Heterogeneity of Nucleotide Frequencies Among Evolutionary Lineages and Phylogenetic Inference

Michael S. Rosenberg and Sudhir Kumar

Department of Biology and Center for Evolutionary Functional Genomics, Arizona State University

A major assumption of many molecular phylogenetic methods is the homogeneity of nucleotide frequencies among taxa, which refers to the equality of the nucleotide frequency bias among species. Changes in nucleotide frequency among different lineages in a data set are thought to lead to erroneous phylogenetic inference because unrelated clades may appear similar because of evolutionarily unrelated similarities in nucleotide frequencies. We tested the effects of the heterogeneity of nucleotide frequency bias on phylogenetic inference, along with the interaction between this heterogeneity and stratified taxon sampling, by means of computer simulations using evolutionary parameters derived from genomic databases. We found that the phylogenetic trees inferred from data sets simulated under realistic, observed levels of heterogeneity for mammalian genes were reconstructed with accuracy comparable to those simulated with homogeneous nucleotide frequencies; the results hold for Neighbor-Joining, minimum evolution, maximum parsimony, and maximum-likelihood methods. The LogDet distance method, specifically designed to deal with heterogeneous nucleotide frequencies, does not perform better than distance methods that assume substitution pattern homogeneity among sequences. In these specific simulation conditions, we did not find a significant interaction between phylogenetic accuracy and substitution pattern heterogeneity among lineages, even when the taxon sampling is increased.

Introduction

Phylogenetic inference plays a major role in the study of evolution. The accuracy of this inference may have a strong effect on comparative studies that make use of phylogenetic information, such as parameter estimation or the comparative method. A large number of factors may lead to phylogenetic inaccuracy; the use of an incorrect evolutionary model is thought to be one of the most severe, because real data may violate the assumptions of the model in rather extreme ways. Evolutionary models may have any number of factors and assumptions, the most common of which include unequal rates of transitional and transversional substitutions (e.g., Kimura 1980; Wakeley 1994), rate heterogeneity among sites (e.g., Sidow and Steel 1992; Yang 1993), and homogeneity of nucleotide frequencies (e.g., Felsenstein 1988). The latter factor is of particular interest because nucleotide frequencies are now known to vary significantly from organism to organism for a large number of genes (Galtier, Tourasse, and Gouy 1999; Kumar and Gadagkar 2001; Kumar and Subramanian 2002) and there is a general belief that divergence of nucleotide frequencies will lead to erroneous phylogenetic inference (Loomis and Smith 1990; Penny et al. 1990; Sidow and Wilson 1990; Lockhart et al. 1992*a*, 1992*b*; Forterre, Benachenhoulafha, and Labedan 1993; Hasegawa and Hashimoto 1993; Sogin, Hinkle, and Lelpe 1993; Mooers and Holmes 2000; Tarrío, Rodríguez-Trelles, and Ayala 2001).

A number of methods have been developed to account for the problem of compositional heterogeneity among sequences, including distance approaches (Lake 1994; Lockhart et al. 1994; Steel 1994; Galtier and Gouy 1995; Gu and Li 1996, 1998; Galtier, Tourasse, and Gouy 1999; Tamura and Kumar 2002), parsimony approaches (Steel, Lockhart, and Penny 1995), and maximum-likelihood approaches (Yang and Roberts 1995; Galtier and Gouy 1998). Although it has been suggested that the problem may be alleviated by using translated amino acid sequences in place of nucleotide sequences (Loomis and Smith 1990; Hasegawa and Hashimoto 1993; Hashimoto et al. 1995), differences in the nucleotide frequency of the coding regions will bias the amino acid composition of the encoded proteins, leading to heterogeneity in the amino acid sequences (Steel, Lockhart, and Penny 1993; Foster, Jermiin, and Hickey 1997; Foster and Hickey 1999; Singer and Hickey 2000).

Heterogeneity of nucleotide frequencies is thought to cause erroneous phylogenetic inference because unrelated clades with similar nucleotide frequencies (due to convergence rather than shared ancestral frequencies) will be more similar and may group together in a phylogenetic analysis, sometimes with strong statistical support. Facultative convergence can also occur when a single lineage evolves divergent nucleotide frequencies from its sister clade, obscuring the shared phylogenetic history; the frequencies of the sister taxa will resemble the ancestral frequencies and may cause the sister clade to group incorrectly with another, more distantly related clade, which has also maintained the ancestral nucleotide frequencies.

Despite the general belief in problems associated with the heterogeneity of nucleotide frequencies, there is actually very little evidence that heterogeneity leads to significant phylogenetic error in real data analysis. Some simulation studies (e.g., Van Den Bussche et al. 1998; Conant and Lewis 2001) have shown that extreme nucleotide frequency changes are necessary before phylogenetic analysis becomes biased toward an incorrect topology. In examining some "classic" examples of phylogenetic error caused by nucleotide frequency convergence, Conant and Lewis (2001) found that the convergence did not explain the phylogenetic errors. Using simulation, they found that nucleotide convergence by itself, or in conjunction with rate heterogeneity, could not explain the observed errors. In a recent study of vertebrate rhodopsin sequences, Chang and Campbell (2000) found

Key words: LogDet distances, nonstationarity, heterogeneous nucleotide composition, phylogenetic inference, taxon sampling.

E-mail: s.kumar@asu.edu.

Mol. Biol. Evol. 20(4):610–621. 2003
DOI: 10.1093/molbev/msg067

that two incorrect branches were consistently strongly supported and concluded that these errors were due to compositional biases at the third base position. They also found that under some circumstances, taxon sampling could help alleviate the problem if the sampled taxa were on intermediate branches and contained intermediate frequencies.

One problem with claims that heterogeneous nucleotide frequencies lead to incorrect phylogenetic inference is that these studies often do not have access to a well-established independent phylogeny with which to compare the results. These studies show that different results are obtained using methods which assume homogeneity versus those that do not, and they assume that the more complex approach must give the correct answer. In only a few cases (e.g., Chang and Campbell 2000) is the true phylogeny known well enough to show definitively that accounting for heterogeneity helps to recover the correct results when the simpler approaches fail.

While the potential problems of the heterogeneity of nucleotide frequencies with respect to phylogenetic inference have often been discussed, little attention has been given to the potential benefits. Groups of sequences may show similar nucleotide frequencies either because of convergence or because they were inherited as the result of a frequency change in a common ancestor (Conant and Lewis 2001). In the latter case, clades which show different nucleotide compositions than other clades should be easier to reconstruct because their taxa will be more alike than would be expected from common ancestry under a homogeneous model of evolution. General heterogeneity may be as likely to increase phylogenetic accuracy as to decrease it. Therefore, we examined the effects of heterogeneity of nucleotide frequencies on phylogenetic inference using computer simulations with realistic evolutionary parameters modeled on real data observed from genetic databases. In addition, we examined the interaction between heterogeneity and taxon sampling, because increased taxon sampling is sometimes thought to counteract the effects of nucleotide frequency heterogeneity among lineages.

Materials and Methods

Simulations

Our model tree (fig. 1) was based on the 66-taxon tree representing the phylogenetic relationships among Eutherian mammals from Murphy et al. (2001) and Eizrik, Murphy, and O'Brien (2001). In this tree, branch lengths represent the number of substitutions per site (see Rosenberg and Kumar (2001*a*) for more details). The tree was divided into 14 clades based on the mammalian orders (fig. 2). In this tree, the Hystricognathi are a subclade of Rodentia. The single species of Dermoptera (*Cynocephalus variegatus*) was included with the Primates. The orders Hyracoidea, Macroscelides, Sirenia, Proboscidea, and Tubulidentata were grouped into a single clade called Afrotheria (see Hedges 2001; Murphy et al. 2001).

To simulate gene evolution under realistic parameters, we based our simulations on evolutionary parameters derived from genetic databases (Rosenberg and Kumar 2001*a*). Using only genes for which data were available from at least four different mammalian orders (448 total genes, table S1 in Supplementary Information online), we calculated the average nucleotide frequency and average number of codons for all available sequences within each order. Using all available sequences for each gene (3rd codon positions only), we estimated the substitution rate and used PAUP* (Swofford 1998) to give the maximum-likelihood estimate of the transition/transversion rate ratio (κ). This estimate was determined for the Neighbor-Joining tree, using the Hasegawa, Kishino, and Yano (1985) (HKY) model of nucleotide substitution.

Simulations were performed using the HKY model of nucleotide substitution. For each gene, the average nucleotide frequency across all orders was used as the ancestral (or starting) frequencies for each simulation. The number of sites used was three times the average number of codons, and the evolutionary rate and κ were those determined from the real data. The rate $\times 10^9$ was used as a branch length multiplier; e.g., for a gene with an observed rate of 2.3×10^{-9}, each branch length on the tree was multiplied by 2.3. Homogeneous simulations (HOM) were conducted by maintaining the ancestral frequencies as the expected frequencies across the entire tree. Heterogeneity (HET) was introduced by changing the expected nucleotide frequency of each order to match the observed averages for that order; orders for which no data were available were assigned the ancestral frequencies (exception: because the Hystricognathi are a subclade of the Rodentia, the Rodentia frequencies were used when no separate data were available for the Hystricognathi). Each of the 448 individual genes was simulated under both homogeneous and heterogeneous scenarios. Each simulation was replicated 100 times, resulting in 89,600 simulated data sets.

Heterogeneity Analysis

There are no well-established methods for estimating the overall degree of nucleotide heterogeneity among lineages within a data set. We therefore employed two different techniques for estimating this quantity. First, we performed a χ^2-test for every pair of taxa (2,145 pairs per data set) and counted the number of pairs which showed significantly different nucleotide frequencies at $P \leqslant 0.05$. The χ^2-test is a simple, overly conservative approach to determining whether members of a pair of sequences differ significantly in their nucleotide frequencies (Kumar and Gadagkar 2001).

A less conservative test which also works on paired sequences is the Disparity Index (Kumar and Gadagkar 2001), I_D. This test, as originally described, is computationally intensive (requiring a Monte Carlo simulation for every pair to be tested = 193 million pairs in this study) so a shortcut to determining the significance of each pair was employed. In the I_D test, the critical value for determining significance (at a fixed P of 0.05) for a pair of sequences is a function of three values: the sequence length, the number of observed site-by-site differences between the sequences (N_d), and the overall nucleotide skew (deviation of the nucleotide frequencies from 25%). We determined the

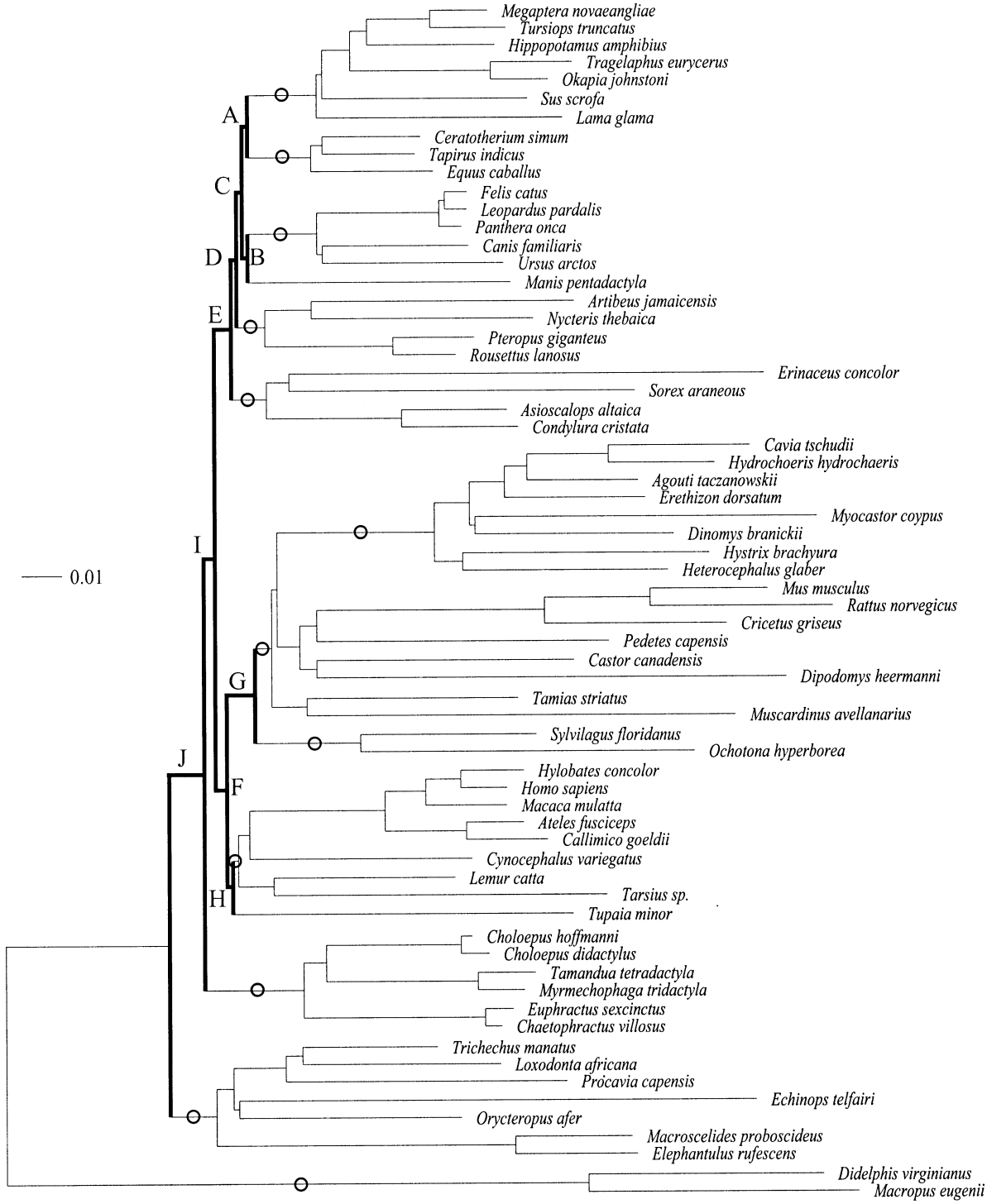

FIG. 1.—Model tree for the simulations based on the Eutherian mammal tree from Murphy et al. (2001). Branch lengths indicate the number of substitutions per site. Ordinal branches (branches which define an order) are marked with circles; interordinal branches (defining relationships among orders) are thick and designated with letters; intraordinal branches (defining relationships within an order) are thin.

critical value of I_D for all possible N_d for sequences of fixed lengths (in multiples of 100 sites) using 10,000 replicate Monte Carlo simulations (Kumar and Gadagkar 2001). We used an extreme nucleotide frequency divergence for these simulations (A = 10%, C = 40%, G = 40%, T = 10%); this is a conservative approach because sequence pairs with more evenly distributed nucleotide frequencies have lower critical values. To determine when a specific pair of simulated sequences showed significantly different nucleotide frequencies using the I_D test, we compared the observed I_D for the pair to the critical value determined from the Monte Carlo simulations. Linear

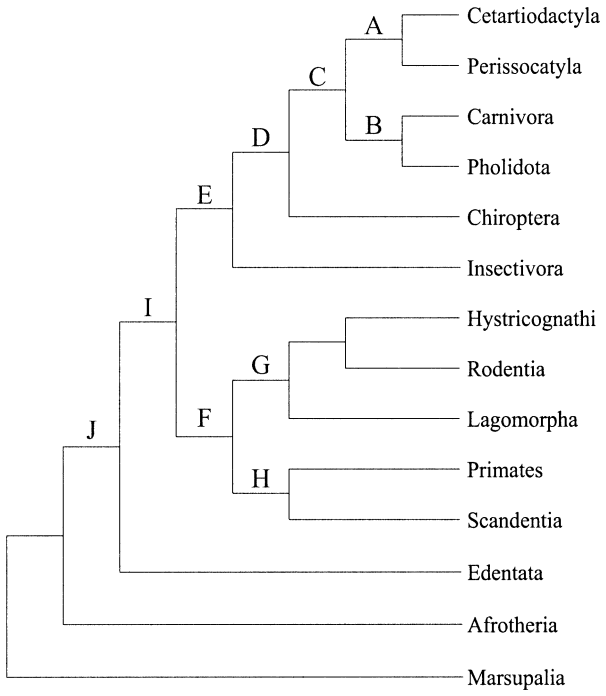

FIG. 2.—Phylogenetic relationships of the mammalian orders present in the assumed model tree. The Hystricognathi are a subclade of the paraphyletic Rodentia. The single species of Dermoptera was included with the Primates. The orders Hyracoidea, Macroscelides, Sirenia, Proboscidea, and Tubulidentata were grouped into the Afrotheria. Letters refer to the same interordinal branches as in figure 1.

extrapolation was used to determine the critical value for sequences with lengths in between those for which the Monte Carlo simulations were performed (this test was restricted to the 370 simulated genes with sequences less than 2,000 sites). As with the χ^2 test, the number of pairs showing significantly different nucleotide frequencies was tallied for each data set.

We have observed that for a fixed sequence length and nucleotide skew, the critical value for significance of the I_D test follows a second-order polynomial regression:

$$D_{C,crit} = b_1 N_D + b_2 N_D^2.$$

The empirically determined regression coefficients (b_1 and b_2) necessary to predict the critical values (both $P = 0.05$ and $P = 0.01$) for the I_D test for a variety of sequence lengths and nucleotide frequency skews can be found in table S2 of the online Supplementary Information.

Phylogenetic Analysis

Phylogenetic analyses were performed in PAUP* (Swofford 1998). A phylogeny was reconstructed for every data set using multiple inference methods, including Neighbor-Joining (NJ), minimum evolution (ME), maximum parsimony (MP), and maximum likelihood (ML). Three distance measures were used for NJ and ME: the Jukes-Cantor (1969) (JC) distance, the Tamura-Nei (1993) (TN) distance, and the LogDet (LD) distance (Lockhart et al. 1994). Although based on a simple model, the JC distance has been shown to lead to more accurate phylogenetic inference than distances based on more complicated (yet correct) models, such as TN and HKY (Takahashi and Nei 2000; Rosenberg and Kumar 2001*b*), for individual gene analysis. LogDet distances were used because they are specifically designed to account for differences in nucleotide frequency among clades. MP analyses were unweighted and ML analyses were conducted using the HKY substitution model. A single heuristic search using nearest-neighbor interchange branch swapping (NNI) was used in the ME, MP, and ML analyses. ME and ML used the NJ tree as the start tree; MP used a stepwise addition method. We did not perform a more thorough search of tree space because the true tree is generally suboptimal and the trees inferred from the simple searches described above are generally already more optimal than the true tree (Nei, Kumar, and Takahashi 1998; Takahashi and Nei 2000). The maximum number of trees that could be saved during the heuristic search procedures was set to 10,000 (most searches never came close to reaching this limit). When multiple trees were found under the ME, MP, and ML procedures, a majority rule consensus tree (retaining all compatible clades even under 50% frequency of occurrence with the LE50 option in PAUP*) was used to create a single resultant tree for each analysis. Because of time constraints, the ML analyses were conducted for only 298 of the 448 genes.

To compare reconstruction accuracy among different branches, the number of times (out of 100 replicates) each internal branch was inferred for a simulated gene was recorded. The internal branches were divided into three categories: ordinal (branches which define an order; open circles), interordinal (branches defining relationships among orders; thick branches), and intraordinal (branches defining relationships within an order; thin branches). These designations are marked on figure 1. The overall accuracy of each reconstruction was tested by calculating the topological distance, d_T (Robinson and Foulds 1981; Penny and Hendy 1985), between the inferred tree and the true (model) tree. This distance is twice the number of interior branches at which the two trees being compared differ.

To test whether increased taxon sampling may improve phylogenetic inference in the face of nucleotide heterogeneity, we subsampled random taxa from each data set and reinferred the phylogeny. There is a common perception that larger taxon samples may mitigate the effect of nucleotide heterogeneity. We stratified the sampling, purposefully spreading the sampled taxa among different clades (Rosenberg and Kumar 2003). One hundred random subsamples of 15, 30, and 45 taxa were constructed, such that each subsample contained at least one species from each of the 14 orders (fig. 2). Each simulation replicate (100 per gene) of each simulated gene (448 genes) used a single random subsample of each size (15, 30, and 45). Each of the 268,800 new data sets was analyzed using all reconstruction methods as above. By stratifying the sampling, we are focusing the analysis toward inferring the relationships among the sampled clades (in our case, mammalian orders): although we may be sampling from the taxa in figure 1, we are interested

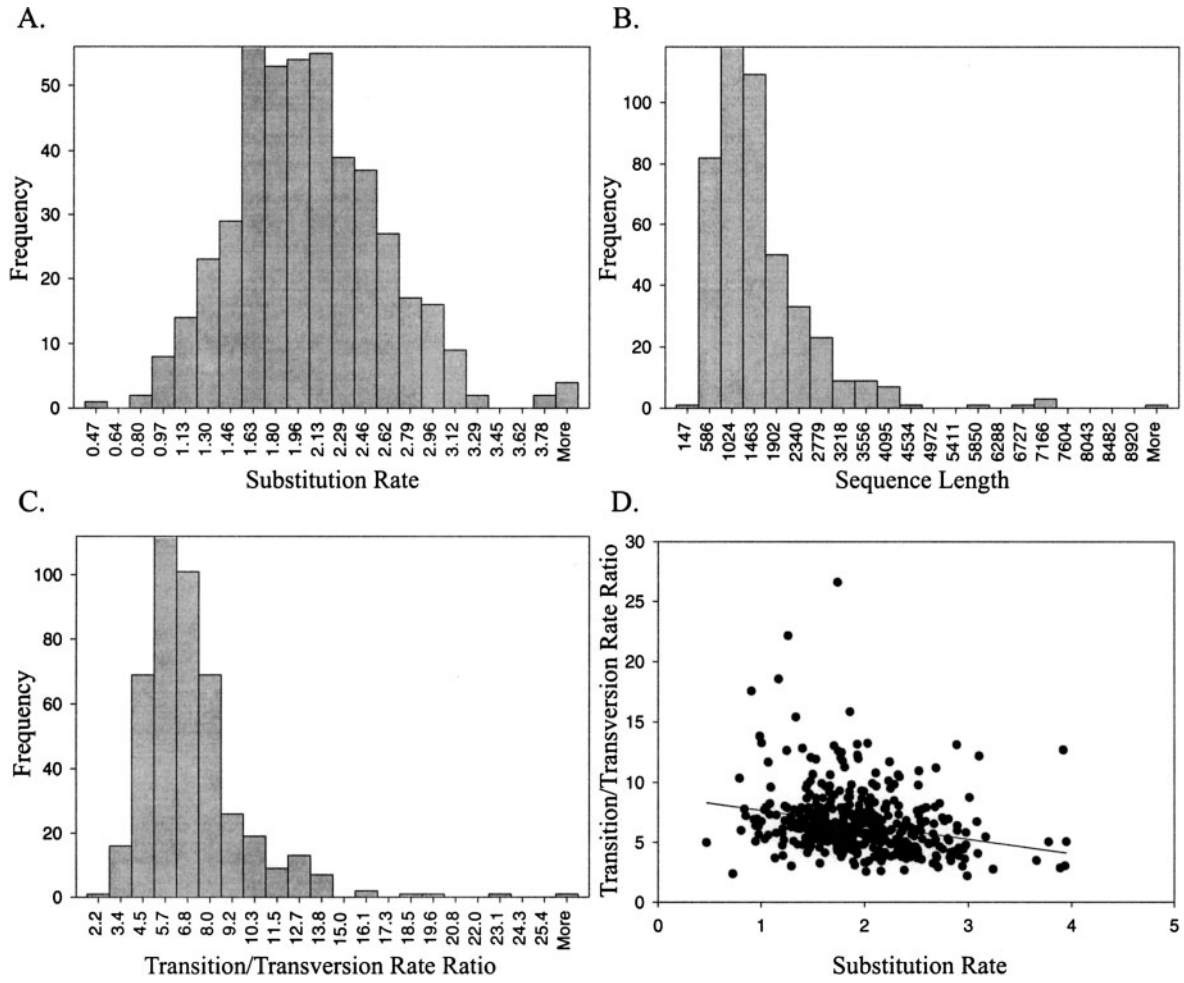

FIG. 3.—Distributions of the input variables for the simulations, derived from data from HOVERGEN and GenBank. (*A*) Histogram of substitution rates. (*B*) Histogram of number of sites. (*C*) Histogram of transition/transversion ratio, κ. (*D*) Scatter plot of κ versus rate.

in recovering the phylogeny in figure 2. Therefore, we evaluate the effects of taxon sampling only with respect to the branches which represent these relationships (the interordinal branches of fig. 1). For the 10 interordinal branches, we simply calculated the percent of replicates in which each branch was reconstructed correctly (the branch connecting the Rodentia and the Hystricognathi was excluded because it is simultaneously an intraordinal and interordinal branch).

Results

There were a median of 1,357 sites per gene (range: 147 to 9,359 sites), a median substitution rate of 1.91 (range: 0.47 to 3.95;), and a median κ of 6.00 (range: 2.19 to 26.61) (fig. 3). There is a negative correlation between the substitution rate and κ ($r = -0.25$), probably because the number of transitions tends to be slightly underestimated for genes with faster substitution rates.

The two measures of heterogeneity (count of significant χ^2 and I_D) are correlated at $r = 0.85$. The χ^2 test found, on average, 5% (maximum 51%) of the sequence pairs with significantly different nucleotide frequencies in the HET data sets, versus only 0.1% in the HOM data sets. In contrast, the I_D tests found 16% (maximum 52%) of the pairs to be significantly different in the HET data versus 6% in the HOM data. The χ^2 test is clearly overly conservative, while the I_D test shows close to the expected type I error. Given the overly conservative nature of the χ^2 test, we will restrict presentation of results to the I_D analyses.

For the phylogenetic analyses of all taxa, the average d_T values for HOM and HET simulations were virtually identical for all methods (table 1 and fig. 4*A*), although the average HET d_T was always slightly smaller than the average HOM d_T. We never find a case where the HET d_T was significantly greater than the HOM d_T (or visa versa). If we examine the worst replicate for each gene (the replicate resulting in the greatest error) rather than the average across all replicates (fig. 4*B*–4*D*), we still find no consistent differences among the HOM and HET data sets. The distributions of genes above and below the equality line are uniform; these distributions also show no particular relationship with sequence length.

All methods showed a strong negative correlation between the number of sites and average d_T—genes with longer sequences were more accurate (table 1). ML, NJ-JC, and ME-JC showed effectively no correlation with

Table 1
Results of Phylogenetic Inference Using All Taxa and Relationship of Phylogenetic Error with Sequence Parameters

	Homogeneous				Heterogeneous			
Method	d_T	r_{sites}[a]	r_{rate}	r_κ	d_T	r_{sites}	r_{rate}	r_κ
NJ-JC	19.6	−0.72	0.02	0.20	19.5	−0.72	0.02	0.20
ME-JC	19.6	−0.72	0.02	0.20	19.5	−0.72	0.01	0.20
NJ-LD	24.2	−0.66	0.13	0.31	24.1	−0.67	0.12	0.30
ME-LD	24.6	−0.63	0.15	0.32	24.4	−0.64	0.14	0.31
NJ-TN	23.4	−0.67	0.14	0.30	23.3	−0.68	0.14	0.29
ME-TN	24.0	−0.63	0.17	0.32	23.9	−0.63	0.17	0.32
MP	20.8	−0.62	0.18	0.29	20.5	−0.63	0.17	0.28
ML[b]	11.8	−0.76	0.04	0.18	11.6	−0.77	0.05	0.18

NOTE.—Values are averages over all replicates and all genes.

[a] The correlations are between d_T and a sequence parameter: number of sites, substitution rate, or κ.

[b] The ML results are based on only 298 of the 448 simulated genes.

Table 2
Correlation Between Phylogenetic Error (d_T) and the Count of Significant I_D Pairs

	HOM	HET	Δ(HET-HOM)[a]
NJ-JC	0.02	−0.09	−0.02
ME-JC	0.02	−0.09	−0.09
NJ-LD	0.03	0.07	−0.29
ME-LD	0.04	0.10	−0.33
NJ-TN	0.04	0.07	−0.15
ME-TN	0.04	0.12	−0.06
MP	0.04	0.03	−0.33
ML[b]	0.00	−0.16	−0.10

NOTE.—Values are averages over all replicates and all genes.

[a] This is the correlation between the difference in d_T and the difference in heterogeneity (measured as the count of paired sequences showing significantly different nucleotide frequencies using the I_D test) between the HET and HOM simulations.

[b] The ML results are based on only 298 of the 448 simulated genes.

substitution rate; all other methods showed a weak positive correlation with substitution rates, slower evolving genes tending to be more accurate. All methods also showed a positive correlation between κ and d_T; higher transition-transversion ratios lead to less accurate phylogenetic reconstructions. This effect was weaker for ML, NJ-JC, and ME-JC. Overall, ML was significantly more accurate than the other methods; the remaining methods were similar in accuracy, with the complex substitution models somewhat less accurate than the simpler substitution models.

We calculated the correlation between d_T and nucleotide heterogeneity in a data set (measured by the count of significant I_D pairs) across all replicates and genes (table 2). There is essentially no correlation between d_T

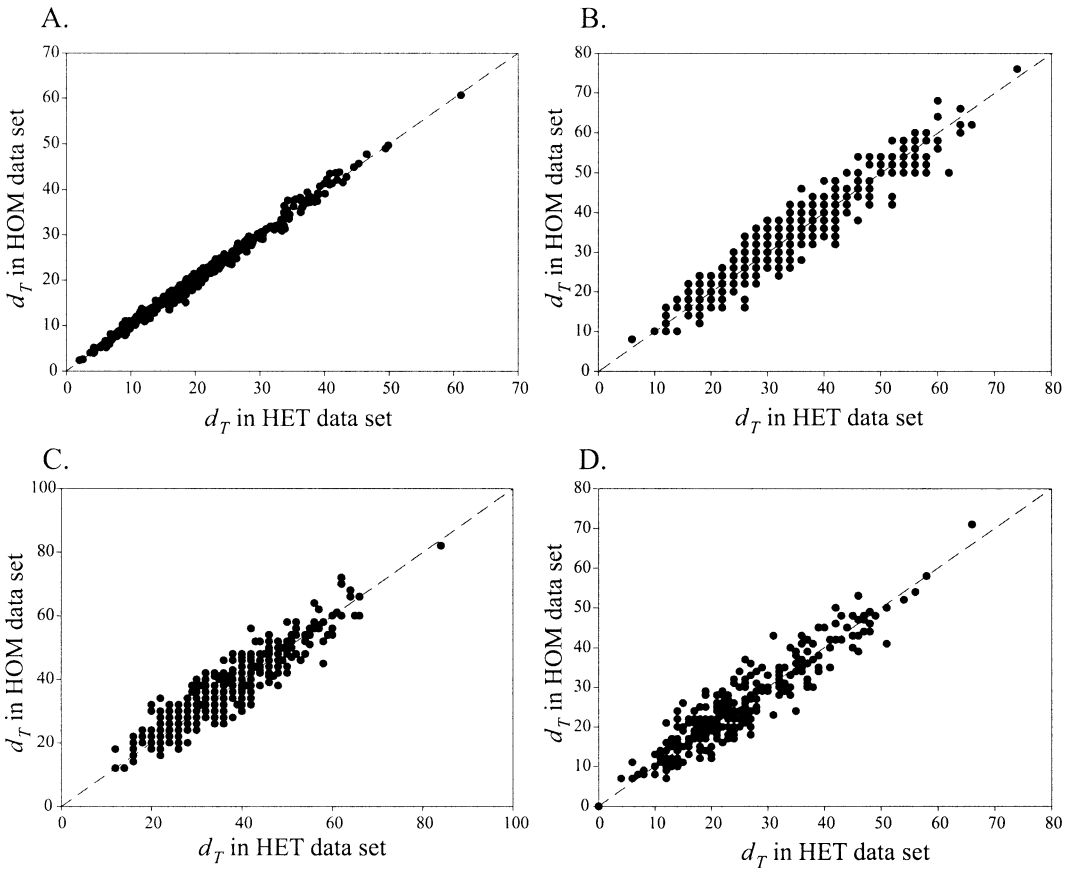

FIG. 4.—Plot of d_T in the HOM simulations versus the HET simulations. (*A*) Average d_T of 100 replicates for a single gene under NJ-JC. The dashed line indicates equality. All other reconstruction methods produced similar results. (*B*) d_T of worst case replicate for each gene under NJ-JC. (*C*) d_T of worst case replicate for each gene under MP. (*D*) d_T of worst case replicate for each gene under ML.

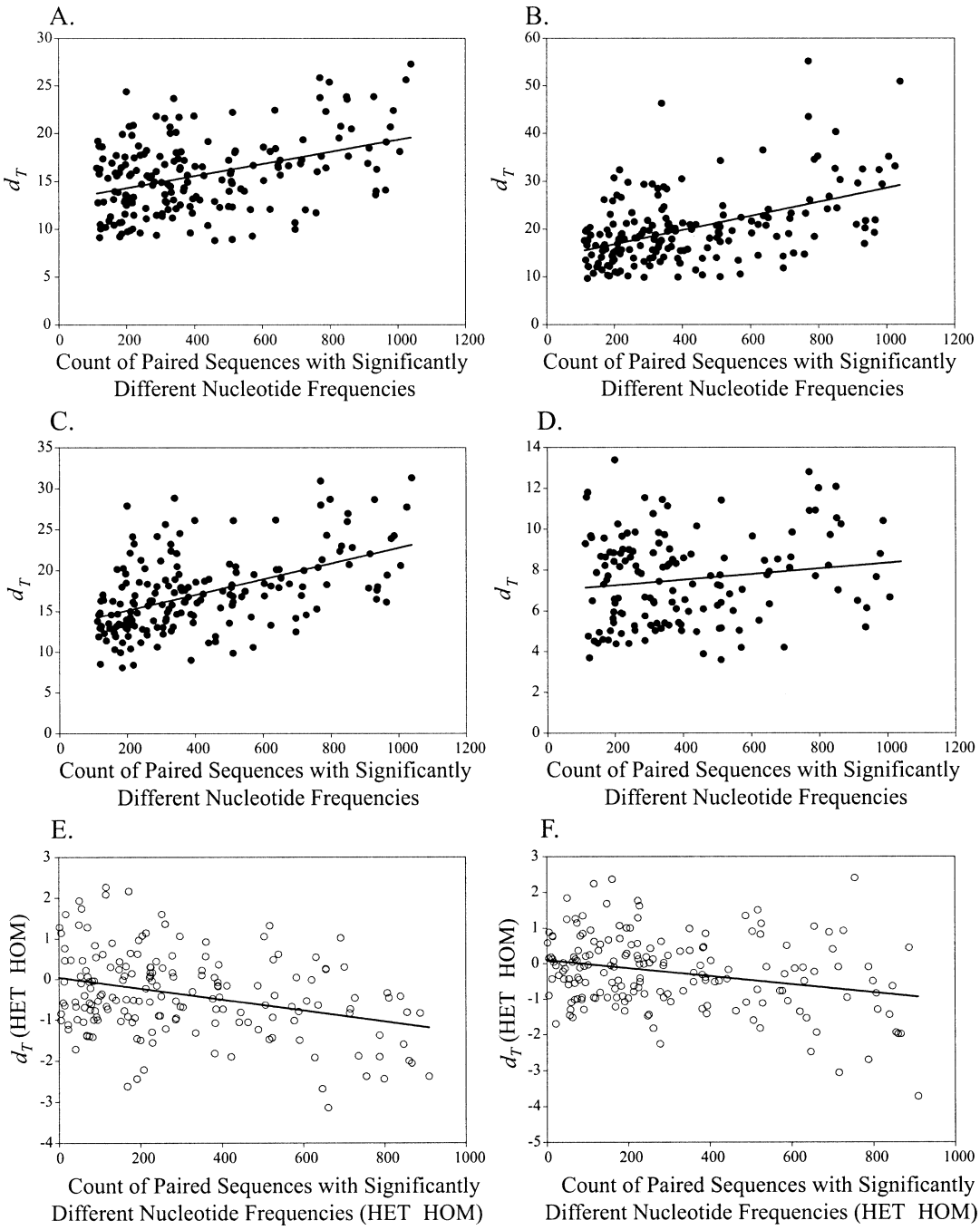

FIG. 5.—Plot of d_T versus the count of paired sequences with significantly different nucleotide frequencies (measured by the I_D test) for genes with sequences between 1,000 and 2,000 sites. (*A–D*) HET simulations: (*A*) NJ-JC, (*B*) NJ-LD, (*C*) MP, (*D*) ML. (*E–F*) Difference in values for the HET and HOM simulations: (*E*) MP, (*F*) NJ-LD.

and the nucleotide heterogeneity for most methods. In the HET simulations, there is a weak negative correlation under ML; this indicates that genes with greater heterogeneity are reconstructed with less error. However, there is a complication with these correlations: they are averaged across genes with different sequence lengths, transition biases, and rates of evolution, all of which show some degree of correlation with phylogenetic accuracy (table 1).

We can partially remove the effect of sequence length by examining genes with a limited range (1,000 to 2,000 sites; 181 genes). Within the HET simulation, there is a relationship between phylogenetic accuracy and nucleotide heterogeneity (fig. 5 *A–D*); genes with more heterogeneity show more error. This effect is particularly strong under the MP method. However, this comparison includes the effects of other variables such as substitution rate and κ, both of which are not only positively correlated with accuracy (table 1) but also with nucleotide heterogeneity (e.g., correlation between the average number of paired sequences with significantly different nucleotide

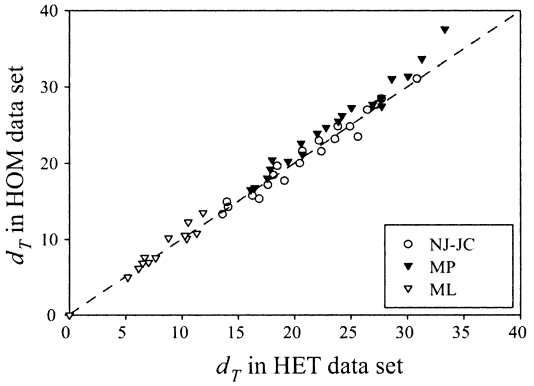

FIG. 6.—Plot of average d_T in the HOM simulations versus the HET simulations for the 20 genes with the largest degree of nucleotide heterogeneity among sequences (measured by I_D). The dashed line indicates equality.

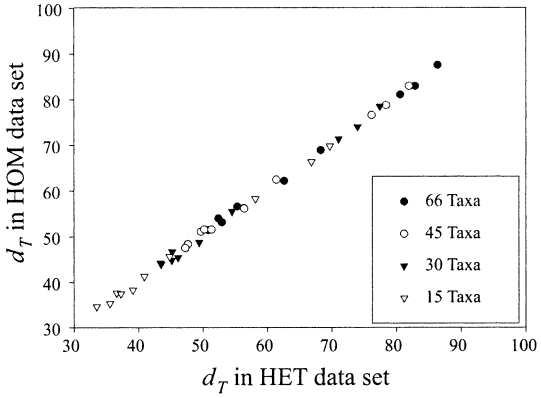

FIG. 7.—Plot of percent accuracy of reconstruction of the 10 interordinal branches in the HOM simulations versus the HET simulations for NJ-JC. Each point represents the average across all genes. All other reconstruction methods produced similar plots.

frequencies and substitution rate for HET simulations, $r = 0.39$). Interestingly, use of LogDet distances does not lead to lower d_T values.

An approach for removing the effects of sequence length and rate (among other factors) is to examine the differences between the HET and HOM simulations, which will allow us to keep all external parameters constant. We determined the correlation between the difference in average d_T for a gene ($d_{T,\mathrm{HET}} - d_{T,\mathrm{HOM}}$) versus the difference in nucleotide heterogeneity for that gene (count $I_{D,HET}$ − count $I_{D,HOM}$). The results (table 2, fig. 5*E–F*) indicate relatively little to moderate negative correlation between the difference in accuracy and the difference in nucleotide heterogeneity among genes simulated under otherwise identical parameters. This indicates that as the amount of nucleotide heterogeneity in a data set increases (while holding all else constant), there was, on average, less error in the phylogenetic reconstruction. These correlations are virtually unchanged whether we include all genes or just those with 1,000 to 2,000 sites (fig. 5*E–F*). The positive association between error and nucleotide heterogeneity illustrated in figure 5*A* through 5*D* appears to be due to correlated factors (e.g., rate of evolution and transition/transversion rate ratios) rather than indicative of causation.

When we examined the 20 genes with the greatest heterogeneity in nucleotide frequencies among simulated HET sequences (i.e., the 20 genes with the highest mean number of significantly different sequences using the I_D test), we again find little difference between the accuracy of the HOM and HET simulations (fig. 6). In fact, for the MP method, there is slightly more error in the HOM simulations than the HET for almost all of the genes; this is consistent with table 1. The average sequence length of these 20 genes was 1,287 sites.

There were virtually no differences among the phylogenetic inferences of the HOM and HET data sets when examined on a branch-by-branch basis. The average difference (HET minus HOM) between branch reconstruction efficiency (percent of replicates inferring a branch correctly) across all genes and branches was $<0.25\%$ for all methods. The magnitudes of the average differences were still essentially zero when subdividing the data according to branch type (ordinal, intraordinal, or interordinal). The branch with the largest accuracy difference between HOM and HET simulations was the ordinal branch leading to rodents; this branch was reconstructed more accurately (2%–5%) in the HET simulations than the HOM simulations for all methods except NJ and ME using LD distances (for these methods the difference was $<1\%$). No other individual branch showed much difference in accuracy with respect to nucleotide homogeneity.

There were no differences between the HOM and HET data sets when their phylogenies were reconstructed using incomplete taxon samples (fig. 7 and fig. 8). We find no evidence that sampling more taxa will affect the accuracy of phylogenetic reconstruction for data sets with heterogeneous nucleotide frequencies in any way different than taxon sampling would affect homogeneous data sets (Rosenberg and Kumar 2001*a*, 2003).

Discussion

Our results reveal that the observed heterogeneity of nucleotide frequencies among mammalian lineages is likely to have little effect on the accuracy of the reconstructed phylogeny. These results were consistent for all methods examined, regardless of overall accuracy. This is not to say that heterogeneity will never cause errors; these errors are simply probably not as common as previously suggested, and other potential problems should be explored in more depth. Although the present results show that heterogeneity of nucleotide frequencies may not be as important a factor in phylogenetic analysis as often believed, accounting for heterogeneity is still an important aspect of many other evolutionary analyses, e.g., in the determination of substitution rates (Tourasse and Li 1999; Kumar and Subramanian 2002; Tamura and Kumar 2002).

As has been shown previously (Graybeal 1998; Poe 1998; Poe and Swofford 1999; Yoder and Irwin 1999; Kumar and Gadagkar 2000; Rosenberg and Kumar 2001*a*, 2001*b*, 2003), sequence length has a strong positive effect

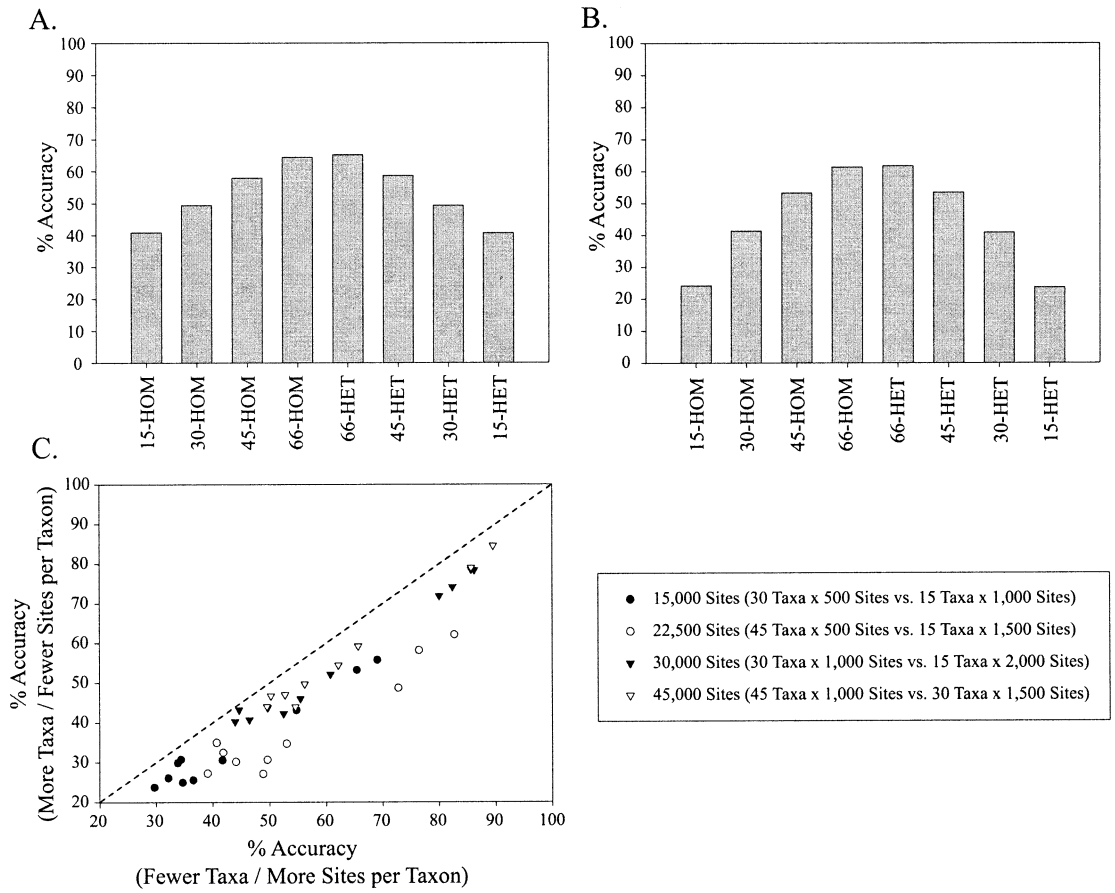

FIG. 8.—Plots of accuracy of reconstruction for specific branches. (*A*) Percent accuracy of reconstruction under MP for branch D for different taxon samples and nucleotide heterogeneity. (*B*) Percent accuracy of reconstruction under MP for branch F for different taxon samples and nucleotide heterogeneity. (*C*) Percent accuracy of reconstruction under NJ-JC of the 10 interordinal branches for different taxon samples, but a fixed number of total bases. Similar plots were produced for all interordinal branches and by all reconstruction methods.

on the accuracy of phylogenetic inference for all inference methods. Not surprisingly, other factors also have significant effects. For instance, increased transition/transversion ratios decrease the accuracy of inferred phylogenies, even when a complex model of nucleotide substitution is used during phylogeny reconstruction. Some methods (MP, and NJ/ME with LD and TN distances) were somewhat more accurate with slower substitution rates, but as a general rule one expects to find the greatest accuracy at an intermediate rate (predicated on the time between speciation events). Very slow substitution rates lead to zero length branches in gene trees which cannot be resolved (Kumar and Gadagkar 2000), whereas very fast substitution rates lead to multiple hits and the saturation of information content. Increasing the number of sites can potentially eliminate the zero branch length problem of slow evolving genes. However, restricting our comparisons of the HOM and HET data sets to those genes with long sequences or slow substitution rates changes none of our conclusions.

It has previously been shown that distance methods of phylogenetic inference seem to work better when simpler distance measures are in use (e.g., *p*-distance or JC) than when more complex distance measures (i.e., TN) are employed (Takahashi and Nei 2000; Rosenberg and Kumar 2001*b*); the present study reconfirms these prior observations. An additional observation is that use of a distance measure explicitly designed to deal with compositional heterogeneity (LogDet) also performed worse than the simpler JC distance. This difference is most likely because the variances associated with JC distance estimates are much smaller than those found with more complicated distance measures, leading to more accurate phylogenetic inferences (Nei and Kumar 2000). Because the variance should decrease with longer sequences, we determined the average d_T in the HET simulations for the longest sequences (more than 3,000 sites) and the shortest sequences (fewer than 500 sites). For the shortest sequences (57 genes), the mean d_T for NJ-JC was 37.2, whereas that of NJ-LD was 45.0; for the longest sequences (30 genes), the mean d_T for NJ-JC was 4.9 and that of NJ-LD was 5.0. If we restrict the comparison to the 5 genes with sequences greater than 5,000 sites, the mean d_T for NJ-JC was 3.8 while that of NJ-LD was 3.3. These values support the idea that LogDet distances were failing to produce accurate results because of the large variances associated with shorter sequences; but, extremely long sequences were needed before LogDet began to outperform the simpler JC measure. LogDet distances are also known to perform poorly when there is

site rate variation (Baake and von Haeseler 1999; Waddell et al. 1999; Mooers and Holmes 2000); this is not an issue in the present study, because our simulations allowed all sites within a gene to mutate with a uniform rate.

Most of the evidence for the problems of heterogeneous nucleotide frequencies was based on the fact that different trees are obtained when using LogDet distances versus distances which do not account for the heterogeneity. As previously mentioned, many of these studies lacked a well established "true" phylogeny with which to compare the results. There is an unfortunate tendency in phylogenetic inference to believe results based on complex models rather than simple models, even when there is no corroborating evidence to support either model (although see Posada and Crandall 2001; Sullivan and Swofford 2001). Our simulation study reveals that while different results were obtained when using LogDet versus simpler distance measures, the simpler measures tended to lead to more accurate results for the nucleotide frequency heterogeneities modeled.

While there are no differences between homogeneous and heterogeneous data sets with respect to taxon sampling, stratified taxon sampling does have an effect on phylogeny reconstruction (fig. 8*A* and *B*). There is a clear increase in accuracy of reconstruction of the interordinal branches when larger taxon samples are used, and this improved accuracy appears to be due to the overall increase in data available for analysis (Rosenberg and Kumar 2003). When the total amount of input data is held constant, however, we find that having more sites for fewer taxa leads to more accurate inference than having fewer sites for more taxa (fig. 8*C*).

This study was modeled on the observed nucleotide heterogeneity found among mammalian orders. Whereas certain clades (e.g., rodents) are known to shown strong differences in nucleotide content relative to other mammals (Kumar and Gadagkar 2001; Kumar and Subramanian 2002), these differences appear to be have little effect on phylogeny reconstruction. Even strong nucleotide convergence among disparate clades appears to seldom cause phylogenetic errors (Conant and Lewis 2001). Therefore, it appears that problematic phylogenies may best be tackled by increasing the total data analyzed.

Appendix

One may compute the significance for the Disparity Index test (Kumar and Gadagkar 2001) using the following shortcut in place of Monte Carlo simulation. For the observed data, the Disparity Index is simply

$$I_D = D_C - N_d,$$

where N_d is the observed number of differences between the sequences and D_C is a measure of the composition difference between the sequences,

$$D_c = \frac{1}{2}\sum_i (x_i - y_i)^2,$$

where x_i and y_i are the counts of state i in sequences x and y, respectively. The frequency skew (F) is measured as

$$F = \Sigma(f_i - 0.25)^2,$$

where f_i is the relative frequency of nucleotide i (averaged across both sequences). The critical value of D_C necessary for significance can be determined by:

$$D_{C,crit} = b_1 N_D + b_2 N_D^2,$$

where b_1 and b_2 are found in table S2. For values of F not found in the table, it is more conservative to use a stronger skew (larger value of F). For sequence lengths (n_i) intermediate to those in the table, we calculate $D_{C,crit}$ for the bracketing sequence lengths ($n_1 < n_i < n_2$) and use linear extrapolation to estimate the critical value for our observed sequence length:

$$D'_{C,crit} = pD_{C,crit-2} + (1-p)D_{C,crit-1},$$

where

$$p = \frac{n_i - n_1}{n_2 - n_i}.$$

Acknowledgments

We thank A. Filipski, S. Gadagkar, and anonymous reviewers for comments on earlier versions of the manuscript, and D. Schwartz for computational help. This research was supported by grants from the National Science Foundation (DBI-9983133), the National Institutes of Health (HG-02096), and the Burroughs Wellcome Fund (BWI 1001311) to S.K., and National Science Foundation grant IBN-9977063 to J. L. Collins.

Literature Cited

Baake, E., and A. von Haeseler. 1999. Distance measures in terms of substitution processes. Theor. Pop. Biol. **55**:166–175.

Chang, B. S. W., and D. L. Campbell. 2000. Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. Mol. Biol. Evol. **17**:1220–1231.

Conant, G. C., and P. O. Lewis. 2001. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. Mol. Biol. Evol. **18**:1024–1033.

Eizrik, E., W. J. Murphy, and S. J. O'Brien. 2001. Molecular dating and biogeography of the early placental mammal radiation. J. Hered. **92**:212–219.

Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Gen. **22**:521–565.

Forterre, P., N. Benachenhou-lafha, and B. Labedan. 1993. Universal tree of life. Nature **362**:795.

Foster, P. G., and D. A. Hickey. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. **48**:284–290.

Foster, P. G., L. S. Jermiin, and D. A. Hickey. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. **44**:282–288.

Galtier, N., and M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA **92**:11317–11321.

———. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA

sequence evolution for phylogenetic analysis. Mol. Biol. Evol. **15**:871–879.
Galtier, N., N. J. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science **283**:220–221.
Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. **47**:9–17.
Gu, X., and W.-H. Li. 1996. Bias-corrected paralinear and LogDet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. Mol. Biol. Evol. **13**:1375–1383.
———. 1998. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc Natl. Acad. Sci. USA **95**:5899–5905.
Hasegawa, M., and T. Hashimoto. 1993. Ribosomal RNA tree misleading? Nature **361**:23.
Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.
Hashimoto, T., Y. Nakamura, T. Kamaishi, F. Nakamura, J. Adachi, K.-i. Okamoto, and M. Hasegawa. 1995. Phylogenetic place of mitochondrion-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2. Mol. Biol. Evol. **12**:782–793.
Hedges, S. B. 2001. Afrotheria: Plate tectonics meets genomics. Proc Natl Acad Sci USA **98**:1–2.
Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.
Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.
Kumar, S., and S. R. Gadagkar. 2000. Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies. J. Mol. Evol. **51**: 544–553.
———. 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. Genetics **158**:1321–1327.
Kumar, S., and S. Subramanian. 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. USA **99**:803–808.
Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91**:1455–1459.
Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. D. Larkum. 1992*a*. Substitutional bias confound inference of cyanelle origins from sequence data. J. Mol. Evol. **34**: 153–162.
Lockhart, P. J., D. Penny, M. D. Hendy, C. J. Howe, T. J. Beanland, and A. W. D. Larkum. 1992*b*. Controversy on chloroplast origins. FEBS Lett. **301**:127–131.
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.
Loomis, W. F., and D. W. Smith. 1990. Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. Proc. Natl. Acad. Sci. USA **87**:9093–9097.
Mooers, A. Ø., and E. C. Holmes. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. **15**:365–369.
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, Oxford.
Nei, M., S. Kumar, and K. Takahashi. 1998. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc. Natl. Acad. Sci. USA **95**:12390–12397.
Penny, D., and M. D. Hendy. 1985. The use of tree comparison metrics. Syst. Zool. **34**:75–82.
Penny, D., M. D. Hendy, E. A. Zimmer, and R. K. Hambry. 1990. Trees from sequences: panacea or Pandora's box. Aust. Syst. Bot. **3**:21–38.
Poe, S. 1998. The effect of taxonomic sampling on accuracy of phylogeny estimation: test case of a known phylogeny. Mol. Biol. Evol. **15**:1086–1090.
Poe, S., and D. L. Swofford. 1999. Taxon sampling revisited. Nature **398**:299–300.
Posada, D., and K. A. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. **50**:580–601.
Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.
Rosenberg, M. S., and S. Kumar. 2001*a*. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA **98**:10751–10756.
———. 2001*b*. Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well. Mol. Biol. Evol. **18**:1823–1827.
———. 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. **52**:119–124.
Sidow, A., and T. P. Steel. 1992. Estimating the fraction of invariable codons with a capture-recapture method. J. Mol. Evol. **35**:253–260.
Sidow, A., and A. C. Wilson. 1990. Compositional statistics: An improvement of evolutionary parsimony and its application to deep branches in the tree of life. J. Mol. Evol. **31**:51–68.
Singer, G. A. C., and D. A. Hickey. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol. Biol. Evol. **17**:1581–1588.
Sogin, M. L., G. Hinkle, and D. D. Lelpe. 1993. Universal tree of life. Nature **362**:795.
Steel, M. A. 1994. Recovering a tree from the leaf colouration it generates under a Markov model. Appl. Math. Lett. **7**:19–23.
Steel, M. A., P. J. Lockhart, and D. Penny. 1993. Confidence in evolutionary trees from biological sequence data. Nature **364**:440–442.
———. 1995. A frequency-dependent significance test for parsimony. Mol. Phyl. Evol. **4**:64–71.
Sullivan, J., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. **50**:723–729.
Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.
Takahashi, K., and M. Nei. 2000. Efficiencies of fast algorthims 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.
Tamura, K., and S. Kumar. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol. Biol. Evol. **19**:1727–1736.
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.
Tarrío, R., F. Rodríguez-Trelles, and F. J. Ayala. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. **18**:1464–1473.

Tourasse, N. J., and W.-H. Li. 1999. Performance of the relative-rate test under nonstationary models of nucleotide substitution. Mol. Biol. Evol. **16**:1068–1078.

Van Den Bussche, R. A., R. J. Baker, J. P. Huelsenbeck, and D. M. Hillis. 1998. Base compositional bias and phylogenetic analyses: A test of the "flying DNA" hypothesis. Mol. Phyl. Evol. **10**:408–416.

Waddell, P. J., Y. Cao, J. Hauf, and M. Hasegawa. 1999. Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid-invariant sites-logdet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo, and elephant. Syst. Biol. **48**:31–53.

Wakeley, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. **11**:436–442.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

Yang, Z., and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. **12**:451–458.

Yoder, A. D., and J. A. Irwin. 1999. Phylogeny of the Lemuridae: effects of character and taxon sampling on resolution of species relationships within *Eulemur*. Cladistics **15**:351–361.

Naruya Saitou, Associate Editor

Accepted December 3, 2002