Sampling Properties of DNA Sequence Data in Phylogenetic Analysis

Michael P. Cummings,[1] *Sarah P. Otto,*[2] *and John Wakeley*[3]

Department of Integrative Biology, University of California at Berkeley

We inferred phylogenetic trees from individual genes and random samples of nucleotides from the mitochondrial genomes of 10 vertebrates and compared the results to those obtained by analyzing the whole genomes. Individual genes are poor samples in that they infrequently lead to the whole-genome tree. A large number of nucleotide sites is needed to exactly determine the whole-genome tree. A relatively small number of sites, however, often results in a tree close to the whole-genome tree. We found that blocks of contiguous sites were less likely to lead to the whole-genome tree than samples composed of sites drawn individually from throughout the genome. Samples of contiguous sites are not representative of the entire genome, a condition that violates a basic assumption of the bootstrap method as it is applied in phylogenetic studies.

Introduction

Much of the work in molecular systematics and evolution rests on the premise that sampled data, such as nucleotide sequences, are representative of the genomes from which they are drawn. Under this assumption, phylogenetic relationships inferred from sampled sequences are viewed as accurate estimates of those that would be obtained from an analysis of the entire genome. In turn, these relationships are presumed to represent those of the organisms involved. In spite of the fundamental importance of this assumption in molecular phylogenetic studies, the sampling properties of DNA sequence data have not been generally assessed.

The DNA sequences of single genes, in part or whole, comprise the most common type of sequence sample used in molecular phylogenetic studies. Although some effort has been made to study the phylogenetic utility of particular genes (Graybeal 1994), the location and length of sequences are generally chosen on the basis of factors other than their ability to accurately represent the genome. These factors include interest in the functional characteristics of a region, its historical use in systematic studies, and technical considerations, which may be unrelated to its ability to reconstruct whole-genome relationships. Of course, how well a gene represents the entire genome is difficult to evaluate a priori.

The purpose of the present study is to investigate the sampling properties of DNA sequence data in phylogenetic analysis using the vertebrate mitochondrial genome as a model system. This system offers several advantages. First, since whole genomes of many organisms are available, we were able to obtain the entire population of sites from which samples can be drawn. Thus, we could study the sampling properties of these sequences, determining the extent to which conclusions based on genes reflect those based on the entire genome. Second, since recombination is rare in vertebrate mitochondria and few within-species polymorphisms are expected to predate the speciation events among these taxa, all portions of these genomes must share a common history. This means that all mitochondrial genes should provide evidence for a single organismal tree. Finally, these sequences constitute empirical data that have resulted from the evolutionary processes among this diverse group of species. They are free of the simplifying assumptions made in simulation studies and embody many of the problems regularly encountered in molecular phylogenetics.

Using the complete mitochondrial genomes of two actinopterygian fish, one amphibian, one bird, and six mammals, we addressed the following specific questions regarding the sampling properties of DNA sequence data

[1] Present address: Department of Botany and Plant Sciences, University of California, Riverside.

[2] Present address: Department of Zoology, University of British Columbia.

[3] Present address: Department of Population Genetics, National Institute of Genetics.

Key words: phylogenetics, neighbor joining, parsimony, maximum likelihood, mitochondrial genome, DNA, bootstrap.

Address for correspondence and reprints: Michael P. Cummings, Department of Botany and Plant Sciences, University of California, Riverside, California 92521. E-mail: mike@uws5.biol.berkeley.edu.

Mol. Biol. Evol. 12(5):814–822. 1995.

0737-4038/95/1205-0010$02.00

in phylogenetic analysis. Do these genes constitute adequate samples to infer relationships that would be obtained using the whole genome? How many sites are needed before we are reasonably sure of inferring the whole-genome tree? How do alternative sampling schemes differ, and what are the implications of the differences?

To address these questions we analyzed samples, both random and nonrandom, drawn from the 10 genomes. The nonrandom samples are the individual genes, and we considered two types of random samples: sets of nucleotide sites sampled individually from throughout the genome without replacement and contiguous blocks of sites. Phylogenetic relationships were inferred from these samples and the results compared to the relationships inferred using the entire genome. We used three methods of phylogenetic inference—maximum likelihood, parsimony, and neighbor joining—not for purposes of comparing methods but rather to determine whether the sampling properties of these DNA sequence data depend on the choice of a particular phylogenetic method. Therefore, we chose commonly applied forms of these methods and made no effort to optimize the performance of each.

Material and Methods

DNA sequences from GenBank (carp, *Cyprinus carpio,* X61010 [Chang et al. 1994]; loach, *Crossostoma lacustre,* M91245 [Tzeng et al. 1992]; frog, *Xenopus laevis,* M10217, X01600, X01601, X02890 [Roe et al. 1985]; chicken, *Gallus gallus,* X52392 [Desjardins and Morais 1990]; mouse, *Mus musculus,* V00711 [Bibb et al. 1981]; rat, *Rattus norvegicus,* X14848 [Gadaleta et al. 1989]; cow, *Bos taurus,* V00654 [Anderson et al. 1982]; whale, *Balaenoptera physalus,* X61145 [Arnason et al. 1991]; seal, *Phoca vitulina,* X63726 [Arnason and Johnson 1992]; and human, *Homo sapiens,* V00662 [Anderson et al. 1981]), exclusive of the control region, were aligned using CLUSTAL V (Higgins et al. 1992) with manual adjustments. Coding regions were aligned by their corresponding amino acid sequences, and nucleotide sequences were made to conform to this. Sites in overlapping genes are duplicated in the alignment, and intergenic regions are eliminated. The resulting data consist of 16,075 sites from each of 10 species.

Maximum-likelihood analyses (Felsenstein 1981) were done using computer code modified from fastDNAml version 1.06 (Olsen et al. 1994), with a transition:transversion ratio of 10:1, empirical base frequencies, one rate class, and global branch swapping. Parsimony analyses (Fitch 1971) were done using code modified from the branch-and-bound algorithm in MEGA (Kumar et al. 1993). For parsimony analyses, gaps were treated as missing data, and all characters and character step changes were weighted equally. Neighbor-joining analyses (Saitou and Nei 1987) were done using computer code modified from neighbor.c, a program in PHYLIP 3.41 (Felsenstein 1991), with Kimura's two-parameter model distances (Kimura 1980), calculated using code modified from CLUSTAL V (Higgins et al. 1992). Consensus trees were then constructed using computer code modified from consense.c, part of PHYLIP 3.5c (Felsenstein 1993).

Results

Whole-Genome Phylogeny

The trees inferred from the entire genome using maximum-likelihood, parsimony, and neighbor-joining methods are identical (fig. 1). Under all three tree-building methods, the bootstrap support for every clade, except one, was greater than 0.996. The remaining clade, whale-cow-seal-human, still had a relatively high bootstrap proportion regardless of method: 0.879 under neighbor joining, 0.952 under parsimony, and 0.996 under maximum likelihood. While there is other evidence supporting the relationships shown in figure 1 (Honeycutt and Adkins 1993), whether this tree depicts true relationships among these species is immaterial for our purposes. We are interested only in evaluating how well samples of the mitochondrial genome represent the genome. Thus, we need only know that this tree is the one inferred when the entire population of sites is used.

Analysis of Individual Genes

If genes are representative samples of the mitochondrial genome, we expect trees based on individual

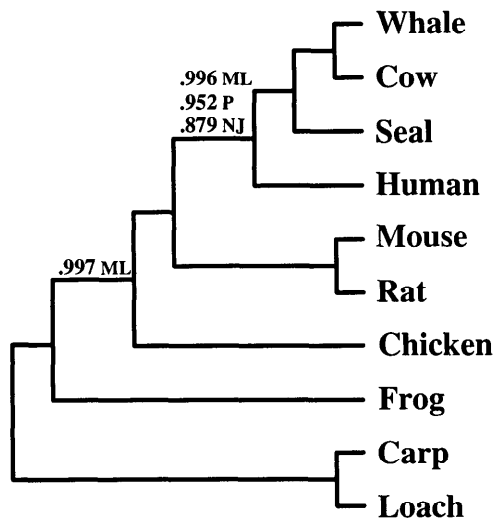

FIG. 1.—Genome tree as inferred using all three methods of analysis. Bootstrap proportions are given for each method of analysis; where no value is given, all 1,024 replicates supported that node. *ML,* maximum likelihood; *P,* parsimony; *NJ,* neighbor joining.

genes to provide accurate estimates of the whole-genome tree. Analyses of all 37 mitochondrial genes demonstrate, however, that trees inferred from single genes are seldom the same as the whole-genome tree. None of the tRNA genes gave a single tree identical to the whole-genome tree, although for two tRNA genes the whole-genome tree was one among many equally parsimonious trees. For the rRNA and protein-coding genes, most of the inferred trees differ from the whole-genome tree in the arrangement of mammalian taxa. However, in five gene trees, the relationship among amniotes is different from that in the whole-genome tree (table 1; fig. 2).

In no case did all three methods produce the same tree for a gene. Further, in the 12 cases where two of the three methods gave the same tree, 8 differed from and only 4 were the same as the whole-genome tree. This observation is in contrast to the suggestion that agreement among analytical methods is more likely for the true topology (Kim 1993). Further, no one particular alternative to the whole-genome tree occurred a large number of times; of the 20 alternative trees found (table 1; fig. 2), the most common one, tree B, occurred only 6 times out of 45 cases (15 genes × 3 phylogenetic methods). For these taxa, single mitochondrial genes appear to be poor samples for estimating the whole-genome tree.

The most commonly used method for measuring support for phylogenetic relationships is the bootstrap (Felsenstein 1985). We examined the relationship between the bootstrap proportion and whether a clade is present on the whole-genome tree. In only one case involving rRNA or protein-coding genes was a bootstrap proportion for a clade that was not on the whole-genome tree greater than 0.95. In that case, the tree inferred from NADH 4L places frog as the sister group to mammals in 980 of the 1,024 bootstrap replicates (0.957) using parsimony. This likely represents stochastic error rather than evidence of a separate history for NADH 4L. For clades not found on the whole-genome tree, we expect bootstrap values greater than 95% to occur about 5% of the time. Here, with 18 such clades, the probability of getting at least one bootstrap proportion greater than or equal to 0.95 is quite high, $1 - (0.95)^{18} = 0.60$.

Figure 3 shows the observed distribution of bootstrap proportions, where each value represents the bootstrap support for a single clade on each of the consensus trees inferred from individual genes. The top portion of the figure gives the distribution when the clade is found on the genome tree, and the bottom shows the distribution for clades not on the genome tree. With the exception of very high values, greater than about 0.9, there is little to distinguish the two distributions. These figures show that there is no correspondence between bootstrap proportions and the probability that any given clade is present on the whole-genome tree.

Analysis of Random Samples

To assess the relationship between the number of sites in a sample and the probability of obtaining the whole-genome tree, random samples of different sizes were drawn, and the proportion of trees identical to the

Table 1
Summary of Analyses Based on Individual Genes, including Bootstrap Consensus Tree (in parentheses)

Gene	Length	Parsimony	Maximum Likelihood	Neighbor Joining
12s rRNA	1,111	* (*)	* (*)	A (A)
16s rRNA	1,786	B (C)	* (*)	* (*)
ATPase6	687	D, E, F (D)	D (D)	G (G)
ATPase8	207	H (H)	H (I)	B (B)
COI	1,560	D (J)	* (*)	J (J)
COII	705	A (A)	K (K)	K (K)
COIII	785	* (*)	* (*)	M (*)
CYTB	1,149	Q (J)	* (*)	K (K)
NADH1	981	L (B)	B (B)	B (B)
NADH2	1,047	B (B)	B (P)	R (R)
NADH3	350	J (J)	M (M)	J (J)
NADH4	1,387	* (*)	T (T)	* (*)
NADH4L	297	Q (Q)	Q (Q)	S (S)
NADH5	1,860	T (T)	T (T)	* (*)
NADH6	561	N (N)	O (N)	A (A)
Identical to genome tree		3/15 (3/15)	5/15 (5/15)	3/15 (4/15)

NOTE.—Letters refer to the toplogies in fig. 2. Trees identical to the whole genome tree are denoted with an asterisk (*).

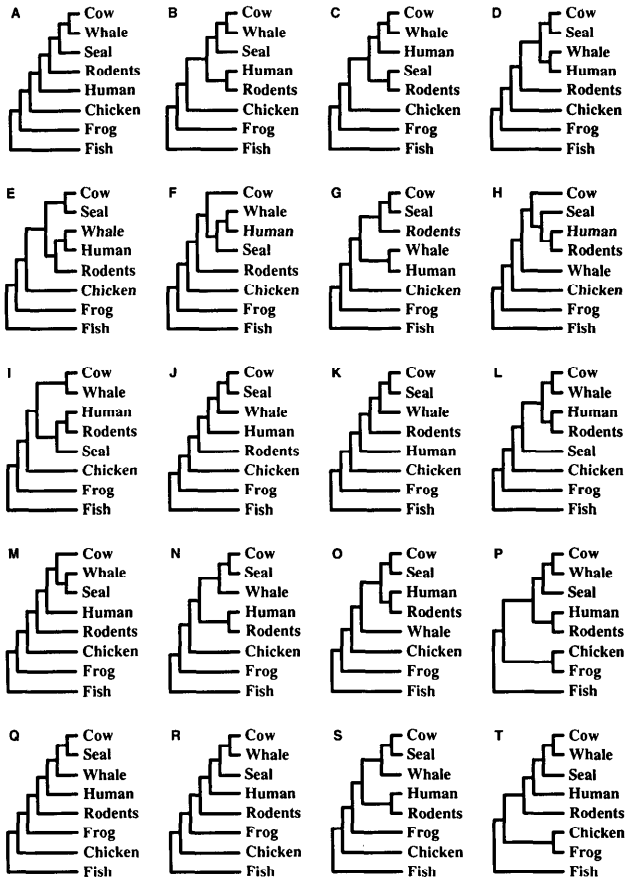

FIG. 2.—The trees inferred according to the protein-coding and rRNA genes. Trees A–O differ from the whole-genome tree in the relationship of the mammalian taxa, and P–T differ from the whole-genome tree in the relationship of the amniote taxa, but some of P–T also differ in the relationships among mammals. Rat and mouse were always sister taxa, as were carp and loach; these clades are denoted Rodents and Fish, respectively.

whole-genome tree was determined. The results of these analyses are shown in figure 4, in which the curves represent the power of a sample to infer the whole-genome tree. These results demonstrate that a large number of sites are required to ensure a high probability of getting a tree identical to the whole-genome tree. The number of sites required depends on how the sites are sampled and how the data are analyzed, but for contiguous sites at least 8,000 sites, or 50% of the genome, are required for a 95% chance of obtaining the whole-genome tree.

While it is true that a large number of sites are required to ensure a high probability of obtaining the whole-genome tree, a smaller number of sites produces trees that are fairly close in topology to the whole-genome tree as measured by the contraction/decontraction metric (Bourque 1978; Robinson and Foulds 1981). This is shown in figure 5. Even relatively small numbers of sites, 1,000–2,000 for contiguous samples, give trees that, on the average, require only one branch contraction and decontraction, or two steps, to convert the tree inferred from the sample to that inferred from the whole genome.

Heterogeneity across the Genome

A fundamental assumption of the bootstrap as applied to sequence data is that sites are independent and identically distributed (Felsenstein 1985). If this assumption is true, samples of contiguous sites should accurately reflect the distribution of sites from the genome as a whole. In terms of phylogenetic reconstruction, samples of contiguous sites should be equivalent to samples of sites dispersed throughout the genome. However, we found that samples of contiguous sites are less likely to lead to the whole-genome tree than are sets of sites chosen from random positions (figs. 4 and 5). The magnitude of this effect depends on the sequence length and the method of analysis, but, on the average, sampling sites contiguously reduced the chance of obtaining the whole-genome tree by 19.8% for maximum likelihood, 19.4% for parsimony, and 11.5% for neighbor joining.

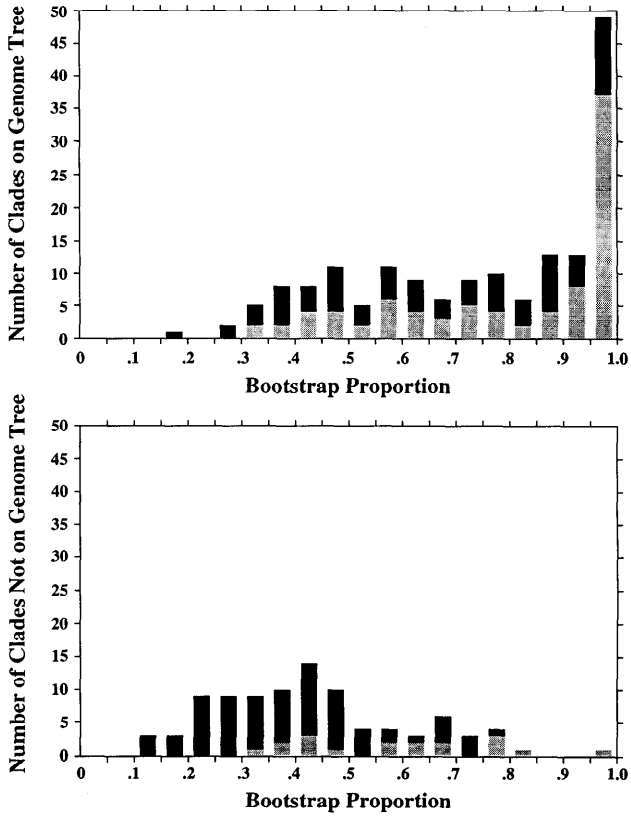

FIG. 3.—Histograms giving the number of times that bootstrap proportions were observed in intervals of 0.05. The distributions are for all clades inferred from all individual genes (rRNA and protein-coding genes—*gray bars*) and tRNAs (*black bars*) as determined by the parsimony method. Distributions for maximum likelihood and neighbor joining are similar.

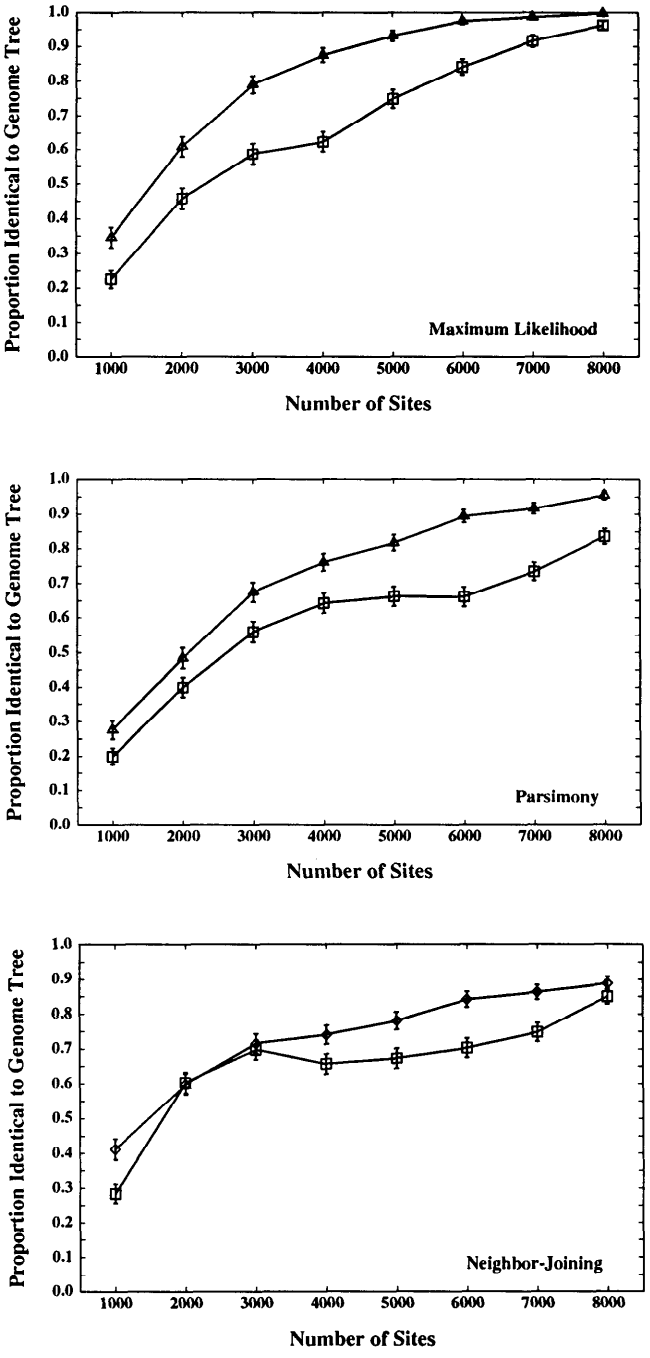

FIG. 4.—Proportion of trees inferred from random samples that are identical to the whole-genome tree. Data points—triangles for sampling without replacement and squares for contiguous sites—represent the mean for 1,024 samples. Error bars denote 95% confidence intervals for the mean.

This result implies that there are location-dependent aspects of DNA evolution, that neighboring nucleotides do not evolve independently. Therefore, resampling with replacement from a sample of contiguous sequence, as done in the bootstrap, is not the same as sampling from the underlying distribution of sites within a genome. The fundamental assumption of the bootstrap that nucleotide sites are independent

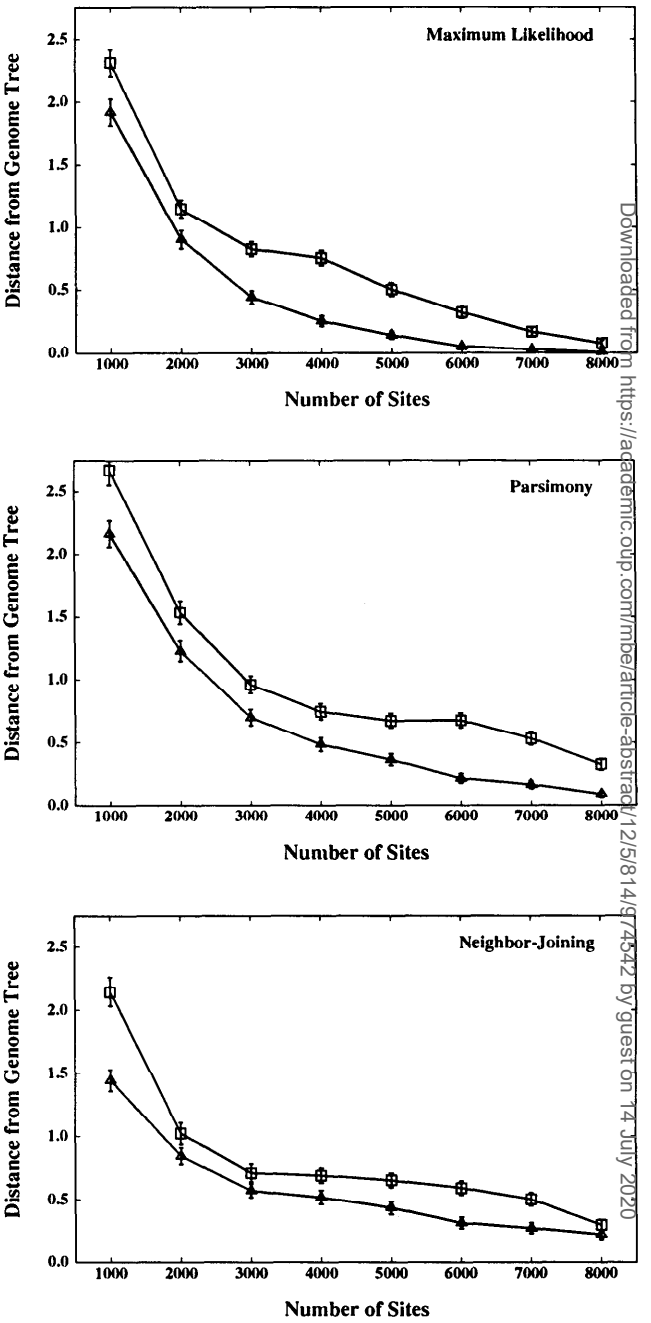

FIG. 5.—Contraction/decontraction metric for distance between trees inferred from random samples and the whole-genome tree. The metric is a measure of the number of branches that must be collapsed plus the number that must be built in order to convert one tree into the other (Bourque 1978; Robinson and Foulds 1981; Penny and Hendy 1985). For 10 taxa, the minimum tree distance is 0 (identical trees), and the maximum is 14. Data points—triangles for sites without replacement and squares for contiguous sites—represent the mean for 1,024 samples. Error bars denote 95% confidence intervals for the mean.

and identically distributed is violated for contiguous DNA sequence data.

The effect of location-dependent DNA evolution can be seen in the distributions of variable sites across the genome. Sites that differ in the 10 taxa are significantly heterogeneous in their distribution as determined using tests based on the sample distribution function: $N = 9{,}609$; Kuiper's $V = 0.039$, $P < 10^{-11}$; Watson's $U^2 = 0.898$, $P < 10^{-8}$. This heterogeneity is depicted using a sliding window in figure 6 for both variable sites and sites informative under parsimony. Similar analyses of the distributions of phylogenetically informative sites, individual nucleotides, purines, pyrimidines, and other classifications of sites give similar results. For example, purines in the human mitochondrial genome exclusive of the control region (fig. 7) are also significantly heterogeneous in their distribution: $N = 7{,}299$; Kuiper's $V = 0.032$, $P < 0.00002$; Watson's $U^2 = 0.651$, $P < 10^{-5}$. These results demonstrate that both variability and base composition are heterogeneously distributed throughout the genome.

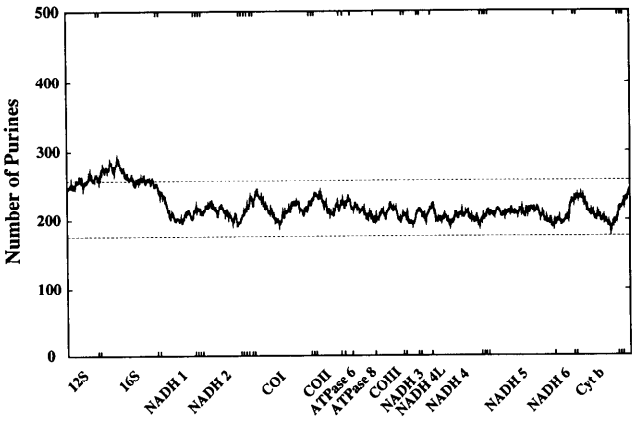

FIG. 7.—Number of purines in a sliding window of 500 sites with step size of 1 site for human mitochondrial genome exclusive of the control region. Dashed lines represent the maximum and minimum of all peaks and troughs observed from similar analyses of 20 randomly permuted genomes.

Codon Positions

It is generally established that the three positions within a codon of a protein-coding gene differ in their rates of nucleotide substitution (see, e.g., Kimura 1980; Nei 1987). These differences in rates constitute the rationale for treating the codon positions differentially in phylogenetic analysis. To assess the sampling properties of each codon position and to compare codon position classes, we constructed separate data sets for first, second, and third codon positions from all the protein-coding genes and sampled sites randomly without replacement as described above. The results, shown in figures 8 and 9, varied with the method, but all three codon positions appear to provide information for phylogenetic analysis. For instance, the performance of maximum likelihood was very similar for all three positions. Interestingly, third positions can provide substantial information about phylogenetic relationships if the number of sites is large and maximum likelihood is used, although this was not the case for parsimony and neighbor joining as employed in this study. The notion that the high rate of nucleotide substitution at third positions precludes their usefulness in phylogenetic analysis is dependent on the analytical method for these taxa.

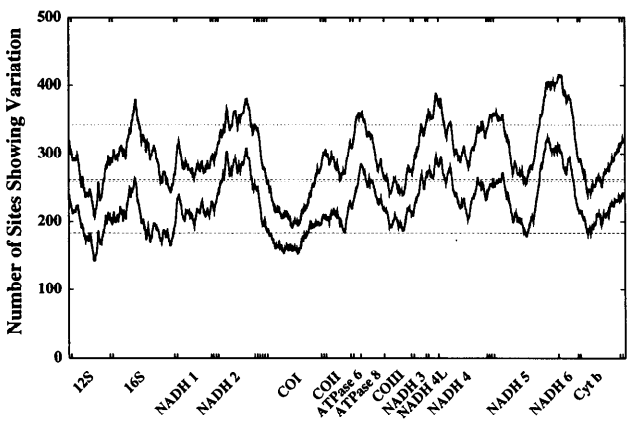

FIG. 6.—Number of sites showing variation (*upper curve*) and the number of sites that are phylogenetically informative under parsimony (*lower curve*) in a sliding window of 500 sites with step size of 1 site. Dotted lines (for variable sites) and dashed lines (for informative sites) represent the maximum and minimum of all peaks and troughs observed from similar analyses of 20 randomly permuted genomes.

Discussion

Our work suggests several possible strategies for improving the power of phylogenetic inference using DNA sequence data. Increasing the number of nucleotides will increase the chance of obtaining the whole-genome tree, and it appears that the number of nucleotides required is substantially greater than that of almost all studies in molecular phylogenetics. Another route to improved power is to sample sites from throughout the genome or from other genomes in the organisms if applicable (e.g., nuclear or plastid), since the effects of location-dependent processes in sequence evolution can be reduced by sampling sites from different regions. We have examined two extremes, contiguous blocks of sites and samples of individual sites from throughout the genome. Any intermediate sampling scheme, for example, sampling several short stretches distributed from different genomic locations, should, on the average, show intermediate performance—better than contiguous sites, but not as good as individual sites. Intermediate sampling

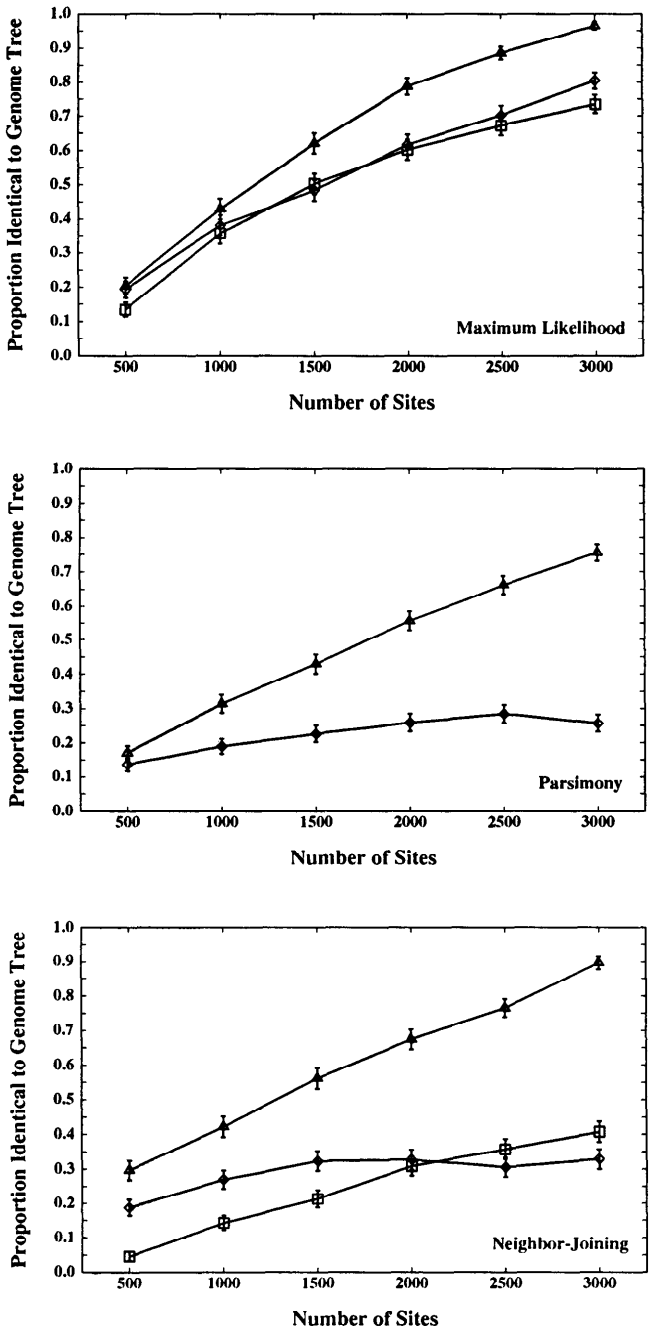

FIG. 8.—Proportion of trees inferred from first, second, and third codon positions sampled without replacement that are identical to the whole-genome tree. Data points—diamonds for first positions, triangles for second positions, and squares for third positions—represent the mean for 1,024 samples. Error bars denote 95% confidence intervals for the mean. Analyses are as described for fig. 1 with the exception of parsimony on third positions, where the branch-and-bound method approaches an exhaustive search due to the high level of homoplasy, and thus required an unacceptable length of time to complete the computations. Preliminary analyses indicated that the number of equally parsimonious trees obtained in these runs was very large, resulting in a very small proportion of trees identical to the genome tree.

can be effected by obtaining sequence data from multiple dispersed regions, such as sequence-tagged sites or expressed sequence tags, through restriction mapping, or, perhaps, other techniques.

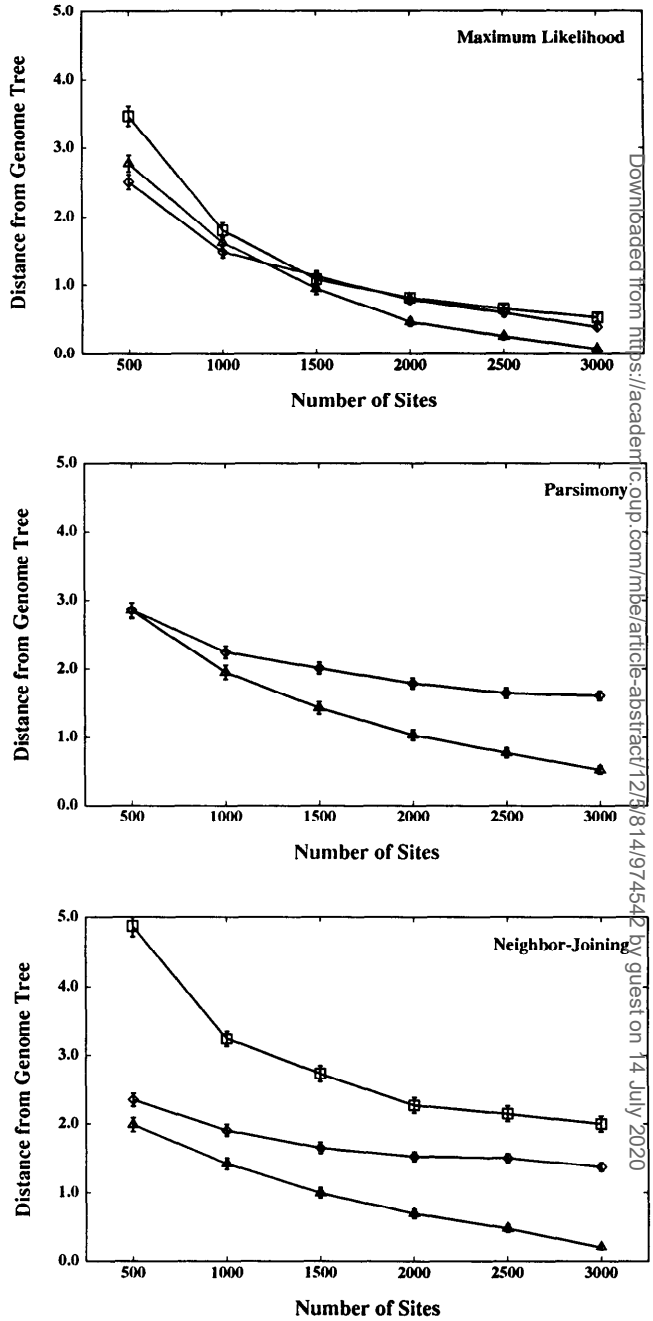

FIG. 9.—Distance, as measured using the contraction/decontraction metric (Bourque 1978; Robinson and Foulds 1981), between trees inferred from first, second, and third codon positions and the tree based on the whole genome. Data points—diamonds for first positions, triangles for second positions, and squares for third positions—represent the mean for 1,024 samples. Error bars denote 95% confidence intervals for the mean.

The generality of our results and their implications for studies involving other taxa and other DNA samples are unknown. Recently, using inferred amino acids sequences and a set of taxa which partially overlaps with those studied here, Cao et al. (1994*a*, 1994*b*) also found variation in the ability of mitochondrial genes to infer entire-coding-region relationships. While the divergence times and the exact features of mitochondrial genome evolution are specific to the taxa studied, nothing about this model system seems particularly unique to it, and we expect the general phenomena uncovered here to be applicable to other taxa and other genomes.

Our results suggest a dual view of phylogenetic relationships inferred in molecular systematic studies. On the one hand, a moderate amount of DNA sequence data can lead to inferred phylogenetic relationships relatively close to those that would be obtained by analyzing the entire genome. On the other hand, a relatively large number of sites are required to fully specify the whole-genome tree. The implications of this depend on the question being addressed. For questions that require the exact relationships of the taxa under study, the probability of success is likely to be very low unless we analyze many thousands of sites. However, if we need know only approximate relationships, then our chances are quite promising with smaller samples.

Acknowledgments

We thank J. Felsenstein, S. Kumar, and G. Olsen, who generously provided computer code; R. Jones, who assisted writing code for parallel computations; R. Guttel for information about ribosomal gene alignments; A. Graybeal, J. Patton, D. Wake, and M. Slatkin for comments on the manuscript; and two anonymous reviewers for their suggestions. M.P.C. was supported by a Alfred P. Sloan Fellowship in Molecular Studies of Evolution, S.P.O. was supported by a Miller Foundation Fellowship, and J.W. was supported by a National Institutes of Health (NIH) predoctoral training grant. This work was supported by a grant from NIH to M. Slatkin, and computer time was provided by Thinking Machines Corporation.

LITERATURE CITED

ANDERSON, S., A. T. BANKIER, B. G. BARRELL, et al. (14 coauthors). 1981. Sequence and organization of the human mitochondrial genome. Nature **290**:457–465.

ANDERSON, S., M. H. L. DE BRUIJNA, A. R. COULSON, I. C. EPERON, F. SANGER, and I. G. YOUNG. 1982. Complete sequence of bovine mitochondrial DNA, conserved features of the mammalian mitochondrial genome. J. Mol. Biol. **156**:683–717.

ARNASON, U., A. GULLBERG, and B. WIDEGREN. 1991. The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenoptera physalus*. J. Mol. Evol. **33**: 556–568.

ARNASON, U., and E. JOHNSON. 1992. The complete nucleotide sequence of the mitochondrial DNA of the harbor seal, *Phoca vitulina*. J. Mol. Evol. **34**:493–505.

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

BOURQUE, M. 1978. Arbres de Steiner et reseaux dont varie l'emplagement de certain sommets. Ph.D. diss., Université de Montréal, Québec, Canada.

CAO, Y., J. ADACHI, and M. HASEGAWA. 1994*a*. Eutherian phylogeny as inferred from mitochondrial DNA sequence data. Jpn. J. Genet. **69**:455–472.

CAO, Y., J. ADACHI, A. JANKE, S. PÄÄBO, and M. HASEGAWA. 1994*b*. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J. Mol. Evol. **39**:519–527.

CHANG, Y.-S., F.-L. HUANG, and T.-B. LO. 1994. The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. J. Mol. Evol. **38**: 138–155.

DESJARDINS, P., and R. MORAIS. 1990. Sequence and gene organization of the chicken mitochondrial genome, a novel gene order in higher vertebrates. J. Mol. Biol. **121**:599–634.

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

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

———. 1991. PHYLIP, version 3.41. University of Washington, Seattle, Washington.

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

FITCH, W. M. 1971. Toward defining the course of evolution: minimal change for a specific tree topology. Syst. Biol. **20**: 406–416.

GADALETA, G., G. PEPE, G. DE CANDIA, C. QUAGLIARIELLO, E. SBISA, and C. SACCONE. 1989. The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: signals revealed by comparative analysis between vertebrates. J. Mol. Evol. **28**:497–516.

GRAYBEAL, A. 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. Syst. Biol. **43**:174–193.

HIGGINS, D. G., A. J. BLEASBY, and R. FUCHS. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comp. Appl. Biosci. **8**:189–191.

HONEYCUTT, R. L., and R. M. ADKINS. 1993. Higher level systematics of eutherian mammals: an assessment of molecular characters and phylogenetic hypotheses. Ann. Rev. Syst. Ecol. **24**:279–305.

KIM, J. 1993. Improving the accuracy of phylogenetic estimation by combining different methods. Syst. Biol. **42**:331–340.

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., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis, version 1.01. Pennsylvania State University, University Park.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

OLSEN, G. J., H. MATSUDA, R. HAGSTROM, and R. OVERBEEK. 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comp. Appl. Biosci. **10**:41–48.

PENNY, D., and M. D. HENDY. 1985. Estimating the reliability of evolutionary trees. Syst. Zool. **34**:75–82.

ROBINSON, D. F., and L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

ROE, B. A., D.-P. MA, R. K. WILSON, and J. F.-H. WONG. 1985. The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J. Biol. Chem. **260**:9759–9774.

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

TZENG, C.-S., C.-F. HUI, S.-C. SHEN, and P. C. HUANG. 1992. The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome: conservation and variations among vertebrates. Nucleic Acids Res. **20**: 4853–4858.

DANIEL L. HARTL, reviewing editor

Received January 16, 1995

Accepted May 1, 1995