

Base Compositional Bias and Phylogenetic Analyses: A Test of the “Flying DNA” Hypothesis

Ronald A. Van Den Bussche,^{*,1} Robert J. Baker,^{*} John P. Huelsenbeck,^{†,2} and David M. Hillis[†]

^{*}Department of Biological Sciences, Texas Tech University, Lubbock, Texas 79409; and [†]Department of Zoology, University of Texas, Austin, Texas 78712

Received November 11, 1997; revised February 25, 1998

Phylogenetic methods can produce biased estimates of phylogeny when base composition varies along different lineages. Pettigrew (1994, *Curr. Biol.* 4:277–280) has suggested that base composition bias is responsible for the apparent support for the monophyly of bats (Chiroptera: megabats and microbats) from several different nuclear and mitochondrial genes. Pettigrew’s “flying DNA” hypothesis makes several predictions: (1) that metabolic constraints associated with flying result in elevated levels of adenine and thymine throughout the genome of both megabats and microbats, (2) that the resulting base compositional bias in bats is sufficient to mislead phylogenetic methods and account for the support for bat monophyly from several nuclear and mitochondrial genes, and (3) that phylogenetic analysis using pairwise distances corrected for compositional bias should eliminate the support for bat monophyly. We tested these predictions by analyzing DNA sequences from two nuclear and three mitochondrial genes. The predicted base compositional bias does not appear to exist in some of the genes, and in other genes the differences in AT content are very small. Analyses under a wide diversity of criteria and models of evolution, including analyses that take base composition into account (using log-determinant distances), all strongly support bat monophyly. Moreover, simulation analyses indicate that even extreme bias toward AT-base composition in bats would be insufficient to explain the observed levels of support for bat monophyly. These analyses provide no support for the “flying DNA” hypothesis, whereas the monophyly of bats appears to be well supported by the DNA sequence data.

Key Words: base compositional bias; phylogenetic analyses; chiropteran; bats. © 1998 Academic Press

Debate over the origin of bats (Chiroptera) has existed since Linnaeus (1758) placed bats within the order Primates. Even with sophisticated techniques for identifying character-state variation and methods for estimating phylogenetic relationships, this debate continues (Smith, 1977; Smith and Madkour, 1980; Wibble and Novacek, 1980; Pettigrew, 1986, 1991a,b; Bennet *et al.*, 1988; Pettigrew *et al.*, 1989; Adkins and Honeycutt, 1991; Baker *et al.*, 1991a,b; Mindell *et al.*, 1991; Thewissen and Babcock, 1991; Simmons *et al.*, 1991; Ammerman and Hillis, 1992; Bailey *et al.*, 1992; Stanhope *et al.*, 1992; Simmons, 1994). Most analyses of chromosomes, morphology, and DNA sequences support the monophyly of bats, including the mostly smaller, echolocating microchiropterans (henceforth microbats) and the mostly larger megachiropterans (henceforth megabats). However, Pettigrew (e.g., 1986, 1991a,b, 1994) has argued that neuroanatomical characters support a relationship between primates and megabats to the exclusion of microbats. This latter relationship has been termed the “flying primate” hypothesis.

The most recent criticism leveled against those characters supporting the monophyly of bats focuses on the DNA sequence data. Pettigrew (1994) argued that phylogenies based on five stretches of DNA, representing both nuclear and mitochondrial genes (Bennet *et al.*, 1988; Adkins and Honeycutt, 1991; Mindell *et al.*, 1991; Ammerman and Hillis, 1992; Bailey *et al.*, 1992; Stanhope *et al.*, 1992), are strongly confounded by base compositional bias towards adenine (A) and thymine (T) in the microbat and megabat lineages. Furthermore, Pettigrew proposed that the higher metabolic rate and smaller nuclei in bats results in mutational biases because of elevated cytosolic ATP concentrations associated with aerobic metabolism “leaking” into the nucleotide precursor pool used for DNA repair and replication (Pettigrew, 1994). Pettigrew argued that because of this AT bias in the genomes of bats, phylogenetic analysis of the DNA sequence data do not accurately reflect the evolutionary history of these taxa; rather, “these studies merely confirm what was already known—that bats share an AT bias in their DNA sequences” (Pettigrew, 1994; pg. 279).

¹ Present address: Department of Zoology, 430 LSW, Oklahoma State University, Stillwater, OK 74078.

² Present address: Department of Biology, University of Rochester, Rochester, NY 14627.

Base compositional variation among different lineages has been shown to bias phylogenetic methods; methods tend to produce estimates in which taxa with similar nucleotide composition are grouped together, irrespective of the actual evolutionary history (Lockhart *et al.*, 1994). Examples of base compositional bias confounding phylogenetic relationships have been reported for studies on the origin of photosynthetic organelles (Lockhart *et al.*, 1992a,b) and trees of life (Loomis and Smith, 1992; Sogin *et al.*, 1993). Here we examine the extent of potential effects of base compositional bias in the estimation of relationships among bats, primates, and other mammals.

HOW STRONG IS THE AT BIAS IN THE CHIROPTERAN LINEAGES?

Compositional bias is a common feature of sequence data (Bernardi *et al.*, 1985; Ikemura, 1985; Jukes and Bhushan, 1986; Sueoka, 1988; Liu and Beckenbach, 1992). Because of this characteristic, it is critical to evaluate the extent to which base composition varies among the taxa under study. When the taxa examined exhibit similar patterns of base composition, these taxa are said to exhibit stationarity or base compositional equilibrium (Saccone *et al.*, 1989; Collins *et al.*, 1994). Stationary patterns are usually observed among closely related taxa. Deviations from stationarity among taxa under study can produce biases in phylogenetic analyses (Loomis and Smith, 1990; Sidow and Wilson, 1990, 1991; Steel *et al.*, 1993; Steel, 1994; Lockhart *et al.*, 1994). Pettigrew (1994) suggested that higher metabolic rates could account for elevated AT content in the first of each of the following comparisons: mitochondrial (mt) DNA relative to nuclear DNA, homeotherm DNA relative to poikilotherm DNA, small mammals relative to large mammals, bats relative to nonflying relatives, and megabats relative to microbats.

Pettigrew's (1994) main argument for the "flying DNA" hypothesis is that bats contain a higher percentage of A and T in their genomes than do other mammals. Pettigrew (1994) cited data on the base composition of the entire nuclear genome presented by Sabeur *et al.* (1993) in support of this conclusion. However, examination of the Sabeur *et al.* (1993) data indicates that the bat genomes (based on single representatives of Megachiroptera and Microchiroptera) do not have elevated levels of A and T relative to the single representative of Primates (Fig. 1). Although base compositional properties of the nuclear genome indicate that the megachiropteran genome has a slightly elevated AT content relative to most other mammals, humans also possess a slightly elevated AT base composition relative to the single representative of the Microchiroptera (Fig. 1). Because base compositional bias groups taxa with similar nucleotide composition, then (contrary to the proposal of Pettigrew, 1994) the observed bias among

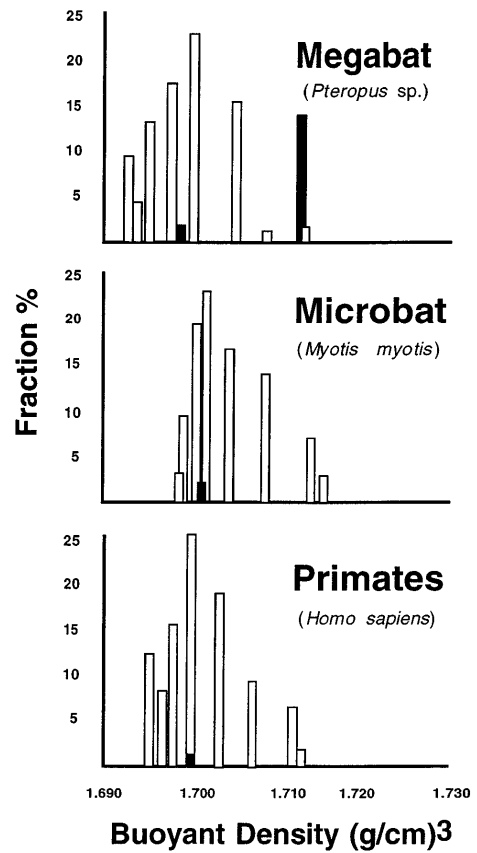


FIG. 1. Compositional distributions of a megabat (*Pteropus sp.*), microbat (*Myotis myotis*), and human mammalian genomes redrawn from Fig. 7 of Sabeur *et al.* (1993). Histogram bars represent different components of the genome and lower buoyant density values reflect high composition of A + T. Black bars represent satellite DNAs.

these three taxa should weaken the association of the megabat and microbat lineages by favoring a megabat-primate association.

Because there is considerable heterogeneity in base composition among mammalian nuclear genomes (Bernardi *et al.*, 1985, 1988; Bernardi, 1993; Mouchiroud and Bernardi, 1993; Sabeur *et al.*, 1993; Cacciò *et al.*, 1994), the relevant issue for molecular systematic studies is the base composition in the segments of DNA that have been sequenced and analyzed for a given set of taxa. Several nuclear and mitochondrial genes from diverse mammalian taxa have been sequenced and Table 1 presents data on nucleotide composition of the five genes under study for representatives of the eutherian orders that were available via GenBank. With the exception of the IRBP gene sequence, however, in many of these comparisons the variation in base composition among taxa is large enough to allow overlap between taxa. Additionally, for all five genes, the base composition between the megabat and microbat lineages are very similar. This observation contradicts the proposal of Pettigrew (1994) that base composition bias towards

TABLE 1

Percent A + T Composition \pm One Standard Deviation for Two Nuclear (IRBP, ϵ -globin) and Three Mitochondrial (12S rRNA, COI, COII) Genes

Taxon	IRBP	ϵ -globin	12S rRNA	COI	COII
Marsupialia	59.8 \pm 0.04, <i>n</i> = 36	58.8	63.1	45.1	n/a
Insectivora	61.7 \pm 0.03, <i>n</i> = 8	63.8	60.4 \pm 0.09, <i>n</i> = 5	38.6 \pm 0.13, <i>n</i> = 3	62.2
Dermoptera	54.5	59.4	55.6	37.1	59.4
Megachiroptera	57.7 \pm 0.02, <i>n</i> = 21	55.1	60.5	40.7 \pm 0.01, <i>n</i> = 2	60.2
Microchiroptera	56.8 \pm 0.03, <i>n</i> = 14	59.8	58.9 \pm 0.02, <i>n</i> = 3	41.1 \pm 0.01, <i>n</i> = 3	60.5
Primates	54.9 \pm 0.01, <i>n</i> = 35	54.1 \pm 0.01, <i>n</i> = 7	58.0 \pm 0.08, <i>n</i> = 38	43.3 \pm 1.70, <i>n</i> = 8	58.3 \pm 0.01, <i>n</i> = 33
Edentata	58.2 \pm 0.01, <i>n</i> = 9	n/a	60.4	36.3	n/a
Pholidota	61.8	n/a	n/a	n/a	62.9
Lagomorpha	60.2	62.9	58.9	34.3	n/a
Rodentia	61.6 \pm 0.03, <i>n</i> = 90	n/a	62.7 \pm 0.03, <i>n</i> = 16	45.8 \pm 0.25, <i>n</i> = 4	n/a
Cetacea	59.8 \pm 0.03, <i>n</i> = 25	57.7	59.5	37.1 \pm 0.01, <i>n</i> = 7	n/a
Carnivora	60.06 \pm 0.03, <i>n</i> = 90	n/a	57.4 \pm 0.09, <i>n</i> = 4	36.3	n/a
Tubulidentata	55.5 \pm 0.74, <i>n</i> = 3	n/a	n/a	43.3	n/a
Proboscidea	57.4 \pm 0.01, <i>n</i> = 3	n/a	n/a	37.8	n/a
Hyracoidea	58.5 \pm 0.01, <i>n</i> = 3	n/a	n/a	40.3	n/a
Sirenia	56.2 \pm 0.01, <i>n</i> = 3	n/a	n/a	37.1	n/a
Perissodactyla	58.7 \pm 0.02, <i>n</i> = 7	56.9 \pm 0.01, <i>n</i> = 4	60.9 \pm 0.04, <i>n</i> = 8	37.6	n/a
Artiodactyla	59.0 \pm 0.02, <i>n</i> = 22	58.4	61.5	39.3	53.7 \pm 0.01, <i>n</i> = 2

AT should be higher in the megabat genome than in the microbat genome.

Pettigrew (1994) also predicted that nonflying mammals should contain lower levels of AT in their genomes than flying mammals due to the higher metabolic requirements associated with flight. However, comparison of percentage of these five genes comprised of AT in comparisons between nonflying mammals and bats indicates that this generalization does not hold in most cases (Table 1). In fact, many nonflying mammals have higher AT compositions in these five genes than found in bats. Therefore, it does not appear that overall base composition of the nuclear or mitochondrial genome is an accurate indicator of the base composition of particular genes. Moreover, there do not appear to be any obvious correlations between metabolic rate, body size, or mode of locomotion and base composition among any of these taxa for these genes. Although much progress is being made concerning our understanding of the organization and evolution of the genome (Bernardi *et al.*, 1988; Bernardi, 1993; Janecek *et al.*, 1993; Baker *et al.*, 1995; Van Den Bussche *et al.*, 1996), our current state of knowledge is far from the point where we can confidently make the broad generalizations suggested by Pettigrew (1994). Nonetheless, we agree with Pettigrew (1994) that it is important to examine the potential for base composition to affect the analyses of mammalian relationships.

HOW STRONG IS THE SUPPORT FOR BAT MONOPHYLY?

All studies that have used DNA sequence data to evaluate the phylogenetic relationships among bats

and primates have, without exception, supported the monophyly of Chiroptera (Adkins and Honeycutt, 1991; Mindell *et al.*, 1991; Ammerman and Hillis, 1992; Bailey *et al.*, 1992; Stanhope *et al.*, 1992). We have reanalyzed all of these data sets using a wide range of criteria (Table 2).

Optimality Criteria

We analyzed trees using maximum parsimony (uniformly and differentially weighted), maximum likelihood (Felsenstein [1981; F81] and Hasegawa-Kishino-Yano [1985; HKY85] models), and minimum evolution (F81 and LogDet distances) (see Swofford *et al.*, 1996). For uniformly weighted maximum parsimony, all sites and character-state transformations were weighted equally; the total number of character changes across each tree was minimized (Fitch, 1971). In the differentially weighted analyses, transversions were weighted twice as heavily as transitions. Maximum likelihood analyses were based on the models of Felsenstein (1981) and Hasegawa *et al.* (1985). The former model assumes the same rate for all nucleotide substitutions, whereas the latter model allows a different rate for transitions and transversions (in every case, we estimated the transition:transversion parameter from the data under the model). In both models, we accounted for different frequencies of the individual nucleotides (we used the empirically observed frequencies for each data set). Phylogenies were also estimated using the minimum evolution criterion (Kidd and Sgaramella-Zonta, 1971; as modified by Rzhetsky and Nei, 1992). The optimal tree under the minimum evolution criterion is that tree with the smallest tree length (where tree length is the sum of the distances across the entire

TABLE 2

Comparison of the Relative Support for Three Hypotheses of Mammalian Phylogeny

Study	Target	No. Taxa	Hypothesis											
			Flying Primate 1				Flying Primate 2				Bat Monophyly			
			ME:F81 (LogDet)	ML:F81 (HKY)	MP (WP)	Rank	ME:F81 (LogDet)	ML:F81 (HKY)	MP (WP)	Rank	ME:F81 (LogDet)	ML:F81 (HKY)	MP (WP)	Rank
Adkins and Honeycutt, 1991	COII	21	2.1662 (2.2182)	-7826.0 (-7458.5)	1529 (2060)	4176 ^a	2.1563 (2.2110)	-7831.2 (-7467.8)	1530 (2061)	6890 ^a	2.1474 (2.1985)	-7802.9 (-7439.8)	1519 (2041)	1
Ammerman and Hillis, 1992	12S rDNA	10	0.8917 (0.8845)	-1287.9 (-1268.9)	211 (307)	1335	0.8718 (0.8638)	-1284.3 (-1266.0)	210 (304)	576	0.8460 (0.8434)	-1272.9 (-1248.6)	204 (293)	1
Bailey <i>et al.</i> , 1992	ε-globin	17	1.4777 (1.5325)	-6573.1 (-6398.4)	1224 (1727)	>242,000 ^b	1.4789 (1.5326)	-6567.1 (-6395.2)	1220 (1722)	>242,000 ^b	1.3738 (1.4242)	-6454.2 (-6273.9)	1165 (1631)	1
Mindell <i>et al.</i> , 1991	12S rDNA	4	0.5723 (0.5867)	-2552.2 (-2517.9)	383 (567)	2	n/a	n/a	n/a	n/a	0.5666 (0.5840)	-2549.8 (-2510.9)	382 (557)	1
Stanhope <i>et al.</i> , 1992	IRBP	13	1.2839 (1.3491)	-6564.6 (-6321.5)	1170 (1607)	>161,000 ^b	1.2816 (1.3435)	-6582.8 (-6333.6)	1174 (1608)	>161,000 ^b	1.2347 (1.2987)	-6480.3 (-6246.6)	1140 (1569)	1

Note. "Flying Primate 1" constrains the analysis to a monophyletic group of primates plus megabats; "Flying Primate 2" corresponds to a monophyletic group of primates, megabats, and flying lemur; and "Bat Monophyly" corresponds to the monophyly of megabats and microbats. In each case the best scores are shown for trees that are consistent with the hypothesis under the minimum evolution (ME), maximum likelihood (ML), and maximum parsimony (MP) criteria; upper scores are for simple models (Felsenstein, 1981 [F81] model or unweighted parsimony); lower scores in parentheses are for more complex models (log-determinant [LogDet] distances, Hasegawa-Kishino-Yano [HKY] model, and weighted parsimony [WP]). "Rank" refers to the rank order of the best parsimony solution for each hypothesis among all possible solutions (where Rank = [the number of better solutions] + 1).

^a Rank based on trees found via branch swapping; the actual rank is probably lower.

^b Analyses were stopped (due to computational constraints) after finding indicated number of better solutions.

tree). Unlike the parsimony criterion, however, a least squares criterion is used to estimate the branch lengths of trees. For the minimum evolution analyses, we used F81 and log-determinant distances. Log-determinant distances were as described by Steel (1994) and Lockhart *et al.* (1994) (except for a scaling factor, this distance transformation is the same as the paralinear distance described by Lake, 1994). Log-determinant distances were chosen because they are robust to changes in base composition among taxa and do not assume stationarity (see Swofford *et al.*, 1996). For each method, we analyzed both simple and complex models because of the common trade-off between efficiency and consistency that is related to model complexity (Hillis *et al.*, 1994a, b; Russo *et al.*, 1996) and to test the results for sensitivity to assumptions of the various models and methods.

Tree Constraints

In each analysis, we sought an optimal solution under each criterion (see below for search strategies). In addition, unless the optimal solution was equivalent to one of the following, we searched for the best tree that satisfied each of the following constraints: (1) flying primate hypothesis 1: monophyly of megabats and primates; (2) flying primate hypothesis 2: monophyly of megabats, primates, and the flying lemur (*Cynocephalus*); and (3) bat monophyly: monophyly of megabats and microbats. The two different versions of the flying primate hypothesis were considered because either would be considered consistent with Petti-grew's (1994) proposals, and different data sets provide

different relative rankings for these two hypotheses (Table 2).

Search Strategies

For the parsimony and minimum evolution criteria, searches for the optimal (unconstrained) trees were exact for all data sets except for that of Adkins and Honeycutt (1991); we used the branch-and-bound algorithm of PAUP* (version 4.0.0d42; written by David Swofford) to find the best solutions. Also, all analyses of the data sets of Mindell *et al.* (1991) and Ammerman and Hillis (1992) were exact (based on either exhaustive or branch-and-bound searches). Best trees for all other analyses were found using heuristic approaches. For maximum likelihood and parsimony, initial trees were found by simple stepwise addition, and near optimal solutions were sought by branch swapping using the tree-bisection and reconnection method of PAUP* (see Swofford *et al.*, 1996). For minimum evolution, initial constrained trees were estimated using the neighbor-joining algorithm by allowing only joinings that were compatible with the constraints tree. Near optimal solutions were then sought using tree-bisection and reconnection.

Ranking

We define the rank of a solution as [the number of known better solutions] + 1. We determined the rank of the best solution under the parsimony criterion for each set of constraints compared to the unconstrained analyses (Table 2).

Results

The support from all studies for bat monophyly over either version of the flying primate hypothesis appears to be extremely strong. Table 2 shows the relative support (under the various criteria defined above) for bat monophyly versus the two versions of the flying primate hypothesis. For every criterion and every data set, the optimal tree supports bat monophyly. For all data sets except that of Mindell *et al.* (1991), the difference in support between bat monophyly and either of the flying primate hypotheses is considerable (Table 2). For instance, the best tree consistent with either flying primate hypothesis is ranked (according to the parsimony criterion) 576th for the data set of Ammerman and Hillis (1992) and 4176th for the data set of Adkins and Honeycutt (1991). For the data sets of Stanhope *et al.* (1992) and Bailey *et al.* (1992), there are over 161,000 and 242,000 solutions better than any that fit either version of the flying primate hypothesis.

Although the analyses of both nuclear and mitochondrial genes strongly support bat monophyly, how can we tell if the level of support is statistically significant? A likelihood-ratio test of monophyly (Hillis *et al.*, 1996; Huelsenbeck *et al.*, 1996a, 1996b) provides a framework for testing the difference in support for two alternative phylogenetic hypotheses. We consider as the null hypothesis the best tree satisfying either of the Pettigrew hypotheses (e.g., Fig. 2a). For the IRBP data of Stanhope *et al.* (1992), the likelihood (HKY85 model) under the null hypothesis is $L_0 = -6316.83$. The alternative hypothesis relaxes the constraint of megabat + primate monophyly. The best tree under the alternative hypothesis is $L_1 = -6246.61$. This tree is consistent with bat monophyly. The likelihood-ratio

test statistic is $\delta = (\log L_1 - \log L_0) = 70.22$. Is this value larger than would be expected under the null hypothesis of megabat + primate monophyly?

Usually, for nested hypotheses of the sort considered here 2δ is χ^2 distributed with q degrees of freedom, where q is the difference in the number of parameters between the null and alternative hypotheses. However, topology is not a standard statistical parameter and 2δ is not χ^2 distributed (Goldman, 1993). Therefore, we used simulation to estimate the appropriate null distribution (Goldman, 1993) given the constraints of the flying primate hypothesis. The simulations of this tree assume the same HKY85 model of evolution that was used to perform the likelihood analyses described above. We simulated 100 data sets of 935 base pairs (the same size as the original data set) under the null hypothesis. For each simulated data set, the likelihood was calculated under the null hypothesis (flying primate) and under the alternative hypothesis (no constraints). The difference in scores between the constrained and unconstrained hypotheses (for likelihood and parsimony) were calculated for each simulated data set to generate the null distributions. These distributions indicate the expected difference between the optimal trees and the best trees consistent with the flying primate hypothesis, if the latter hypothesis is assumed to be correct.

Figure 3 shows the estimated null distributions of the test statistic δ . The greatest difference in log-likelihood scores in the null distribution is <4 , whereas the observed difference in the real data is 70.22 log-likelihood units (Fig. 2). Therefore, using this test, we can reject the null hypothesis that the flying primate tree is an adequate description of the data at $P \ll 0.01$. Although we have not conducted similar analyses for

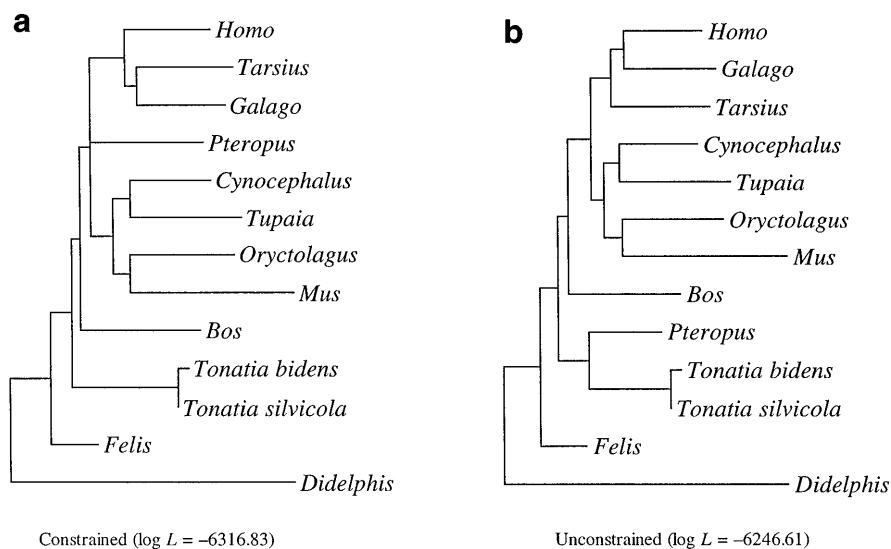


FIG. 2. (a) The model tree (null hypothesis) for simulations. This tree is the best tree that supports the flying primate hypothesis for the original IRBP data of Stanhope *et al.* (1992). *Pteropus* (the megabat) is weakly united with the primates *Homo*, *Tarsius*, and *Galago*. (b) The best tree supported by the IRBP data. Bats (*Pteropus* and *Tonatia*) are supported as a monophyletic group.

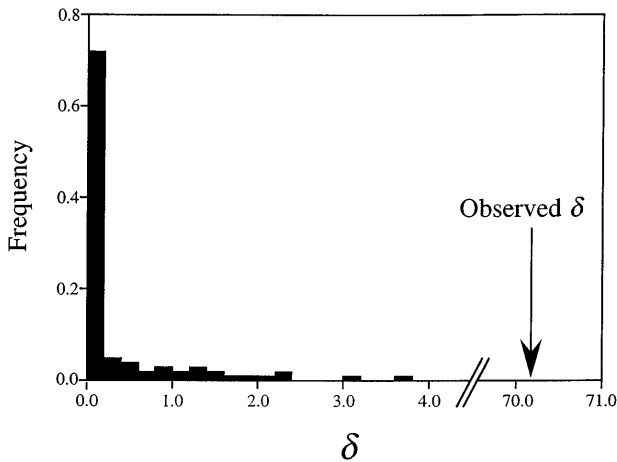


FIG. 3. Distribution of δ (the likelihood ratio test statistic) comparing the best trees that support the null hypothesis (flying primates) and the unconstrained analyses for simulated data sets based on the IRBP data, when the flying primate tree (Fig. 2a) is used as a model for simulations. The observed value of δ for the actual data is much greater than any values sampled from the null distribution, so the null hypothesis is rejected ($P < 0.01$) as an adequate explanation of the data.

the other data sets, they are all consistent with the findings of Stanhope *et al.* (1992) with regard to bat monophyly, so we take the overall level of support for this finding to be highly significant. However, a question remains: is the strong support for bat monophyly a result of phylogenetic signal, or could it be the result of base compositional bias?

WHAT LEVEL OF BASE COMPOSITIONAL BIAS WOULD PRODUCE THE "FLYING DNA" EFFECT?

Extreme base compositional biases have been shown to confound phylogenetic relationships for some molecular phylogenetic studies (Penny *et al.*, 1990; Sidow and Wilson, 1990; Loomis and Smith, 1992; Forterre *et al.*, 1993; Hasegawa and Hashimoto, 1993; Lockhart *et al.*, 1992a,b, 1994). Log-determinant (=paralinear) distances were developed to account for changes in base composition among taxa in a phylogenetic analysis (Steel, 1994; Lockhart *et al.*, 1994; Lake, 1994). Such corrections should prove useful for taking base compositional differences into account in phylogenetic analyses. However, because many molecular data sets have been generated, analyzed, and published, it is useful to evaluate the level of base compositional bias required to produce the effect proposed by Pettigrew (1994). If small levels of base composition bias, such as those shown in Table 1, can confound phylogenetic analyses (as suggested by Pettigrew, 1994), this would have serious implications concerning the validity of previous phylogenetic hypotheses based on sequence data. However, if such an effect is seen only at extreme base composition differences, then the problems associated

with base compositional bias are likely to be more limited.

Lockhart *et al.* (1994) tested the log-determinant transformation procedure on three empirical and one simulated data set and found that the data transformation provided a reliable correction for this potential source of error. However, the details of those simulations were considerably different from the present case, so we conducted simulations that are more applicable to the question of bat monophyly. To evaluate what level of base composition bias would be required to produce the effect proposed by Pettigrew (1994), and whether log-determinant distances can correct for such a bias, we performed simulations in which the AT base frequency in two separated bat lineages was higher than in the other lineages. Simulations were performed with parameters estimated using maximum likelihood (K80 model) for the IRBP data set of Stanhope *et al.* (1992). This data set was chosen because it contains a relatively large number of characters (935 bp after all positions aligned with gaps or questionable nucleotides are eliminated) as well as diverse taxa. The simulated data were produced under the following assumptions: (1) the topology was consistent with the "flying primate" hypothesis (Fig. 2a); (2) the base composition of all lineages except that leading to the megabat and the microbat was equal (50:50 AT:GC), and 50% of the substitutions were to A + T; and (3) the megabat and microbat lineages had the same substitution probabilities as each other, but varied from the other lineages in the tree. Simulations were performed in which the AT composition of the megabat and microbat lineages was varied from 0.5 to 1.0 in 0.05 increments. Each simulated data set was evaluated using the parsimony criterion and the minimum evolution criterion with log-determinant distances (Lockhart *et al.*, 1994).

Figure 4 shows the results of the simulations testing the effect of base composition bias on phylogenetic accuracy. AT composition of the bat lineages was plotted on the x -axis of Fig. 4b and the percentage of the time that bats were incorrectly found to be monophyletic (based on the model flying-primate tree) is plotted on the y -axis. In these simulations, the probability of finding a bat monophyly tree under the parsimony criterion was less than 0.05 in all simulations in which AT composition of the bat lineages was less than or equal to 80%. Only at the most extreme levels of base composition bias (>85%) would one expect to see estimated trees that strongly supported bat monophyly if the modeled flying primate tree were true. Furthermore, analyses based on log-determinant distances appear to be relatively unbiased by base compositional differences at even the most extreme levels in the simulations (Fig. 4). Given the minor differences in actual base composition (Table 1), together with the consistency of the results between log-determinant and other analyses (Table 2), we conclude that base compo-

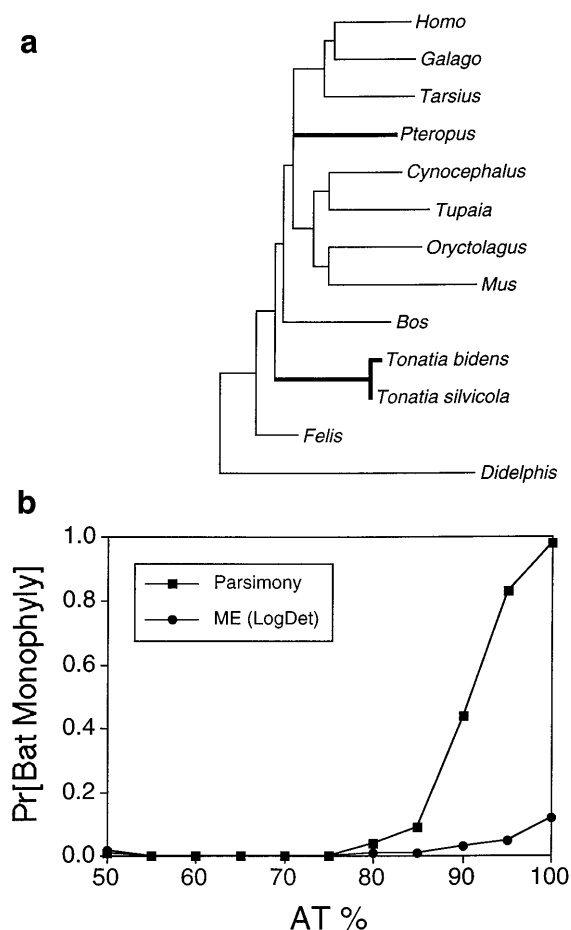


FIG. 4. Effect of base compositional bias for the simulations of the IRBP data. (a) The model tree. (b) The relationship between base composition of the bat lineages and the probability of finding bat monophyly (an incorrect result according to the model). Each plotted point represents the results from 100 data sets; black squares represent uniformly weighted parsimony analyses, and black circles represent minimum evolution analyses with log-determinant distances. Even at the most extreme conditions of parallel base-composition bias in the megabat (*Pteropus*) and microbat (*Tonatia*) lineages, few of the simulated trees incorrectly (according to the model) group the two bat lineages together. The effect is especially small for the LogDet analyses.

sitional bias has relatively small effects on phylogenetic estimates among these taxa.

CONCLUSIONS

Many different factors can affect phylogenetic trees reconstructed from DNA sequence data. Some of these factors include heterogeneity in rate of substitutions among lineages, mutation rate, symmetry of nucleotide substitutions, and composition of the four nucleotides. Although the effects of heterogeneity, mutation rate, and symmetry of nucleotide substitutions have been explored over the past few years, only recently has it been shown that base compositional bias can affect the

results of phylogenetic analyses (Loomis and Smith, 1990; Saccone *et al.*, 1990; Sidow and Wilson, 1990; Lockhart *et al.*, 1992, 1994; Collins *et al.*, 1994). Our study using empirical and simulated nucleotide data sheds light on base compositional bias and whether this potential source of error is affecting our understanding of mammalian evolution and other studies of evolutionary history. In particular, if the observed low levels of variation in base composition among mammals (Table 1) is sufficient to significantly bias the phylogenetic results as suggested by Pettigrew (1994), then most previous phylogenetic analyses of DNA sequences would be called into question. However, our empirical analyses and simulation studies indicate that base composition is not producing the effect proposed by Pettigrew (1994). Although we do not question the importance of considering base compositional bias in phylogenetic analysis, it appears that the bias must be much more extreme than the observed bias among mammals to produce the kind of extreme results suggested by the flying DNA hypothesis.

Although this study has added to our understanding of the effect of base compositional biases in phylogenetic analyses, it is not possible to model all potential variables at once. Therefore, these results must be viewed in light of the variables we simulated, the results from the empirical data sets, and the results of other simulation studies that have examined similar variables in phylogenetic analyses (Lockhart *et al.*, 1994; Collins *et al.*, 1995). As more studies of this type are performed, we should gain a better understanding of the factors affecting phylogenetic analyses and in turn develop better models for reconstructing phylogenetic relationships based on nucleotide sequence data.

ACKNOWLEDGMENTS

J. A. DeWoody provided criticisms on earlier versions of this manuscript. This work was supported in part by NSF grants to R.J.B. and D.M.H.

REFERENCES

- Adkins, R. M., and Honeycutt, R. L. (1991). Molecular phylogeny of the superorder Archonta. *Proc. Natl. Acad. Sci. USA* **88**: 10317–10321.
- Ammerman, L. K., and Hillis, D. M. (1992). A molecular test of bat relationships: Monophyly or diphyly? *Syst. Biol.* **41**: 222–232.
- Bailey, W. J., Slightom, J. L., and Goodman, M. (1992). Rejection of the "Flying Primate" hypothesis by phylogenetic evidence from the ϵ -globin gene. *Science* **256**: 86–89.
- Baker, R. J., Honeycutt, R. L., and Van Den Bussche, R. A. (1991a). Examination of monophyly of bats: Restriction map of the ribosomal DNA cistron. In "Contributions to Mammalogy in Honor of Karl F. Koopman" (T. A. Griffiths and D. Klingener, Eds.) *Bull. Am. Mus. Nat. Hist.* **206**: 42–53.
- Baker, R. J., Longmire, J. L., and Van Den Bussche, R. A. (1995). Organization of repetitive elements in the upland cotton genome (*Gossypium hirsutum*). *J. Hered.* **86**: 178–185.

- Baker, R. J., Novacek, M. J., and Simmons, N.B. (1991b). On the monophyly of bats. *Syst. Zool.* **40**: 216–231.
- Bennet, S., Alexander, L. J., Crozier, R. H., and Mackinlay, A. G. (1988). Are megabats flying primates? Contrary evidence from a mitochondrial DNA sequence. *Aust. J. Biol. Sci.* **41**: 327–332.
- Bernardi, G. (1993). Genome organization and species formation in vertebrates. *J. Mol. Evol.* **37**: 331–337.
- Bernardi, G., Mouchiroud, D., Gautier, C., and Bernardi, G. (1988). Compositional patterns in vertebrate genomes: Conservation and change in evolution. *J. Mol. Evol.* **28**: 7–18.
- Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cunny, G., Meunier-Rotival, M., and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228**: 953–957.
- Cacciò, S., Perani, P., Saccone, S., Kadi, F., and Bernardi, G. (1994). Single-copy sequence homology among the GC-richer isochores of the genomes from warm-blooded vertebrates. *J. Mol. Evol.* **39**: 331–339.
- Collins, T. M., Wimberger, P. H., and Naylor, G. N. P. (1994). Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.* **43**: 482–496.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**: 368–376.
- Fitch, W. M. (1971). Toward defining the course of evolution: Minimal change for a specific tree topology. *Syst. Zool.* **20**: 406–416.
- Forster, P., Benachou-lafha, N., and Labedan, B. (1993). Universal tree of life. *Nature* **362**: 795.
- Goldman, N. (1993). Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**: 182–198.
- Hasegawa, M., and Hashimoto, T. (1993). Ribosomal RNA trees misleading? *Nature* **361**: 23.
- Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**: 160–174.
- Hillis, D. M., Huelsenbeck, J. P., and Cunningham, C. W. (1994a). Application and accuracy of molecular phylogenies. *Science* **264**: 671–677.
- Hillis, D. M., Huelsenbeck, J. P., and Swofford, D. L. (1994b). Hobgoblin of phylogenetics? *Nature* **369**: 363–364.
- Hillis, D. M., Mable, B. K., and Moritz, C. (1996). Applications of molecular systematics: The state of the field and a look to the future. In "Molecular Systematics" (D. M. Hillis, C. Moritz, and B. K. Mable, Eds.), pp. 515–543. Sinauer, Sunderland, MA.
- Huelsenbeck, J. P., Hillis, D. M., and Jones, R. (1996a). Parametric bootstrapping in molecular phylogenetics: Applications and performance. In "Molecular Zoology: Strategies and Protocols" (J. D. Ferraris and S. R. Palumbi, Eds.), pp. 19–45. Wiley, New York.
- Huelsenbeck, J. P., Hillis, D. M., and Nielsen, R. (1996b). A likelihood-ratio test of monophyly. *Syst. Biol.* **45**: 546–558.
- Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**: 13–34.
- Janacek, L. L., Longmire, J. L., Wichman, H. A., and Baker, R. J. (1993). Genome organization of repetitive elements in the rodent, *Peromyscus leucopus*. *Mamm. Genome* **4**: 374–381.
- Jukes, T. H., and Bhushan, V. (1986). Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes. *J. Mol. Evol.* **24**: 39–44.
- Kidd, K. K., and Sgaramella-Zonta, L. A. (1971). Phylogenetic analysis: Concepts and methods. *Am. J. Human Genet.* **23**: 235–252.
- Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralineal distances. *Proc. Natl. Acad. Sci. USA* **91**: 1455–1459.
- Linnaeus, C. (1758). *Systema Nature*. 10th ed. Stockholm, Laurentii Salvii.
- Liu, H., and Beckenbach, A. T. (1992). Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. *Mol. Phylogenet. Evol.* **1**: 41–52.
- Lockhart, P. J., Howe, C. J., Bryant, D. A., Beanland, T. J., and Larkum, A. W. D. (1992a). Substitutional bias confound inference of cyanelle origins from sequence data. *J. Mol. Evol.* **34**: 153–162.
- Lockhart, P. J., Penny, D., Hendy, M. D., Howe, C. J., Beanland, T. J., and Larkum, A. W. D. (1992b). Controversy on chloroplast origins. *FEBS Lett.* **301**: 127–131.
- Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **11**: 605–612.
- Loomis, W. F., and Smith, D. W. (1992). Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *Proc. Natl. Acad. Sci. USA* **87**: 9093–9097.
- Mindell, D. P., Dick, C. W., and Baker, R. J. (1991). Phylogenetic relationships among megabats, microbats, and primates. *Proc. Natl. Acad. Sci. USA* **88**: 10322–10326.
- Mouchiroud, D., and Bernardi, G. (1993). Compositional properties of coding sequences and mammalian phylogeny. *J. Mol. Evol.* **37**: 109–116.
- Penny, D., Hendy, M. D., Zimmer, E. A., and Hambry, R. K. (1990). Trees from sequences: Panacea or Pandora's box. *Aust. Syst. Bot.* **3**: 21–38.
- Pettigrew, J. D. (1986). Flying primates? Megabats have the advanced pathway from eye to midbrain. *Science* **231**: 1304–1306.
- Pettigrew, J. D. (1991a). Wings or brain? Convergent evolution in the origin of bats. *Syst. Zool.* **40**: 199–216.
- Pettigrew, J. D. (1991b). A fruitful wrong hypothesis? Response to Baker, Novacek, and Simmons. *Syst. Zool.* **40**: 231–239.
- Pettigrew, J. D. (1994). Flying DNA. *Curr. Biol.* **4**: 277–280.
- Pettigrew, J. D., Jamieson, B. G. M., Robson, S. K., Hall, L. S., McNally, K. I., and Cooper, N. M. (1989). Phylogenetic relations between microbats, megabats and primates (Mammalia: Chiroptera and Primates). *Philos. Trans. R. Soc. Lond. Ser. B.* **325**: 489–559.
- Russo, C. A. M., Takezaki, N., and Nei, M. (1996). Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.* **13**: 525–536.
- Rzhetsky, A., and Nei, M. (1992). A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.* **9**: 945–967.
- Sabeur, G., Macaya, G., Kadi, F., and Bernardi, G. (1993). The isochores patterns of mammalian genomes and their phylogenetic implications. *J. Mol. Evol.* **37**: 93–108.
- Saccone, C., Lanave, C., Pesole, G., and Preparata, G. (1990). Influence of base composition on quantitative estimates of gene evolution. *Methods Enzymol.* **183**: 570–583.
- Sidow, A., and Wilson, A. C. (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.
- Sidow, A., and Wilson, A. C. (1991). Compositional statistics evaluated by computer simulations. In "Phylogenetic Analysis of DNA Sequences" (M. M. Miyamoto and J. Cracraft, Eds.), pp. 129–146. Oxford Univ. Press, New York.
- Simmons, N. B. (1994). The case for chiropteran monophyly. *Am. Mus. Novit.* **3103**: 1–54.
- Simmons, N. B., Novacek, M. J., and Baker, R. J. (1991). Approaches, methods, and the future of the chiropteran monophyly controversy: A reply to J. D. Pettigrew. *Syst. Zool.* **40**: 239–243.
- Smith, J. D. (1977). Chiropteran evolution. In "Biology of Bats of the New World Family Phyllostomatidae, Part I" (R. J. Baker, J. K. Jones, Jr., and D. C. Carter, Eds.), pp. 49–69. Spec. Publ. The Mus. Texas Tech. Univ., Lubbock.
- Smith, J. D., and Madkour, G. (1980). Penial morphology and the question of chiropteran phylogeny. In "Proceedings of the Fifth

- International Bat Research Conference" (D. E. Wilson and A. L. Gardner, Eds.), pp. 347–365. Texas Tech Univ. Press, Lubbock.
- Sogin, M. L., Hinkle, G., and Lelpe, D. D. (1993). Universal tree of life. *Nature* **362**: 795.
- Stanhope, M. J., Czelusniak, J., Si, J.-S., Nickerson, J., and Goodman, M. (1992). A molecular perspective on mammalian evolution from the gene encoding Interphotoreceptor Retinoid Binding Protein, with convincing evidence for bat monophyly. *Mol. Phylogenet Evol.* **1**: 148–160.
- Steel, M. (1994). Recovering a tree from the Markov leaf colourations it generates under a Markov model. *Appl. Math. Lett.* **7**: 19–23.
- Steel, M. A., Lockhart, P. J., and Penny, D. (1993). Confidence in evolutionary trees from biological sequence data. *Nature* **366**: 440–442.
- Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**: 2653–2657.
- Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference. In "Molecular Systematics" (D. M. Hillis, C. Moritz, and B. K. Mable, Eds.), pp. 407–514. Sinauer, Sunderland, MA.
- Thewissen, J. G. M., and Babcock, S. K. (1991). Distinctive cranial and cervical innervation of wing muscles: New evidence for bat monophyly. *Science* **251**: 934–936.
- Van Den Bussche, R. A., Longmire, J. L., and Baker, R. J. (1995). How bats achieve a small C-value: Frequency of repetitive DNA in *Macrotus*. *Mamm. Genome* **6**: 521–525.
- Wibble, J. R., and Novacek, M. J. (1988). Cranial evidence for the monophyletic origin of bats. *Am. Mus. Novit.* **2911**: 1–19.