Sullivan, J. & D. L. Swofford. 1997. J. Mamm. Evol., 4:77-86

Are Guinea Pigs Rodents? The Importance of Adequate Models in Molecular Phylogenetics

Jack Sullivan and David L. Swofford

Laboratory of Molecular Systematics, MSC, Smithsonian Institution, MRC-534
Washington, DC. 20560, USA

Address for Correspondence:
Jack Sullivan
Laboratory of Molecular Systematics
Smithsonian Museum Support Center
4210 Silver Hill Rd.
Suitland, MD 20746
E-mail: sullivan@onyx.si.edu
Phone: (301) 238-3444 χ102
Fax: 301-238-3059

The monophyly of Rodentia has repeatedly been challenged based on several studies of molecular sequence data. Most recently, D'Erchia et al. (1996) analyzed complete mtDNA sequences of sixteen mammals and concluded that rodents are not monophyletic. We have reanalyzed these data using maximum-likelihood methods. We use two methods to test for significance of differences among alternative topologies, and show that: 1) models that incorporate variation in evolutionary rates across sites fit the data dramatically better than models used in the original analyses, 2) that the mtDNA data fail to refute rodent monophyly, and 3) the original interpretation of strong support for nonmonophyly results from systematic error associated with an oversimplified model of sequence evolution. These analyses illustrate the importance of incorporating recent theoretical advances into molecular phylogenetic analyses, especially when results of these analyses conflict with classical hypotheses of relationships.

KEY WORDS: Inconsistency, Maximum likelihood, Molecular systematics, Rodents, Rate Heterogeneity.

INTRODUCTION

The assertions made in several molecular phylogenetic studies (Graur et al., 1991; Li et al., 1992; Ma et al., 1993) have led to the growing acceptance of the conclusion that the order Rodentia is not monophyletic, in spite of the facts that these data sets essentially provide no significant refutation of the classical hypothesis (e.g., Hasegawa et al., 1992; Cao et al., 1994), and other molecular studies actually support rodent monophyly (Martignetti and Brosius, 1993; Porter et al., 1996). Recently, D'Erchia et al. (1996) suggested that their phylogenetic analyses of complete mtDNA sequences of 16 species firmly establish that the guinea pig is not a rodent based on its placement as a sister taxon to a clade containing Lagomorpha, Carnivora, Primates, Perissodactyla, and Artiodactyla (including cetaceans), rather than in a clade with mouse and rat. They claim that this placement is both consistent across phylogenetic reconstruction methodologies and is supported by “very significant” bootstrap values. Because nonmonophyly of the rodents would imply a remarkable amount of convergence in morphology (including the masticatory apparatus, cranial and post-cranial skeletal characters, and placentation patterns; reviewed in Luckett and Hartenberger, 1993), the possibility that the results of D'Erchia et al. (1996) stem from systematic error in the mtDNA analyses must be explored.

The purpose of this paper is to assess the degree to which the conclusions of D'Erchia et al. (1996) are supported by their data. These authors estimated evolutionary trees using parsimony, distance, and maximum-likelihood methods. A frequent criticism of parsimony methods is their susceptibility to systematic error associated with "long-branch attraction" and related phenomena (e.g., Felsenstein, 1978; Hendy and Penny, 1989). Distance and likelihood methods have the advantage of being based on explicit models of evolutionary change, but recent studies have illustrated that even model-based methods are not immune to being inconsistent estimators of phylogeny when their assumptions are strongly violated (Gaut and Lewis, 1995; Waddell, 1995; Yang, 1996). The inconsistency of likelihood analysis under an oversimplified evolutionary model is demonstrated in Figure 1. Sequences were simulated on the tree shown under a Jukes-Cantor model (Jukes and Cantor, 1969) with 50% invariable sites (JC + I). The upper curve represents the probability of inferring the true tree with increasing sequence length when the reconstruction model fits the data, and the lower curve represents the probability of inferring the true tree with increasing sequence lengths using likelihood under an assumption of an equal-rates Jukes-Cantor model. Clearly, maximum likelihood is inconsistent when all sites are incorrectly assumed to be free to vary.

The model-based analyses that D'Erchia et al. (1996) conducted on nucleotide sequences used an evolutionary model that allows both unequal base frequencies and different probabilities for each of the six possible transformations (A⇔C, A⇔G, A⇔T, C⇔G, C⇔T, G⇔T). However their model assumes all sites are variable and evolve at the same rate. This assumption is clearly violated in mammalian mtDNA (e.g., Yang et al., 1994; Sullivan et al., 1995). Our reanalysis of their “protein super-gene” data set (combined data from all mitochondrial protein genes) reveals that the evidence for excluding the guinea pig from the rodent clade is strongly overstated and is attributable D'Erchia et al.'s use of an oversimplified model.

METHODS

We chose to focus on the "protein supergene" data set because it provided stronger apparent refutation of rodent monophyly than the "rRNA supergene" data set. We omitted third-codon positions because these sites strongly violate the assumption of stationary base frequencies made by nearly all phylogenetic methods (either explicitly or implicitly); this data set is therefore very similar to the data set containing only non-synonymous substitutions analyzed by D'Erchia et al. (1996). Because the MOLPHY program (Adachi and Hasegawa, 1996) used in D'Erchia et al.'s maximum-likelihood analysis does not allow for among-site rate heterogeneity, we performed maximum-likelihood analyses at the nucleotide (rather than amino acid) level. All analyses were conducted using test versions of the PAUP* computer program (4.0d46 - 4.0d53) written by one of us (DLS).

Our emphasis on maximum-likelihood methods is motivated by two considerations. First, it has been shown that maximum likelihood is a consistent estimator of phylogeny over a larger set of conditions than is parsimony (e.g., Huelsenbeck, 1995). Second, a major advantage of likelihood relative to parsimony or distance methods is that the likelihood score provides an objective criterion of goodness-of-fit between model and data that is comparable across models. This property provides a means for choosing an appropriate reconstruction model for phylogenetic analysis. Under parsimony, the optimality criterion (tree length) is not directly comparable across weighting schemes; this makes choice of weighting schemes (including the choice of equal weights) under the parsimony framework somewhat arbitrary.

We examined four substitution models: Jukes-Cantor (JC; Jukes and Cantor, 1969), Kimura two-parameter (K2P; Kimura, 1980), Hasegawa-Kishino-Yano (HKY-85; Hasegawa et al., 1985), and general time-reversible (GTR, equals REV of Yang, 1994a). The JC model assumes that all six transformations (A⇔C, A⇔G, A⇔T, C⇔G, C⇔T, G⇔T) have equal probability, and that all four nucleotides are present in equal frequencies. The K2P model also assumes equal base frequencies, but allows different probabilities for transitions and transversions (i.e., a transition bias). The HKY-85 model also allows for a transition bias and, further, relaxes the assumption of equal base frequencies. The GTR model allows unequal base frequencies and allows a unique probability for each of the six possible transformations.

In addition, four models of among-site rate heterogeneity were examined: 1) equal rates assumed at all sites; 2) a proportion of sites estimated to be invariable, with equal rates assumed at variable sites ("I", Hasegawa et al., 1985); 3) rates at all sites assumed to follow a discrete approximation of the gamma distribution ("Γ", Yang, 1994b); 4) and some sites assumed to be invariable, with gamma-distributed rates at variable sites ("I+Γ"; Gu et al., 1995). Thus, 16 models of sequence evolution were examined (four substitution models, each with four rate-heterogeneity models), each of which is a special case of the most parameter-rich model, GTR+I+Γ. The assumptions made by each of these models are compared in the appendix (see Swofford et al., 1996, for a more detailed description of models).

Separate heuristic tree-searches (stepwise addition of taxa, 10 random input orders, and TBR branch swapping) were conducted under the equal-rates GTR model (equivalent to the model used by D'Erchia et al., 1996) and under the heterogeneous-rates model with the best fit to the data (GTR+I+Γ; see below). An initial search was conducted with model parameters fixed to values estimated using the topology of D'Erchia et al. (1996). These parameters were then reoptimized on the resulting tree to refine the model further for subsequent tree searches. Both unconstrained searches and searches constrained for rodent monophyly were conducted, and the significance of differences in likelihood scores of alternative topologies was examined using the test of Kishino and Hasegawa (1989) and a simulation method similar to the parametric bootstrap (Huelsenbeck et al., 1996).

The Kishino-Hasegawa test uses the standard error of differences in single-site likelihoods between two trees to estimate the significance of an observed difference between them (under the assumption that the distribution of single-site likelihood differences approximates a normal distribution; Kishino and Hasegawa, 1989). The simulation method estimates the probability of observing a particular result under a model of sequence evolution that is estimated from the data, assuming some hypothesis is true. In this case, it is important to ask how often a tree on which rodents are monophyletic would be expected to generate data that appear to reject rodent monophyly. The best fit tree (under the likelihood criterion) supporting rodent monophyly was used as the model (true) tree and 100 replicate data sets were simulated under the best-fit (GTR+I+Γ; see below) model estimated from the original data. Each of these simulated data sets was then subjected to heuristic searches under parsimony (10 random input orders) and likelihood (simple addition sequence) using the GTR+I+Γ and GTR equal rates model (see below). The proportion of replicates in which rodents are non-monophyletic represents the probability of incorrectly inferring non-monophyly if the best rodent monophyly tree were the true tree.

RESULTS AND DISCUSSION

Model Choice

The log-likelihood score of the tree of D'Erchia et al. (1996) under the GTR model, (equal rates; equivalent to their model) is –49028.14, whereas its score under the GTR+I+Γ is –45593.28 (see appendix for likelihood of this tree under all models tested). This very large difference in likelihood score (3457.36 log-likelihood units) demonstrates that the fit between the data and the reconstruction model is dramatically improved by assuming that a proportion of sites are not free to vary and that rates at the remaining sites follow a gamma distribution (with shape parameter estimated from the data). The significance of this improvement can be evaluated using a likelihood-ratio test. The test statistic is twice the difference in log-likelihood and this can be compared to the χ^2-distribution, with degrees of freedom equal to the difference in the number of free parameters between the two models (two in this case: the proportion of sites that are invariable and the gamma-distribution shape parameter). The value of the test statistic is 6914.72, whereas the critical value (at the 0.001 significance level) is 13.82. Although the assumptions of this test may not be strictly met (Goldman, 1993; but see Yang et al., 1995), it nevertheless highlights the dramatic improvement in fit associated with allowing for heterogeneous rates.

Phylogenetic Analyses

Maximum-Likelihood Analyses Using Different Models Produce Different Trees

Under the equal-rates GTR model, the tree shown by D'Erchia et al. (1996) is indeed the maximum-likelihood tree, but the maximum-likelihood tree under the more appropriate heterogeneous-rates GTR model (Fig. 2A) no longer supports the basal placement of the hedgehog. Furthermore, the hedgehog branch is remarkably long, longer even than the branch leading to the outgroup (opossum). The possibility of artifactual results due to long-branch attraction is well-known in parsimony analysis (Felsenstein, 1978), but the same problem can affect distance and maximum-likelihood analyses if the amount of change in long branches is systematically underestimated. This will be the case when among-site rate variation is ignored (Fig. 1; Waddell, 1995; Yang, 1996), and very likely explains the basal position of hedgehog in Cao et al.'s (1997) analysis of the mtDNA data (those authors also assumed equal rates for all amino-acid positions within each gene, but allowed a different uniform rate for each gene). Even when heterogeneous rates are accommodated, the hedgehog can be placed on the tree in several locations without changing the likelihood significantly (Table I; Fig. 2A). The hedgehog therefore appears to represent a “rogue” taxon that cannot be placed reliably with these data, and which possibly confounds attempts to estimate the relationships among the remaining taxa.

With the hedgehog/opossum long-branch attraction broken up (Fig. 2A), the unrooted tree for placental mammals (the ingroup topology obtained by pruning the opossum lineage from Fig. 2A) is consistent with rodent monophyly. The issue reduces to how the ingroup topology of placental mammals is rooted, that is, the reliability of the placement of the outgroup (the opossum sequence). In other molecular studies (e.g., Stanhope et al., 1992), variable sites in opossum sequences have been shown to be essentially randomized relative to placental mammals, rendering the opossum sequences unlikely to provide a reliable root for those data. To examine this possibility specifically with respect to these data, we rooted the ingroup topology of Figure 2A with 100 random sequences under the parsimony criterion (after exclusion of the hedgehog, to avoid the effect of long-branch attraction). These sequences were generated with MacClade (Maddison and Maddison, 1992), and were constrained to have the same base frequencies as the opossum sequence. Ninety-eight of those random outgroups rooted the tree within Rodentia, and many (24) rooted the tree at the same location as in D'Erchia et al.’s (1996) analyses. The observation that random sequences nearly always root the tree at or near the same location as the opossum sequence raises the possibility of a spurious rooting in D'Erchia et al's analyses. This tendency of the opossum sequence towards random rooting behavior, combined with the long-branch attraction between hedgehog and opossum sequences, contributed to their apparent strong refutation of rodent monophyly.

Rodent Monophyly is Not Significantly Refuted

More importantly, with all taxa included and the analysis conducted under the more appropriate (GTR+I+Γ) model, the likelihood score of the best rodent monophyly tree (Fig. 2B) is only 0.02% worse than that of the maximum-likelihood tree (Fig. 2A). This is not a significant difference as judged by the Kishino-Hasegawa test ($P > 0.1$). Interestingly, when the Kishino-Hasegawa test is applied to the constrained (for rodent monophyly) and unconstrained trees under the (inappropriate) equal rates GTR model (used by D'Erchia et al., 1996), the best rodent monophyly tree is significantly worse than their tree ($P = 0.012$). Thus, the apparent strong support for non-monophyly of rodents reported by D'Erchia et al. (1996) is attributable to their use of an oversimplified model, that is, their incorrect assumption of equal rates across sites.

The results of the simulation analyses are similar. When the best rodent monophyly tree (Fig. 2B) is used as the model (true) tree and 100 data sets are simulated on that tree, maximum-likelihood analysis of the resulting data sets (under the model used to generate the data, GTR+I+Γ) supports rodent monophyly only 75% of the time. That is, in 25% of the simulations, a non-monophyletic Rodentia is supported, even though rodents are monophyletic on the model tree. We therefore cannot reject the null hypothesis ($P = 0.25$) of no significant difference between the topologies presented in Figure 2. Interestingly, when the simulated data sets are analyzed under parsimony, the probability of incorrectly inferring non-monophyly of Rodentia increases to 0.68, and when analyzed using an equal rates likelihood model (GTR with equal rates) that probability increases further to 0.85. In the oversimplified analyses, there is actually a higher probability of inferring the wrong tree than the model (true) tree. This analysis further demonstrates that the strong apparent support for rodent nonmonophyly reported by D'Erchia et al (1996) resulted from ignoring rate heterogeneity across sites.

The Utility of Models

Our analyses using the more appropriate heterogeneous-rates model (GTR+I+Γ) not only break up the long-branch attraction between the hedgehog and opossum sequences, but demonstrate that the statistical support (relatively high bootstrap values) that D'Erchia et al. (1996) reported for rodent non-monophyly is an artifact of systematic error associated with ignoring among-site rate variation. This illustrates the point that large data sets (with respect to the number of bases) are not immune to systematic error. In fact, because inconsistent phylogenetic methods will (by definition) ascribe increasing confidence to incorrect estimates of topology as sequence length increases, the match between model and data becomes more critical for very long sequences rather than less so, as intuition might suggest. Thus, future analyses of complete mitochondrial genomes that fail to accommodate rate heterogeneity explicitly (e.g., Janke et al., 1996; Cao et al., 1997) will be susceptible to the same systematic error that misled D'Erchia et al. (1996).

Thus, these results demonstrate that rodent monophyly clearly is not refuted by the mtDNA genome data. In spite of the large number of base pairs, a more thorough sampling of taxa will be required to adequately test the hypothesis of rodent monophyly. In particular, more rodent and insectivoran sequences will be required (e.g., Nedbal et al. 1996), and inclusion of xenarthran sequences (e.g., armadillo or sloth), the probable outgroup to the rest of the placental mammals (e.g., McKenna, 1975; Novacek, 1990), would divide the long branch leading to the opossum and thereby possibly provide a more reliable root for non-xenarthrous eutherians. Additional complete mtDNA sequences will soon be available. Analyses of these new data under appropriate models may or may not support D'Erchia et al.'s conclusions; for the moment the molecular data do not support the dissolution of Rodentia.

ACKNOWLEDGMENTS

We thank Mark Hershkovitz, John Huelsenbeck, Patrick Luckett, Luis Ruedas Scott Steppan, and anonymous reviewers for advice, discussion, and/or comments. While this work was being conducted, JS was supported by the Smithsonian Institution Molecular Evolution Fellowship (Office of Fellowships and Grants).

LITERATURE CITED

Adachi, J., and Hasegawa, M. (1996) MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood. *Computer Science Monographs,* No. **28**. Institute of Statistical Mathematics, Tokyo.

Cao, Y., Adachi, J., Yano, T., and Hasegawa, M. (1994). Phylogenetic place of guinea pigs: no support of the rodent-polyphyly hypothesis from maximum-likelihood analyses of multiple protein sequences. *Mol. Biol. Evol.*, **11**: 593-604.

Cao, Y., Okada, N., and Hasegawa, H. (1997). Phylogenetic position of guinea pigs revisited. *Mol. Biol. Evol.*, **14**:461-464.

D'Erchia, A. M., Gissi, C., Pesole, G., Saccone, C. and Arnason, U. (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 misleading. *Syst. Zool.*, **27**: 401-410.

Gaut, B. S., and Lewis, P. O. (1995). Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* **12**: 152-162.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**: 182-198.

Graur, D., Hide, W. A., and Li , W.-H. (1991). Is the guinea-pig a rodent? *Nature* **315**; 649-652.

Gu, X., Fu, Y.-X, and Li, W-H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol*. **12**: 546-557.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating the human-ape split by a molecular clock of mitochondrial DNA. *J. Mol. Evol*. **22**: 160-174.

Hasegawa, M., Cao, Y., and Adachi, J. (1992). Rodent polyphyly? *Nature* **255**: 595.

Hendy, M. D., and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Syst. Zool*. **20**: 406-416.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**:17-48.

Huelsenbeck, J. P., Hillis, D. M., and Jones, R. (1996). Parametric bootstrapping in molecular phylogenetics: Applications and performance. In. *Molecular Zoology: Advances, Strategies, and Protocols*, J. D. Ferraris and S. R. Palumbi, eds. pp. 19-45, Wiley-Liss, New York.

Janke, A, Xu, X., and Arnason, U. (1996) 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.

Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. In: *Mammalian protein metabolism,* H. N. Munro, ed., pp. 21-132. Academic Press, New York.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**: 111-120.

Kishino, H., and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of Hominoidea. *J. Mol. Evol.* **29;** 170-179 .

Li, W.-H., Hide, W. A., Zharkikh, A., Ma, D.-P., and Graur, D. (1992) The molecular taxonomy and evolution of the guinea pig. *J. Hered.* **83**:174-181.

Luckett, W. P., and Hartenberger, J.-L. (1993). Monophyly or polyphyly of the order Rodentia: Possible conflict between morphological and molecular interpretations. *J. Mammal. Evol.* **2:** 127-147.

Ma, D.-P., Zharkikh, A., Graur, D., VandeBerg, J. L., and Li, W.-H. (1993) Structure and evolution of opossum, guinea pig, and porcupine cytochrome *b* genes. *J. Mol. Evol.* **36**:327-334.

Maddison, W. P., and Maddison, D. R. (1992). MacClade--analysis of phylogeny and character evolution, version 3.5. Sinauer, Sunderland, Mass.

Martignetti, J. A., and J. Brosius. (1993) Neural BC1 RNA as an evolutionary marker: guinea pig remains a rodent. *Proc. Nat. Acad. Sci. USA* **90**:9698-9702.

McKenna, M. C. (1975). Toward a phylogenetic classification of the Mammalia. In Phylogeny of the primates: *An interdisciplinary approach,* W. P. Luckett and F. S. Szalay, eds. pp. 21-46, Plenum Press, New York.

Novacek, M. J. (1990). Morphology, paleontology, and the higher clades of mammals. In: *Current Mammalogy,* H. H. Genoways, ed. pp. 507-543, Plenum Press, New York.

Porter, C. A., Goodman, M , and Stanhope. M. J. (1996). Evidence on mammalian phylogeny from sequences of exon 28 of the von Willebrand factor gene. *Mol. Phyl. Evol.* **5**: 89-101.

Stanhope, M. J., Czelusniak, J., Si, J.-S., Nickerson, J., and Goodman, M. (1992). A molecular perspective on mammalian evolution from the gene encoding Interophotoreceptor Retinoid Binding Protein, with convincing evidence for bat monophyly. *Mol. Phyl. Evol.* **1:** 148-160 .

Sullivan, J., Holsinger, K. E., and Simon, C. (1995). Among-site rate variation and phylogenetic analysis of 12S rRNA data in Sigmodontine rodents. *Mol. Biol. Evol.* **12**: 988-1001.

Swofford, D. L., Olsen, G. P., Waddell, P. J. and Hillis, D. M. (1996). Phylogenetic inference. In: *Molecular Systematics,* 2nd ed., D. M. Hillis, C. Moritz, and B. K. Mable, eds. pp. 407-514, Sinauer, Sunderland, MA

Waddell, P. (1995). Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms, and maximum likelihood. Ph.D. Dissert., Massey University

Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *J. Mol. Evol*. **39**:105-111.

Yang, Z. (1994b.). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**: 306-314.

Yang, Z. (1996). Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol*., **42** :294-307.

Yang, Z., Goldman, N. and Friday, A. E. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11:** 316-324 .

Yang, Z., Goldman, N. and Friday, A. E. (1995). Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. 44:384-399

APPENDIX. Comparison of models of sequence evolution examined here. All these models asusme stationarity of each parameter. Likelihood scores are calculated on the tree of D'Erchia et al. (1996).

	Base Frequencies	Substitution Types	Rate Heterogeneity	lnL
JC	Equal	One	None	-50999.9171
JC+I	Equal	One	Invariable Sites	-47911.44192
JC+Γ	Equal	One	Gamma-distributed Rates	-47607.6553
JC+I+Γ	Equal	One	Invariable Sites + Gamma-distributed Rates	-47605.3659
K2P	Equal	Two	None	-50016.04069
K+I	Equal	Two	Invariable Sites	-46833.56109
K+Γ	Equal	Two	Gamma-distributed Rates	-46471.74497
K+I+Γ	Equal	Two	Invariable Sites + Gamma-distributed Rates	-46446.36463
HKY-85	Empirical	Two	None	-49433.62471
HKY+I	Empirical	Two	Invariable Sites	-46201.59995
HKY+I+Γ	Empirical	Two	Gamma-distributed Rates	-45803.54994
HKY+I+Γ	Empirical	Two	Invariable Sites + Gamma-distributed Rates	-45773.48647
GTR	Empirical	Six	None	-49028.13801
GTR+I	Empirical	Six	Invariable Sites	-45965.65841
GTR+Γ	Empirical	Six	Gamma-distributed Rates	-45620.12374
GTR+I+Γ	Empirical	Six	Invariable Sites + Gamma-distributed Rates	-45593.28162

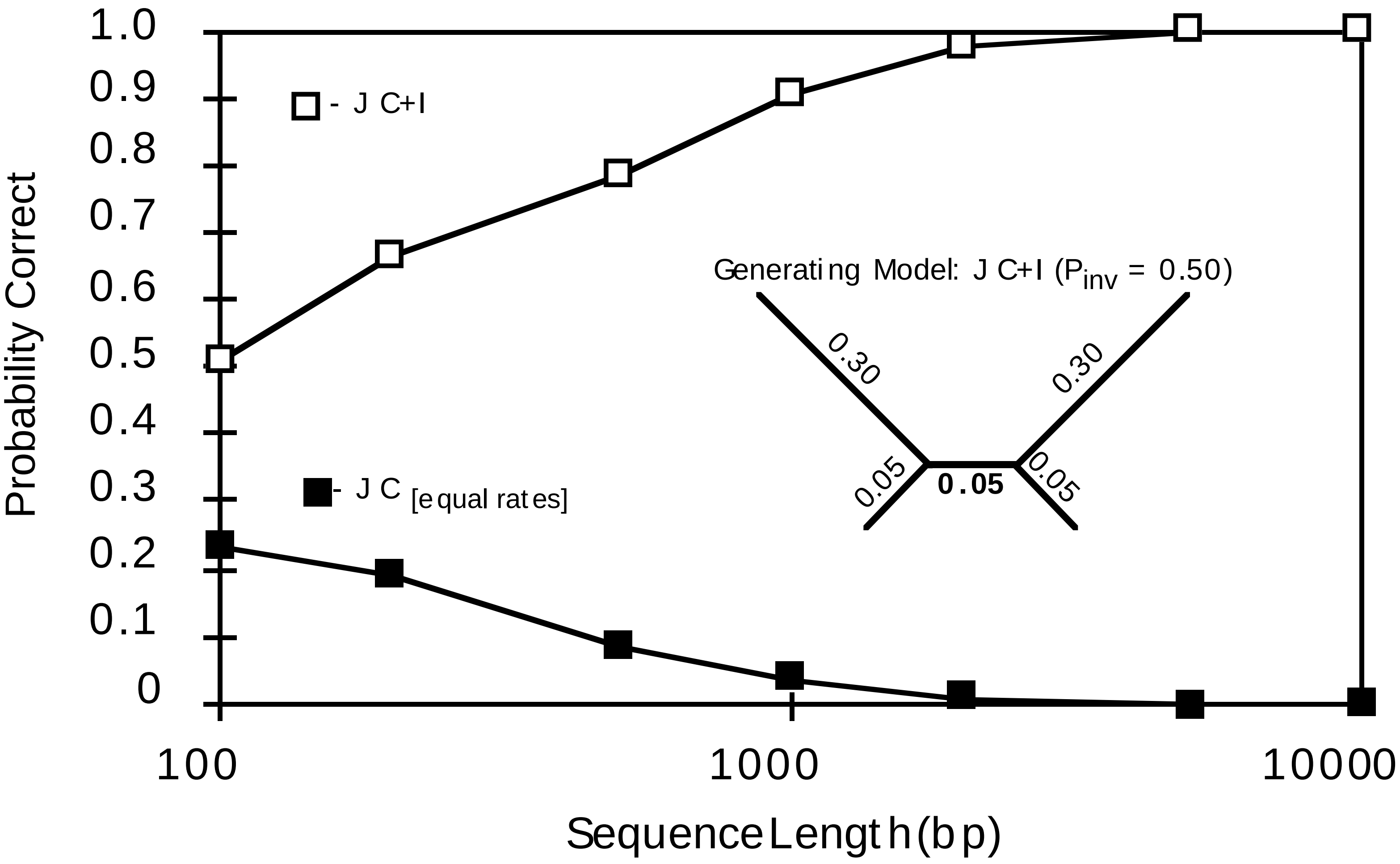
Table I. Cost (decrease in log-likelihood) associated with alternative placements of hedgehog on tree A of Fig. 1 and alternative attachments of opossum outgroup on tree B of Fig. 1. Table shows rearrangements that are not significantly different based on Kishino-Hasegawa test.

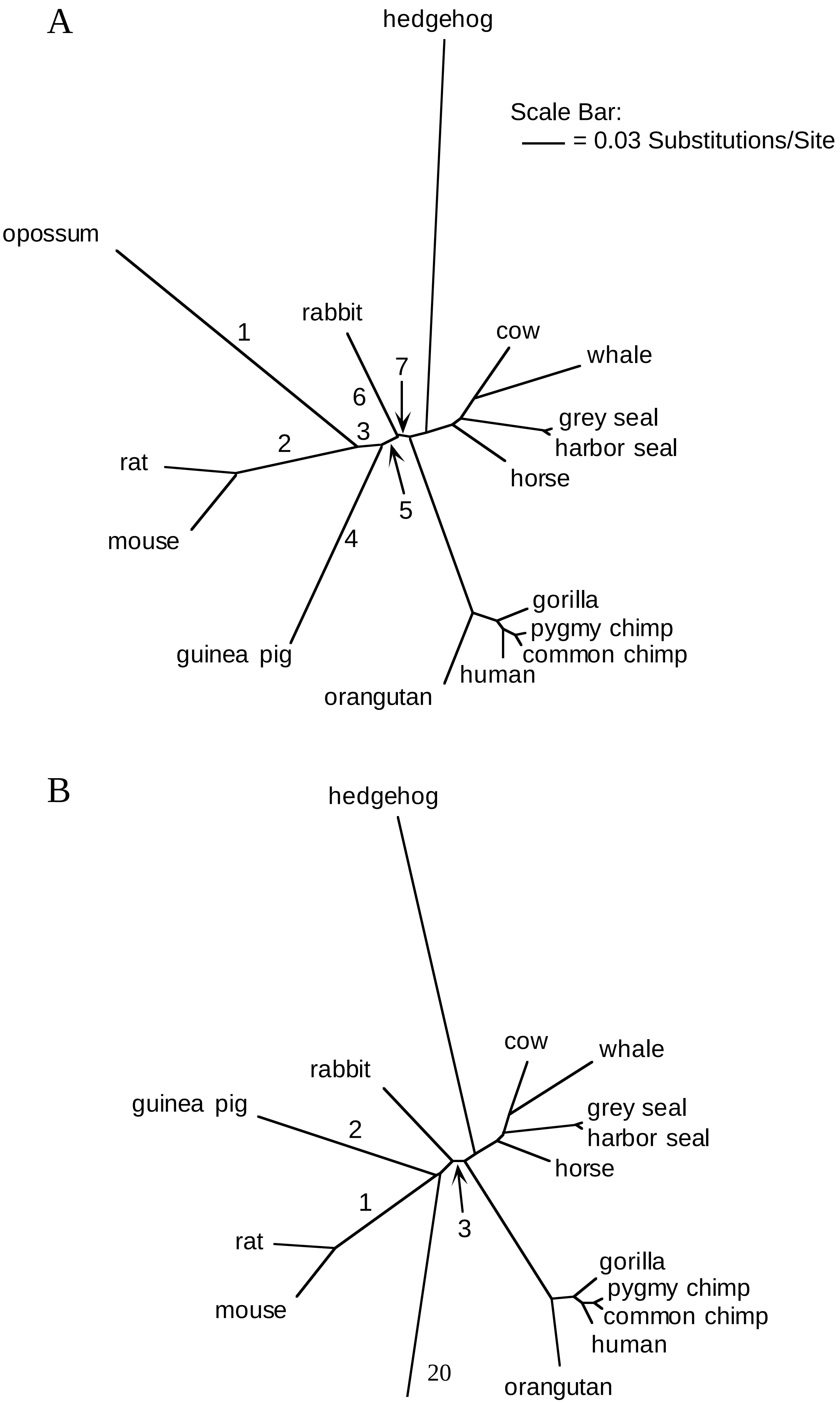
Tree A			Tree B		
Branch	Cost	*P*	Branch	Cost	*P*
1	7.41	0.42	1	best[1]	—
2	12.82	0.38	2	8.26	0.19
3	6.23	0.63	3	24.54	0.10
4	2.29	0.86			
5	5.93	0.56			
6	7.79	0.42			
7	6.40	0.40			

[1] Attachment of the opossum sequence to branch 1 of Tree B produces Tree A

Fig. 1. Maximum-likelihood can be an inconsistent estimator of phylogeny. One thousand replicate data sets of several different sequence lengths were simulated on the tree shown under a JC+I model of sequence evolution with 50 % invariable sites (Pinv = 0.5). Each data set was then analyzed under both the appropriate JC+I model (open squares) and an incorrect equal rates JC model (closed squares). Under the incorrect assumption of equal rates, the probability of inferring the correct tree is zero for long sequences (> 2000 bp).

Fig. 2 (A). Maximum-likelihood tree (lnL = -45570.78) for first and second codon positions for the “protein supergene” data of D'Erchia et al. (1996) estimated using a pre-release version of the PAUP* 4.0 (GTR+I+Γ model of evolution, p_{inv} = 0.4103, α = 0.8239). The hedgehog branch can be reconnected to each of the numbered branches with little change in likelihood score (see Table 1). (B) The most likely tree under the constraint of rodent monophyly (lnL = -45579.04). The root (attachment point of the opossum sequence) could be placed along the three numbered branches with insignificant change in likelihood (see Table I); attachment to the branch labeled one results in the identical topology as in (A) . The likelihood score of tree (B) does not differ significantly from that of the most likely tree (A).


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Probability Correct
□ - JC+I
■ - JC [equal rates]
Generating Model: JC+I (Pinv = 0.50)
0.30
0.30
0.05
0.05
0.05
0.05
100
1000
10000
Sequence Length (bp)


A
hedgehog
Scale Bar:
= 0.03 Substitutions/Site
opossum
1
rabbit
7
6
3
2
cow
whale
grey seal
harbor seal
horse
rat
mouse
5
4
gorilla
pygmy chimp
common chimp
human
guinea pig
orangutan
B
hedgehog
cow
whale
rabbit
guinea pig
2
grey seal
harbor seal
horse
1
3
rat
mouse
gorilla
pygmy chimp
common chimp
human
orangutan