JOURNAL OF EXPERIMENTAL ZOOLOGY (MOL DEV EVOL) 304B:64–74 (2005)

Inferring Species Phylogenies From Multiple Genes: Concatenated Sequence Tree Versus Consensus Gene Tree

SUDHINDRA R. GADAGKAR,[1]* MICHAEL S. ROSENBERG,[2,3] AND SUDHIR KUMAR[2,3]*
[1]*Department of Biology, University of Dayton, Dayton, Ohio 46469-2320*
[2]*School of Life Sciences, Arizona State University, Tempe, Arizona 85287-4501*
[3]*Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, Arizona 85287-5301*

ABSTRACT Phylogenetic trees from multiple genes can be obtained in two fundamentally different ways. In one, gene sequences are concatenated into a super-gene alignment, which is then analyzed to generate the species tree. In the other, phylogenies are inferred separately from each gene, and a consensus of these gene phylogenies is used to represent the species tree. Here, we have compared these two approaches by means of computer simulation, using 448 parameter sets, including evolutionary rate, sequence length, base composition, and transition/transversion rate bias. In these simulations, we emphasized a worst-case scenario analysis in which 100 replicate datasets for each evolutionary parameter set (gene) were generated, and the replicate dataset that produced a tree topology showing the largest number of phylogenetic errors was selected to represent that parameter set. Both randomly selected and worst-case replicates were utilized to compare the consensus and concatenation approaches primarily using the neighbor-joining (NJ) method. We find that the concatenation approach yields more accurate trees, even when the sequences concatenated have evolved with very different substitution patterns and no attempts are made to accommodate these differences while inferring phylogenies. These results appear to hold true for parsimony and likelihood methods as well. The concatenation approach shows $>95\%$ accuracy with only 10 genes. However, this gain in accuracy is sometimes accompanied by reinforcement of certain systematic biases, resulting in spuriously high bootstrap support for incorrect partitions, whether we employ site, gene, or a combined bootstrap resampling approach. Therefore, it will be prudent to report the number of individual genes supporting an inferred clade in the concatenated sequence tree, in addition to the bootstrap support. *J. Exp. Zool.(Mol. Dev. Evol.) 304B:64–74, 2005.*

INTRODUCTION

Owing to the rapid advances in DNA sequencing, a considerable amount of sequence data is now available to molecular systematists for inferring the evolutionary history of species. Consequently, multiple gene and genome sequence datasets can now be used to reconstruct more robust evolutionary relationships (e.g., Delsuc et al., 2003; Hoofer et al., 2003; Teeling et al., 2003; Hedges et al., 2004; Wolf et al., 2004). There are many ways of inferring phylogenetic trees from multiple genes for the same set of species (de Queiroz et al., '95; Huelsenbeck et al., '96; Yang, '96; Nei et al., 2001; Suchard et al., 2003), but two fundamentally different ways are considered most often. In one, phylogenetic reconstruction is done after the gene sequences are concatenated head-to-tail to form a super-gene alignment—the *concatenation* (*Ct*) approach. In the other, phylogenies are inferred separately for each gene and

Grant sponsor: University of Dayton; Grant sponsor: National Institutes of Health; Grant sponsor: National Science Foundation
*Correspondence to: Sudhir Kumar, School of Life Sciences and Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, AZ 85287-5301. E-mail: s.kumar@asu.edu; or S.R. Gadagkar, Department of Biology, University of Dayton, Dayton, OH 46469-2320.
E-mail: gadagkar@notes.udayton.edu
Received 28 July 2004; Accepted 30 September 2004
Published online 18 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.21026

the resulting gene trees are used to generate a *consensus* phylogeny (the *Cn* approach).

The *Ct* approach has been used for its presumed statistical advantages: greater phylogenetic accuracy conferred by the increased sample size (number of sites) for the given set of taxa. This increased sample size works in different ways towards improving phylogenetic accuracy (e.g., Barrett et al., '91). For example, Olmstead and Sweere ('94) found that combining data from any two out of three different molecular datasets produced a more resolved phylogeny than when a single gene was used (also see Rokas et al., 2003). Today, many multigene studies routinely use the concatenation approach (e.g., Amrine-Madsen et al., 2003; Delsuc et al., 2003; Hoofer et al., 2003; Rokas et al., 2003; Teeling et al., 2003; Hedges et al., 2004; Wolf et al., 2004).

The *Cn* approach, on the other hand, summarizes congruence among individual gene trees and produces high resolution in the branching pattern only when there is at least a majority consensus among the different data sets. Thus, it gives a ''conservative'' or ''safe'' estimate of the phylogeny (Hillis, '87). The argument in favor of the *Cn* approach includes the fact that it accounts for extensive differences in evolutionary rates and substitution patterns among genes in a gene-specific manner (Bull et al., '93; de Queiroz, '93; Rodrigo et al., '93; Huelsenbeck et al., '94). While methods have now been developed for modeling different substitution parameters onto subsets of the concatenated sequence set as part of a phylogenetic analysis (e.g., Yang, '96), most investigators continue to simply concatenate sequences and apply a single substitution model to the entire alignment (e.g., Rokas et al., 2003; Hedges et al., 2004; Wolf et al., 2004).

Interestingly, there have even been suggestions that neither the *Ct* nor the *Cn* approach should be used in case individual gene phylogenies are not identical (e.g., Bull et al., '93; de Queiroz et al., '95).

To resolve some of these issues, we conducted a computer simulation study to examine the relative merits of the *Ct* and *Cn* approaches, using biologically realistic evolutionary parameters obtained from a set of 448 mammalian genes for which extensive data exists in the public databases (see Rosenberg and Kumar, 2003). In this study, we have particularly focused on the comparison between the *Ct* and *Cn* approaches for the case where individual gene trees resolve the phylogeny poorly. We examined the relative merits of the two approaches principally using the neighbor-joining (NJ) method of phylogenetic inference (Saitou and Nei, '87).

MATERIALS AND METHODS

Computer simulations to generate datasets

Figure 1 shows the model tree topology with relative branch lengths selected for computer simulations. This phylogeny is based on an independent analysis of 66 mammalian species using a variety of methods for phylogenetic analysis (Eizirik et al., 2001; Murphy et al., 2001; Rosenberg and Kumar, 2003). This tree topology and relative branch lengths were used to generate sequence data by computer simulations using 448 sets of realistic evolutionary parameters available from Rosenberg and Kumar (2003). In brief, Rosenberg and Kumar (2003) obtained all available mammalian DNA sequences for 448 nuclear genes from the HOVERGEN database (Duret et al., '94). The average base frequencies and the average number of codons were estimated for each gene and the third codon positions in these sequences were used to estimate the mean evolutionary rate (r) and the transition–transversion rate ratio (κ) for a given gene; see Rosenberg and Kumar (2003) for details and distribution of evolutionary parameters. For each set of evolutionary parameters (448 different sets), the branch lengths of the model tree were estimated using the corresponding evolutionary rate and 100 replicate datasets were generated under the HKY model of nucleotide substitution. This yielded a total of 44,800 datasets.

Phylogenetic analysis

The neighbor-joining (NJ) method of phylogenetic inference (Saitou and Nei, '87) was used with the Jukes–Cantor (JC) ('69) and Tamura–Nei (TN) ('93) methods of estimating pairwise distances. Phylogenetic analysis was carried out using PAUP* 4.0b10 (Swofford, 2001). For a given dataset (whether containing an alignment of one gene or the concatenation of multiple genes), a single model of nucleotide substitution was used in the phylogenetic analysis and no attempts were made to model differences in evolutionary parameters among genes. This was done to ensure that our procedures were similar to the approach employed in many empirical studies. Finally, the NJ method was the primary focus of analysis because it is known for its speed and efficiency in

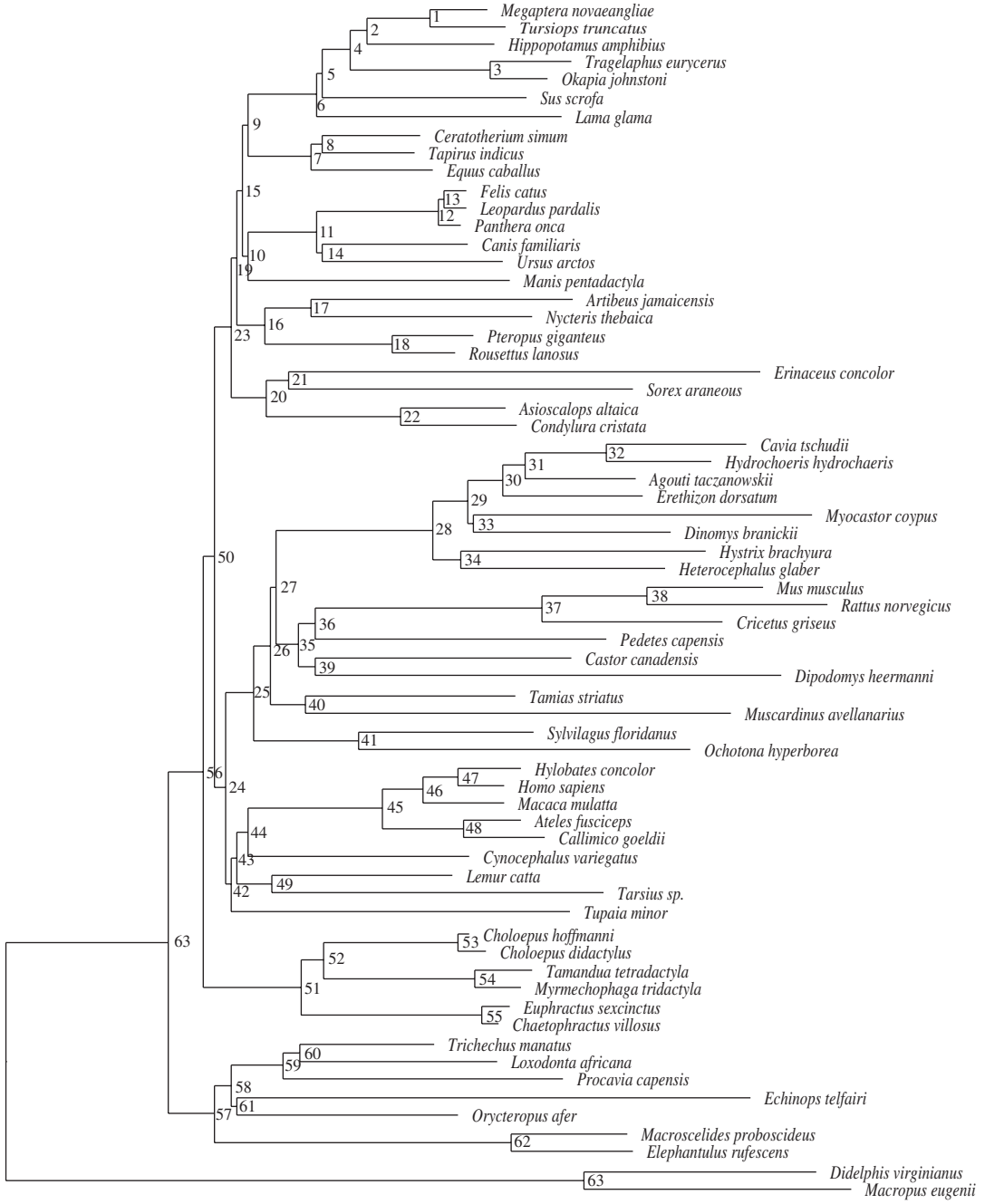

Fig. 1. The 66-taxa model tree used in computer simulations (Eizirik et al., 2001; Murphy et al., 2001; Rosenberg and Kumar, 2003). All branches are drawn to relative scale, and each clade is identified by a number.

inferring small as well as large phylogenies (Kumar and Gadagkar, 2000; Rosenberg and Kumar, 2001; Tamura et al., 2004), although some phylogenetic analyses were also conducted using the maximum parsimony and maximum likelihood methods (see Table 1 for a description) in order to assess the generality of the conclusions reached from the NJ analysis.

Assessing the accuracy of the inferred phylogeny

The accuracy of the phylogenetic trees inferred was measured by the percentage of clades reconstructed correctly (P_C). This was obtained by $P_C = 100\ [1 - d_T/(2m - 6)]$, where d_T is the topological distance between the inferred and model trees and

TABLE 1. Change in phylogenetic accuracy when a gene sequence is concatenated to an existing gene[1]

	Worst-case scenario		Random-case scenario	
Method	Overall	Poorer gene is added	Overall	Poorer gene is added
NJ–JC	96 (19)	91 (9)	89 (9)	82 (5)
NJ–TN	98 (26)	95 (12)	84 (9)	75 (4)
MP[2]	98 (31)	97 (21)	74 (7)	61 (3)
ML[3]	97 (17)[4]	93 (9)[4]	86 (11)[5]	74 (3)[5]

Note: All comparisons are made with the randomly chosen intial gene. "Poorer gene is added" indicates the performance when the tree from the added gene contained more errors than the tree from the intial gene.

[1]The numbers in the columns refer to the percentage of multigene sets (from 10,000 trials) that showed improved accuracy after concatenation, with the average amount of improvement (change in P_C) over all trials shown in parentheses in each case. $P_C=100\ [1-d_T/(2m-6)]$, where d_T is the topological distance between the inferred and model trees and m is the number of sequences in the phylogeny (see text).

[2]MP analyses were done using heuristic searches with nearest neighbor interchange (NNI) branch swapping, with the initial tree obtained by a stepwise addition procedure. All sites were uniformly weighted.

[3]For the ML method, a heuristic search with NNI branch swapping was conducted (NJ tree as the initial tree) and the HKY model of nucleotide substitutuion was used. (In the case of MP and ML analyses, a 50 percent majority-rule consensus tree was obtained when there were multiple equally parsimonious or equally likely trees for a given analysis).

[4]Based on 658 trials.

[5]Based on 207 trials.

m is the number of sequences in the phylogeny (Robinson and Foulds, '81; Penny and Hendy, '85). All comparisons were made between the inferred trees and the model tree given in Figure 1.

Construction of multigene datasets

For comparing the performance of the *Ct* and *Cn* approaches, 100 simulated datasets for each of 448 sets of evolutionary parameters (genes) were available (see above). In the construction of multigene datasets, we needed to select one dataset from the 100 simulation replicate datasets for each gene (parameter set). This selection was done in two ways. In one (the random-case scenario, RS), all analyses were conducted using randomly chosen replicates to represent individual genes. In the other, we specifically selected the simulation replicate for a given gene that produced a phylogeny with the lowest P_C, i.e., with the highest number of incorrect clades when compared to the model tree. This is referred to as the worst-case scenario (WS). Throughout this paper the RS and WS simulation replicates will also be referred to as RS and WS genes.

The *Ct* approach consisted of a simple head-to-tail concatenation of the gene sequences, which was followed by phylogenetic analysis. In the *Cn* approach, individual gene trees were first inferred and then the majority-rule consensus tree was generated to represent the multigene phylogeny. We used the "LE-50" option in PAUP* to fully resolve all multifurcations in this consensus tree. We preferred using the majority rule consensus approach for combining unrooted gene trees because other applicable approaches (e.g., strict consensus) would yield many polytomies, which would make it difficult to compare consensus and concatenation approaches.

RESULTS

Concatenating two-gene datasets

First, we investigated if adding a single gene to an existing one improves the accuracy of phylogenetic reconstruction. Results in Table 1 (in the "Overall" columns) show that concatenated alignments containing two genes, in general, produced more accurate phylogenies than those inferred from the single, initial gene ($\geq$96% for WS genes and $\geq$74% for RS genes). The percentage of correctly inferred clades per tree improved after concatenation by 17–31% for WS genes and 7–11% for RS genes. Greater improvement in accuracy is seen for NJ–TN as compared to NJ–JC. This could be attributed to at least two facts. First, the initial NJ–TN trees are known to contain more errors than NJ–JC [see Rosenberg and Kumar (2001) and Nei and Kumar (2000)]. This means that there is more room for improvement in the NJ–TN case (see also the difference in the amount of improvement in WS and RS replicates for the same tree making method). Second, the concatenation procedure leads to longer sequences and thus reduced variance in the distance estimates. Because the reduction in variance occurs for both TN and JC distances, the absolute amount of

reduction in variance of TN distances is much larger than that for JC distances, as the multi-parameter-based multiple-hit correction in TN distance leads to a manifold higher variance.

We also examined whether phylogenetic accuracy decreases if a worse performing gene is added to the initial gene. In Table 1 the column ''Poorer gene is added'' provides the percentage of concatenated datasets in which the *Ct* tree was more accurate, even in cases where the tree from the added gene had more errors than the tree from the initial gene. When the WS genes were used, a vast majority of cases ($\geq$91%) showed improvement in phylogenetic accuracy, with a minimum improvement of 9%. The improvement is slightly lower for the RS genes, partly because there was much less room for improvement as random, rather than worst-performing, replicates were used.

We investigated the effect of the relative accuracy of the added gene on the improvement (or decline) in the correctness of the inferred *Ct*-phylogeny, as compared to the accuracy of the initial gene tree. Figure 2 shows the relationship between the difference in accuracy of the initial and the added gene trees ($\Delta P_C = P_{C,added} - P_{C,initial}$) with the extent of improvement in the *Ct* tree ($\Delta P_{C,Ct} = P_{C,Ct} - P_{C,initial}$). Each point in Fig. 2 is an average of 100 gene pairs, and these averages are arranged in the order of increasing percent difference in ΔP_C (difference in accuracy between the added and initial gene trees). Negative values indicate cases in which the NJ–TN tree for the added gene had more errors than those observed in the initial gene tree. It is clear from Fig. 2 that even if the correct branches in the tree from the added gene are only 50% of that found in the initial gene tree, the concatenation of these genes still (on average) leads to improvement in phylogenetic accuracy. As expected, the improvement in the *Ct* tree increases as the accuracy of the added gene tree increases. This is true in both WS and RS cases, although, as seen earlier (Table 1) the extent of improvement is higher in the case of the WS genes.

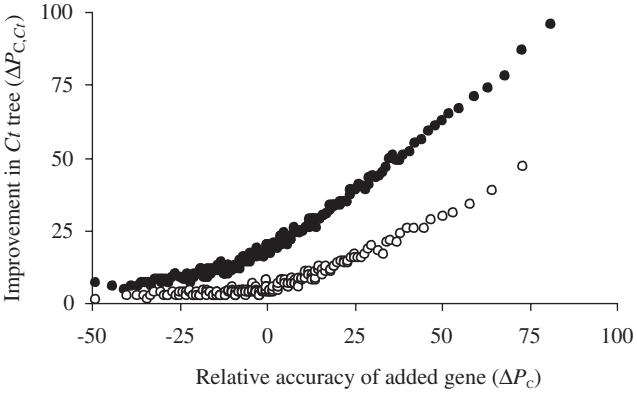

Fig. 2. Effect of the accuracy of the tree from the added gene on the performance of the *Ct* tree (relative to the accuracy of the tree from the initial gene in both cases). The percent difference in P_C (=100 $[1 - d_T/(2m - 6)]$, where d_T is the topological distance between the inferred and model trees and m is the number of sequences in the phylogeny; see text) between the tree from the added gene and the tree from the initial gene (ΔP_C) is plotted on the X axis, and the percent difference in P_C between the *Ct* tree and the tree from the initial gene ($\Delta P_{C,Ct}$) is plotted on the Y axis. Negative values for ΔP_C indicate cases in which the tree from the added gene had more errors than that from the initial gene. A total of 10,000 trials were carried out, and each plotted point represents an average from 100 trials. All phylogenetic analyses were done using NJ–TN. The filled and open circles refer to results from WS (simulation replicates with the largest phylogenetic error) and RS (randomly chosen simulation replicates), respectively.

These results clearly show that concatenating a second gene improves the accuracy of phylogenetic inference substantially over that from the initial gene. This is in spite of the fact that we simply concatenated the sequences without making an attempt to accommodate gene-specific differences in substitution pattern during phylogenetic inference. Our investigations into the effect of the observed differences in substitution parameters (between the genes that were concatenated) on the change in the phylogenetic accuracy of the *Ct* tree over the initial gene tree, did not yield any significant patterns, except that, as expected, the addition of a longer gene (sequence) produced greater improvement (results not shown). We did not apply the *Cn* approach to these two-gene datasets, as each clade will occur with either 100% or 50% frequency in the consensus tree, making it impossible to generate an unequivocal *Cn* phylogeny in most cases.

Progressive addition of genes

Next, we investigated the effect of a progressive addition of genes on phylogenetic accuracy when using *Ct* and *Cn* approaches. For this purpose, we started with six randomly chosen genes and added other randomly chosen genes one-by-one until the concatenated dataset contained 20 genes. This procedure was repeated five times and the average P_C computed for both WS and RS datasets. Figure 3 shows the results for NJ–TN trees. It is clear that an increase in the number of genes improves the accuracy of phylogenetic inference

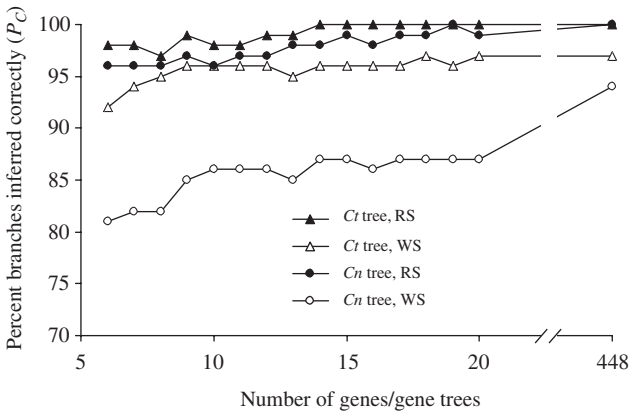

Fig. 3. Effect of a progressive addition of genes on the accuracy of the inferred species tree. A randomly chosen simulation replicate was progressively added to an initial set of six randomly chosen genes until the dataset contained a total of 20 genes. With each addition, the analysis was done using the concatenation (*Ct*) and consensus (*Cn*) approaches. For the trees inferred using the *Ct* approach, the P_C values ($P_C = 100\ [1 - d_T/(2m - 6)]$, where d_T is the topological distance between the inferred and model trees and m is the number of sequences in the phylogeny; see text) show the percent branches inferred correctly in the NJ–TN tree (neighbor-joining tree with Tamura–Nei distances). For the *Cn* approach, the P_C value was computed from the majority-rule consensus tree that was constructed such that partitions with frequency $<50\%$ were also retained to obtain a fully resolved tree if they were not in conflict with the rest of the tree (''LE50'' option in PAUP*). Each *P*C value plotted is the average from five rounds of progressive gene addition. Results are shown for the worst-case (WS) genes and random-case (RS) genes.

regardless of the approach used (*Cn* or *Ct*) or the type of dataset available (WS or RS). Figure 3 also shows that, in general, the *Ct* approach consistently produced more accurate trees than the *Cn* approach, within each scenario (WS or RS). The extent of improvement with gene addition is different for the *Cn* and *Ct* approaches as it depends upon whether RS or WS genes are used. In the *Cn*–WS case (open circles), $\sim 80\%$ partitions were inferred correctly with six genes in the dataset. This increased rapidly and reached a plateau at $\sim 86\%$ but needed more than 20 genes (between 20 and 448 genes) to reach an accuracy of around 94%. In *Ct*–WS case (open triangles), the level of accuracy was more than 90% with six genes and it leveled off at $\sim 97\%$ when only 10–20 genes were used. Interestingly, this level of accuracy remained unchanged even when all 448 genes were concatenated. The *Ct* approach was also better for the analysis of RS datasets, where it reached 100% accuracy with <15 genes, whereas the *Cn* approach required >20 genes (Fig. 3).

In the WS case, neither approach (*Ct* or *Cn*) produced the correct tree even when all 448 genes were used. The tree inferred using the *Ct* approach was missing two correct clades, whereas the *Cn* approach produced a tree in which four correct clades were missing. Two errors (in nodes #33 and #61, both of which are shallow nodes with two taxa each) were common to both approaches, whereas the *Cn* approach produced two additional errors concerning deeper nodes with multiple taxa (#50 and #56). Figure 4 shows the percentage of gene trees in which each of the clades in the true tree was recovered in the WS (above branches) and RS (below branches) datasets, when the *Cn* approach was used. These numbers indicate the number of gene trees in which the given correct phylogenetic partition was observed. In a consensus gene tree, this may be referred to as the gene-support-frequency (GSF). In *Cn*–WS analysis, GSF is 26%, 11%, 27%, and 21%, for clades #33, #50, #56, and #61, respectively. These four clades also show relatively low GSF in *Cn*–RS analysis: 50%, 72%, 47%, and 44%, respectively.

DISCUSSION

Our results indicate that the use of multiple genes produces more accurate phylogenies whether we use consensus or concatenation approaches when using the NJ method. We find that the concatenation approach produces a greater increase in accuracy as compared to the consensus approach. Table 1 also contains results from limited Maximum Parsimony (MP) and Maximum Likelihood (ML) analysis, where the trees resulting from concatenation of two genes are compared to the tree from the single initial gene. They are consistent with those obtained using the NJ method. For instance, adding a gene to a one-gene initial dataset increases the accuracy by $\sim 10\%$ for MP and ML analyses of the WS genes. (Of course, there are differences in phylogenetic accuracy of NJ–JC, NJ–TN, MP, and ML methods; see Rosenberg and Kumar (2001) and Nei and Kumar (2000) for discussion.)

Part of the increase in accuracy afforded by concatenating multiple genes is contributed by the fact that many branches in individual gene trees may have experienced no substitutions (or only a few), due to reasons such as low evolutionary rate for the gene, short elapsed time after divergence, and small gene sequences, resulting in multifurcating internal branches. These gene-specific

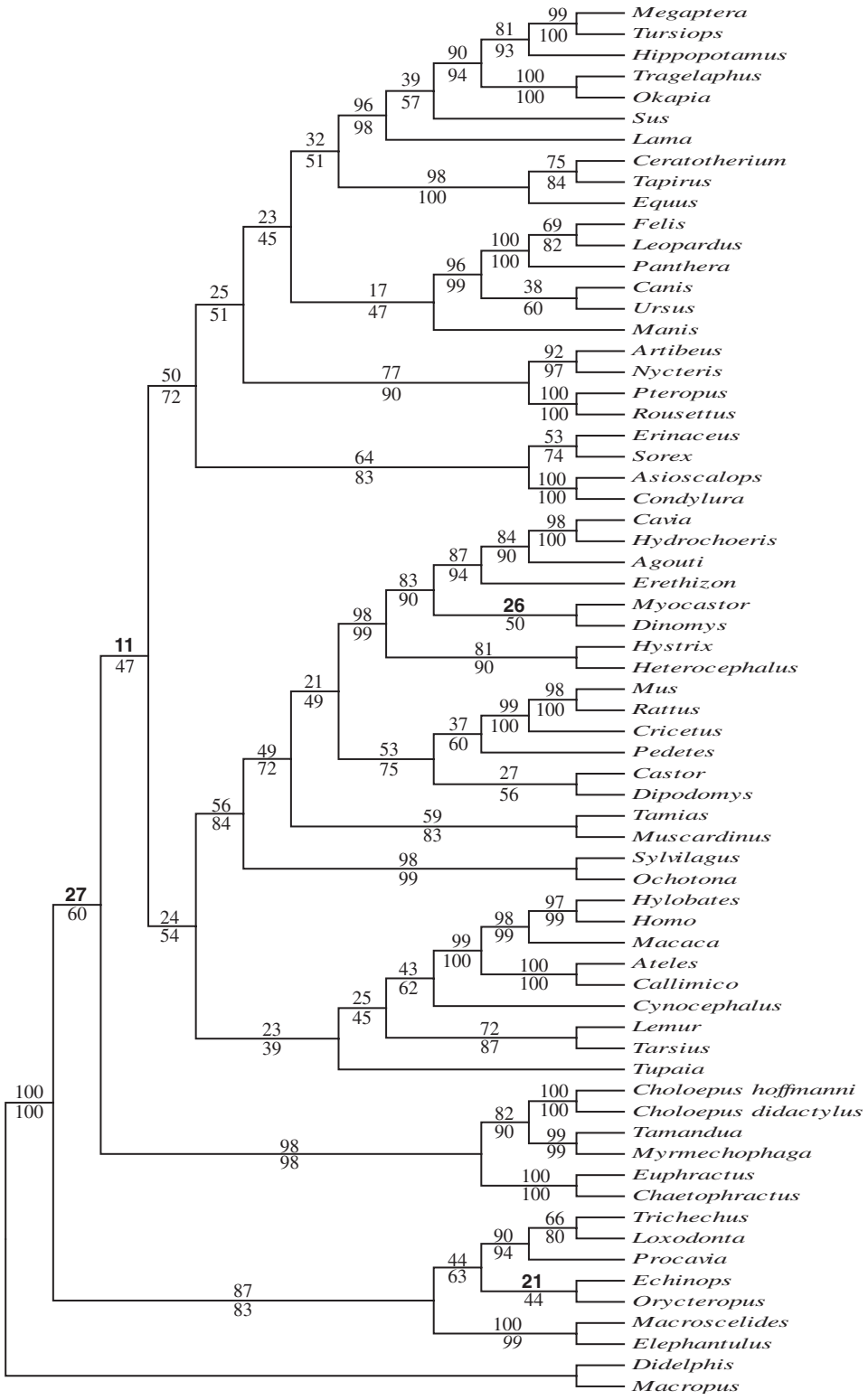

Fig. 4. Majority rule consensus tree from 448 gene trees, each obtained using the neighbor-joining method with Tamura–Nei distances (NJ–TN). Percent frequency with which individual clades were inferred for worst-case (above branch) and random-case (below branch) scenarios are shown. For incorrectly inferred partitions, percent frequencies are shown in bold. The majority-rule consensus tree was constructed using PAUP* using the LE50 option, which also retains all compatible partitions with frequency <50%, if they are not in conflict with the rest of the tree.

polytomies are not resolvable (see Kumar and Gadagkar, 2000) using a single gene, even if the species tree is truly bifurcating. Adding genes to a dataset by concatenation increases the absolute number of evolutionary changes on such branches and makes it possible to infer them with greater accuracy. Furthermore, an overall increase in sequence length would lead to smaller variances for evolutionary distances and other parameters in model based methods (e.g., NJ and ML). There is evidence in the literature that this may be the case when multiple gene sequence alignments produce trees with higher bootstrap support (e.g., Barrett et al., '91; Doyle et al., '94; Olmstead and Sweere, '94; Baldauf, '99).

The multigene concatenated datasets performed better despite the fact that we did not make an effort to account for large variations in evolutionary rate, sequence length, transition–transversion rate ratio, and base composition (G+C content) among the sequences concatenated. This indicates that the increase in phylogenetic signal due to the concatenation is much higher than any bias introduced by using a single substitution pattern applied to the entire concatenated sequence. It is possible that the use of gene-specific substitution parameters (Yang, '96) may improve the accuracy of concatenated sequence analysis, but the *Ct* approach produces more accurate phylogenies than the *Cn* approach even when they are not used (Fig. 3). This observation ameliorates some concerns such as the following: because different genes are likely to give independent estimates of the phylogeny, it is unlikely that the same (wrong) phylogeny will be supported by all the genes when analyzed individually, whereas if they are concatenated, some genes may dominate and sweep the signal (e.g., Doyle, '92; de Queiroz, '93; Miyamoto and Fitch, '95). Sometimes the congruence of datasets in terms of the phylogenetic trees produced is advocated prior to data combination (e.g., Bull et al., '93; Rodrigo et al., '93; de Queiroz et al., '95). In our studies, we did not test for congruence among the genes that were combined but still saw significant improvements, especially with WS genes that individually produced vastly different trees. Therefore, it may be better not to discard genes producing incongruent phylogenies, as they may provide additional information for resolving some short branches (also see Shevchuk and Allard, 2001; Rokas et al., 2003). On the other hand, if individual gene trees contain systematic errors that may result in similar (but wrong) phylogenies, then the use of congruent phylogenies may actually result in a more severe reinforcement of this error.

Recently, Rokas et al. (2003) conducted an empirical study with seven species of *Saccharomyces* and an outgroup, and showed that concatenated datasets with subsets of 106 genes produced the same tree as the concatenated set of all 106 genes [but see Phillips et al. (2004) for a different treatment of the same dataset]. Rokas et al. (2003) suggested that a set of only $\sim$20 genes is needed to be sampled to obtain a species phylogeny that would be similar to that obtained with a much larger number of genes. Their results are consistent with ours (see Fig. 3), except for one difference. Rokas et al. (2003) assumed that the concatenation of all genes yielded the true phylogeny because different phylogenetic methods produced the same tree, whereas our *Ct* analysis of 448 WS genes resulted in an identical, but incorrect, tree when different methods were used. Hence, this assumption may not always hold. Our results show that although the *Ct* approach is better than the *Cn* approach, it is not perfect, especially with WS datasets. We found that the *Ct* analysis of WS datasets yielded a phylogenetic tree with two true clades missing (#33, #61; Fig. 1) in our simulation investigations. Instead, two other (incorrect) clades appeared: *Myocastor coypus* (nutria) and *Echinops telfairi* (the lesser hedgehog tenrec) became basal to *Dinomys branickii* (pacarena) and *Orycteropus afer* (aardvark), respectively, rather than sister taxa (see clade #33 and #61 in Fig. 1). The difficulty observed with correct inference of clades #33 and #61 in WS simulations also cannot be attributed to shortness of their branch length because there are 10 other internal branches with the same length. Also, clades #33 and #61 do not stand out as the only shallow clades of two taxa each, as other clades of similar depth (#13 and #14) were inferred correctly. This led us to explore the relationship between the accuracy of reconstruction of an internal branch in the tree and the lengths of the neighboring branches relative to that of the internal branch (see schematic in Fig. 5A). Figure 5B shows the proportion of gene trees in which branches of various lengths (x) were inferred correctly in the worst-case replicates. As expected, the frequency of correct reconstruction is higher for longer internal branches. The problem branches (#33 and #61) belong to a set of 12 shortest internal branches in the tree, all of which have an identical length of 3.71 substitutions per gene (see Fig. 1). For these 12 branches, the

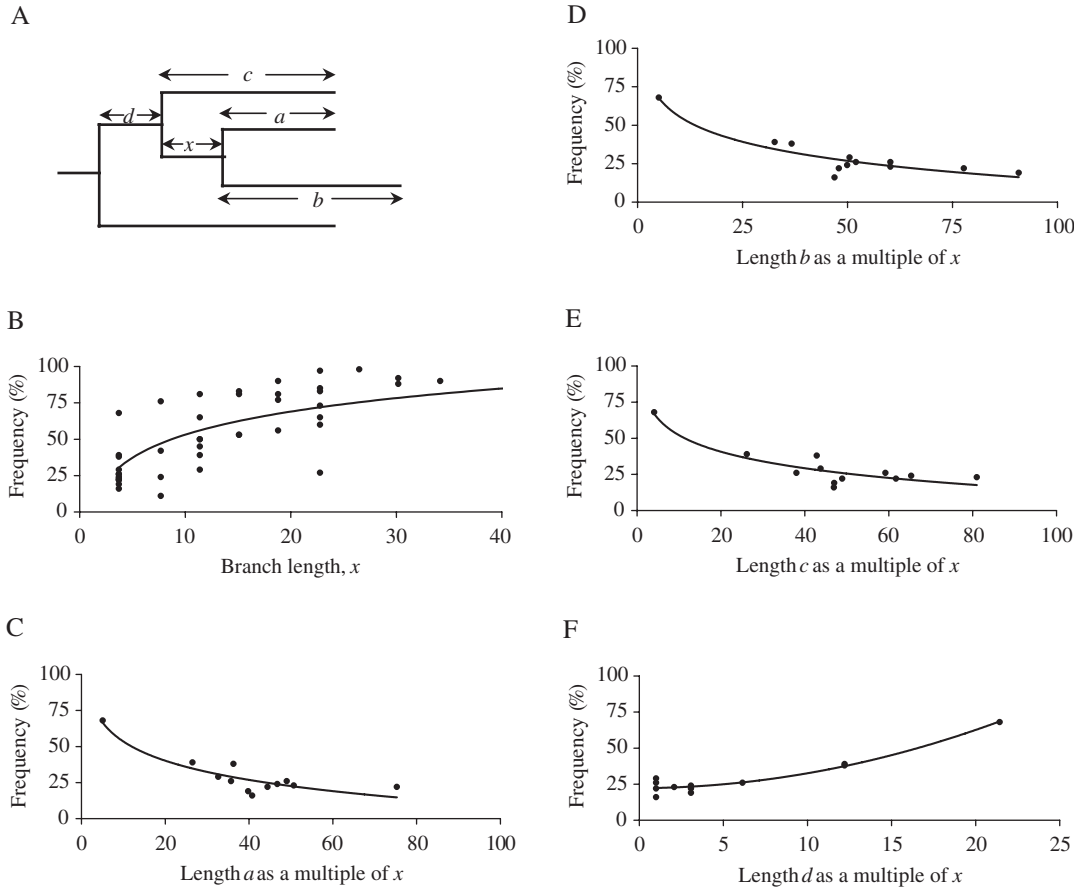

Fig. 5. Accuracy of inference of individual branches in different gene trees as a function of their length (x) and the relative lengths of their neighbors (a, b, c, d). (A) Schematic showing the definition of branch lengths x, a, b, c, d. Branches a and b were defined such that $a \leq b$. When neighboring branches led to multitaxa subtrees (such as for branch #10), we computed the branch length to be the average distance from that node to all the tip nodes in the given subtree. (B) Effect of the expected internal branch length (x_i; $1 \leq i \leq 63$) on the accuracy of inferring the corresponding clade in gene trees. (C–F) Accuracy of inferring the 12 shortest branch lengths as a function of the neighboring branch lengths a to d, respectively. All results were obtained from trees inferred for the 448 worst-case simulation replicates (WS), using NJ–TN (neighbor-joining method with Tamura–Nei distances).

influence of the lengths, a, b, c, and d, of the neighboring branches on the accuracy of their inference is shown, respectively, in panels C–F of Fig. 5. It is clear that the longer the lengths a, b, and c, as compared to x, the lower the probability that the phylogenetic partition defined by branch x will be inferred correctly. That is, an internal branch may fail to be recovered when it is much shorter than its neighboring branches (or the average branch lengths of the different clusters). In contrast, a larger value of d (ancestral branch) translates into a higher accuracy.

In order to ascertain if the incorrect inference of certain short branches was due to long (short)-branch attraction (Felsenstein, '78; Nei and Kumar, 2000), we examined the relationship between the percent difference between the lengths a and b and the accuracy of recovery of the corresponding internal branches. Our analyses for the 12 shortest branches showed a slight negative relationship between the number of gene trees that correctly recovered a given branch and the difference between the lengths a and b for that branch (not shown). We find that branches #33 and #61, which are the most difficult to infer, also show the largest length difference between the descendent taxa in the true tree (69% and 128% for clades #33 and #61, respectively). It is tempting to speculate that this is the reason for the incorrect inference of these two clades, but another branch (#9) where the descendent taxa show a large (55%) difference does not suffer from a similar problem. Furthermore, both of these clades are inferred perfectly in the RS case.

Our study shows additional evidence that the concatenation approach is not necessarily a panacea for the accurate recovery of the species tree. This was revealed when we conducted a bootstrap test by resampling sites in order to determine if the two incorrectly inferred nodes (#33 and #61) in the *Ct* tree of all the 448 WS genes represented significant systematic errors (Felsenstein, '85). If the incorrect inferences made (instead of clades #33 and #62) were by chance alone, then their frequency in the bootstrap consensus tree is expected to be low. (We refer to the frequency of occurrence of a clade in the bootstrap replicate tree as the bootstrap support frequency [BSF].) Our results showed that the two correct clades (#33 and #61) did not appear in *any* of the bootstrap replicate tree, that is, BSF = 0% for both clades. Instead, we found that the two incorrect clades (that appeared instead) were supported with increasingly higher BSF as the number of genes concatenated increased, ultimately reaching 100% for the case of 448 WS genes. In fact, BSF for all branches in the *Ct*–WS tree was 100%!

Recently, Nei et al. (2001) suggested that the genes, rather than sites [as used in the original proposal by Felsenstein ('85)], be used as the unit of resampling in the bootstrap estimation. In our study, we resampled 448 WS genes with replacement and generated *Ct* trees for 100 bootstrap replicates. Still, all branches (including the two incorrect ones) were supported with a BSF = 100%. A more comprehensive bootstrap resampling regime is to actually resample genes and to then resample sites within each gene in order to account for site as well as gene sampling errors. This combined procedure also produced BSF = 100% for concatenation analysis of 448 WS genes. These results are particularly disturbing because the incorrect clades inferred (and supported with BSF = 100%) were found in only 34% and 33% individual gene trees, respectively, as compared to the correct clades (#33 and #61), which appeared in 26% and 21% gene trees. Therefore, it is clear that weak phylogenetic signal can be substantially reinforced when sequences are concatenated. Most often, this signal amplification happens for the correct partitions (see Fig. 4), but in some cases it can also boost support for the erroneous inferences.

In conclusion, a simple concatenation appears to be better than consensus for phylogenetic reconstruction when multigene data are available, especially when the individual genes yield inaccurate trees. However, neither approach may guarantee a completely accurate species phylogeny even when a large number of genes are used, apparently due to the effect of certain systematic biases. Indeed, there is no magic number of genes that, when concatenated, will yield the correct tree in all instances, because addition of genes does not add to the accuracy of the tree in the presence of systematic biases. In the future, it will be important to examine if these biases can be overcome by using gene-specific patterns of substitution in the concatenation approach. However, because these systematic errors can lead to very high bootstrap values, we recommend that the number of gene trees supporting a given partition (gene support frequency, GSF) be presented along with its bootstrap support frequency (BSF) to guard against conveying spurious high confidence in evolutionary inferences based on the analysis of multiple genes.

ACKNOWLEDGMENTS

We thank S. Subramanian and Balaji Ramanujam for help with computational analyses. This research was supported by start-up support from the University of Dayton (to S.R.G.) and grants from the National Institutes of Health (to S.K.) and National Science Foundation (to James P. Collins, ASU, the Principal Investigator for the grant, which provided partial funding for this work).

LITERATURE CITED

Amrine-Madsen H, Koepfli KP, Wayne RK, Springer MS. 2003. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. Mol Phylogenet Evol 28:225–240.

Baldauf SL. 1999. A search for the origins of animals and fungi: comparing and combining molecular data. Am Nat 154:S178–S188.

Barrett M, Donoghue MJ, Sober E. 1991. Against consensus. Syst Zool 40:486–493.

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ. 1993. Partitioning and combining data in phylogenetic analysis. Syst Biol 42:384–397.

de Queiroz A. 1993. For consensus (sometimes). Syst Biol 42:368–372.

de Queiroz A, Donoghue MJ, Kim J. 1995. Separate versus combined analysis of phylogenetic evidence. Annu Rev Ecol Syst 26:657–681.

Delsuc F, Stanhope MJ, Douzery EJP. 2003. Molecular systematics of armadillos (Xenarthra, Dasypodidae): contribution of maximum likelihood and Bayesian analyses of mitochondrial and nuclear genes. Mol Phylogenet Evol 28:261–275.

Doyle JA. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. Syst Botany 17:144–163.

Doyle JA, Donoghue MJ, Zimmer EA. 1994. Integration of morphological and rRNA data on the origin of angiosperms. Ann Missouri Bot Gard 81:419–450.

Duret L, Mouchiroud D, Gouy M. 1994. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res 22:2360–2365.

Eizirik E, Murphy WJ, O'Brien SJ. 2001. Molecular dating and biogeography of the early placental mammal radiation. J Hered 92:212–219.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410.

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

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular time scale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 4:2.

Hillis DM. 1987. Molecular versus morphological approaches to systematics. Annu Rev Ecol Syst 18:23–42.

Hoofer SR, Reeder SA, Hansen EW, Van den Bussche RA. 2003. Molecular phylogenetics and taxonomic review of noctilionoid and vespertilionoid bats (Chiroptera: Yangochiroptera). J Mammal 84:809–821.

Huelsenbeck JP, Bull JJ, Cunningham CW. 1996. Combining data in phylogenetic analysis. Trends Ecol Evol 11:152–158.

Huelsenbeck JP, Swofford DL, Cunningham CW, Bull JJ, Waddell PJ. 1994. Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? Syst Biol 43:288–291.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p 21–132.

Kumar S, Gadagkar SR. 2000. Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies. J Mol Evol 51(6):544–553.

Miyamoto MM, Fitch WM. 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst Biol 44:64–76.

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryderk OA, O'Brien SJ. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Nei M, Xu P, Glazko G. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proc Natl Acad Sci U S A 98:2497–2502.

Olmstead RG, Sweere JA. 1994. Combining data in phylogenetic systematics—an empirical approach using 3 molecular data sets in the Solanaceae. Syst Biol 43:467–481.

Penny D, Hendy MD. 1985. The use of tree comparison metrics. Syst Zool 34:75–82.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455–1458.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci 53(1–2):131–147.

Rodrigo AG, Kellyborges M, Bergquist PR, Bergquist PL. 1993. A randomization test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. New Zealand J Bot 31:257–268.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Rosenberg MS, Kumar S. 2001. Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well. Mol Biol Evol 18:1823–1827.

Rosenberg MS, Kumar S. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol Biol Evol 20:610–621.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

Shevchuk NA, Allard MW. 2001. Sources of incongruence among mammalian mitochondrial sequences: COII, COIII, and ND6 genes are main contributors. Mol Phylogenet Evol 21:43–54.

Suchard MA, Kitchen CMR, Sinsheimer JS, Weiss RE. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. Syst Biol 52:649–664.

Swofford DL. 2001. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sunderland, MA: Sinauer Associates.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A 101(30):11030–11035.

Teeling EC, Madsen O, Murphy WJ, Springer MS, O'Brien J. 2003. Nuclear gene sequences confirm an ancient link between New Zealand's short-tailed bat and South American noctilionoid bats. Mol Phylogenet Evol 28:308–319.

Wolf YI, Rogozin IB, Koonin EV. 2004. Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res 14:29–36.

Yang ZH. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42: 587–596.