognize either a monophyletic Glires or (Primates, Glires) clade, those groups would also have to include *Erinaceus*. Since it was the only topology considered with a monophyletic Glires or (Primates, Glires) clade, the Waddell *et al.* (1999a) phylogeny was chosen (from among the five topologies in Table 4) to estimate the 'stemminess' of those clades. Each of the alternative mammal phylogenies is similar in terms of their degree of 'stemminess.'

The mammal COI phylogeny (Figure 1C), like the mussel COI phylogeny (Figure 1A), is 'leafy.' In addition, there is no branch support for either Glires or (Primates, Glires) (Table 5). This is not surprising, however, since neither of these hypotheses is supported by any analysis of the mammal COI character set (Table 4). The low 'stemminess' values for Glires and (Primates, Glires) clades on the mammal COI phylogeny (and low branch support) are not improved by the addition of more mtDNA characters (Table 5).

DISCUSSION

Both the mammal and mussel COI phylogenies are similarly 'leafy' (Figure 1). The inappropriateness of a 'leafy' tree— one with long terminal branches— for accurately recovering phylogeny should be obvious (see recent review and simulations by Salisbury, 1999). For any clade, informative characters are the synapomorphies (derived homologies) on the internal branch leading to it. Autapomorphic characters on the 'True' tree are at best uninformative for recovering phylogeny, and in all other cases they are misleading. Among the 'leafy' phylogenies considered (*i.e.*, mussel COI, mammal COI, and mammal combined), no less that two-thirds of the parsimony steps occur on the terminal branches. Of these terminal steps, the overwhelming majority are misinformative (Table 6). We consider this to be root of all long-branch evils.

In the case of the mammal phylogenies, these misinformative characters completely overshadow the informative characters on the internal branches. The more change that has occurred within any clade, the more the value of the characters along its supporting internal branches are eroded by increasing levels of homoplasy (Lanyon, 1988; Naylor and Brown, 1998). This is evidenced here by the inability of either mammalian character set to recover the phylogeny considered the "best estimate of placental relationships based on... extensive analyses of published and unpublished data" (Waddell *et al.*, 1999a). This 'leafiness' is also reflected in the low CI, low branch support, and low %PC. However, our analyses also indicate that it is not impossible to generate well-supported phylogenies for the same assemblages of taxa.

In the case of the freshwater mussels, the COI and 28S character sets support the same phylogeny (Figure 1; mussel topology I, Table 4), but this should not be taken as an endorsement for the value of COI in recovering family-level relationships within Unionoida. If a single terminal taxon is removed from each of the clades represented in the COI analysis (for example, remove *Amblema*, *Cafferia*, *Lasmigona*, and *Velesunio*), the phylogeny changes and supports neither a monophyletic Unioninae nor Unionidae [*i.e.*, (*Cumberlandia*, ((*Unio*, *Hyridella*), (*Pyganodon*, *Lampsilis*)))]. Similarly, if taxa are added, as was the case in the analysis of the same gene fragment by Graf and Ó Foighil (2000), Unionini again moves— this time to a position sister to (Anodontini, Ambleminae) (Graf and Ó Foighil, 2000: their figures 2 and 3; mussel topology II, Table 4). While the mussel COI phylogeny is sensitive to both the deletion and addition of taxa, the 28S phylogeny is well-supported, 'stemmy,' and is relatively insensitive to the addition or deletion of taxa from the clades in question (analyses not shown). We consider the 'stemminess' of the more-conservative 28S data set to be an indication of how much data has been eroded from the internal branches of the COI phylogeny due to much higher levels of homoplasy.

Many (if not all) of the whole-mitochondria phylogenetic studies suffer from poor taxon sampling. This is not surprising given the difficulties and expense associated with sequencing the entire mitochondrial genome for even a single organism. It has been suggested by Graybeal (1998) and Naylor and Brown (1998) that adding taxa at the expense of additional characters can significantly increase phylogenetic accuracy. Although this is an on-going discussion (*e.g.*, Hillis, 1998), we find it interesting that this approach appears to have been overlooked as an alternative to substantial additional data acquisition for the studies in question. We do acknowledge that comparison of our results with those of Graf and Ó Foighil (2000)

Table 4. — Alternative Tree Topologies for Unionoida and Mammalia. Unionoida topologies are rooted between Hyriidae and Unionoidea (= Margaritiferidae, Unionidae) following Graf (2000). Mammalia topologies are rooted between Laurasiatheria and (Primates, Glires) following Waddell *et al.* (1999a).

DATASET	ANALYSIS	UNIONOIDA TOPOLOGY
COI, 28S	MP, ML	I. (Hyriidae, (Margaritiferidae, (Ambleminae, Unioninae))).
Graf & Ó Foighil (2000)		II. (Hyriidae, (Margaritiferidae, (Unionini, (Anodontini, Ambleminae)))).

DATASET	ANALYSIS	MAMMALIA TOPOLOGY
Waddell *et al.* (1999a)		I. ((Cetartiodactyla, *Erinaceus*), (Primates, (Lagomorpha, Rodentia))).
COI	MP	II. (Cetartiodactyla, (Primates, (*Rattus*, (*Mus*, (*Erinaceus*, (*Cavia*, *Oryctolagus*))))).
COI	ML	III. (Cetartiodactyla, ((*Oryctolagus*, *Cavia*), (Primates, (*Erinaceus*, Muridae)))).
Combo	MP	IV. (Cetartiodactyla, (Primates, (Muridae, (*Cavia*, (*Oryctolagus*, *Erinaceus*).
Combo	ML	V. (Cetartiodactyla, (*Cavia*, (*Erinaceus*, *Oryctolagus*), (Primates, Muridae))).

Table 5. — 'Stemminess' and branch support statistics for Unionidae and (Primates, Glires). See text for explanation of the various metrics. The values given parenthetically with the BDI are the percentage of the total tree length each BDI represents. Total tree lengths are given in Table 6.

Clade	Data Set	Stemminess	BDI	%JK	%PC
Unioninae	COI	0.295	2 (0%)	60	37
	28S	2.410	3 (2%)	88	77
Unionidae	COI	0.285	6 (1%)	86	24
	28S	2.184	20 (13%)	100	94
Glires	COI	0.297	<0	<50	4
	Combo	0.310	<0	<50	3
(Glires, Primates)	COI	0.231	<0	<50	12
	Combo	0.251	<0	<50	10

demonstrates that adding taxa can break up long branches within a clade (*i.e.*, shortening the terminal branches and reducing the number of misinformative characters). However, the addition does nothing to improve the 'stemminess' of the clade or the quality of characters supporting it.

With regard to the evolution of Rodentia, Glires, and (Primates, Glires), the inconsistency among the analyses, we think, clearly demonstrates the unsuitability of mitochondrial, protein-coding DNA for the task of recovering the phylogeny of mammalian orders. Mindell and Honeycutt (1990) argued that under certain conditions mtDNA ribosomal genes could be used to recover divergences up to 300 million years old (*i.e.*, if homoplasy within the data set is minimized by partitioning, Mindell, pers. comm.). However, a similarly 'leafy' topology (matching none of the five listed in Table 4) was recovered from analyses of mammal 12S mtDNA sequences, arguably the most conserved region of the mitochondrial genome, from the same set of taxa (not shown).

The reasoning behind the whole-mitochondria methodology implicitly follows from alternative #2 for dealing with 'leafy' phylogenies— improve the 'stemminess' by making the internal branches longer. This can be achieved not only by employing explicit, model-based optimality criteria (*e.g.*, maximum likelihood) that correct for loss of phylogenetic signal

Table 6. — Tree statistics Indicating Overall Character Congruence Based on Parsimony & ML Analyses. Terminal steps and misinformative changes are given as both absolute values and as percentages of the total steps and terminal steps, respectively.

	Characters		MP Steps				ML
Phylogeny	Total	Info	Total	Terminal	Misinformative	CI	-ln L
Unionoida COI	650	161	499	333 (66.7%)	236 (70.9%)	0.608	2852.13
Unionoida 28S	452	69	151	71 (47.0%)	11 (15.5%)	0.907	1376.51
Mammalia COI	650	209	759	551 (72.6%)	460 (83.5%)	0.535	3584.63
Mammalia Combo	2580	914	3184	2327 (73.1%)	1921 (82.6%)	0.535	15059.56

resulting from multiple substitutions (reviewed in Swofford *et al.*, 1996) but also by adding character data. Our analyses demonstrate that the central flaw of the whole-mitochondria approach to improve on a 'leafy' topology is the same as pointed out by Felsenstein (1978, 1981:368).

> "When amounts of evolutionary change in different lineages are sufficiently unequal [*i.e.*, long branches], it can be shown that parsimony methods make an inconsistent estimate of the evolutionary tree, *converging to the wrong tree with increasing certainty as more sequences are considered for the same set of species*" (our emphasis).

Thus, even though branch support and resolution may increase by adding nucleotide character data under the conditions above, it is unlikely to lead one to the 'True' tree. 'Leafy' topologies are inherently more likely to exhibit long-branch attraction, the erroneous association of long branches that are separated by short internal branches. By adding characters in an attempt to increase character support on the internal branches, the whole-mitochondria methodology simultaneously extends the terminal branches as well and potentially increases the number of misinformative terminal characters. In other words, long-branch problems can not be rectified by making branches longer.

Our results suggest that recovery of the interordinal relationships of 'higher' vertebrates using nucleotide data is not a lost cause. We demonstrate the potential suitability of nuclear rDNA for this particular problem based on the similarities of the mammal and freshwater mussel COI phylogenies. Although amniote systematists have been slow to employ nuclear rDNA (*i.e.*, 18S and 28S), these genes have been used extensively to recover phylogeny for a wide range of other taxa. For example, a cursory search of the recent literature recovers nuclear rDNA phylogenies for echinoderms (Smith *et al.*, 1995), lophophorates (Cohen *et al.*, 1998), arthropods (Whiting *et al.*, 1997; Dobson and Barker, 1999; Giribet *et al.*, 1999), mollusks (Park and Ó Foighil, 2000), annelids (Trontelj *et al.*, 1999), flatworms (Jondelius, 1998; Noren and Jondelius, 1999; Litvaitis and Rohde, 1999), sponges (Chombard *et al.*, 1998), plants (Bremer *et al.*, 1998; Stefanovic *et al.*, 1998), algae (Sorhannus *et al.*, 1995), fungi (Percudani *et al.*, 1999), chaetognaths (Telford and Holland, 1997), and 'lower' vertebrates (Hedges *et al.*, 1990; Hillis *et al.*, 1991; Zardoya and Meyer, 1996; Mallat and Sullivan, 1998), among others.

As with any experimental system, there are tradeoffs to using nuclear genes, and nuclear rDNA in particular. With almost any nuclear gene, at minimum, there is the complication of multiple, intragenomic copies due to diploidy. With regard specifically to nuclear rDNA, repeated arrays occur at numbers on the order of 10^2 (Hillis and Dixon, 1991). Recombination and heterozygosity are also contributing factors that may limit the usefulness of nuclear genes as historical markers. While some intragenomic variation is possible, rDNA arrays of individual genomes tend to be homogenized by concerted evolution (reviewed in Hillis and

Dixon, 1991). Other potential concerns include the non-independence of substitutions related to secondary structure (Wheeler and Honeycutt, 1988; but see Hillis and Dixon, 1991) and greater-than-expected intraspecific variation (*e.g.*, Telford and Holland, 1997; Park and Ó Foighil, 2000; Graf and Ó Foighil, in review). In the latter case, the cause of this lineage-specific variation is often introgression or gene duplication events (see Park and Ó Foighil, 2000).

What we find most troubling with regard to the entire mitochondrial genome approach to 'deep' phylogenetic inference is not so much the poor quality of the resulting hypotheses, but the fact that its practitioners seem unaware of (or uninterested in) the problems and limitations that we have described above. Phylogenetic reconstructions based on whole mitochondrial genomes have led to a number of highly controversial hypotheses (*e.g.*, Janke *et al.*, 1997; Rasmussen *et al.*, 1998; Mindell *et al.*, 1999; Rasmussen and Arnason, 1999; Waddell *et al.*, 1999b). These hypotheses are controversial in the sense that they are incongruent with hypotheses of relationships based on other classes of data, such as morphological evidence (Waddell *et al.*, 1999a). When widely accepted phylogenies are challenged, it seems obvious that further study is warranted via the analysis of alternative, independent data sets (see Curole and Kocher, 1999; and Gatesy *et al.*, 1999). As pointed out by Allard *et al.* (1999: 216), many of the aforementioned studies have failed to consider a substantial amount of paleontological and morphological evidence which has formed the basis for study of many of these 'higher' vertebrate groups. However, the methodology employed or the applicability of the data analyzed (as it relates to the question addressed) in these analyses is hardly ever suggested as a potential contributing factor to these controversial findings (but see Mindell *et al.*, 1999). Simply stated, the data utilized must not only retain phylogenetic information but also minimize the number of misinformative sites for the level of divergence that is being investigated. Throwing more similarly homoplastic data at a 'leafy' phylogeny will result in the extension of terminal branches, whereas the goal should be to increase the length of internal branches relative to the terminals for the clades of interest.

There are occasions when the addition of more characters might improve a weak phylogeny. For example, when there are simply too few character changes to resolve the nodes of interest. As demonstrated by our COI analysis of (Glires, Primates), the problem is actually too much change. As practicing molecular systematists, we are all too aware that a fundamental (although often tacit) assumption of molecular phylogenetic analyses is that 'it will all come out in the wash.' At the same time, this raises the question, "If 759 character transformations can not resolve a phylogeny of 9 mammals, how will 1000 or even 10,000?" Rather than trying to recover the White Whale 'signal' from a 'noisy' (*sensu* Wenzel and Siddall, 1999) character set by adding more genes, applying *a priori* weighting schemes, or likelihood methods, *we advocate abandoning such data* a posteriori *in favor of characters more appropriate to the questions being asked*. Some may consider such a position in conflict with 'total' or 'combined' evidence philosophies (Kluge, 1989). We simply find the assumption that there are no bad data untenable, and counterproductive that such data, once obtained and analyzed, must be retained (DeSalle and Brower, 1997).

As we have demonstrated, direct mitochondrial DNA sequence comparisons (in contrast to gene-order *a la* Boore *et al.*, 1995) may be inappropriate to robustly test the phylogeny of taxa that diverged prior to the early Tertiary (and perhaps even later). Furthermore, we have shown that analysis of nuclear ribosomal DNA, specifically 28S, is more appropriate for recovering divergences occurring within this timeframe (also see Hillis and Dixon, 1991). In addition, the utility of nuclear protein-coding genes for investigating divergences spanning approximately 50 to 200 million years ago has been discussed by Graybeal (1994). It is our hope that our commentary will stimulate exploration of novel molecular character sets, and encourage systematists using molecular sequence data to become (more) aware of its inherent limitations in phylogeny reconstruction.

ACKNOWLEDGMENTS

We are grateful to David Mindell and Diarmaid O'Foighil for many helpful comments on an earlier draft of this manuscript. Important feedback was also provided by several past and present UMMZ students, Jennifer Ast, Sharon Jansa, Ben Salisbury, and Tom DiBenidetto, and the attendees of the 2000 Hennig and SSB meetings. Additional comments were provided by Alan de Queiroz, Diana Lipscomb, and an anonymous reviewer. JSS was supported by the Carl L. and Laura C. Hubbs fellowship. DLG was supported by NSF grant 9617689 to Diarmaid Ó Foighil and the Department of Biology, University of Michigan.

REFERENCES

Allard, M.W., R.L. Honeycutt, and M.J. Novacek. 1999. Advances in higher level mammalian relationships. *Cladistics* **15**, 213–219.

Alroy, J. 1999. The fossil record of North American mammals: Evidence for a Paleocene evolutionary radiation. *Syst. Biol.* **48**, 107–118.

Anderson, S., A.T. Bankier, B.G. Barrell, M.H.L. de Bruijn, A.R. Coulson, J. Drouin, I.C. Eperon, D.P. Nierlich, B.A. Roe, F. Sanger, P.H. Schreier, A.J.H. Smith, R. Staden, and I.G. Young. 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.

Anderson, S., M.H.L. de Bruijn, A.R. Coulson, I.C. Eperon, F. Sanger, and I.G. Young. 1982. Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. *J. Mol. Biol.* **156**, 683–717.

Arnason, U., and A. Gullberg. 1993. Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. *J. Mol. Evol.* **37**, 312–322.

Arnason, U., A. Gullberg, and X. Xu. 1996. A complete mitochondrial DNA molecule of the white-handed gibbon, *Hylobates lar*, and comparison among individual mitochondrial genes of all hominoid genera. *Hereditas* **124**, 185–189.

Bibb, M.J., R.A. Van Etten, C.T. Wright, M.W. Walberg, and D.A. Clayton. 1981. Sequence and gene organization of mouse mitochondrial DNA. *Cell* **26**, 167–180.

Boore, J.L., T.M. Collins, D. Stanton, L.L. Daehler, and W.M. Brown. 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature* **376**, 163–165.

Bremer, K. 1995. Branch support and tree stability. *Cladistics* **10**, 295–304.

Bremer, K., M.W. Chase, P.F. Stevens, A.A. Anderberg, *et al.* 1998. An ordinal classification for the families of flowering plants. *Ann. Mo. Bot. Gard.* **85**, 531–553.

Chombard, C., N. Boury-Esnault, and S. Tillier. 1998. Reassessment of homology of morphological characters in tetractinellid sponges based on molecular data. *Syst. Biol.* **47**, 351–366.

Cohen, B.L., A. Gawthrop, and T. Cavalier-Smith. 1998. Molecular phylogeny of brachiopods and phoronids based on nuclear-encoded small subunit ribosomal RNA gene sequences. *Philos. Trans. R. Soc. Lond. B* **353**, 2039–2061.

Curole, J. P., and T. D. Kocher. 1999. Mitogenomics: digging deeper with complete mitochondrial genomes. *Trends Ecol. Evol.* **14**, 394–398.

DeSalle, R., and A.V.Z. Brower. 1997. Process partitions, congruences, and the independence of characters: Inferring relationships among closely related Hawaiian *Drosophila* from mutiple gene regions. *Syst. Biol.* **46**, 751-764.

Dobson, J.S., and S.C. Barker. 1999. Phylogeny of the hard ticks (Ixodidae) inferred from 18S rRNA indicates that the genus *Aponomma* is paraphyletic. *Mol. Phylogenet. Evol.* **11**, 288–295.

D'Erchia, A.M., C. Gissi, G. Pesole, C. Saccone, and U. Arnason. 1996. The guinea-pig is not a rodent. *Nature* **381**, 597–600.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively mislead. *Syst. Biol.* **27**, 401–410.

Felsenstein, J. 1981. Evolutionary trees from DNA Sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.

Fiala, K.L., and R.R. Sokal. 1985. Factors determining the accuracy of cladogram estimation: Evaluation using computer simulation. *Evolution* **39**, 609–622.

Folmer, O., M. Black, W. Hoeh, R. Lutz & R. Vrijenhoek. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* **3**, 294–299.

Gadaleta, G., G. Pepe, G. De Candia, C. Quagliariello, E. Sbisa, and C. Saccone. 1989. The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: cryptic signals revealed by comparative analysis between vertebrates. *J. Mol. Evol.* **28**, 497–516.

Gatesy, J., P. O'Grady, and R.H. Baker. 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics* **15**, 271–313.

Giribet, G., M. Rambla, S. Carranza, J. Baguna, M. Riutort, and C. Ribera. 1999. Phylogeny of the arachnid order Opiliones (Arthropoda) inferred from a combined approach of complete 18S and partial 28S ribosomal DNA sequences and morphology. *Mol. Phylogenet. Evol.* **11**, 296–307.

Gissi, C., A. Gullberg, and U. Arnason. 1998. The complete mitochondrial DNA sequence of the rabbit, *Oryctolagus cuniculus*. *Genomics* **50**, 161–169.

Graf, D.L. 2000. The Etherioidea revisited: a phylogenetic analysis of hyriid relationships (Mollusca: Bivalvia: Paleoheterodonta: Unionoida). *Occ. Pap. UMMZ* (729), 1-21.

Graf, D.L., and D. Ó Foighil. 2000. The evolution of brooding characters among the freshwater pearly mussels (Mollusca: Bivalvia: Unionoidea) of North America. *J. Mollusc. Stud.* **66**, 157-170.

Graf, D.L., and D. Ó Foighil. In review. Molecular phylogenetic analysis of 28S rDNA supports a Gondwanan origin for Australasian Hyriidae (Mollusca: Bivalvia: Unionoida). *Vie et Milieu*.

Graybeal, A. 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. *Syst. Biol.* **43**, 174–193.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**, 9–17.

Haas, F. 1969. Superfamily Unionacea. pp. N411–N470. *In* Treatise on Invertebrate Paleontology, Part N, Mollusca, 6. Volume 1: Bivalvia (R.C. Moore, ed.). University of Kansas and Geological Society of America, Lawrence.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondiral DNA. *J. Mol. Evol.* **21**, 160–174.

Hedges, S.B., K.D. Moberg, and L.R. Maxson. 1990. Tetrapod phylogeny inferred from 18S-ribosomal and 28S-ribosomal RNA sequences and a review of the evidence for amniote relationships. *Mol. Biol. Evol.* **7**, 607–633.

Hillis, D.M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* **47**, 3–8.

Hillis, D.M., and M.T. Dixon. 1991. Ribosomal DNA: Molecular evolution and phylogenetic inference. *Q. Rev. Biol.* **66**, 411–453.

Hillis, D.M., M.T. Dixon, and L.K. Ammerman. 1991. The relationships of the coelacanth *Latimeria chalumnae*— evidence from sequences of vertebrate 28S ribosomal-RNA genes. *Environ. Biol. Fishes* **32**, 119–130.

Hillis, D.M., C. Moritz, and B.K. Mable (eds.). 1996. Molecular Systematics, 2nd Edition. Sinaur Associates, Inc., Sunderland, Massachusetts.

Hoeh, W.R., D.T. Stewart, B.W. Sutherland, and E. Zouros. 1996. Multiple origins of gender-associated mitochondrial DNA lineages in bivalves (Mollusca: Bivalvia). *Evolution* **50**, 2276–2286.

Janke, A., X. Xu, and U. Arnason. 1997. The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupalia, and Eutheria. *Proc. Nat. Acad. Sci. USA* **94**, 1276–1281.

Jondelius, U. 1998. Flatworm phylogeny from partial 18S rDNA sequences. *Hyrdobiologia* **383**, 147–154.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst. Zool. **38**, 7-25.

Krettek, A., A. Gullberg, and U. Arnason. 1995. Sequence analysis of the complete mitochondrial DNA molecule of the hedgehog, *Erinaceus europaeus*, and the phylogenetic position of the Lipotyphla. *J. Mol. Evol.* **41**, 952–957.

Lanyon, S.M. 1988. Stochastic mode of molecular evolution: what consequences for systematic investigations. *Auk* **105**, 565–573.

Litvaitis, M.K., and K. Rohde. 1999. A molecular test of platyhelminth phylogeny: Inferences from partial 28S rDNA sequences. *Invertebr. Biol.* **118**, 42–56.

Mallat, J., and J. Sullivan. 1998. 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. *Mol. Biol. Evol.* **15**, 1706–1718.

Mindell, D.P., and R.L. Honeycutt. 1990. Ribosomal RNA in vertebrates: Evolution and phylogenetic implications. *Annu. Rev. Ecol. Syst.* **21**, 541–566.

Mindell, D.P., M.D. Sorenson, D.E. Dimcheff, M. Hasegawa, J.C. Ast, and T. Yuri. 1999. Interordinal relationships of birds and other reptiles based on whole mitochondrial genomes. *Syst. Biol.* **48**, 138–152.

Naylor, G. J. P., and W. M. Brown. 1998. Amphioxus mitochocndrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* **47**, 61–76.

Noren, M., and U. Jondelius. 1999. Phylogeny of the Prolecithophora (Platyhelminthes) inferred from 18S rDNA sequences. *Cladistics* **15**, 103–112.

Palumbi, S.R. 1996. Nucleic Acids II: the polymerase chain reaction. pp. 205–247. *In* Molecular Systematics, 2nd Edition (D.M. Hillis, C. Moritz, and B.K. Mable, eds.). Sinauer Associates, Inc. Sunderland, Massachusetts.

Park, J.-K., and Ó Foighil, D. 2000. Sphaeriids and corbiculids clams represent separate heterodont bivalve radiations into freshwater environments. *Mol. Phylogenet. Evol.* **14**, 75–88.

Penney, D., M. Hasegawa, P.J. Waddell, and M.D. Hendy. 1999. Mammalian evolution: Timing and implications from using the LogDeterminant Transformation for proteins of differing amino acid composition. *Syst. Biol.* **48**, 76–93.

Percudani, R., A. Trevisi, A. Zambonelli, and S. Ottonello. 1999. Molecular phylogeny of truffles (Pezizales: Terfeziaceae, Tuberaceae) derived from nuclear rDNA sequence analysis. *Mol. Phylogenet. Evol.* **13**, 169–180.

Rasmussen, A.S., and Arnason. 1999. Phylogenetic studies of complete mitochondrial DNA molecules place cartilaginous fish within the tree of bony fish. *J. Mol. Evol.* **48**, 118–123.

Rasmussen, A.S., A. Janke, and U. Arnason. 1998. The mitochondrial DNA molecule of the hagfish (*Myxine glutinosa*) and vertebrate phylogeny. *J. Mol. Evol.* **46**, 382–388.

Salisbury, B. 1999. Misinformative characters and phylogeny shape. *Syst. Biol.* **48**, 153–169.

Simpson, G.G. 1944. Tempo and Mode in Evolution. Columbia University Press, New York.

Smith, A.B., G.L.J. Paterson, and B. Lafay. 1995. Ophiuroid phylogeny and higher taxonomy—morphological, molecular and paleontological perspectives. *Zool. J. Linn. Soc.* **114**, 213–243.

Sorhannus, U., F. Gasse, R. Perasso, and A.B. Tourancheau. 1995. A preliminary phylogeny of diatoms based on 28S ribosomal-RNA sequence data. *Phycologia* **34**, 65–73.

Springer, M.S., H.M. Amirine, A. Burke, and M.J. Stanhope. 1999. Additional support for Afrotheria and Paenungulata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition. *Syst. Biol.* **48**, 65–75.

Stefanovic, S., M. Jager, J. Deutsch, J. Broutin, and M. Masselot. 1998. Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences. *Am. J. Bot.* **85**, 688–697.

Swofford, D.L. 1998. PAUP*: Phylogenetic Analysis Using Parsimony, version 4.0b3. Sinauer Associates, Inc., Sunderland, Massachusetts.

Swofford, D.L., G.J. Olsem, P.J. Waddell, and D.M. Hillis. 1996. Phylogenetic inference. pp. 407–514. *In* Molecular Systematics, 2nd Edition (D.M. Hillis, C. Moritz, and B.K. Mable, eds.). Sinauer Associates, Inc. Sunderland, Massachusetts.

Telford, M.J., and P.W.H. Holland. 1997. Evolution of 28S ribosomal DNA in chaetognaths: Duplicate genes and molecular phylogeny. *J. Mol. Evol.* **44**, 135–144.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.

Trontelj, P., B. Sket, and G. Steinbruck. 1999. Molecular phylogeny of leeches: Congruence of nuclear and mitochondrial rDNA data sets and the origin of blood sucking. *J. Zool. Syst. Evol. Res.* **37**, 141–147.

Waddell, P.J., N. Okada, and M. Hasegawa. 1999a. Towards resolving the interordinal relationships of placental mammals. *Syst. Biol.* **48**, 1–5.

Waddell, P.J., Y. Cao, J. Hauf, and M. Hasegawa. 1999b. Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid-invariant sites-LogDet plus site stripping, to detect internal

conflicts in the data, with special reference to the positions of hedgehog, armadillo, and Elephant. *Syst. Biol*. **48**, 31–53.

Waddell, P.J., Y. Cao, M. Hasegawa, and D.P. Mindell. 1999c. Assessing the Cretaceous superordinal divergence times within birds and placental mammals by using whole mitochondrial protein sequences and an extended statistical framework. *Syst. Biol.* **48**, 119–137.

Wenzel, J.W., and M.E. Siddall. 1999. Noise. *Cladistics* **15**, 61-64.

Wheeler, W.C., and R.L. Honeycutt. 1988. Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. *Mol. Biol. Evol.* **5**, 90–96.

Whiting, M.F., J.C. Carpenter, Q.D. Wheeler, and W.C. Wheeler. 1997. The Strepsiptera problem: Phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. *Syst. Biol.* **46**, 1–18.

Zardoya, R., and A. Meyer. 1996. Evolutionary relationships of the coelacanth, lungfish, and tetrapods based on the 28S ribosomal RNA gene. *Proc. Nat. Acad. Sci. USA* **93**, 5449–5454.

Received 17 July 2000; accepted 17 July 2000.