Among-Site Rate Variation and Phylogenetic Analysis of 12S rRNA in Sigmodontine Rodents

Jack Sullivan, Kent E. Holsinger, and Chris Simon
Department of Ecology and Evolutionary Biology, University of Connecticut

We analyze sequences from two mitochondrial genes, cytochrome *b* (cyt *b*) and 12S rRNA (12S), for a group of sigmodontine rodents among which phylogenetic relationships are well understood based on concordance of morphological, chromosomal, allozyme, and other DNA data sets. Because these two genes are physically linked on the nonrecombining mitochondrial genome, they necessarily share the same history. Phylogenetic analysis of the cyt *b* gene recovers the well-corroborated relationships, generally with strong support. None of the methods that we employed, including variously weighted parsimony, neighbor joining on both single-rate and Γ-corrected distances, and maximum likelihood, were able to recover these relationships for the 12S gene. Parsimony analyses of the 12S data resulted in a relatively strongly supported placement of *Peromyscus eremicus* that conflicts with that suggested by cyt *b* and all other data. There is extreme among-site rate variation in the 12S sequences and moderate levels in the cyt *b* sequences. This highly skewed distribution of rates in the 12S gene makes phylogenetic analyses of these sequences particularly susceptible to the misleading effects of nonindependence and other nonrandom noise, suggesting that phylogenetic analyses of data sets that contain a great deal of among-site rate variation be interpreted with caution.

Introduction

The widespread use of molecular systematics across a broad range of biological disciplines makes it imperative to elucidate the factors that influence the estimation of molecular phylogenies. Important advances have demonstrated the effects of biased nucleotide composition and substitution (see Lockhart et al. 1994) and among lineage rate variation (Felsenstein 1978; Huelsenbeck and Hillis 1993). Although the effect of among-site rate variation has also received attention, most studies have addressed codon positions in protein coding genes (Beckenbach et al. 1993). Few studies have attempted to correct for among-site rate variation due to structural and functional constraints in rRNA genes (exceptions include Manske and Chapman 1987; Van de Peer et al. 1993).

Ribosomal RNA genes have been used commonly over a wide range of divergence levels including recent (see Allard and Honeycutt 1992), intermediate (Hedges et al. 1990), and deep (Field et al. 1988) divergences. The functional products of these genes are single-stranded RNA molecules that exhibit extensive secondary structure and bind with ribosomal proteins to form the ribosomal subunits involved in the assembly of proteins. These diverse molecular interactions result in a complex hierarchy of functional constraints governing the process of nucleotide substitution in rRNA genes, and we would therefore expect rRNA genes to exhibit a high degree of among-site rate variation.

Despite the complexities governing the evolution of rRNA genes, some constraints, primarily having to do with secondary structure, have been identified, and methods have been proposed to accommodate them in phylogenetic analyses. One class of weighting schemes emphasizes down-weighting stem sites because of the nonindependence involved in compensatory substitutions (Wheeler and Honeycutt 1988; Kraus et al. 1992; Dixon and Hillis 1993). Conversely, Miyamoto et al. (1994) recommended heavier weighting of stem positions based on the observation that stem sites tend to evolve more slowly than loop sites. In spite of this general tendency, however, it now appears that there is a continuum of rates in rRNA genes and that a bipartite classification of stems versus loops may be an oversimplification (see Van de Peer et al. 1993).

In this paper, we provide a maximum-likelihood method for estimating the distribution of evolutionary

Key words: rate variation, phylogeny, rRNA, sigmodontine rodents.

Address for correspondence and reprints: Jack Sullivan, Department of Ecology and Evolutionary Biology, U-43 University of Connecticut, Storrs, CT 06269-3043, E-mail: jsulliv@uconnvm.uconn.edu.

Mol. Biol. Evol. 12(6):988–1001. 1995.

0737-4038/95/1206-0003$02.00

rates across nucleotide sites. We also examine the evolution of the 12S gene in the context of an extremely well-corroborated phylogeny of sigmodontine rodents (sensu Musser and Carleton 1993), ranging from very closely related species to older divergences (approximately 20 million years before present, MYBP). In addition, we explore the consequences of differing rate distributions on phylogenetic analysis of these data.

Material and Methods

Well-Corroborated Relationships

The group we have chosen is murid rodents in the subfamily Sigmodontinae (Musser and Carleton 1993), emphasizing members of the genus *Peromyscus.* This genus comprises approximately 58 species allocated into 13 species groups (Carleton 1989) and has been the subject of intensive phylogenetic study since Osgood's (1909) revision. Although application of morphological, chromosomal, allozyme, mtDNA restriction site, and DNA-DNA hybridization analyses have not resolved relationships among all of the species in the genus, the *P. maniculatus* and *P. leucopus* species groups exhibit remarkable congruence among these data sets (reviewed in Carleton 1989; Hogan et al. 1993; C. W. Kilpatrick, personal communication). The close relationship between these two species groups is supported by allozyme (Avise et al. 1979), chromosomal (Stangl and Baker 1984), and DNA-DNA hybridization (C. W. Kilpatrick, personal communication) data, and the well-corroborated relationships among the taxa examined here are shown in figure 1.

An additional advantage of these rodents is that there are a number of progressively distantly related genera that also have been studied extensively. The genus *Onychomys* diverged from the *Peromyscus* lineage approximately 6 MYBP (Hibbard 1968) and comprises three species (fig. 1). *Neotoma* split from the *Peromyscus/Onychomys* lineage approximately 7.5 MYBP; and the genus *Sigmodon,* approximately 12 MYBP (Catzeflis et al. 1992). The sigmodontine/murine split probably occurred approximately 20 MYBP (reviewed in Catzeflis et al. 1992; but see O'hUigan and Li 1992). The fossil record for sigmodontines is compatible with DNA-DNA hybridization data (Brownell 1983; C. W. Kilpatrick, personal communication) for these taxa (reviewed in Catzeflis et al. 1992).

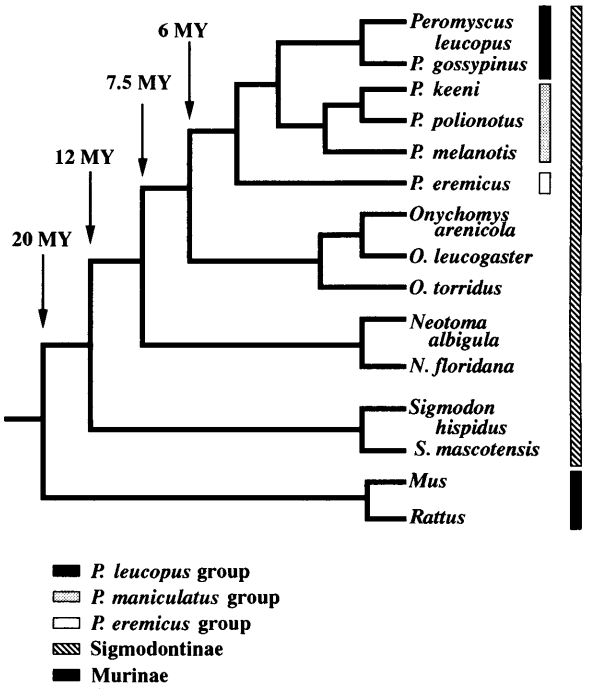

FIG. 1.—The well-corroborated relationships among the species included in this study. The topology and divergence dates are consistent with paleontology, morphology, allozymes, chromosomal data, and DNA-DNA hybridization and restriction site data (see text).

Samples

The following specimens were examined (sample sizes in parentheses): *Peromyscus leucopus,* Connecticut, Tolland Co., Union, Yale Forest (3), Texas, Brown Co., 15 mi south of Cross Cut (2), Texas, Hall Co., 1 mi north of Estelline (2); *P. gossypinus,* Florida, Alachua Co., 5 mi southwest of Gainesville (2); *P. keeni,* Washington, Grays Harbor Co., 5 mi northeast of Grisdale (3), Washington, Grays Harbor Co., 2 mi west of Grisdale (3); *P. polionotus,* Florida, Walton Co., Grayton Beach State Park (5); *P. melanotis,* Mexico, Durango, 3 mi east of La Ciudad (3), Mexico, Durango, Hacienda Coyote (1); *P. eremicus,* Mexico, Sonora (2); *Onychomys leucogaster,* Texas, Winkler Co. (1), *O. torridus,* Utah, Washington Co. (1); *O. arenicola,* New Mexico, Eddy Co. (1); *Neotoma floridana,* Texas, Montague Co., 4 mi west of Nocona (2), Kansas (1); *N. albigula,* Texas, Knox Co., 4 mi east of Benjamin (2); *Sigmodon hispidus,* Texas, Wichita Co., Wichita Falls (2); *S. mascotensis,* Mexico, Jalisco, 8 mi southwest of Cocula (1), Mexico, Michoacan, 6 mi east of Zamora (1). The specimens of *O. leucogaster, O. torridus,* and *O. arenicola* examined here are individuals 13, 27, and 24, respectively, examined by Riddle and Honeycutt (1990) for restriction site variation in the mitochondrial genome.

Data Collection

In most cases, genomic DNA was extracted from 10 mg of liver tissue following the method of Gustincich et al. (1991). The *Onychomys* samples were aliquots of CsCl-

purified mtDNA examined by Riddle and Honeycutt (1990) for restriction site variation and were diluted 1/10 prior to use. One μl of all samples was used for polymerase chain reaction (PCR) amplification with the following 12S rRNA primers to generate an approximately 800-bp fragment (*L* and *H* refer to heavy and light strand, respectively; sequence numbers refer to the *Mus* sequence; Bibb et al. 1981): (1) L82 5′-CATAAACACAAGGTTTGGTCC-3′, (2) L316 5′-GGTAAATTTCGTGCCAGCCAC-3′, (3) H484 5′-ATAGTGGGGTATCTAATCCCAGTT-3′, (4) L509 5′-AACTGGGATTAGATACCCCACTAT-3′, and (5) H901 5′-AAGAGCGACGGGCGATGTGT-3′ (Allard and Honeycutt 1992). Primers L316, H484, L509, and H901 correspond to SR-N-14756, SR-J-14612, SR-N-14588, and SR-J-14233 of the Simon et al. (1994) compilation, whereas primer L82 broadly overlaps primer SR-N-14922 of Simon et al. (1994). To examine the possibilities of gene tree/species tree conflicts, a 350-bp fragment of the cytochrome *b* (cyt *b*) gene also was amplified and sequenced for *Peromyscus* and *Onychomys* samples using the following primers: L14115 5′-GATATGAAAAACCATCGTTA-3′ and H14541 5′-CAGAATGATATTTGTCCTCA-3′ (here, sequence numbers refer to human sequence; Kocher et al. 1989).

In most cases single-stranded template was generated for sequencing by asymmetric amplification using the one-primer method (Allard et al. 1991). The products of asymmetric amplifications were purified by ultrafiltration (Ultrafree-MC 30,000 NMWL Filter Units, Millipore Inc.) and used as templates for dideoxy sequencing. In some cases the double-stranded products of symmetric amplifications were excised from low melting point agarose gels, cleaned using standard glassmilk procedures, and sequenced directly.

The resulting sequences were aligned by eye using ESEE (Cabot and Beckenbach 1989), and the sigmodontine 12S sequences were aligned to published *Mus* and *Rattus* sequences (Bibb et al. 1981; Kobayashi et al. 1981). In almost all cases, indels were consistent with the relationships among these taxa, and in all cases the placement of indels was consistent with secondary structural considerations (Kjer et al. 1994; Hickson et al., 1996). Two small regions of alignment ambiguity (16 bp) were omitted from phylogenetic analyses.

Data Analysis

Parsimony Analyses

All parsimony analyses were conducted using PAUP (version 3.1.1; Swofford 1993). Analyses of the 12S data were conducted on the full data set (15 taxa), with *Mus* and *Rattus* designated as a monophyletic outgroup, and on a restricted data set comprising only those individuals for which cyt *b* data were available (nine taxa). The analyses of the restricted 12S data set are therefore directly comparable to the cyt *b* data. Gaps were treated either as missing or as a fifth character state, and separate analyses were conducted using the following weighting schemes: equal weights for all changes, stem-weighting procedures (Wheeler and Honeycutt 1988; Dixon and Hillis 1993; Miyamoto et al. 1994), step matrices to down-weight transitions by factors of zero (equals transversion parsimony), 0.33, and 0.20, and successive approximations (Farris 1969). Branch-and-bound analyses were conducted on the complete 12S data set, and exhaustive searches were done on the restricted 12S and cyt *b* data sets. Bootstrap analyses were conducted using 500 replicates, and decay indexes (Bremer 1988) were calculated.

Distance Analyses

Neighbor-joining analyses (Saitou and Nei 1987) were conducted using MEGA (Kumar et al. 1993). Because of the existence of nucleotide bias in these sequences (not shown), genetic distances were estimated following both Kimura (1980) and Tamura and Nei (1993). Again, separate analyses were conducted on the complete and restricted 12S data sets, and bootstrap analyses were conducted using 500 replicates. Separate analyses also were conducted assuming no among-site rate variation and incorporating Γ-corrected distances (Tamura and Nei 1993) where the Γ-distributed rates model could not be rejected (see below).

Maximum-Likelihood Analysis

Maximum-likelihood analyses were conducted on the restricted 12S data using the DNAML program of the PHYLIP package (Felsenstein 1993) and the discrete gamma program of Baseml (Yang 1994) using 40 rate categories and the F84 model. Because Baseml is computationally intensive, it was used only to compare likelihoods of alternative topologies. This program simultaneously estimates among-site rate variation and the transition bias. The estimate of Ts/Tv derived from the Baseml analyses was used in the DNAML analysis.

Analysis of Rate Variation

Among-site rate variation was examined two ways. MacClade (Maddison and Maddison 1992) was used to reconstruct character state changes on the well-corroborated phylogeny of these taxa and to chart the number of substitutions inferred at each site. In the absence of among-site rate variation, the distribution of the number of changes at each site should conform to the Poisson distribution. If evolutionary rates are gamma distributed across sites, the number of changes at each site is ex-

pected to conform to the negative binomial distribution. We fit the resulting histograms to the negative binomial and simultaneously found maximum-likelihood estimates of both the shape (α) and scale (p) parameters (see the appendix; DOS program written in C available on request). The shape parameter (α) is inversely related to the coefficient of variation, and low values ($\alpha < 0.5$) suggest extreme rate heterogeneity (Tateno et al. 1994). Because the gamma distribution converges on the Poisson distribution when α becomes large, the Poisson model of no among-site rate variation is a special case of the Γ-distributed rates model. This allowed us to use likelihood-ratio tests to determine whether use of the Γ-distributed rates model led to an improved fit relative to the Poisson model (Goldman 1993). Chi-squared tests were used to test for goodness-of-fit to the negative binomial and classes in which the expected number of substitutions per site were less than five were pooled.

The use of parsimony estimates of the distribution of the number of changes at each site ignores transition and nucleotide biases. In addition, there is a conservative bias inherent in the use of parsimony which has been shown to cause an underestimation of the degree of among-site rate variation (i.e., over-estimation of α; Wakeley 1993). To examine the effect of these biases on the parsimony-based estimates of α, we also estimated shape parameters using Baseml (Yang 1994). This program uses sequence data directly and accounts for transition and nucleotide biases. Likelihood-ratio tests were conducted to ascertain whether accounting for these biases led to significantly different estimates of α.

Results

Levels of Variation

In the 350-bp fragment of cyt *b* there were 96 variable sites, 61 of which were potentially informative in parsimony analyses. Almost all variation was synonymous. Divergence estimates (Tamura and Nei 1993) ranged from 0.006 within *Peromyscus leucopus* to 0.223 between *Onychomys leucogaster* and *P. polionotus* (Table 1).

Alignment of the 12S sequences resulted in 775 aligned sites (fig. 2). In the restricted 12S data set, there were 87 variable sites (94 including indels), 48 of which were potentially informative in parsimony analyses (almost five times the number of taxa). In the complete 12S data set there were 172 variable sites (210 including indels), 126 of which were informative in parsimony analyses. Intraspecific variation in the 12S data was <0.006 in all samples and was highest in *P. leucopus* and *Sigmodon mascotensis,* and divergences ranged up

Table 1
Divergence in the 12S rRNA Gene (above the diagonal) and a Portion of the Cyt *b* Gene (below the diagonal for *Peromyscus* and *Onychomys*) Calculated Following Tamura and Nei (1993)

	Ple1	*Ple2*	*Pgo*	*Pke*	*Ppo*	*Pme*	*Per*	*Ole*	*Oar*	*Oto*	*Nal*	*Nfl*	*Shi*	*Sma1*	*Sma2*
Ple1	...	0.006	0.018	0.032	0.025	0.021	0.037	0.061	0.061	0.070	0.095	0.095	0.119	0.117	0.113
Ple2	0.006	...	0.018	0.031	0.022	0.018	0.034	0.058	0.058	0.067	0.093	0.093	0.119	0.117	0.113
Pgo	0.046	0.039	...	0.024	0.027	0.022	0.041	0.065	0.065	0.078	0.096	0.096	0.120	0.123	0.120
Pke	0.111	0.111	0.116	...	0.014	0.013	0.043	0.059	0.054	0.066	0.096	0.096	0.122	0.120	0.116
Ppo	0.114	0.114	0.118	0.056	...	0.007	0.035	0.047	0.045	0.057	0.087	0.087	0.112	0.113	0.110
Pme	0.103	0.103	0.127	0.064	0.071	...	0.030	0.049	0.044	0.059	0.087	0.087	0.110	0.108	0.105
Per	0.156	0.156	0.150	0.144	0.164	0.169	...	0.066	0.068	0.079	0.092	0.092	0.131	0.128	0.125
Ole	0.188	0.188	0.213	0.220	0.223	0.215	0.201	...	0.018	0.025	0.079	0.079	0.102	0.107	0.105
Oar	0.145	0.145	0.165	0.171	0.172	0.175	0.150	0.145	...	0.025	0.079	0.077	0.111	0.114	0.112
Oto	0.138	0.146	0.166	0.196	0.187	0.171	0.135	0.133	0.053	...	0.085	0.085	0.114	0.117	0.115
Nal	...	...	...	...	...	...	...	...	...	...	...	0.013	0.105	0.121	0.118
Nfl	...	...	...	...	...	...	...	...	...	...	...	...	0.103	0.121	0.118
Shi	...	...	...	...	...	...	...	...	...	...	...	...	...	0.031	0.034
Sma1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	0.006
Sma2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

NOTE.—Abbreviations are as follows: *Ple1*, *Peromyscus leucopus* northern; *Ple2*, *P. leucopus*, southern; *Pgo*, *P. gossypinus*; *Pke*, *P. keeni*; *Ppo*, *P. polionotus*; *Pme*, *P. melanotis*; *Per*, *P. eremicus*; *Ole*, *Onychomys leucogaster*; *Oar*, *O. arenicola*; *Oto*, *O. torridus*; *Nal*, *Neotoma albigula*; *Nfl*, *N. floridana*; *Shi*, *Sigmodon hispidus*; *Sma*, *S. mascotensis*.

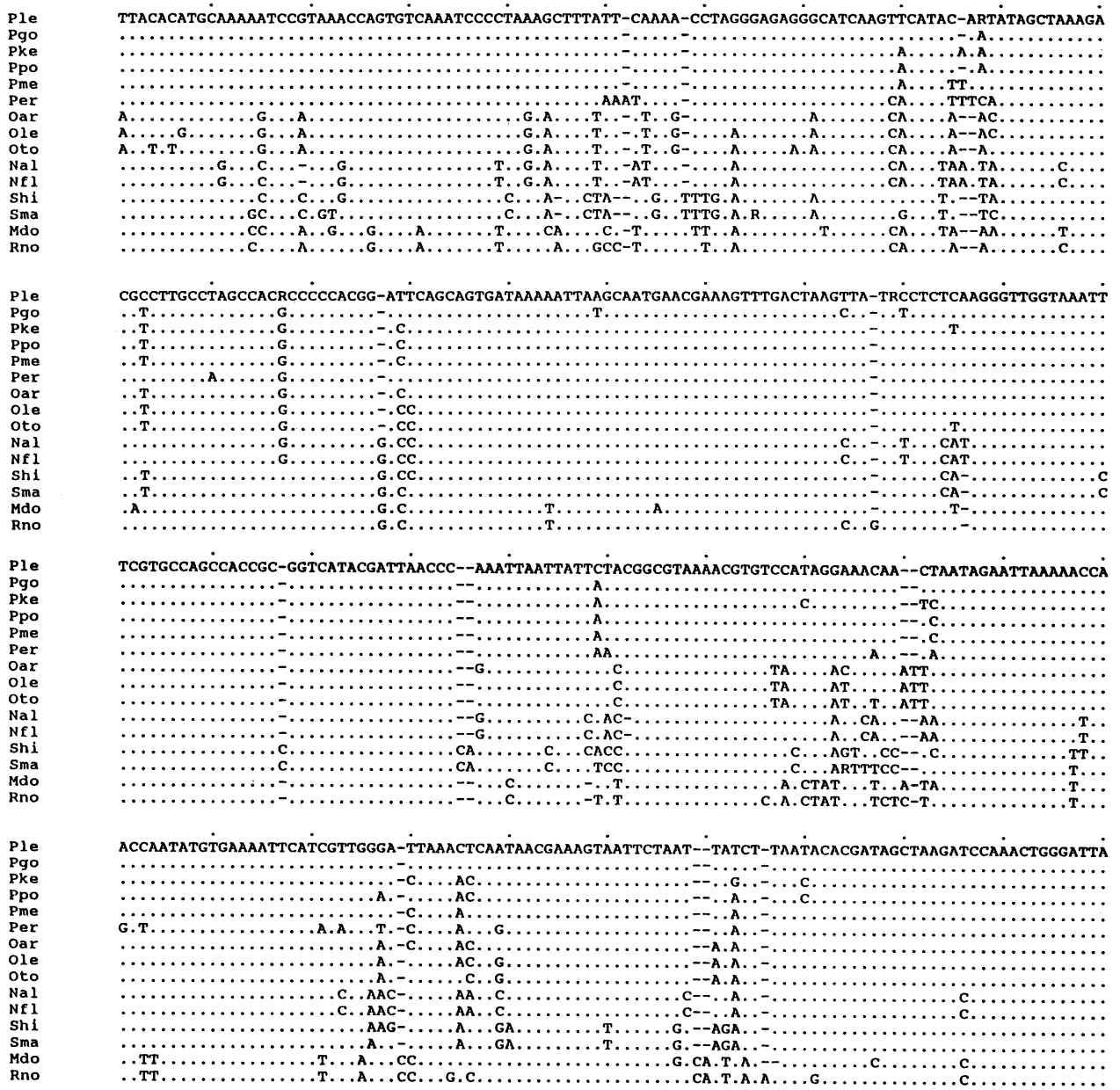

FIG. 2.—Alignment of the 12S rRNA sequences examined here. Taxon abbreviations as in table 1 with the following additions: *Mdo, Mus domesticus; Rno, Rattus norvegicus.*

to 0.128 between *P. eremicus* and *S. mascotensis* (table 1).

Phylogenetic Analyses

Parsimony Analyses

Parsimony analysis of the cyt *b* data with equal weights produced one tree (fig. 3*A;* CI = 0.63, RC = 0.50) that was congruent with the well-corroborated relationships among these taxa. Bootstrap analysis and decay indexes showed strong support for all nodes, with the exception of the node indicating that *P. keeni* and *P. polionotus* share an ancestor not shared by *P. melanotis.* The relationships suggested by the cyt *b* data were the same regardless of method of phylogenetic analysis employed.

Parsimony analysis of the restricted 12S data set with equal weights resulted in two equally parsimonious trees (fig. 3*B;* CI = 0.77, RC = 0.72). Bootstrap analysis suggested relatively strong support for the monophyly of the *P. leucopus* group (88% bootstrap value), relatively

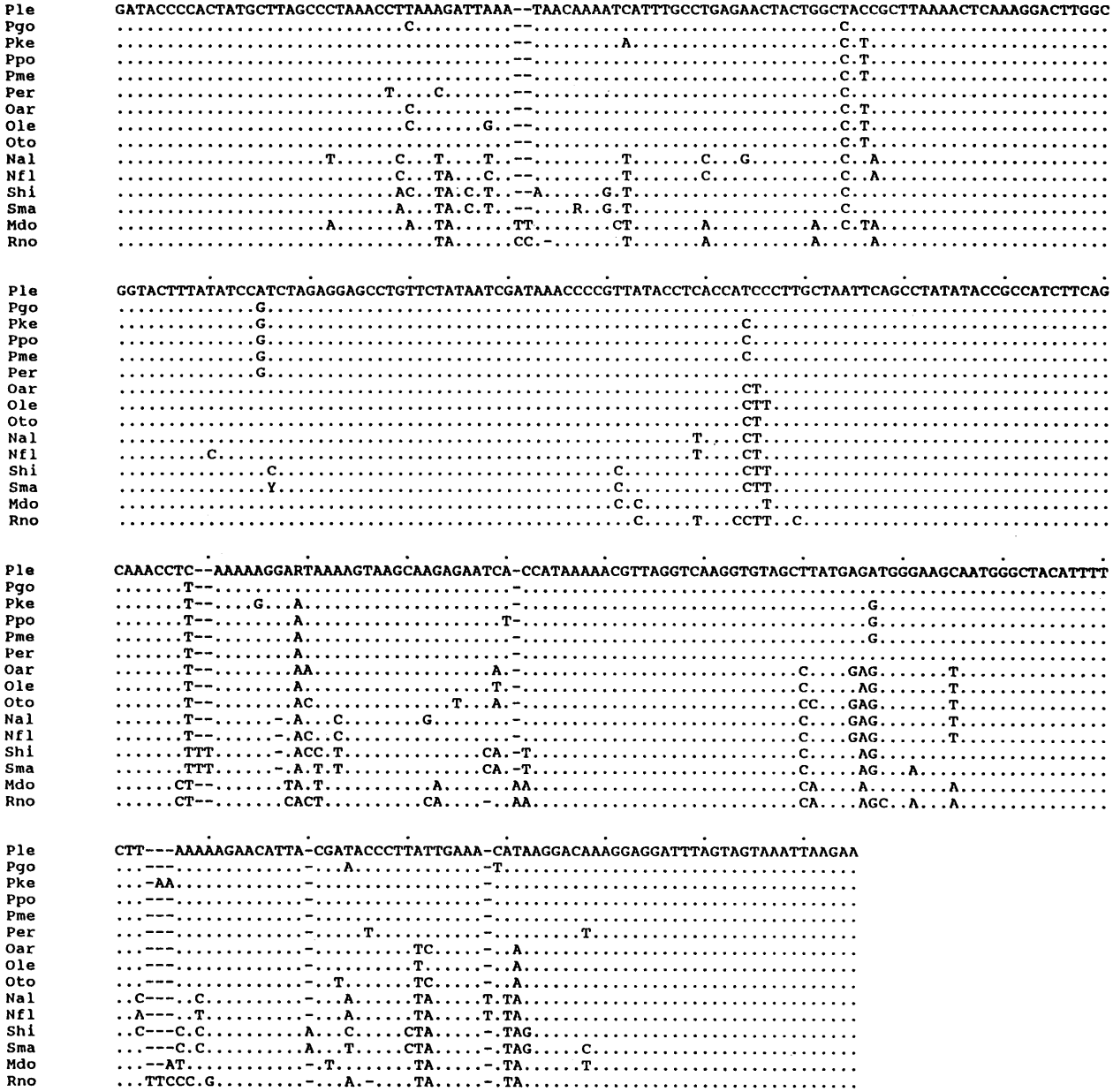

FIG. 2 (*Continued*)

strong support for a node uniting *P. eremicus* with the *P. leucopus* group (90% bootstrap value), and moderate support for paraphyly of the *maniculatus* group (65% bootstrap value). The node uniting *P. eremicus* with the *leucopus* group was supported by four unambiguous putative synapomorphies (sites 129, 476, 564, and 677; fig. 2), but, as indicated by the decay index of three, there was conflict in the data. Two sites (79 and 339) supported the well-corroborated relationships (fig. 2). None of the weighting schemes employed led to an increase in congruence with the cyt *b* data; all resulted in a *P. eremicus*/*P. leucopus* group clade.

Parsimony analysis of the complete 12S data set with equal weights resulted in six equally parsimonious trees (fig. 3*C*; CI = 0.73, RI = 0.59). There was strong support for nodes that united genera but little resolution among genera. There was less support for the node uniting *P. eremicus* with the *P. leucopus* group (60% bootstrap value) than in the restricted 12S data set, indicating that some of the sites that united these taxa in the re-

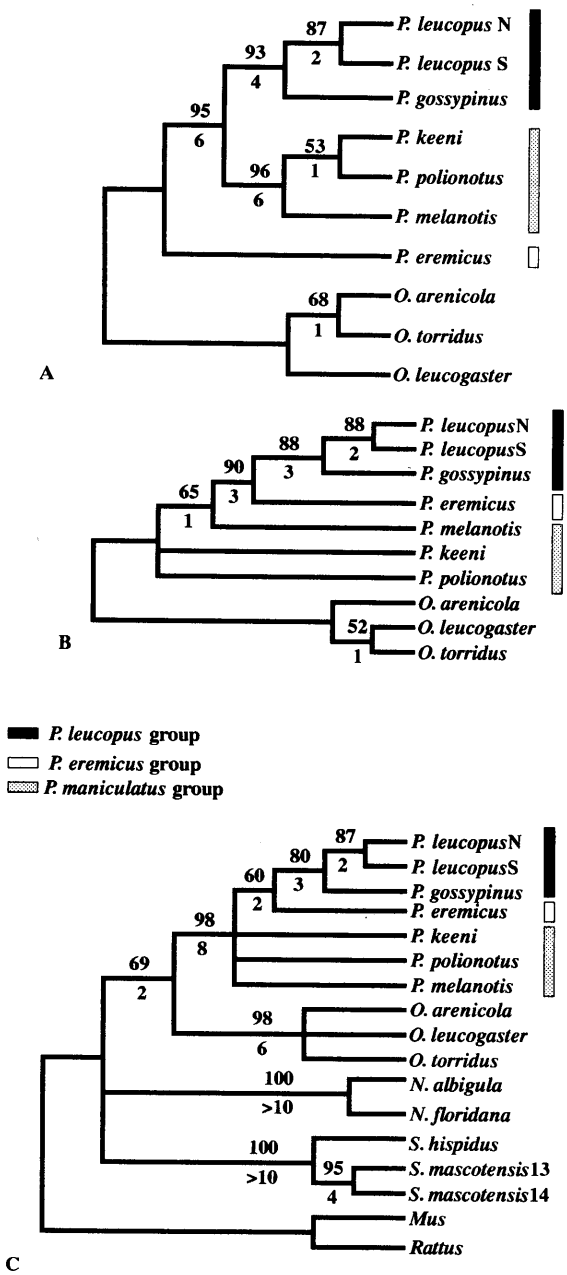

FIG. 3.—Results of equally weighted parsimony analysis. *A*, The single MP tree from cyt *b* data (total length = 161, CI = 0.63, RC = 0.50); *B*, strict consensus of two trees from the restricted 12S data set (total length = 119, CI = 0.77, RC = 0.72); and *C*, strict consensus of six equally parsimonious trees from the full 12S data set (total length = 379, CI = 0.73, RC = 0.59). Numbers above the branches indicate bootstrap values (500 replicates) and numbers below, decay indexes.

stricted 12S data set were variable among the additional taxa (e.g., site 476). Again, none of the weighting schemes employed produced results congruent with the cyt *b* data.

Distance Analyses

Results of the neighbor-joining bootstrap analyses for the 12S data sets under the equal rates model were similar to parsimony analyses for both Kimura (1980) and Tamura-Nei (1993) distances (fig. 4). There was moderate support for a sister group relationship between *P. eremicus* and the *P. leucopus* species group (67% bootstrap value) in the restricted 12S data set (fig. 4*B*), again conflicting with the relationships suggested by cyt *b* and all other data. Incorporation of among-site rate variation into the analysis by using Γ-corrected distances (Tamura and Nei 1993) did not result in an increase in congruence between the 12S data and the cyt *b* data. In fact, support for the relationship between *P. eremicus* and the *P. leucopus* group increased (not shown).

Maximum Likelihood

Maximum-likelihood analysis of the restricted 12S data set using DNAML yielded results identical in to-

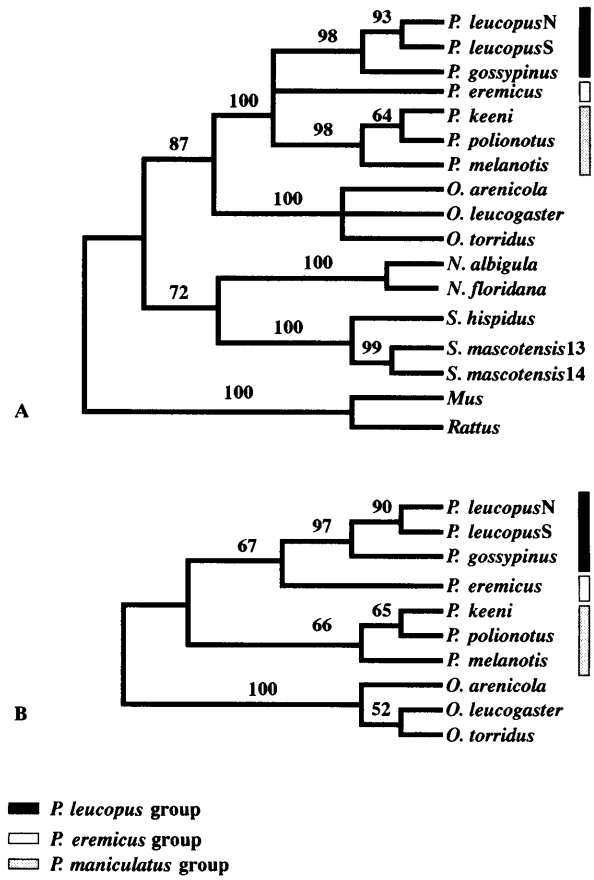

FIG. 4.—Results of neighbor-joining analysis of single-rate distances corrected following Tamura and Nei (1993) for (*A*) the complete 12S data set and (*B*) the restricted 12S data set. Numbers above branches indicate bootstrap values, and all nodes with values <50% have been collapsed.

pology to the results of the parsimony analyses (fig. 3*B*). Although we did not use the tree-searching algorithm of Baseml, the (incorrect) maximum-parsimony tree (fig. 3*B*) had a higher likelihood score than the well-corroborated topology (fig. 3*A*) even when among-site rate variation was incorporated into the analysis.

Analysis of Rate Variation

Likelihood ratio tests indicated that, for all the data sets (complete 12S, restricted 12S, and cyt *b*) there was a significant improvement in likelihood scores when the Γ-distributed rates model was used relative to the Poisson model. This suggests that significant among-site rate variation exists in all three data sets (fig. 5).

The maximum-likelihood estimate of the shape parameter for the cyt *b* data was 0.649, suggesting moderate among-site rate variation. However, the Γ-distributed rates model could be rejected (fig. 5*A*) because of a preponderance of both invariant sites and two-change sites. Therefore, neither the Γ-distributed rates model nor the single-rate model could accommodate the pattern of rate variation observed in this gene.

The maximum-likelihood estimate of the shape parameter for the restricted 12S data set was 0.281, suggesting extreme among-site rate variation. In this case the Γ-distributed rates model appeared to accommodate the large amount of rate variation present (fig. 5*B*), and the model could not be rejected.

The maximum-likelihood estimate of the shape parameter was 0.347 for the complete 12S data set, suggesting substantial rate heterogeneity. The Γ-distributed rates model could be rejected for these data, however (fig. 5*C*), due to a preponderance of high rate sites and two-change sites. The estimates of α for this data set and the restricted 12S data set therefore are not directly comparable.

To further explore the nature of the among-site rate variation, we removed sites that were not free to vary (Fitch and Margoliash 1967; Fitch 1971; Steele et al. 1991; Reeves 1993) from the data sets and examined the rate distributions as above. In the case of the complete 12S data set with invariable sites removed, the estimated shape parameter of the Γ-distributed rates model was 16. The likelihood-ratio test indicated that use of the Γ-distributed rates model did not significantly improve the fit relative to the single-rate model. Similarly, in the restricted 12S data set with invariable sites removed, the single-rate and Γ-distributed rates models could not be distinguished as the estimated gamma shape parameter was 22. In the cyt *b* data, the Γ-distributed rates model also converged to the Poisson model after removal of sites not free to vary. Thus, in each case the observed among-site rate variation could be accommodated by

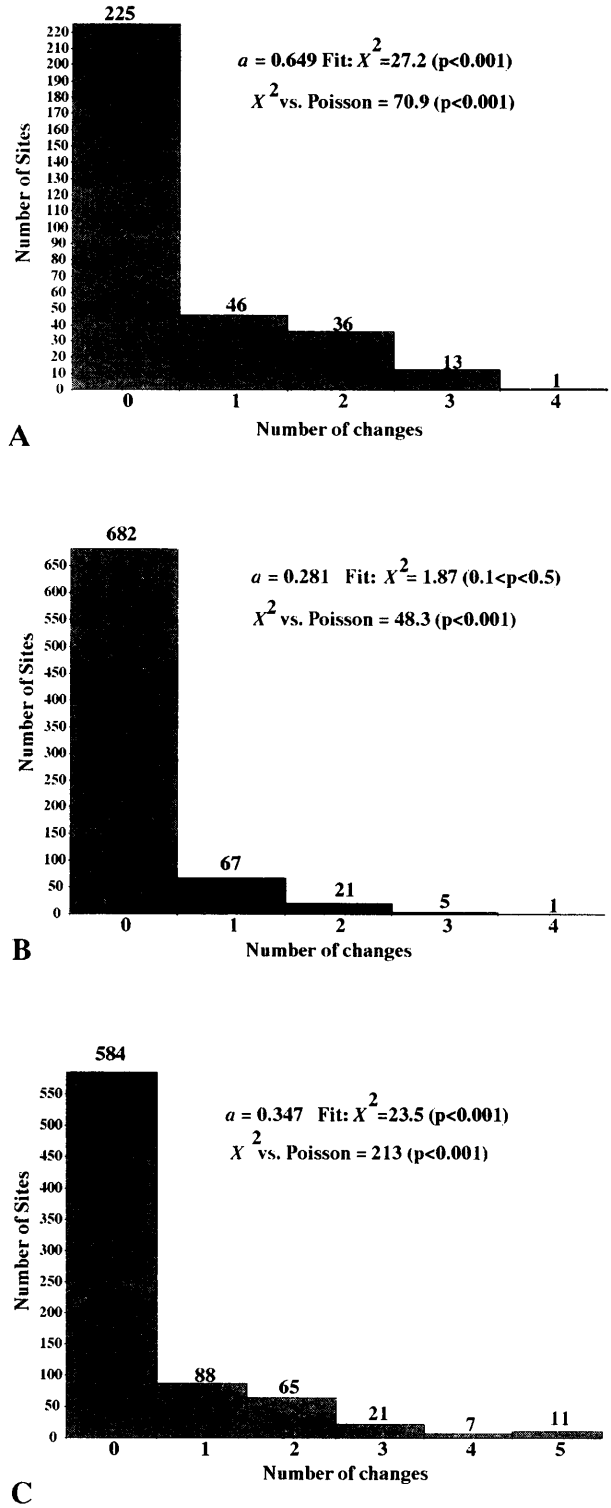

FIG. 5.—Analysis of among-site rate variation in (*A*) the cyt *b* data, (*B*) the restricted 12S data set, and (*C*) the complete 12S data set. The value *a* is the maximum-likelihood estimate of the shape parameter, the two χ^2 values are goodness-of-fit to the Γ-distributed rates model and likelihood-ratio tests relative to the single rate (Poisson) model, respectively.

assigning a constant rate at those sites free to vary. The differences in the skewness of the raw distribution of rates, as measured by α (fig. 5), was caused by a marked difference in the proportion of sites estimated to be free to vary in the 12S data sets versus the cyt *b* data set. In the restricted 12S data set, 26% of all sites were free to vary, whereas 32% and 44% were free to vary in the complete 12S and cyt *b* data sets, respectively.

Discussion

Phylogenetic Analyses

Phylogenetic analysis of the cyt *b* data recovered the well-corroborated phylogeny for *Peromyscus* taxa with very strong support for all but one node (fig. 3*A*). The exception was the node that resolved the relationships among the *maniculatus* species group, but the basal position of *Peromyscus melanotis* within the *maniculatus* group is strongly supported by ND-3, ND-4, and ND-4L gene sequences (S. Davis, personal communication).

The results of the phylogenetic analyses of the restricted 12S data set (fig. 3*B*) provide strong support for a node incongruent with the well-corroborated relationships. This placement of *P. eremicus* was also indicated by neighbor-joining (fig. 4*B*) and maximum-likelihood analyses. Because phylogenetic analysis of the cyt *b* data recovers these relationships with generally strong support, the difficulties with the 12S data cannot be related to a gene tree/species tree conflict; the 12S and cyt *b* genes are physically linked on the nonrecombining mitochondrial genome and must therefore share the same history. In addition, the number of variable sites in the restricted 12S data set (87) was comparable to the number in the cyt *b* data set (96), as were the number of sites potentially informative to parsimony analyses (48 in 12S versus 61 in cyt *b*). Therefore, the lack of congruence between the 12S and cyt *b* (as well as all other) data is not simply a result of too little variation in the 12S gene; rather, the incongruence most likely results from differences in the pattern of variation in the two genes.

Patterns of Variation

The results of the analysis of rate variation within 12S genes indicated a high degree of heterogeneity. In the restricted data set (fig. 5*B*), where the distribution of number of inferred substitutions at each site conformed to the Γ-distributed rates model, the shape parameter indicated extreme rate variation ($\alpha = 0.281$). In the complete 12S data set (fig. 5*C*), there appeared to be slightly less extreme rate variation as indicated by the significantly higher shape parameter ($\alpha = 0.347$); however, this data set did not fit the Γ-distributed rates model. To explore the cause of the lack of fit, we examined the variance-to-mean ratio. Under the Γ-distributed rates model, this value is expected to be

$$s^2/\bar{x} = 1 + \bar{x}/\alpha$$

(Johnson and Kotz 1969). In the complete 12S data set however, the observed ratio was 2.070, and the expected was 2.434. The observed lack of fit to the Γ-distributed rates model was caused by smaller variance than expected, relative to the mean, even though the ratio was twice that expected in the absence of among-site rate variation. Similarly, a smaller than expected variance caused the lack of fit to the Γ-distributed rates model for the cyt *b* data; the expected ratio was 1.771, whereas the observed value was 1.481. By comparison, the expected variance-to-mean ratio was 1.487 in the restricted 12S data, where the Γ-distributed rates model could not be rejected, whereas the observed value was 1.491.

Several methods of estimating the shape parameter for the Γ-distributed rates model have now been proposed. The simplest of these is the method-of-moments (Johnson and Kotz 1969), which uses parsimony to estimate the distribution of the number of changes at each site along a tree. The shape parameter is then found by

$$\alpha = \bar{x}^2/(s^2 - \bar{x})$$

(Uzzell and Corbin 1971; Larson and Wilson 1989; Kocher and Wilson 1991). However, the method-of-moments estimate has a higher mean-squared error than maximum-likelihood estimation (Anraku and Yanagimoto 1990). The maximum-likelihood method presented here also relies on parsimony estimates of the number of changes at each site, and this has been shown to lead to an underestimation of among-site rate variation (Wakeley 1993). In addition, Collins et al. (1994) have shown that compositional bias can lead to an underestimation in the number of changes inferred by parsimony-based character reconstructions. Both these factors will introduce a conservative bias in any method that relies on parsimony to infer the distribution of rates. Yang (1994) presented a maximum-likelihood method (Baseml) that estimates α directly from sequence data and that accounts for compositional and transition biases. Use of the maximum-likelihood method presented here can lead to significantly better estimates of α than the method-of-moments, and use of Baseml can lead to significantly better estimates than either our method or the method-of-moments (table 2). Baseml is computationally intensive, however, and cannot be used for more than 12–15 taxa. We are currently ex-

Table 2
Comparison of Estimates of α Derived by the Method-of-Moments (α_1), the Method Presented Here (α_2), and the Method of Yang (α_3; 1994)

Data Set	α_1	α_2	α_3
12S full	0.438*	0.347	...
12S restricted	0.334**	0.281**	0.157
Cyt *b*	0.985*	0.649**	0.290

NOTE.—Likelihood-ratio tests were used to test for significant differences.
* Significantly different from α_2 at $P < 0.05$.
** Significantly different from α_3 at $P < 0.01$.

amining the effect of taxon sampling density on estimates of alpha.

The Effect of Rate Heterogeneity on Phylogenetic Analyses

Because shape parameters of 0.5 are typically the lower limit of those examined in simulations (Tateno et al. 1994), the behavior of data with the degree of heterogeneity observed in our 12S sequences (fig. 5*B*) has not previously been explored. Phylogenetic analyses of the data set with the most among-site rate variation, the restricted 12S data set, resulted in strong support (90% bootstrap value) for an incorrect placement of *Peromyscus eremicus* (fig. 3*B*). However, because the phylogenetic analyses that accommodate among-site rate variation (neighbor joining on Γ-corrected distances and maximum-likelihood analyses using Baseml) also produced the incorrect resolution of *P. eremicus,* it appears that this problem is not solely the result of among-site rate variation.

There are four sites (sites 129, 476, 564, and 677) that unambiguously support the incorrect placement of *P. eremicus* (fig. 2). Two of these sites (564 and 677) are stem sites, and the substitutions are clearly compensatory (fig. 6). The nonindependence of complementary stem sites has been shown to lead to falsely inflated values of nodal support, and Tillier and Collins (1995) have suggested treatment of two sites that pair-bond as one character in maximum-likelihood and neighbor-joining analyses. The other two sites that unite *P. eremicus* with the *leucopus* group (129 and 476) are loop sites and therefore are not structurally interdependent (fig. 6). If, as is indicated by phylogenetic analysis of the cyt *b* data, the gene tree for the mtDNA is the same as the well-corroborated relationships among these taxa, there may be some unidentified substitution bias leading to the distribution of character states at the sites that supported the placement of *P. eremicus* with the *P. leucopus* group. The nonindependence of sites 564 and 677, coupled with the unidentified nonrandom noise, apparently leads to the relatively strong support for this incorrect placement of *P. eremicus.*

It has been shown that biases in nucleotide sequences (i.e., compositional biases) tend to accumulate at rapidly evolving sites (Hancock et al. 1988). Data sets in which there is a great deal of among-site rate variation (low values of α) will be especially susceptible to the misleading effects of nonrandom noise because a high proportion of the variable sites in such data sets will be high-rate sites. This will allow for the possibility of any unidentified bias that may have accumulated at the high-rate sites to outweigh phylogenetic signal present at more slowly evolving sites. Phylogenetic analyses based on data sets with much among-site rate variation should therefore be interpreted cautiously, especially if they result in topologies that conflict with other data. The relationship between among-site rate variation and susceptibility to potentially misleading effects of nonrandom noise should be explored by simulation studies.

Miyamoto et al. (1994) proposed that phylogenetic analysis of rRNA genes should incorporate heavier weight for stem sites than loop sites because this approach improved their ability to discriminate among alternative trees and led to improved congruence among mitochondrial data sets for five artiodactyl taxa. This is just the opposite of down-weighting stem sites as suggested by Wheeler and Honeycutt (1988) and Dixon and Hillis (1993). The location of the sites that have changed at least three times in our data provides some support for the heavy weighting of stems as most of those sites are indeed in loops (fig. 6). There are, however, several high-rate sites that occur in stem regions, and this explains why heavy weighting of stem sites did not improve the congruence of the 12S data with the cyt *b* data. Thus, it appears that models and weighting schemes that only incorporate secondary structure are inadequate to accommodate the among-site rate heterogeneity observed in these 12S data.

Conclusions

The results of this and other studies (Larson and Wilson 1989; Van de Peer et al. 1993; Yang et al. 1994) suggest that extreme among-site rate variation can exist in rRNA sequences. Phylogenetically informative sites will be a subset of those that change few times in the history of the taxa examined. Thus, those data sets with rate distributions in which a large portion of the variable sites fall into the region that corresponds to one or two changes (data sets with a high value of α) will contain much more information

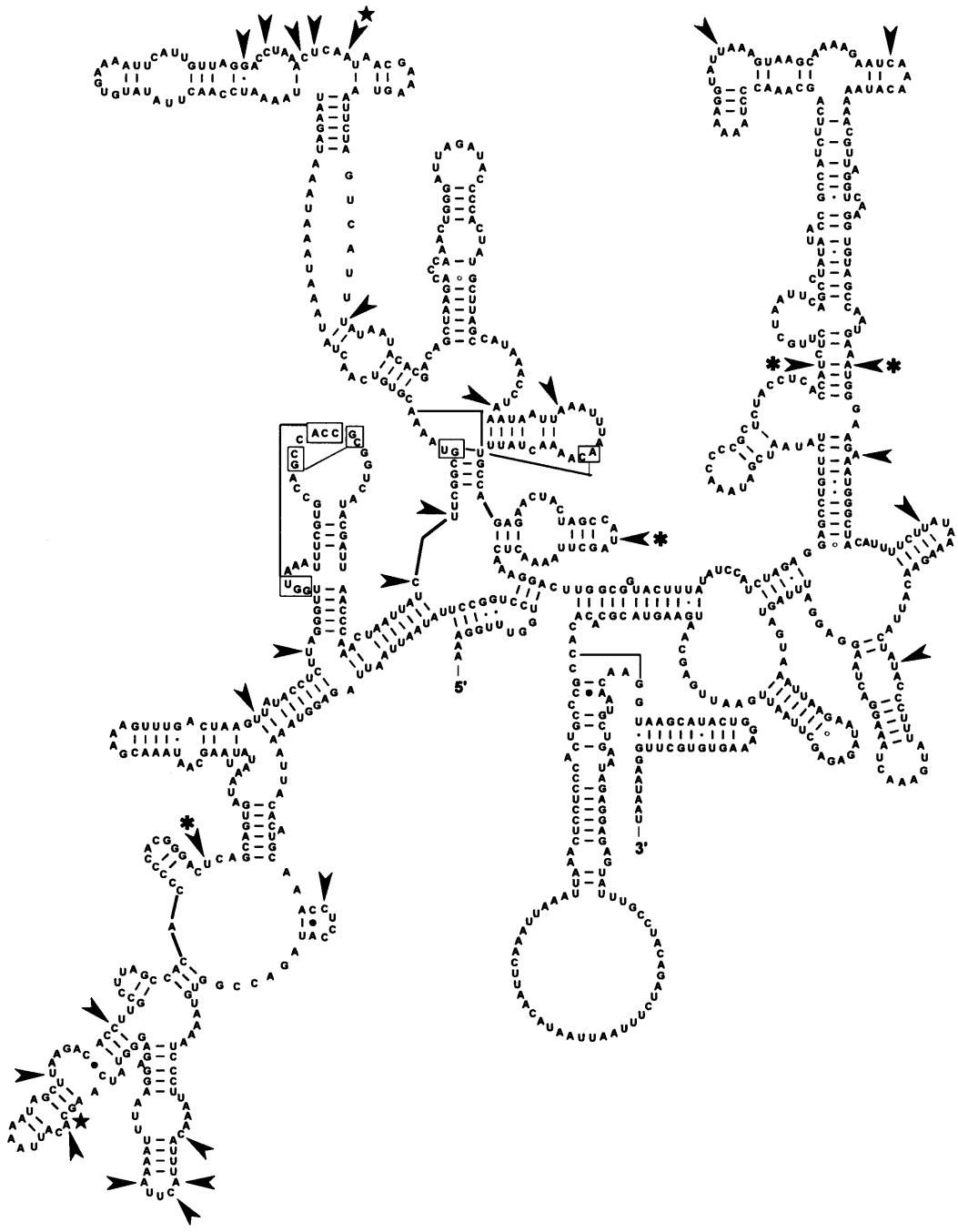

FIG. 6.—The location of high rate sites (those that are observed to change at least three times) identified in the complete 12S data set. Arrows indicate high-rate sites that are unambiguously aligned, asterisks indicate sites that unite *Peromyscus eremicus* with the *P. leucopus* species group, and stars indicate sites that support the accepted resolution of these taxa. The secondary structure is the model of S. Damberger and R. Gutell (unpublished manuscript) for *Mus*.

than data sets with rate distributions in which proportionally fewer of the variable sites fall into that portion of the distribution (data sets with a low value of α). Furthermore, data sets in which there is a great deal of among-site rate variation (low α) will be particularly susceptible to the possibility of misleading nonrandom noise that tends to accumulate at rapidly evolving sites.

Sequence Availability

The nucelotie sequences reported here are available from the GenBank/EMBL databases under the accession numbers X89780–X89799, X89888–X89889.

Acknowledgments

We would like to thank C. W. Kilpatrick (University of Vermont), M. Wooten (Auburn University), and N. R. Holler (U.S. Fish and Wildlife Service) for providing tissues; Brett Riddle (University of Nevada at Las Vegas) for providing aliquots of purified mtDNA; Z. Yang (The Pennsylvania State University) for providing Baseml; and S. Damberger and R. Gutell for providing the unpublished 12S secondary structure for *Mus.* D. Wagner, C. W. Kilpatrick, D. Les, and G. Wallis examined various versions of the paper. R. Honeycutt and an anonymous reviewer provided comments that led to significant improvements. This work was supported by a grant from the University of Connecticut Research Foundation and National Science Foundation (NSF) grant BSR 88-22710, both to C.S. Much of this work was conducted while J.S. was supported by an NSF Graduate Research Traineeship in the evolution, ecology, and conservation of biodiversity (BIR-9256616).

APPENDIX

If substitutions at each nucleotide site in a sequence occur according to a Poisson process but the rate at each site is drawn at random from a gamma distribution with mean μ and shape parameter α, the number of substitutions per site will follow the negative binomial distribution

$$P\{K = k\} = \binom{k + \alpha - 1}{k}\left(\frac{1}{1 + \mu/\alpha}\right)^{\alpha}\left(\frac{\mu/\alpha}{1 + \mu/\alpha}\right)^{k}$$

(Johnson and Kotz 1973). The mean number of substitutions per site is μ, and its variance is $\mu(1 + \mu/\alpha)$. Observations on the number of substitutions per site can be summarized as a set of numbers N_k, where N_k is the number of sites in the sample with k substitutions. Both the maximum-likelihood estimator and the method-of-moments estimator for μ is the mean number of substitutions observed per site, $\bar{x} = \sum_k kN_k / \sum_k N_k$ (Johnson and Kotz 1973).

Molecular evolutionists have generally used a method-of-moments estimator for the shape parameter α ($\hat{\alpha} = \bar{x}^2/[s^2 - \bar{x}]$, where s^2 is the variance of the number of nucleotide substitutions per site), but Anraku and Yanagimoto (1990) showed that a maximum-likelihood estimate of α is preferable because of its smaller mean-squared error, especially when α is small. Unfortunately, no closed-form expression for the maximum-likelihood estimate of α exists, but it can be found numerically by maximizing the log of the likelihood function for the number of substitutions per site:

$$\log(L) = \sum_k N_k\left[\log\binom{k + \alpha - 1}{k} + \alpha \log\left(\frac{1}{1 + \mu/\alpha}\right) + k \log\left(\frac{\mu/\alpha}{1 + \mu/\alpha}\right)\right].$$

LITERATURE CITED

ALLARD, M. D., and R. L. HONEYCUTT. 1992. Nucleotide sequence variation in the mitochondrial 12S rRNA gene and phylogeny of African mole-rats (Rodentia: Bathyergidae). Mol. Biol. Evol. **8**:71–84.

ALLARD, M. D., D. L. ELLSWORTH, and R. L. HONEYCUTT. 1991. The production of single-stranded DNA suitable for sequencing using the polymerase chain reaction. BioTechniques **10**:24–26.

ANRAKU, K., and T. YANAGIMOTO. 1990. Estimation for the negative binomial distribution based on the conditional likelihood. Commun. Statist. Simula. **19**:771–786.

AVISE, J. C., M. H. SMITH, and R. K. SELANDER. 1979. Biochemical polymorphism and systematics in the genus *Peromyscus.* VII. Geographic variation in the *truei* and *manoculatus* species groups. J. Mammal. **60**:177–192.

BECKENBACH, A. T., Y. W. WEI, and H. LIU. 1993. Relationships in the *Drosophila obscura* species group, inferred from mitochondrial cytochrome oxidase II sequences. Mol. Biol. Evol. **10**:619–634.

BIBB, M. J., R. A. VANETTEN, C. T. WRIGHT, M. W. WALBERG, and D. A. CLAYTON. 1981. Sequence and gene organization of mouse mitochondrial DNA. Cell **26**:167–180.

BREMER, K. 1988. The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. Evolution **42**:795–803.

BROWNELL, E. 1983. DNA/DNA hybridization studies of muroid rodents: symmetry and rates of molecular evolution. Evolution **37**:1034–1051.

CABOT, E. L. and A. T. BECKENBACH. 1989. Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. Comput. Appl. Biosci. **5**:233–234.

CARLETON, M. D. 1989. Systematics and evolution. Pp. 7–141 *in* G. L. KIRKLAND and J. N. LAYNE, eds. Advances in the study of *Peromyscus.* Texas Tech University Press, Lubbock.

CATZEFLIS, F. M., J. AGUILAR, and J. JEAGER. 1992. Muroid rodents: phylogeny and evolution. Trends Ecol. Evol. **7**: 122–127.

COLLINS, T. M., F. KRAUS, and G. ESTABROOK. 1994. Compositional effects and weighting of nucleotide sequences for phylogenetic analysis. Syst. Biol. **43**:482–496.

DIXON, M. T., and D. M. HILLIS. 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. **10**:256–267.

FARRIS, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool. **18**:374–385.

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

———. 1993. PHYLIP—phylogeny inference package, version 3.5c University of Washington, Seattle.

FIELD, K. G., G. J. OLSEN, D. J. LANE, S. J. GIOVANNONI, M. T. GHISELIN, E. C. RAFF, N. R. PACE, and R. A. RAFF. 1988. Molecular phylogeny for the animal kingdom. Science **239**:748–753.

FITCH, W. M. 1971. The non-identity of variable positions in the cytochrome *c* of different species. Biochem. Genet. **5**: 231–241.

FITCH, W. M., and E. MARGOLIASH. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case. Biochem. Genet. **1**:65–71.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

GUSTINCICH, S., G. MANFIOLETTI, G. DEL SAL, C. SCHNEIDER, and P. CARNICHI. 1991. A fast method for high quality genomic DNA extraction from whole human blood. BioTechniques **11**:298–301.

GUTELL, R. R., B. WEISER, C. R. WOESE, and H. F. NOELLER. 1985. Comparative anatomy of 16S-like ribosomal RNA. Prog. Nucleic Acid Res. Mol. Biol. **32**:155–216.

HANCOCK, J. M., D. TAUTZ, and G. A. DOVER. 1988. Evolution of the secondary structure and compensatory mutations of the ribosomal RNAs of *Drosophila.* Mol. Biol. Evol. **5**:393–414.

HEDGES, S. B., K. D. MOBERG, and L. R. MAXSON. 1990. Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. Mol. Biol. Evol. 7:607–633.

HIBBARD, C. W. 1968. Paleontology. Pp. 6–26 *in* J. A. KING, ed. Biology of *Peromyscus.* Special Publication no. 2, American Society Mammalologists, Stillwater, Oklahoma.

HICKSON, R. E., C. SIMON, A. COOPER, G. SPICER, J. SULLIVAN, D. PENNY. 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S r RNA. Mol. Biol. Evol. (in press).

HOGAN, K. M., M. C. HEDIN, H. S. KOH, S. K. DAVIS, and I. F. GREENBAUM. 1993. Systematic and taxonomic implications of karyotypic, electrophoretic, and mitochondrial-DNA variation in *Peromyscus* from the Pacific Northwest. J. Mammal. **74**:819–831.

HUELSENBECK, J. P., and D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. **42**: 247–264.

JOHNSON, N. L., and S. KOTZ. 1969. Discrete distributions. Houghton Mifflin, Boston.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KJER, K. M., G. D. BALDRIDGE, and A. M. FALLON. 1994. Mosquito large subunit ribosomal RNA: simultaneous alignment of primary and secondary structure. Biochemica et Biophysica Acta **1217**:147–155.

KOBAYASHI, M., T. SEKI, K. YAGINUMA, and K. KOIKE. 1981. Nucleotide sequences of small ribosomal RNA and adjacent transfer RNA genes in rat mitochondrial DNA. Gene **16**: 297–307.

KOCHER, T. D., W. K. THOMAS, A. MEYER, S. V. EDWARDS, S. PÄÄBO, F. X. VILLABLANCA, and A. C. WILSON. 1989. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proc. Nat. Acad. Sci. USA **86**:6196–6200.

KOCHER, T. D., and A. C. WILSON. 1991. Sequence evolution of mitochondrial DNA in human and chimpanzees: control region and protein coding region. Pp. 391–413 *in* S. OSAWA, and T. HONJO, eds. Evolution of life: fossils, molecules and culture. Springer, Tokyo.

KRAUS, F., L. JARECKI, M. M. MIYAMOTO, S. M. TANHAUSER, and P. J. LAPIS. 1992. Mispairing and compensational changes during the evolution of mitochondrial ribosomal RNA. Mol. Evol. Biol. **9**:770–774.

KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA—Molecular evolutionary genetics analysis, version 1.0. Pennsylvania State University, University Park.

LARSON, A., and A. C. WILSON. 1989. Patterns of ribosomal RNA evolution in salamanders. Mol. Biol. Evol. **6**:131–154.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, and D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

MADDISON, W. P., and D. R. MADDISON. 1992. MacClade—analysis of phylogeny and character evolution, version 3.0. Sinauer, Sunderland, Mass.

MANSKE, C. L., and D. J. CHAPMAN. 1987. Nonuniformity of nucleotide substitution rates in molecular evolution: computer simulation and analysis of 5S ribosomal RNA sequences. J. Mol. Evol. **26**:226–251.

MIYAMOTO, M. M., M. W. ALLARD, R. M. ADKINS, L. L. JANACEK, and R. L. HONEYCUTT. 1994. A congruence test of reliability using linked mitochondrial DNA sequences. Syst. Biol. **43**:236–249.

MUSSER, G. G., and M. D. CARLETON. 1993. Family Muridae. Pp. 501–755 *in* D. E. WILSON and D. M. REEDER, eds. Mammal species of the world: a taxonomic and geographic reference. 2d ed. Smithsonian Institution Press, Washington, D.C.

O'HUIGIN, C., and W.-H. LI. 1992. The molecular clock ticks regularly in muroid rodents and hamsters. J. Mol. Evol. **35**:377–384.

OSGOOD, W. H. 1909. Revision of the mice of the American genus *Peromyscus.* N. Amer. Fauna **28**:1–285.

REEVES, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. **35**:17–31.

RIDDLE, B. R., and R. L. HONEYCUTT. 1990. Historical biogeography in North American arid regions: an approach using mitochondrial-DNA phylogeny in grasshopper mice (genus *Onychomys*). Evolution **44**:1–15.

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

SIMON, C., F. FRATI, A. BECKENBACH, B. CRESPI, H. LIU, and P. FLOOK. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Ann. Ent. Soc. Amer. **87**:651–701.

STANGL, F. B., and R. J. BAKER. 1984. Evolutionary relationships in *Peromyscus:* congruence among chromosomal, genic, and classical data sets. J. Mammal. **65**:643–654.

STEELE, K. P., K. E. HOLSINGER, R. K. JANSEN, and D. W. TAYLOR. 1991. Assessing the reliability of 5S rRNA sequence data for phylogenetic analysis in green plants. Mol. Biol. Evol. **8**:240–248.

SWOFFORD, D. L. 1993. PAUP—phylogenetic analysis using parsimony, version 3.1.1. Illinois Natural History Survey, Champaign.

TAMURA, K., and M. NEI. 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.

TATENO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum likelihood, neighbor joining, and maximum parsimony methods when substitution rate varies with site. Mol. Biol. Evol. **11**:261–277.

TILLIER, E. R. M., and R. A. COLLINS. 1995. Neighbor joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. Mol. Biol. Evol. **12**:7–15.

UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science **172**: 1089–1096.

VAN DE PEER, Y., J.-M. NEEFS, P. DERIJK, and R. DE WACHTER. 1993. Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock. J. Mol. Evol. **37**:221–232.

WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. J. Mol. Evol. **36**:613–623.

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

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

YANG, Z., N. GOLDMAN, and A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. **11**:316–324.

RODNEY L. HONEYCUTT, reviewing editor

Received August 22, 1994

Accepted May 23, 1995