Combining Multiple Data Sets in a Likelihood Analysis: Which Models are the Best?

Tal Pupko, Dorothée Huchon,† Ying Cao,* Norihiro Okada,† and Masami Hasegawa**

*The Institute of Statistical Mathematics, Tokyo, Japan; and †Molecular Evolution Laboratory, Faculty of Bioscience and Biotechnology, Tokyo Institute of Technology, Japan

Until recently, phylogenetic analyses have been routinely based on homologous sequences of a single gene. Given the vast number of gene sequences now available, phylogenetic studies are now based on the analysis of multiple genes. Thus, it has become necessary to devise statistical methods to combine multiple molecular data sets. Here, we compare several models for combining different genes for the purpose of evaluating the likelihood of tree topologies. Three methods of branch length estimation were studied: assuming all genes have the same branch lengths (concatenate model), assuming that branch lengths are proportional among genes (proportional model), or assuming that each gene has a separate set of branch lengths (separate model). We also compared three models of among-site rate variation: the homogenous model, a model that assumes one gamma parameter for all genes, and a model that assumes one gamma parameter for each gene. On the basis of two nuclear and one mitochondrial amino acid data sets, our results suggest that, depending on the data set chosen, either the separate model or the proportional model represents the most appropriate method for branch length analysis. For all the data sets examined, one gamma parameter for each gene represents the best model for among-site rate variation. Using these models we analyzed alternative mammalian tree topologies, and we describe the effect of the assumed model on the maximum likelihood tree. We show that the choice of the model has an impact on the best phylogeny obtained.

Introduction

In the last 30 years, vast improvements in DNA-sequencing methods have effected an exponential increase in the number of gene entries in databases worldwide (e.g., International Human Genome Sequencing Consortium 2001) and given scientists access to entire genome sequences of many organisms. Access to large amounts of sequence data has spawned a revolution in the understanding of biological diversity (see Graur and Li 1999, pp. 217–247). For example, in a recent analysis Murphy et al. (2001) examined variations among 18 homologous gene segments (nearly 10,000 base pairs) to infer the mammalian evolutionary tree. The sequence data explosion is not without its caveats, however, because it brings with it the need for development of methods to combine efficiently information from multiple molecular data sets. Examples of multiple data sets are (1) several genes, (2) the three-codon positions of a protein-coding sequence, and (3) different parts of the protein-coding sequence that correspond to different secondary structures (e.g., alpha-helix, beta-sheet). Because different genes likely have different evolutionary constraints, the evolution of different genes might be best described by different sets of parameters. However, when statistically analyzing a data set, adding new parameters is not always justified and can lead to erroneous conclusions (for examples, see Burnham and Anderson 1998). Thus, when modeling sequence evolution in a maximum likelihood (ML) framework of tree reconstruction (Felsenstein 1981), investigators must strive to determine a median set of parameters that neither assumes too few parameters nor results in ''over fitting'' by assuming too many. An example of a significant improvement in sequence evolution models is the parameterization of among-site rate variation. By adding only one parameter to describe the site rate variation distribution, a substantial increase in the log-likelihood is typically gained (Yang 1996*a*). Not only is this addition of a single parameter reasonable in a statistical and biological sense, but it has also been shown to affect the resulting phylogenetic conclusions (Sullivan and Swofford 1997). Cases in which the addition of new parameters is statistically unjustified are rarely published, but such pitfalls are discussed by Nei and Kumar (2000, pp. 154–155) and Takahashi and Nei (2000).

When analyzing a multiple sequence alignment, one must assume an underlying evolutionary model. This model includes tree topology, branch lengths, rate heterogeneity among sites, and substitution probabilities. All these parameters may change from gene to gene. For example, the tree topology of various genes may not be the same because of horizontal gene transfer. Similarly, the substitution model may vary from gene to gene because of different evolutionary constraints or because of differences in GC content. In this article we focus on the fitting of different branch length models and rate heterogeneity models to several protein data sets.

The Number of Branch Length Parameters

For an unrooted tree topology T, with n sequences, the number of branches is $2n - 3$. We denote branches by $t_1, \ldots, t_{2n-3}$. Assume now that we have two data sets (e.g., two different genes), each one with n homologous sequences. One way to analyze a combination of these data sets is to concatenate the sequences and evaluate the resultant branch lengths. We refer to this model as

Key words: combining data sets, phylogeny, maximum likelihood, Mammalia, molecular evolution.

Address for correspondence and reprints: Tal Pupko, The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan. E-mail: tal@ism.ac.jp.

Mol. Biol. Evol. 19(12):2294–2307. 2002

Table 1
The Number of Parameters in Each of the Nine Models, Where *g* is the Number of Genes and *n* is the Number of Sequences in Each Gene

	Homogenous	1-GAM	N-GAM
Concatenation . . .	2n − 3	2n − 2	2n + g − 3
Proportional	2n + g − 4	2n + g − 3	2n + 2g − 4
Separate	2ng − 3g	2ng − 3g + 1	2ng − 2g

NOTE.—For example, in the case of the proportional–1–GAM model, the number of branches is 2n − 3; there are g − 1 gene-specific rate parameters (one to each gene, with an average equal to 1), hence 2n − 3 + g − 1 or 2n + g − 4; adding the one gamma rate parameter yields a total of 2n + g − 3 parameters.

the ''concatenate model.'' In such a scenario the joint probability would be

$$P(data1 \,\&\, data2/T, t_1, \ldots, t_{2n-3}) = P(data1/T, t_1, \ldots, t_{2n-3}) \cdot P(data2/T, t_1, \ldots, t_{2n-3})$$

The number of free parameters here is $2n - 3$. (This model assumes that both genes have the same branch length.) Another approach is to assume that branch lengths for the two genes are independent (i.e., each gene is analyzed separately). We refer to this model as the ''separate model.'' In this case the joint probability would be

$$P(data1 \,\&\, data2/T, t_1^1, \ldots, t_{2n-3}^1, t_1^2, \ldots, t_{2n-3}^2) = P(data1/T, t_1^1, \ldots, t_{2n-3}^1) \cdot P(data2/T, t_1^2, \ldots, t_{2n-3}^2)$$

where the superscripts denote the data set attributes. In our present work, we study these two alternative models and explore a third alternative for combining data sets, namely, the ''proportional branch lengths'' approach first suggested by Yang (1996*b*).

The Proportional Branch Lengths Approach

The proportional branch lengths approach assumes that branch lengths for two trees are the same, up to a scaling factor r. Thus, if $t_1, \ldots, t_{2n-3}$ are the branches of the first gene, the branch lengths of the second gene would be $rt_1, \ldots, rt_{2n-3}$. This scaling factor r corresponds to a gene-specific rate that is assigned to each gene, and for n genes we have n gene-specific rate factors $r_1, \ldots, r_n$. The average r should be equal to 1.0. We refer to this model as the ''proportional model.'' For two data sets the joint probability is

$$P(data1 \,\&\, data2/T, t_1, \ldots, t_{2n-3}, r_1, r_2) = P(data1/T, t_1, \ldots, t_{2n-3}, r_1) \times P(data2/T, t_1, \ldots, t_{2n-3}, r_2)$$

Consequently, the total number of parameters for two genes under this proportional model is $2n - 3 + 1 = 2n - 2$. The number of parameters for n genes for each of the three models is summarized in table 1.

The biological meaning of the proportional model relies on the assumption that the rate in each branch is a multiplication of two rates: the rate of the specific gene multiplied by the rate of the specific lineage. It is the rate of the specific lineage that is common to all genes under the proportional model. In contrast, the concatenate model assumes that all genes' rates and all lineages' rates are the same, whereas the separate model assumes that the rate in each lineage is independent among genes.

The Number of Among-Site Rate Variation Parameters

We consider three possible models of among-site rate variation. The first model assumes that all sites have the same rate of evolution (''homogenous'' model), the second model assumes one gamma rate parameter for all genes (''1-GAM'' model), and the third model assumes a separate gamma parameter for each gene (''N-GAM'' model).

We compare all nine combinations of models with respect to likelihood (concatenate-homogenous, concatenate–1-GAM, concatenate–N-GAM, proportional-homogenous, proportional–1-GAM, proportional–N-GAM, separate-homogenous, separate–1-GAM, separate–N-GAM). With respect to branch lengths, we show that the proportional and separate models are always better than the concatenate model. Selecting between these two models depends on the specification of the data set under study. For some data sets the proportional model represents the best model, whereas for others the separate model is the best. With respect to the number of gamma parameters, the N-GAM model is the best model for all the data sets included in our study.

Material and Methods

Sequences

Computations were based on three protein alignments: those given by Madsen et al. (2001) and Murphy et al. (2001), and an updated mitochondrial data set of Nikaido et al. (2001).

Madsen Data Set

The Madsen nucleotide alignment includes 28 species for four independent nuclear genes: the alpha-2B adrenergic receptor (A2AB, 344 sites), the breast cancer susceptibility gene (BRCA1, 557 sites), the interphotoreceptor retinoid-binding protein (IRBP, 301 sites), and the von Willebrand factor (vWF, 338 sites). The sequences of the golden mole (*Amblysomus hottentotus*) and the Madagascar hedgehog (*Echinops telfairi*) were not included because sequence data were not available for IRBP. The BRCA1 sequence of the thick-tailed opossum (*Lutreolina crassicaudata*; accession number: AY057826) was added manually to the Madsen alignment.

Murphy Data Set

Among the 18 genes considered in the nucleotide alignment of Murphy et al. (2001), we excluded the seven noncoding genes. Because we required that all genes share the same species sampling (i.e., no missing sequences), three more genes for which marsupial sequences were unavailable were excluded. Two genes

with a poor species sampling were also excluded. This exclusion allows our analysis to maintain a large and diversified species sampling, where all sequences are available for all genes. The final alignment includes 46 species for six nuclear genes: adenosine A3 receptor (ADORA3, 107 sites), the Menkes disease gene (ATP7A, 220 sites), the brain-derived neurotrophic factor (BDNF, 182 sites), the cannabinoid receptor 1 (CNR1, 219 sites), the sphingolipid G-protein–coupled receptor 1 (EDG1, 199 sites), and the zinc finger protein X-linked (ZFX, 67 sites).

Mitochondrial Data Set

The mitochondrial data set included 56 species comprising the 43 complete mitochondrial coding sequences analyzed by Nikaido et al. (2001), together with the following sequences: (1) Asiatic shrew, *Soriculus fumidus*, AF348081; (2) long-tailed bat, *Chalinolobus tuberculatus*, AF321051; (3) little red flying fox, *Pteropus scapulatus*, AF321050; (4) northern brown bandicoot, *Isoodon macrourus*, AF358864; (5) gymnure, *Echinosorex gymnura*, AF348079; (6) American pika, *Ochotona princeps*, AF348080; (7) barbary ape, *Macaca sylvanus*, AJ309865; (8) slow loris, *Nycticebus coucang*, AJ309867; (9) white-fronted capuchin, *Cebus albifrons*, AJ309866; (10) cane rat, *Thryonomys swinderianus*, AJ301644; (11) vole, *Volemys kikuchii*, AF348082; (12) tree shrew, *Tupaia belangeri*, AF217811; and (13) small Madagascar hedgehog, *E. telfairi*, AJ400734. The 12 H-strand mitochondrial protein-coding genes are ND1 (313 sites), ND2 (313 sites), COX1 (512 sites), COX2 (225 sites), ATP8 (32 sites), ATP6 (201), COX3 (259), ND3 (104 sites), ND4L (94 sites), ND4 (438 sites), ND5 (526 sites), and Cyt*b* (375 sites). The overlapping regions between ATP6 and ATP8 and between ND4 and ND4L were excluded.

All nucleotide alignments were translated into amino acid alignments. To agree with the reading frame, some minor changes were made to the alignments of Madsen et al. (2001) and Murphy et al. (2001). For all genes, gap positions were excluded from the analysis. If data for certain positions were missing in >5% of the species studied, then such positions were excluded from the analysis. All the protein alignments and accession numbers are attached as supplementary material at http://www.molbiolevol.org/.

Tree Topologies

For each data set four different topologies were considered: a morphological tree, a mitochondrial tree, and two nuclear trees (Madsen and Murphy topologies). Because species sampling differed among the four data sets, the four trees were slightly different with respect to the data set used. For the mitochondrial data set the morphological and mitochondrial trees are presented in figures 1 and 2, respectively. For the Murphy data set the Murphy tree is given in figure 3, and for the Madsen data set the Madsen tree is given in figure 4. All 12 trees are attached as supplementary material at http://www.molbiolevol.org/.

Morphological Tree

For all data sets, morphological trees were based on the phylogeny of McKenna and Bell (1997) (e.g., fig. 1). The topology of Novacek (1992) was adopted for relationships between clades that were not determined by McKenna and Bell (i.e., among the grandorders of Epitheria). Any relationships that were not fully resolved by the above criteria were subsequently chosen based on the nuclear topology of Murphy et al. (2001).

Mitochondrial Tree

The mitochondrial tree is based on Cao et al. (2000). However, the position of the rodents was chosen in agreement with Reyes, Pesole, and Saccone (2000), and Mouchaty et al. (2001). The position of the vole among the rodents was chosen in agreement with morphological data (McKenna and Bell 1997). Among cetartiodactyls, the alpaca and the pig were sister clades in agreement with Arnason et al. (2000). The relationships among bats were chosen in agreement with Nikaido et al. (2001) and McKenna and Bell (1997). The shrews and moles were placed as a sister clade of the bats, in agreement with Nikaido et al. (2001). The Afrotheria phylogeny was in agreement with Murphy et al. (2001). Xenarthra was placed as a sister clade of Afrotheria, in agreement with Reyes, Pesole, and Saccone (2000). The position of the rabbit, tree shrew, primate, and hedgehog was in agreement with Schmitz, Ohme, and Zischler (2000). The lagomorphs were considered monophyletic. The relationship among primates followed McKenna and Bell (1997). Hedgehog and gymnure were placed together. Finally, relationships among marsupials were taken from Phillips et al. (2001). For the other relationships we followed Murphy et al. (2001). Figure 2 presents the mitochondrial tree for the mitochondrial data set.

The Murphy Tree

This tree was based on Murphy et al. (2001). McKenna and Bell (1997) were followed for relationships that were not determined by Murphy et al. Figure 3 presents the Murphy tree for the Murphy data set.

The Madsen Tree

This tree is based on Madsen et al. (2001, fig. A). Murphy et al. (2001) and McKenna and Bell (1997) were followed for relationships that were not determined by Madsen et al. Figure 4 presents the Madsen tree for the Madsen data set.

Model

In this study, models based on amino acid sequences were used. The replacement probabilities among amino acids were calculated with the JTT matrix (Jones, Taylor, and Thornton 1992) for nuclear genes and the REV model (Adachi and Hasegawa 1996) for mitochondrial genes. However, the approach presented here is also valid for nucleotide sequences and for any sub-

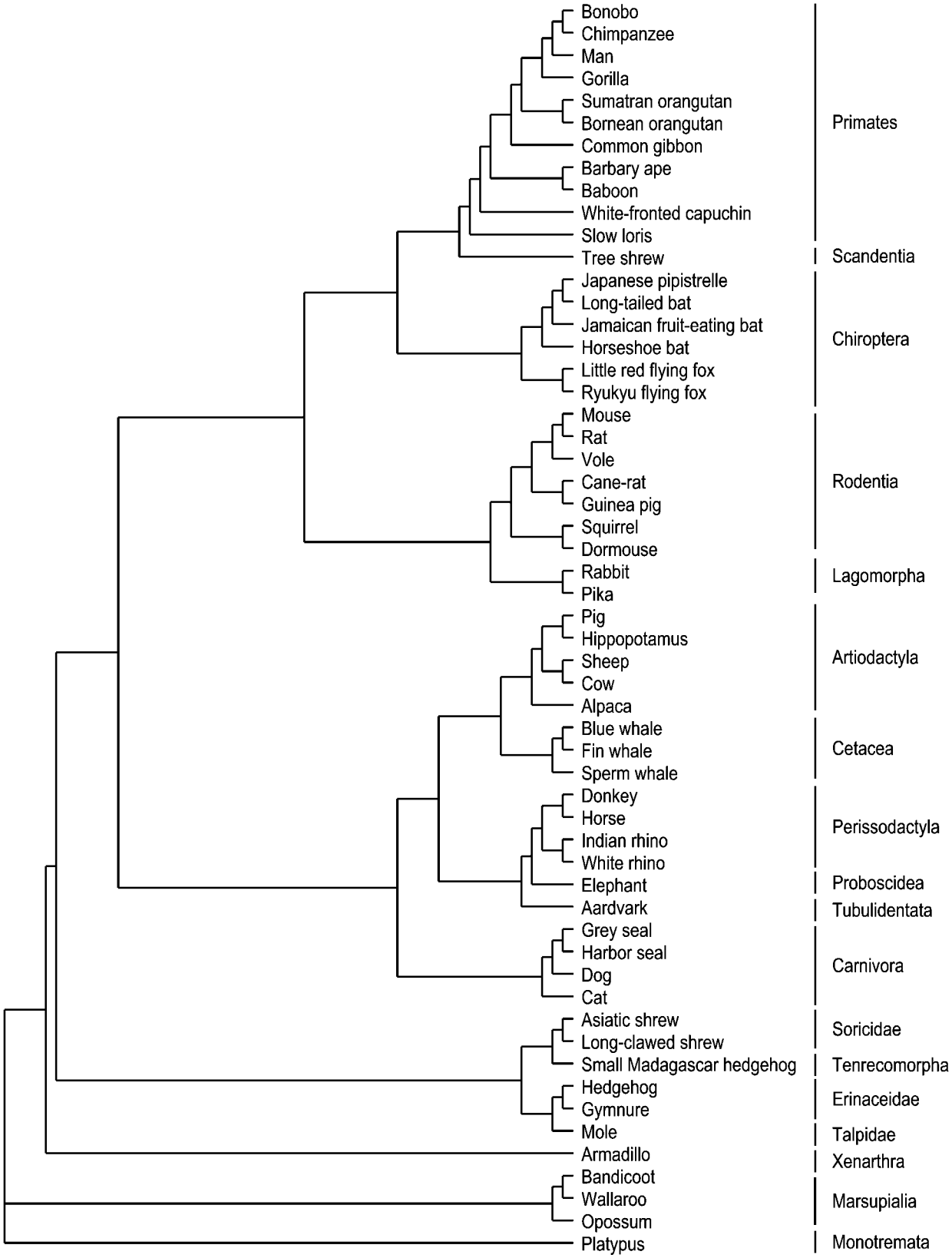

FIG. 1.—The morphological tree topology for the mitochondrial data set.

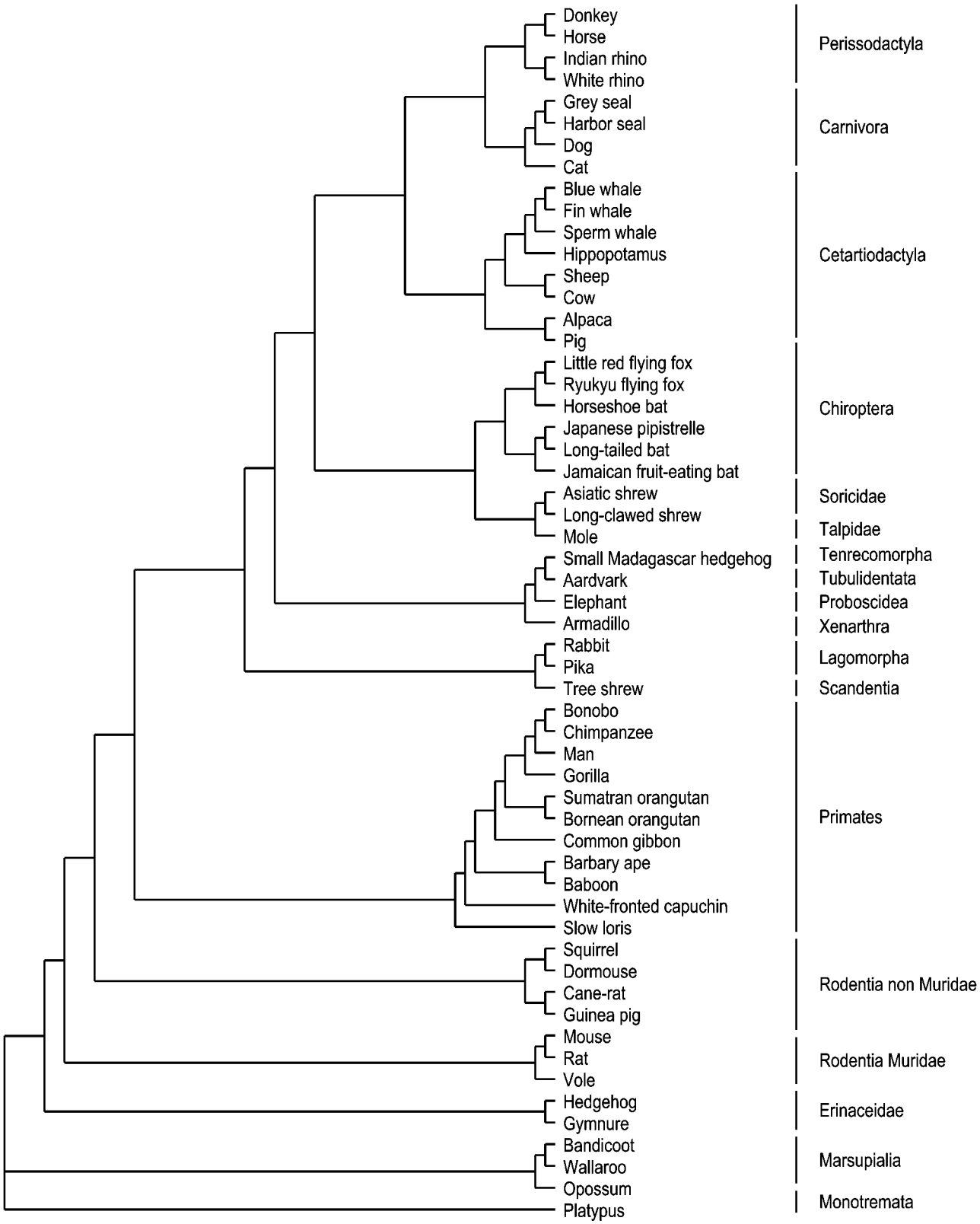

FIG. 2.—The mitochondrial tree topology for the mitochondrial data set.

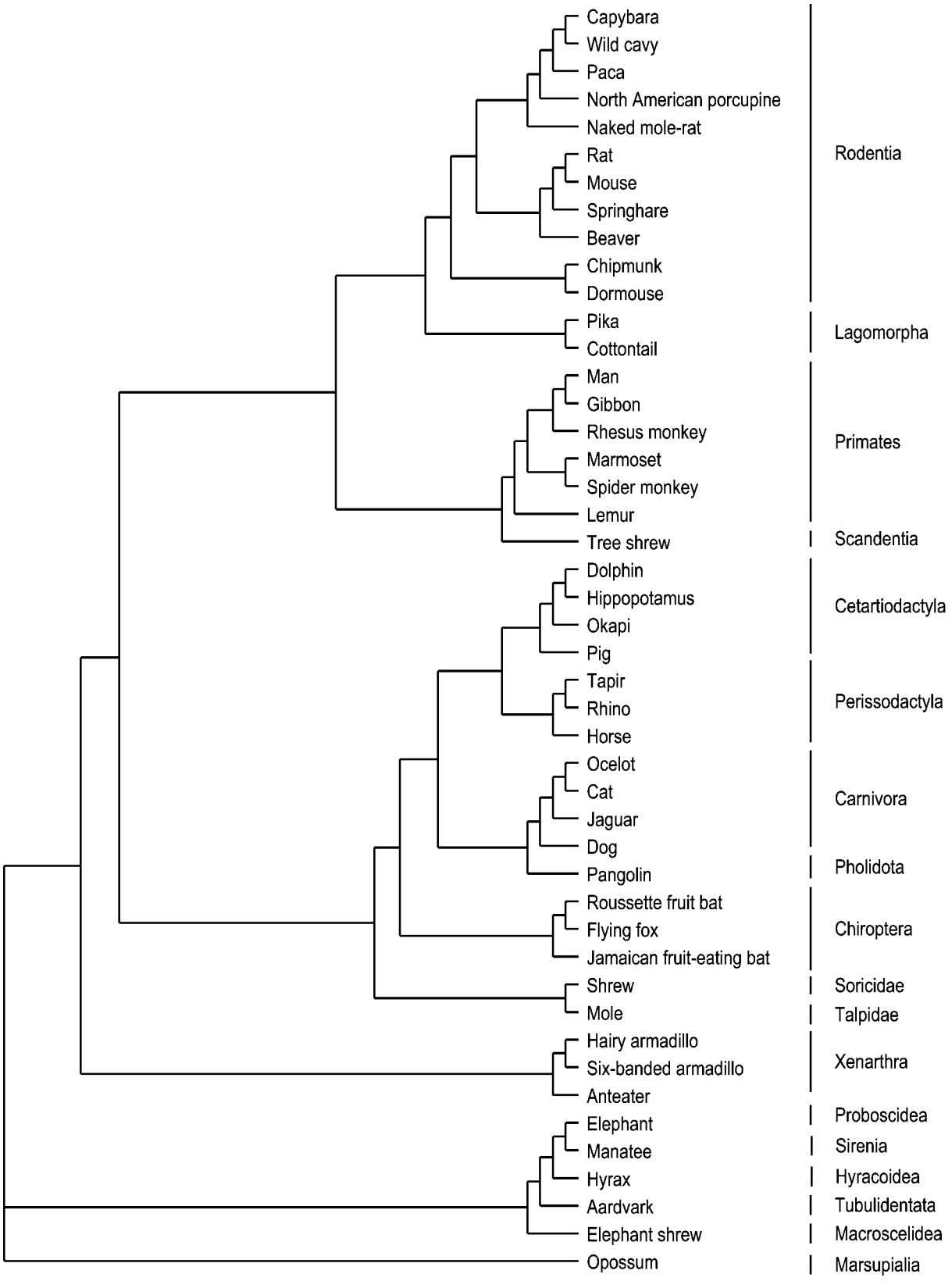

FIG. 3.—The Murphy tree topology for the Murphy data set.

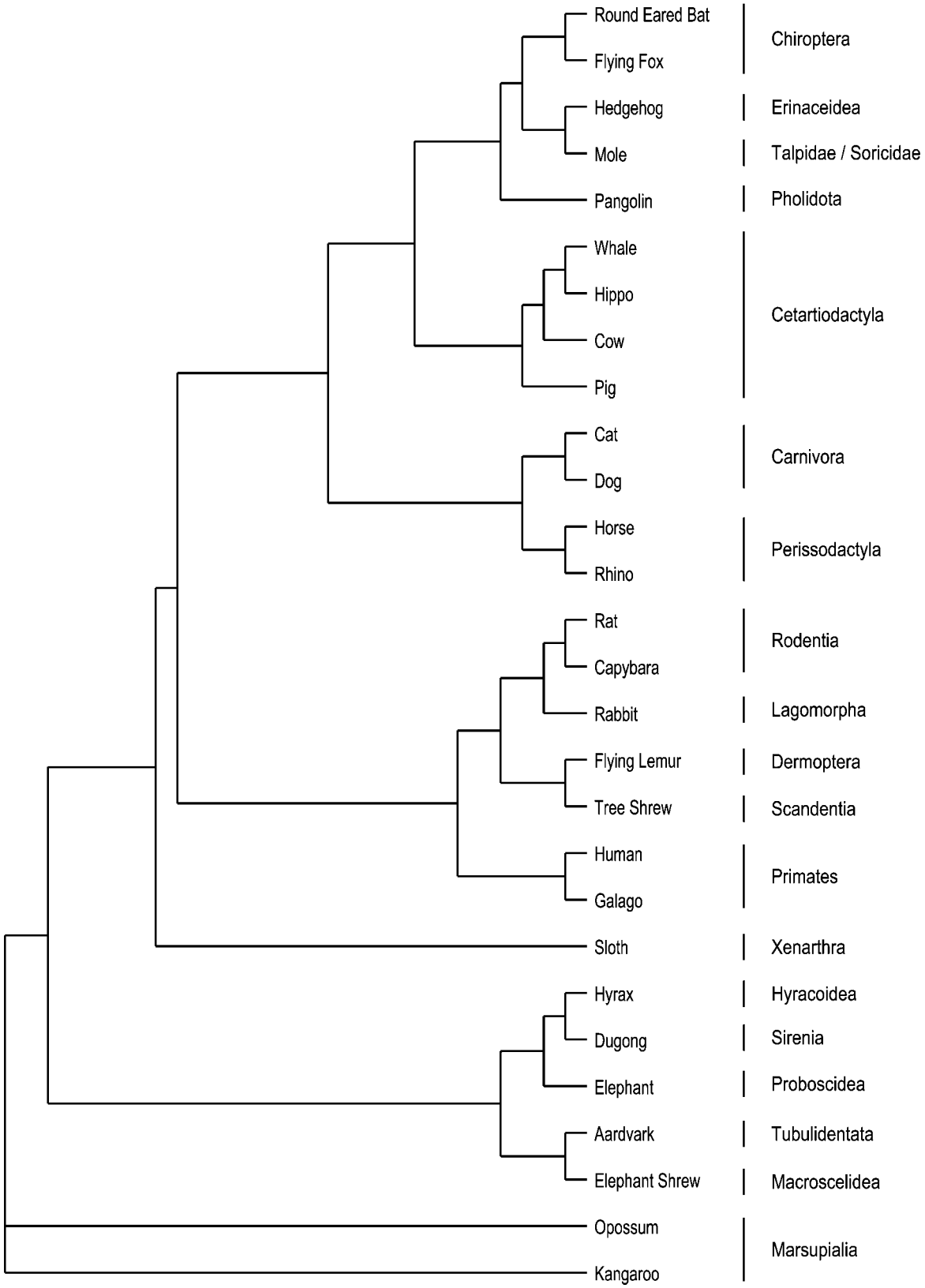

FIG. 4.—The Madsen tree topology for the Madsen data set.

Table 2
Results of the Mitochondrial Data Set, Assuming the N-GAM model and the Mitochondrial Tree (ML tree for this data set)

Model: Gene (positions)	Concatenate		Proportional			Separate	
	Log-likelihood	Alpha Parameter	Log-likelihood	Gene-Specific Rate	Alpha Parameter	Log-likelihood	Alpha Parameter
nd1(313). . . .	−8,157.72	0.55	−8,151.33	0.82	0.51	−8,079.67	0.52
nd2(313). . . .	−13,881.80	0.68	−13,857.77	1.43	0.83	−13,780.25	0.82
co1(512). . . .	−5,916.21	0.25	−5,840.83	0.29	0.26	−5,738.85	0.26
co2(225). . . .	−4,331.77	0.41	−4,323.39	0.71	0.48	−4,155.21	0.43
atp8(32)	−1,479.19	0.73	−1,471.97	1.48	0.98	−1,398.87	0.85
atp6(201) . . .	−5,527.05	0.55	−5,520.59	0.82	0.52	−5,407.84	0.52
co3(259). . . .	−4,922.87	0.37	−4,902.91	0.73	0.31	−4,827.74	0.31
nd3(104). . . .	−3,165.27	0.48	−3,163.01	1.20	0.49	−3,116.25	0.46
nd4l(94)	−3,307.18	0.70	−3,305.66	0.93	0.65	−3,218.65	0.60
nd4(438). . . .	−13,618.68	0.50	−13,606.69	1.27	0.50	−13,536.09	0.49
nd5(526). . . .	−17,624.30	0.53	−17,613.26	1.48	0.55	−17,519.18	0.55
Cytb(375). . .	−9,256.67	0.40	−9,241.90	0.83	0.37	−9,143.17	0.38
Sum(3392) . .	−91,188.71	—	−90,999.30	—	—	−89,921.78	—

stitution model. The alpha parameter of the gamma distribution was estimated using the ML method. The discrete gamma distribution with four categories was used (Yang 1994). A program implementing all the nine models described above was written in C++ and is attached as supplementary material at: http://www.molbiolevol.org/.

Procedures for calculation of the likelihood functions were adapted from the SEMPHY program (Friedman et al. 2001).

To compare the different models, the Akaike Information Criterion (AIC), defined as AIC = $-2 \times$ log-likelihood + $2 \times$ number of free parameters, was used (Sakamoto, Ishiguro, and Kitagawa 1986). A model with a lower AIC is considered more appropriate (Sakamoto, Ishiguro, and Kitagawa 1986). To evaluate if the AIC values of two models are significant, the test of Linhart (1988) was used. When comparing different tree topologies for the same model, the one-tailed Kishino-Hasegawa test was used (Kishino and Hasegawa 1989).

Results

We examined the effect of different branch length models as well as the number of gamma parameters (alpha) when combining different genes for phylogenetic analysis. For all data sets and all trees, the lowest AIC values (indicative of the best model) were achieved by assuming a different gamma parameter for each gene (the N-GAM model). This result held for all categories of branch length models. Further, AIC values were always lower when one gamma parameter for all genes was assumed (the 1-GAM model) than in the model that assumed no among-site rate variation (the homogenous model). Hence, we present detailed results (i.e., log-likelihoods, alpha parameter, and gene-specific rate factors) only for the best trees under the N-GAM model (tables 2–4). For all other cases, only the sums of the log-likelihoods and the AIC values are presented (tables 5–7).

Mitochondrial Data Set

The proportional method assigns a specific rate for each gene. This gene-specific rate can be used to rank the evolutionary rate among different genes. For example, in the mitochondrial tree, the N-GAM model yielded a gene-specific rate of 1.48 for the ATP8 gene, whereas that for cytochrome oxidase subunit 1 (the ''slowest'' gene) was 0.29 (table 2). Only minor changes in the gene-specific rates were observed when other tree topologies were assumed.

Table 3
Results of the Murphy Data Set, assuming the N-GAM model and the Madsen Tree (ML tree for this data set)

Model: Gene (positions)	Concatenate		Proportional			Separate	
	Log-likelihood	Alpha Parameter	Log-likelihood	Gene-Specific Rate	Alpha Parameter	Log-likelihood	Alpha Parameter
ADO3 (107) . .	−2,459.36	0.39	−2,435.45	2.59	0.59	−2,362.59	0.58
ATP7 (220) . . .	−4,433.48	0.86	−4,431.60	1.59	0.93	−4,388.23	0.90
BDNF (182) . .	−1,623.36	0.29	−1,608.58	0.57	0.32	−1,528.31	0.32
CNR1 (219). . .	−1,245.59	0.13	−1,228.21	0.31	0.12	−1,181.36	0.13
EDG1 (199). . .	−1,441.12	0.17	−1,435.97	0.69	0.17	−1,361.04	0.17
ZFY (67)	−415.76	0.16	−404.06	0.26	0.20	−370.58	0.20
Sum (994)	−11,618.67	—	−11,543.87	—	—	−11,192.12	—

Table 4
Results of the Madsen Data Set, Assuming the N-GAM model and the Murphy Tree (ML tree for this data set)

	CONCATENATE		PROPORTIONAL			SEPARATE	
MODEL: GENE (positions)	Log-likelihood	Alpha Parameter	Log-likelihood	Gene-Specific Rate	Alpha Parameter	Log-likelihood	Alpha Parameter
A2AB (344) . .							
BRCA1	−4,097.60	0.34	−4,028.88	0.47	0.29	−3,978.66	0.29
(557).	−15,655.40	2.56	−15,636.23	1.61	2.95	−15,575.42	2.82
IRBP (301) . . .	−5,178.58	0.61	−5,163.13	0.90	0.64	−5,079.71	0.64
VWF (338) . . .	−6,587.52	0.75	−6,578.57	1.03	0.77	−6,519.49	0.79
Sum (1540) . . .	−31,519.10	—	−31,406.81	—	—	−31,153.28	—

For the 12 mitochondrial genes studied (table 5), the AIC values obtained using the proportional model were significantly lower than those obtained with either the concatenate model or the separate model. For example, the N-GAM model for the mitochondrial tree yielded an AIC value of 182,262.6 for the proportional model, whereas the separate analysis and concatenate models gave AIC values of 182,483.55 and 182,619.42, respectively. This difference of 221 between the proportional model and the separate model is significant ($P < 0.05$). Thus, the proportional model is significantly better than both the separate model and the concatenate model. However, the AIC difference of 136 between the separate model and the concatenate model (table 5) is not statistically significant.

Among the four different tree topologies, for all models the most likely tree was the mitochondrial tree. For the N-GAM model with proportional branch length, the log-likelihood of the mitochondrial tree was −90,999.3, whereas the Murphy tree was second best at −91,022.96. This difference of 23.66 ± 47.84 corresponds to a P value of 0.31, using the Kishino-Hasegawa test (Kishino and Hasegawa 1989). Hence, when using the proportional model, the mitochondrial tree is not significantly different from the Murphy tree. Similar results were obtained when comparing the mitochondrial tree and the Madsen tree (P value of 0.21 using the Kishino-Hasegawa test). However, the morphological tree was rejected when compared with all other trees (log-likelihood difference > 742; $P < 0.001$).

Murphy Data Set

For the six nuclear genes studied, again the AIC values obtained using the proportional model were significantly lower than those obtained with either the concatenate model or the separate model. For example, the N-GAM model for the Madsen tree yielded an AIC value of 23,287.7 for the proportional model, whereas the separate analysis and concatenate models gave 23,464.2 and 23,427.3, respectively (table 6). This difference of 176.5 between the proportional model and the separate model is significant ($P < 0.05$). As for the mitochondrial data set, the AIC differences between the separate model and the concatenate model were not significant. The most likely tree topology for all models is the Madsen topology, in contrast with the observations of Murphy et al. (2001), except for the concatenate–N-GAM model, where the best tree is the Murphy tree. This discrepancy most likely arises from our use of amino acid data sets instead of nucleotide data sets, as well as from differences in alignment. However, the difference in log-likelihood between the Madsen topology and the Murphy topology is very small and nonsignificant in each case (i.e., log-likelihood difference < 10; table 6). For example, for the proportional–N-GAM model the log-likelihood difference between the Madsen and the Murphy tree topologies is 1.87 ± 8.22, corresponding to a P value of 0.41 by the Kishino-Hasegawa test. Assuming this model, both the mitochondrial tree and the morphological tree are significantly worse than the Madsen tree (log-likelihood difference > 44; $P < 0.01$).

Madsen Data Set

Unlike the results obtained for the other two data sets, the lowest AIC values for the Madsen data set were obtained with the separate model regardless of the number of gamma parameters assumed. For example, the N-GAM model for the Murphy tree yielded an AIC value of 62,738.6 for the separate model, whereas the proportional analysis and concatenate models gave 62,933.6 and 63,152.2, respectively. This difference of 195 between the proportional model and the separate model is significant ($P < 0.01$). Here, the proportional model is also significantly better than the concatenate model ($P < 0.01$). The most likely tree topology was obtained for the Murphy topology for all the models considered. Surprisingly, the Murphy tree appears to be significantly better than all other tree topologies. For example, for the N-GAM–separate model the log-likelihood difference between the Madsen (the second best tree) and Murphy tree topologies is 53.19 ± 20.62, corresponding to a P value of <0.05 by the Kishino-Hasegawa test.

Tree Search

In the above analyses, we investigated the differences between the supports of four predetermined tree topologies under nine different models. As can be seen from table 6, the model can have an impact on the best topology found. For the Murphy data set under the concatenate–N-GAM model, the Murphy tree appears to be

Table 5
AIC Values for the Mitochondrial Data set, Where df is the Number of Degrees of Freedom, LogL the Log-likelihood Value and AIC the AIC Value

Tree		Homogeneous			1-GAM			N-GAM		
		Concatenate	Proportional	Separate	Concatenate	Proportional	Separate	Concatenate	Proportional	Separate
	df	109	120	1,308	110	121	1,309	121	132	1,320
Morphological	LogL	−102,027.42	−100,246.07	−99,327.81	−92,081.43	−91,879.39	−90,869.66	−91,973.62	−91,786.99	−90,778.76
	AIC	204,272.84	200,732.14	201,271.63	184,382.86	184,000.78	184,357.31	184,189.24	183,837.98	184,197.52
Mitochondrial	LogL	−100,722.51	−98,998.68	−98,009.91	−91,292.16	−91,094.30	−90,014.84	−91,188.71	−90,999.30	−89,921.78
	AIC	201,663.03	198,237.37	198,635.83	182,804.31	182,430.61	182,647.69	182,619.42	182,262.60	182,483.55
Madsen	LogL	−100,949.52	−99,192.90	−98,199.89	−91,332.05	−91,138.47	−90,056.74	−91,225.67	−91,044.51	−89,966.39
	AIC	202,117.04	198,625.80	199,015.77	182,884.11	182,518.94	182,731.47	182,693.34	182,353.02	182,572.79
Murphy	LogL	−100,886.44	−99,142.72	−98,152.00	−91,306.67	−91,118.49	−90,019.26	−91,193.30	−91,022.96	−89,927.73
	AIC	201,990.87	198,525.44	198,920.00	182,833.34	182,478.97	182,656.53	182,628.61	182,309.91	182,495.45

the best, whereas under all the other models the Madsen tree is the best.

To further determine the effect of the model on tree topology, we implemented a tree-search algorithm to find the most likely tree under each of the models. Because of computational limitations, the tree search was conducted on 14 taxa representing the main mammalian clades, using a subset of the mitochondrial data set (fig. 5). Starting with various starting points (a neighbor-joining tree, a tree based on the mitochondrial topology, and a tree based on the Murphy topology), we searched the tree space for better trees through the nearest neighbors interchange (NNI) algorithm. We also limited our searches to the two best models found above (i.e., the proportional–N-GAM model and the separate–N-GAM model). The alpha parameters and the gene-specific rates for this search were based on the corresponding N-GAM analysis of the complete mitochondrial data set.

We found different best trees under these two models. The ML tree under the N-GAM–proportional is given in figure 5*A*, whereas the ML tree under the N-GAM–separate is given in figure 5*B*. Under the proportional–N-GAM model the glires (rodents and rabbit) are monophyletic. The glires are related to a man + tree shrew clade. In the separate model the tree-shrew clusters with rabbit rather than with man; hence, the glires are not monophyletic. Another difference concerns the Laurasiatheria. In the N-GAM–proportional model the Laurasiatheria are divided into two clades. The first includes mole and bat. The second includes whale, horse, and cat. In this second clade, horse and whale cluster together. On the contrary, in the N-GAM–separate model, whale is the first diverging taxa of the Laurasiatheria. However, in each model the differences in log-likelihood between these two topologies are not significantly different: under the proportional model the log-likelihood difference between the two trees is 13.3, whereas under the separate model the difference in log-likelihood between the two trees is 2.15. These differences are not significant based on the Kishino-Hasegawa test ($P > 0.05$; results not shown). Nevertheless, these results demonstrate that choosing among alternative models can lead to different best trees. Thus, choosing the best model is important not only to model the molecular evolutionary process better but also for inferring phylogenetic relationships among taxa in general.

Discussion

Assigning Gene-Specific Rates for Genes

In the proportional model a specific rate is assigned to each gene. These rates can be used to classify genes, and estimating relative rates of sequence evolution can be used to determine whether a gene is relevant for a specific phylogenetic analysis. Conserved genes (i.e., genes with lower evolutionary rates) are expected to give better estimates for deeper evolutionary divergences, whereas fast-evolving genes can be used to resolve recent divergences. For example, Corneli and Ward (2000) used sequence similarity plots to determine empirically the relative rates of evolution of mitochondrial

Table 6
AIC Values for the Murphy Data set, Where df is the Number of Degrees of Freedom, LogL the Log-likelihood Value and AIC the AIC Value

Tree		Homogeneous			1-GAM			N-GAM		
		Concatenate	Proportional	Separate	Concatenate	Proportional	Separate	Concatenate	Proportional	Separate
	df	89	94	534	90	95	535	95	100	540
Morphological	LogL	−12,832.62	−12,280.51	−11,946.67	−11,840.86	−11,744.52	−11,398.65	−11,759.04	−11,685.26	−11,338.58
	AIC	25,843.24	24,749.02	24,961.33	23,861.72	23,679.03	23,867.29	23,708.08	23,570.51	23,757.16
Mitochondrial	LogL	−12,676.01	−12,150.79	−11,806.49	−11,742.81	−11,646.79	−11,294.09	−11,658.77	−11,587.42	−11,233.06
	AIC	25,530.02	24,489.58	24,680.98	23,665.63	23,483.59	23,658.17	23,507.54	23,374.83	23,546.12
Madsen	LogL	−12,617.49	−12,089.84	−11,747.83	−11,700.02	−11,600.78	−11,250.58	−11,618.67	−11,543.87	−11,192.12
	AIC	25,412.98	24,367.68	24,563.65	23,580.05	23,391.56	23,571.16	23,427.33	23,287.74	23,464.23
Murphy	LogL	−12,620.83	−12,096.87	−11,756.97	−11,700.70	−11,603.23	−11,255.29	−11,618.24	−11,545.75	−11,195.93
	AIC	25,419.66	24,381.75	24,581.94	23,581.4	23,396.46	23,580.57	23,426.48	23,291.49	23,471.86

genes. The advantages of our method for determining gene-specific evolutionary rates lie in its statistical robustness (it is a maximum-likelihood estimate of the gene-specific rate) and its ability to combine specific model assumptions (the gamma distribution, the substitution model) into the estimation scheme. Furthermore, our method takes tree topology into consideration, whereas sequence similarity plots do not.

Model Selection

Currently, the method of concatenating sequences is most frequently used for analyzing multiple data sets (Reyes, Pesole, and Saccone 2000; Madsen et al. 2001; Murphy et al. 2001). Conversely, our results show that both the proportional model and the separate model yield results that are superior to the concatenate model. This is in agreement with the result of Cao et al. (1998), who showed that for mitochondrial data, separate analysis is superior to the concatenate method. However, selecting the superior of the separate and proportional models is a more complicated issue. The proportional model is more appropriate for analysis of the mitochondrial and Murphy data sets, whereas the separate model is favored for the Madsen data set. The separate analysis is preferable when each gene is of sufficient length to allow accurate determination of a separate set of branch lengths. In our study the sequence lengths in the Madsen data set were longer than those in the Murphy or mitochondrial data sets. When only the first 100 positions in each gene of the Madsen data set were analyzed, the proportional model was the best model (data not shown). Another factor that can affect model selection is inconsistent branch length among different genes in a data set. If, for example, significant rate acceleration occurred for only one gene in a specific clade, then analysis by the separate model would be justified. Thus, it is expected that the proportional model would be most appropriate when analyzing closely related sequences. Yang (1996*b*) showed that the proportional model was superior to the separate analysis for a data set composed of four parts: the first-, second-, and third-codon positions of six mitochondrial genes, and a fourth part composed of 11 tRNA sequences. We speculate that this proportional model preference is due to the fact that Yang's analysis considered only primate sequences.

In our analyses it proved more fruitful to assume a different gamma parameter for each gene rather than a single gamma parameter for all genes. Each gamma distribution involves a shape parameter alpha. This alpha parameter determines the shape of the gamma distribution and is inversely related to the extent of rate variation among sites. Some genes show substantial rate variance, whereas others exhibit a more homogenous distribution of the rate of different positions. For example, for the Murphy data set, assuming the best model, the alpha parameter ranges from 0.93 for the ATP7 gene to 0.12 for the CNR1 gene (see table 3). We conclude that the substantial difference in rate distribution among genes is sufficient to justify a separate gamma parameter for each gene.

Table 7
AIC Values for the Madsen Data set, Where df is the Number of Degrees of Freedom, LogL the Log-likelihood Value and AIC the AIC Value

Tree		Homogeneous			1-GAM			N-GAM		
		Concatenate	Proportional	Separate	Concatenate	Proportional	Separate	Concatenate	Proportional	Separate
	df	53	56	212	54	57	213	57	60	216
Morphological	LogL	−33,712.48	−33,200.38	−32,974.87	−32,390.67	−32,297.30	−32,057.79	−32,272.40	−32,156.03	−31,921.08
	AIC	67,530.96	66,512.76	66,373.74	64,889.34	64,708.60	64,541.58	64,658.80	64,432.07	64,274.16
Mitochondrial	LogL	−32,977.19	−32,473.50	−32,231.85	−31,767.63	−31,676.91	−31,423.53	−31,644.52	−31,530.83	−31,283.07
	AIC	66,060.38	65,059.00	64,887.70	63,643.26	63,467.82	63,273.06	63,403.05	63,181.67	62,998.14
Madsen	LogL	−32,871.70	−32,375.32	−32,133.17	−31,697.03	−31,601.83	−31,346.64	−31,565.68	−31,457.24	−31,206.50
	AIC	65,849.40	64,862.64	64,690.33	63,502.06	63,317.66	63,119.27	63,245.35	63,034.47	62,845.00
Murphy	LogL	−32,811.40	−32,314.62	−32,070.52	−31,647.93	−31,550.48	−31,292.80	−31,519.10	−31,406.81	−31,153.28
	AIC	65,728.80	64,741.24	64,565.04	63,403.86	63,214.96	63,011.60	63,152.21	62,933.63	62,738.56

Regarding the effect of the model on tree selection, our results show that the model chosen has an effect on tree topology. It is expected that the model would also affect bootstrap support for different clades, and molecular date estimation based on several genes. More simulation studies and improvements in computational techniques are required to explore fully the effect of these different models on phylogeny reconstruction.

Before selecting a model that combines different genes, one must consider whether there is a basis for combining the genes of interest in the first place. To address this issue, Huelsenbeck and Bull (1996) proposed a likelihood ratio test designed to detect conflicting phylogenetic signals among genes. Regarding the genes used in our study, we followed Cao et al. (2000), Madsen et al. (2001), and Murphy et al. (2001) and assumed that there is agreement between the gene tree and the species tree. Of course, before any analysis of a new data set, such an assumption should be verified (for review see Huelsenbeck, Bull, and Cunningham 1996).

Mammalian Phylogeny

For all the models and data sets considered in our study, the morphological tree exhibited significantly lowest log-likelihood values (results of the Kishino-Hasegawa test not shown). Many traditional morphological clades are not supported by molecular phylogeny analysis (see Springer et al. 1997, 1999; Murphy et al. 2001), as exemplified by the clades Archonta (bats and primates), Anagalida (elephant shrew and glires), and Ungulata (aardvark, horses, cows, whales, elephants, dugongs, and hyraxes). Interestingly, the McKenna tree (McKenna and Bell 1997) has also been challenged by recent morphological discoveries. For example, Thewissen et al. (2001) confirmed a close relationship between Cetacea and Artiodactyls, whereas Cetacea was previously considered as a sister clade of Mesonychia.

Both the mitochondrial and the nuclear data sets support their respective trees for all the models considered. Our results for the mitochondrial data set show that there is no significant difference between the mitochondrial tree and the nuclear tree with regard to likelihood when using the 1-GAM or the N-GAM models ($P > 0.05$; results of the Kishino-Hasegawa test not shown). However, with the homogenous models the mitochondrial tree was found to be significantly better than both the Madsen and the Murphy trees ($P < 0.03$; results of the Kishino-Hasegawa test not shown). This is in agreement with Sullivan and Swofford (1997), who showed that simplified models could lead to systematic errors.

Our results for the two nuclear data sets reject the mitochondrial tree for all the models considered ($P < 0.05$; results of the Kishino-Hasegawa test not shown). Thus, the nuclear data sets discriminate more than the mitochondrial data set between alternative topologies. Hence, it is apparent that there is more ''phylogenetic signal'' in the nuclear genes (e.g., Springer et al. 2001). The main differences between the mitochondrial tree and the nuclear trees are that (1) Eulipotyphla insecti-

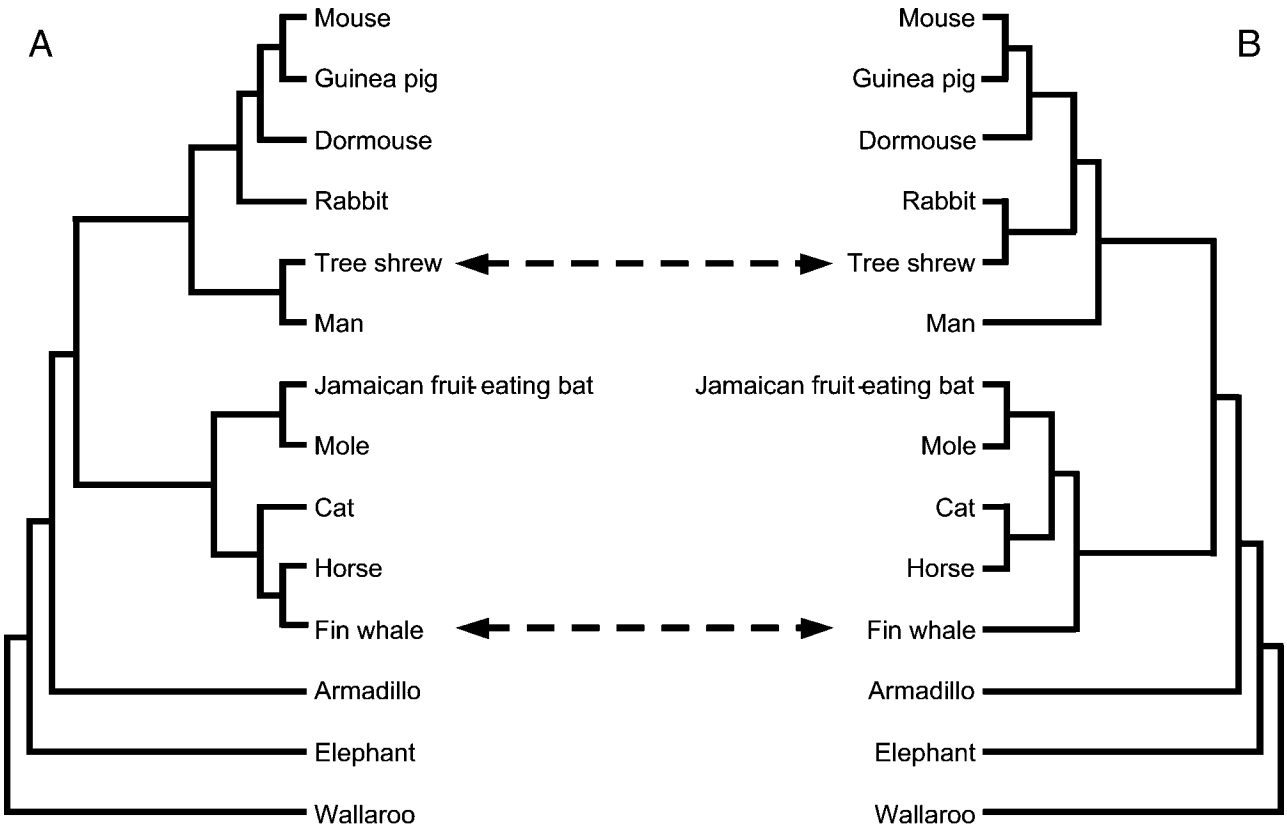

FIG. 5.—ML tree obtained after NNI search on a subset of the mitochondrial data set. *A*, The best tree obtained under the proportional–N-GAM model. *B*, The best tree obtained under the separate–N-GAM model. Arrows indicate taxa with different positions in the two trees.

vores (hedgehogs, moles, shrews) are paraphyletic in the mitochondrial tree and the Erinaceidae (hedgehogs) are the most basal mammalian taxa in the mitochondrial tree; (2) glires and Euarchonta (primates, flying lemur, tree shrews) do not cluster in a single clade (the Euarchontoglires) in the mitochondrial tree but appear paraphyletic at the base of the placental tree; (3) rodents are paraphyletic in the mitochondrial tree and monophyletic in the nuclear tree; and (4) consequently, Afrotheria (armadillos, anteaters, and sloths) and Xenarthra are at the base of the placental trees but have a more internal position in the mitochondrial tree.

When comparing the two nuclear trees, the Madsen data set supports the Murphy tree, and the Murphy data set supports the Madsen tree (for eight out of the nine models). For the Murphy data set the differences are not significant; however, the Madsen data set significantly supports the Murphy tree. Both trees support the same topology between the four main clades, Laurasiatheria, Euarchontoglires, Xenarthra, and Afrotheria, and any differences concern only the relationships among these four clades. It is worth noting that the full NNI tree search on the subset of the mitochondrial data set led to a tree supporting these four main clades as well as the rodent monophyly. Our results suggest that the Murphy tree is probably closer to the ''true tree'' than is the Madsen tree. However, we speculate that the true tree lies between these two alternative nuclear trees, and additional gene sequences and the development of better models will help to address these questions.

Acknowledgments

We thank Nir Friedman and Nicolas Galtier for helpful discussions. Ross Crozier and two anonymous referees provided helpful comments on this paper. T.P. is supported by a grant from the Japanese Society for the Promotion of Science (JSPS), and D.H. is supported by a Lavoisier grant from the French Ministry of Foreign Affairs. This work was partially supported by grants from the JSPS and Monbusho to M.H.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. **28**:1–150.

ARNASON, U., A. GULLBERG, S. GRETARSDOTTIR, B. URSING, and A. JANKE. 2000. The mitochondrial genome of the sperm whale and a new molecular reference for estimating eutherian divergence dates. J. Mol. Evol. **50**:569–578.

BURNHAM, K. P., and D. R. ANDERSON. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York.

CAO, Y., M. FUJIWARA, M. NIKAIDO, N. OKADA, and M. HASEGAWA. 2000. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. Gene **259**:149–158.

CAO, Y., A. JANKE, P. J. WADDELL, M. WESTERMAN, O. TAKENAKA, S. MURATA, N. OKADA, S. PÄÄBO, and M. HASEGAWA. 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. **47**:307–322.

CORNELI, P. S., and R. H. WARD. 2000. Mitochondrial genes and mammalian phylogenies: increasing the reliability of branch length estimation. Mol. Biol. Evol. **17**:224–234.

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

FRIEDMAN, N., M. NINIO, I. PE'ER, and T. PUPKO. 2001. A structural EM algorithm for phylogenetic inference. Pp. 132–140 *in* T. LENGAUER, D. SANKOFF, S. ISTRAIL, P. PEVZNER, and M. WATERMAN, eds. Proceedings of the Fifth Annual International Conference on Computational Biology. ACM Press, New York.

GRAUR, D., and W. H. LI. 1999. Fundamentals of molecular evolution. 2nd edition. Sinauer Press, Sunderland, Mass.

HUELSENBECK, J. P., and J. J. BULL. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. Syst. Biol. **45**: 92–98.

HUELSENBECK, J. P., J. J. BULL, and C. W. CUNNINGHAM. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. **11**:152–157.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. **29**:170–179.

LINHART, H. 1988. A test whether two AIC's differ significantly. S. Afr. Stat. J. **22**:153–161.

MADSEN, O., M. SCALLY, C. J. DOUADY, D. J. KAO, R. W. DEBRY, R. ADKINS, H. M. AMRINE, M. J. STANHOPE, W. W. DE JONG, and M. S. SPRINGER. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature **409**:610–614.

MCKENNA, M. C., and S. K. BELL. 1997. Classification of mammals above the species level. Columbia University Press, New York.

MOUCHATY, S. K., F. M. CATZEFLIS, A. JANKE, and U. ARNASON. 2001. Molecular evidence of an African Phiomorpha-South-American Caviomorpha clade and support for Hystricognathi based on the complete mitochondrial genome of cane rat (*Thryonomys swinderianus*). Mol. Phylogenet. Evol. **18**:127–135.

MURPHY, W. J., E. EIZIRIK, W. E. JOHNSON, Y. P. ZHANG, O. A. RYDER, and S. J. O'BRIEN. 2001. Molecular phylogenetics and the origins of placental mammals. Nature **409**: 614–618.

NEI, M., and S. KUMAR. 2000. Molecular evolution and phylogenetics. Oxford University Press, New York.

NIKAIDO, M., K. KAWAI, Y. CAO, M. HARADA, S. TOMITA, N. OKADA, and M. HASEGAWA. 2001. Maximum likelihood analysis of the complete mitochondrial genomes of eutherians and a reevaluation of the phylogeny of bats and insectivores. J. Mol. Evol. **53**:508–516.

NOVACEK, M. J. 1992. Mammalian phylogeny: shaking the tree. Nature **356**:121–125.

PHILLIPS, M. J., Y.-H. LIN, G. HARRISON, and D. PENNY. 2001. Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials. Proc. R. Soc. Lond. B **268**:1533–1538.

REYES, A., G. PESOLE, and C. SACCONE. 2000. Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. Gene **259**:177–187.

SAKAMOTO, Y., M. ISHIGURO, and G. KITAGAWA. 1986. Akaike information criterion statistics. Reidel, Dordrecht, The Netherlands.

SCHMITZ, J., M. OHME, and H. ZISCHLER. 2000. The complete mitochondrial genome of Tupaia belangeri and the phylogenetic affiliation of Scandentia to other Eutherian orders. Mol. Biol. Evol. **17**:1334–1343.

SPRINGER, M. S., H. M. AMRINE, A. BURK, 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.

SPRINGER, M. S., A. BURK, J. R. KAVANAGH, V. G. WADDELL, and M. J. STANHOPE. 1997. The interphotoreceptor retinoid binding protein gene in therian mammals: implications for higher level relationships and evidence for loss of function in the marsupial mole. Proc. Natl. Acad. Sci. USA **94**: 13754–13759.

SPRINGER, M. S., R. W. DEBRY, C. DOUADY, H. M. AMRINE, O. MADSEN, W. W. DE JONG, and M. J. STANHOPE. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. Mol. Biol. Evol. **18**: 132–143.

SULLIVAN, J., and D. L. SWOFFORD. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. **4**:77–86.

TAKAHASHI, K., and M. NEI. 2000. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol. Biol. Evol. **17**:1251–1258.

THEWISSEN, J. G. M., E. M. WILLIAMS, J. L. ROE, and S. T. HUSSAIN. 2001. Skeletons of terrestrial cetaceans and the relationships of whales to artiodactyls. Nature **413**:277–281.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

YANG, Z. 1996*a*. Among-site rate variation and its impact on phylogenetics analysis. Trends Ecol. Evol. **11**:367–372.

YANG, Z. 1996*b*. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. **42**:587–596.

ROSS CROZIER, reviewing editor

Accepted August 22, 2002