Patterns of Nucleotide Substitution in Mitochondrial Protein Coding Genes of Vertebrates

Sudhir Kumar

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802

Manuscript received August 7, 1995
Accepted for publication January 18, 1995

ABSTRACT

Maximum likelihood methods were used to study the differences in substitution rates among the four nucleotides and among different nucleotide sites in mitochondrial protein-coding genes of vertebrates. In the 1st+2nd codon position data, the frequency of nucleotide *G* is negatively correlated with evolutionary rates of genes, substitution rates vary substantially among sites, and the transition/transversion rate bias (R) is two to five times larger than that expected at random. Generally, largest transition biases and greatest differences in substitution rates among sites are found in the highly conserved genes. The 3rd positions in placental mammal genes exhibit strong nucleotide composition biases and the transitional rates exceed transversional rates by one to two orders of magnitude. Tamura-Nei and Hasegawa-Kishino-Yano models with gamma distributed variable rates among sites (gamma parameter, α) adequately describe the nucleotide substitution process in 1st+2nd position data. In these data, ignoring differences in substitution rates among sites leads to largest biases while estimating substitution rates. Kimura's two-parameter model with variable-rates among sites performs satisfactorily in likelihood estimation of R, α, and overall amount of evolution for 1st+2nd position data. It can also be used to estimate pairwise distances with appropriate values of α for a majority of genes.

MITOCHONDRIAL DNA (mtDNA) sequences are widely used in molecular evolutionary studies. These sequences have proven useful for estimating times of species and population divergences, comparison of relative rates of evolution, and phylogenetic inference within and between species of vertebrates (NEI 1987; AVISE 1994; GILLHAM 1994). In such studies, information about the differences in the probability of change among four nucleotides (pattern of nucleotide substitution) and the variability of substitution rates among sites is necessary for obtaining reliable results. For example, the knowledge that transitions occur more often than transversions in primate mtDNA (BROWN *et al.* 1982) and that the probability of substitution varies among sites in the control D-loop region of human mtDNA has been useful in obtaining better estimates of the age of the "mitochondrial Eve" (KOCHER and WILSON 1991; VIGILANT *et al.* 1991; NEI 1992; TAMURA and NEI 1993; HORAI *et al.* 1995).

The tempo and mode of amino acid and nucleotide sequence evolution in animal mitochondrial genes have been studied previously (LANAVE *et al.* 1984; HASEGAWA and KISHINO 1989; REEVES 1992; ADACHI *et al.* 1993; LYNCH and JARRELL 1993; SIMON *et al.* 1994; HONEYCUTT *et al.* 1995). However, the actual pattern of nucleotide substitution and the extent of variability of substitution rates among sites in vertebrate mitochondrial genes have yet to be determined. The availability of nucleotide sequences of the complete mitochondrial genome from diverse vertebrate species allows us to examine these aspects of sequence evolution in various protein-coding genes.

Rates of substitution between the four nucleotides and among different sites can be estimated by the likelihood and parsimony methods. In the likelihood analysis, the parameters of a given substitution model are estimated by maximizing the likelihood function. For this purpose, various models of nucleotide substitution and a gamma distribution of variable substitution rates among sites can be assumed (GOLDING 1983; HOLMQUIST *et al.* 1983; TAMURA and NEI 1993; WAKELEY 1993; YANG 1994a). Unlike the likelihood method, parsimony analysis does not account for uneven nucleotide frequencies, transition-transversion rate bias, differences in substitution rates among sites, and unequal branch lengths of a tree (*e.g.*, COLLINS *et al.* 1994; PERNA and KOCHER 1995). Improved parsimony-based methods have been developed (YANG and KUMAR 1996), but likelihood methods are preferable if computationally feasible, and, thus, used in the present study.

Most commonly used models of nucleotide substitution are special cases of the general reversible (REV) model (YANG 1994a), and this model is expected to fit a given data set better than simpler models. However, it is desirable to use a simple, but adequate, model of substitution in the analysis because they give, for

Corresponding author: Sudhir Kumar, Department of Biology, Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802. E-mail: imeg@psuvm.psu.edu

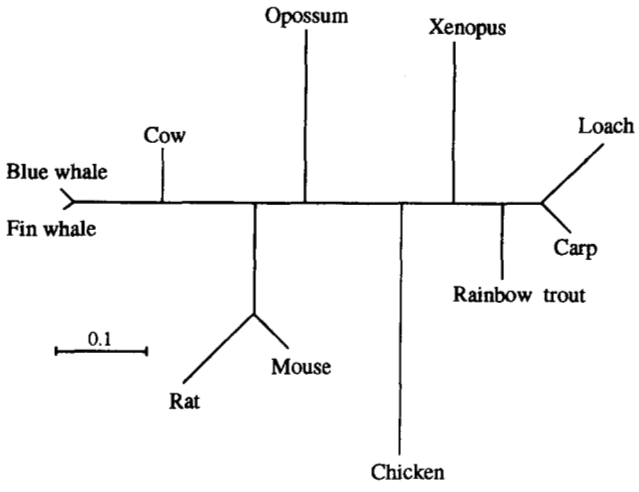

FIGURE 1.—The biological tree of 11 vertebrates species. The data set included complete mitochondrial DNA sequence (accession numbers in parentheses) of a fin whale (*Balaenoptera physalus*, X61145), blue whale (*B. musculus*, X72704), cow (*Bos taurus*, V00654 and J01394), rat (*Rattus norvegicus*, X14848), mouse (*Mus musculus*, V00711), opossum (*Didelphis virginiana*, Z29573), chicken (*Gallus gallus*, X52392), African clawed frog (*Xenopus laevis*, X02890, M10217, X01600, and X01601), carp (*Cyrinus carpio*, X61010), loach (*Crossostoma lacustre*, M91245), and rainbow trout (*Oncorhynchus mykiss*, L29771). The 1st+2nd codon position data of *Ndh*5 were used for estimating branch lengths by the maximum likelihood method; a general reversible model of nucleotide substitution and gamma-distributed substitution-rates among sites were assumed.

example, distance estimates with smaller variances. Several statistical tests are available to evaluate and compare the fit of models to a given data set (RITLAND and CLEGG 1987; KISHINO and HASEGAWA 1990; SACCONE *et al.* 1990; NAVIDI *et al.* 1991; REEVES 1992; GOLDMAN 1993; TAMURA 1994; RZHETSKY and NEI 1995). By using such tests, models that contain fewer parameters than the general reversible model can be identified. However, for large data sets, all simple models are likely to be rejected because even small departures from the complicated model examined can be detected in the statistical tests. Moreover, the choice of a model for different evolutionary analyses does not appear to be universal. For example, a general model is preferred for obtaining unbiased estimates of the branch lengths of a tree (and thus the time of divergence), but it may not be best suited for inferring phylogenetic relationships (NEI 1991; TATENO *et al.* 1994; YANG 1995a; RUSSO *et al.* 1996). Therefore, we have studied the performance of simple (and overly simple) models, rejected by the statistical tests, in estimating several useful evolutionary quantities. Based on these results, we have made an attempt to understand the importance of different features of the nucleotide substitution process in the estimation of evolutionary parameters and to identify simple models that may be suitable for a variety of evolutionary analyses in which the protein-coding genes of mtDNA are used.

MATERIALS AND METHODS

The sequence data: Complete mtDNA sequences of 11 vertebrate species were extracted from GenBank and EMBL databanks. The data set contained a fin whale (ARNASON *et al.* 1991), blue whale (ARNASON and GULLBERG 1993), cow (ANDERSON *et al.* 1982), rat (GADALETA *et al.* 1989), mouse (BIBB *et al.* 1981), opossum (JANKE *et al.* 1994), chicken (DESJARDINS and MORAIS 1990), African clawed toad (ROE *et al.* 1985), carp (CHANG *et al.* 1994), loach (TZENG *et al.* 1992), and rainbow trout (R. ZARDOYA, J. M. BAUTISTA, and A. GARRIDO-PERTIERRA, unpublished data). These species were chosen because their phylogenetic relationships are well established by studies of morphological characters and fossil records (Figure 1) (STORER *et al.* 1971; CARROL 1988; GINGERICH *et al.* 1990, 1994). This was done to eliminate systematic errors that may be introduced if the phylogeny is inferred from the data itself. The mitochondrial genomes of these species contain 13 protein coding genes: subunits 6 and 8 of ATP synthase (*Atp*6, *Atp*8), subunits 1–3 of cytochrome *c* oxidase (*Cox*1–3), cytochrome *b* (*Cyt*b), and seven subunits of NADH dehydrogenase (*Ndh*1–6 and *Ndh*4L). With the exception of *Ndh*6, these genes are encoded on the heavy strand (H strand) of the mitochondrial genome. The relative positions of these genes are identical for all the vertebrate species studied, with one exception: the positions of *Ndh*6 and *Cyt*b are reversed in the chicken.

For each gene, first the amino acid sequences were aligned by using the default option of CLUSTAL V (HIGGINS *et al.* 1992), and then the nucleotide sequences were adjusted to reflect those alignments. All codons containing alignment gaps in one or more species were removed from the analysis, and the data were analyzed at the nucleotide sequence level. The first (1st) and second (2nd) codon positions were pooled together for analysis. Even though these two positions are under different selective constraints, these data were combined to avoid working with sequences with small lengths. The analysis of 3rd position data was

TABLE 1

Substitution rate matrix (Q) for the general reversible (REV) model

	To			
From	T	C	A	G
T	$-(a\pi_C + b\pi_A + c\pi_G)$	$a\pi_C$	$b\pi_A$	$c\pi_G$
C	$a\pi_T$	$-(a\pi_T + d\pi_A + e\pi_G)$	$d\pi_A$	$e\pi_G$
A	$b\pi_T$	$d\pi_C$	$-(b\pi_T + d\pi_C + f\pi_G)$	$f\pi_G$
G	$c\pi_T$	$e\pi_C$	$f\pi_A$	$-(c\pi_T + e\pi_C + f\pi_A)$

$\mathbf{Q}_{ij}\Delta t$ is the probability that nucleotide i will change into nucleotide j in an infinitesimally short time interval Δt. a–f are the rate parameters and π_T, π_C, π_A, and π_G are the frequency parameters. Reversibility restriction requires that $\pi_i\mathbf{Q}_{ij} = \pi_j\mathbf{Q}_{ji}$.

TABLE 2

Models of nucleotide substitution

Model[a]	Rate parameters[b]	Frequency parameters[c]	Gamma parameter	Total free parameters
Uniform substitution rates among sites				
REV	a, b, c, d, e, f	$\pi_T, \pi_C, \pi_A, \pi_G$	∞	9
TN93	$a, f; b = c = d = e$	$\pi_T, \pi_C, \pi_A, \pi_G$	∞	6
HKY85	$a = f; b = c = d = e$	$\pi_T, \pi_C, \pi_A, \pi_G$	∞	5
F81	$a = f = b = c = d = e$	$\pi_T, \pi_C, \pi_A, \pi_G$	∞	4
K80	$a = f; b = c = d = e$	$1/4, 1/4, 1/4, 1/4$	∞	2
JC69	$a = f = b = c = d = e$	$1/4, 1/4, 1/4, 1/4$	∞	1
Variable substitution rates among sites				
REV+G	a, b, c, d, e, f	$\pi_T, \pi_C, \pi_A, \pi_G$	α	10
TN93+G	$a, f; b = c = d = e$	$\pi_T, \pi_C, \pi_A, \pi_G$	α	7
HKY85+G	$a = f; b = c = d = e$	$\pi_T, \pi_C, \pi_A, \pi_G$	α	6
F81+G	$a = f = b = c = d = e$	$\pi_T, \pi_C, \pi_A, \pi_G$	α	5
K80+G	$a = f; b = c = d = e$	$1/4, 1/4, 1/4, 1/4$	α	3
JC69+G	$a = f = b = c = d = e$	$1/4, 1/4, 1/4, 1/4$	α	2

[a] REV (TAVARE 1986; YANG 1994a), TN93 (TAMURA and NEI 1993), HKY85 (HASEGAWA *et al.* 1995), F81 (FELSENSTEIN 1981), K80 (KIMURA 1980), and JC69 (JUKES and CANTOR 1969). Models with gamma distributed substitution rates among sites have suffix "+G".

[b] As in Table 1.

[c] One of the frequency parameters is redundant.

restricted to the placental mammal sequences (whales, cow, mouse, and rat) because of very high sequence divergences between these and other vertebrate species (BROWN 1985; JANKE *et al.* 1994). Furthermore, the analysis of the 3rd position data is tentative because the substitution parameters estimated are likely to have large stochastic errors due to small sequence lengths and extremely fast rates of evolution (*e.g.*, IRWIN 1991). The 1st+2nd position data of placental mammal genes were also analyzed.

Estimating the pattern of nucleotide substitution: The pattern of nucleotide substitution was estimated by using the likelihood methods implemented in the program package PAML (YANG 1995c). For this computation, independently known tree and the general reversible (REV) model with variable substitution rates among sites were assumed (Figure 1, Table 1). Furthermore, it was assumed that the pattern of nucleotide substitution has remained the same in different parts of the tree (homogeneous Markov process) and that this process is at equilibrium (stationary Markov process).

In the general reversible model, two different transitional rates (a: $C \leftrightarrow T$; f: $A \leftrightarrow G$) and four different transversional rates (b: $A \leftrightarrow T$; c: $G \leftrightarrow T$; d: $A \leftrightarrow C$; e: $G \leftrightarrow C$) are considered, and frequencies of four nucleotides (π_T, π_C, π_A, π_G) are not assumed to be equal (Table 1) (TAVARE 1986; YANG 1994a). The variable substitution rates over sites were assumed to follow a gamma distribution with the shape parameter α (called gamma parameter) with its mean fixed to be one. For computational efficiency, the discrete gamma model with eight categories was used in the likelihood analysis (YANG

TABLE 3

Average nucleotide compositions observed

	1st+2nd positions					3rd positions					3rd positions in chicken			
Gene	n	π_T	π_C	π_A	π_G	n	π_T	π_C	π_A	π_G	π_T	π_C	π_A	π_G
Atp6	438	31.0	29.7	24.2	15.0	219	25.1	27.9	42.7	4.4	8.2	45.6	41.1	5.0
Atp8	104	29.7	30.7	29.9	9.7	52	20.3	25.2	49.3	5.2	17.3	38.5	40.4	3.9
Cox1	1022	32.1	23.5	22.1	22.3	511	25.0	30.3	29.3	5.6	35.0	19.8	41.7	3.5
Cox2	448	29.1	24.8	27.0	19.1	224	23.2	29.5	41.6	5.7	13.0	46.0	37.0	4.0
Cox3	518	31.9	25.4	21.9	20.8	259	22.9	33.1	40.5	3.5	8.5	45.6	40.2	6.8
*Cyt*b	754	32.7	25.2	23.7	18.4	377	17.8	38.5	40.0	3.7	10.6	51.2	35.0	3.2
Ndh1	626	31.7	28.7	22.3	17.3	313	18.8	31.0	46.0	4.2	13.7	43.5	37.1	5.8
Ndh2	684	30.1	29.8	27.3	12.8	342	17.8	30.8	47.5	3.9	8.8	44.7	42.4	4.1
Ndh3	224	34.2	28.2	20.3	17.3	112	21.4	30.7	43.9	4.0	17.9	35.7	42.9	3.6
Ndh4	913	32.0	27.6	25.4	15.0	457	20.3	31.5	43.5	4.7	11.4	45.1	40.3	3.3
Ndh4L	193	34.5	26.9	20.7	17.8	97	19.4	34.1	42.3	4.2	12.4	42.3	43.3	2.1
Ndh5	1164	30.7	25.4	28.6	15.3	582	20.3	36.0	40.1	3.7	13.9	44.9	39.0	2.2
Ndh6[a]	276	40.6	12.2	16.2	31.0	138	47.2	4.5	20.2	28.1	44.2	1.5	10.9	43.5

n is the number of nucleotides analyzed. Likelihood estimates of πs were within 1–2% of the observed averages.

[a] Encoded on the L strand.

TABLE 4

Variation in substitution rates among sites in the 1st+2nd position data

Gene	All species				Mammals
	$2\Delta l$	α	α_{MP}	α'_{MP}	α
*Atp*6	164.6[a]	0.49 ± 0.07	1.66	0.91	0.30 ± 0.07
*Atp*8	38.8[a]	0.86 ± 0.24	6.78	3.24	0.66 ± 0.31
*Cox*1	396.8[a]	0.13 ± 0.02	0.32	0.14	0.07 ± 0.02
*Cox*2	130.2[a]	0.33 ± 0.06	0.87	0.51	0.16 ± 0.05
*Cox*3	171.6[a]	0.23 ± 0.04	0.51	0.32	0.13 ± 0.04
*Cyt*b	401.2[a]	0.23 ± 0.03	0.59	0.33	0.16 ± 0.03
*Ndh*1	349.8[a]	0.27 ± 0.03	0.84	0.43	0.21 ± 0.04
*Ndh*2	294.0[a]	0.66 ± 0.07	3.07	1.50	0.67 ± 0.13
*Ndh*3	87.2[a]	0.39 ± 0.09	1.59	0.84	0.51 ± 0.19
*Ndh*4	440.0[a]	0.45 ± 0.04	1.47	0.79	0.37 ± 0.06
*Ndh*4L	42.8[a]	0.87 ± 0.23	5.42	2.63	0.55 ± 0.20
*Ndh*5	610.4[a]	0.46 ± 0.04	1.41	0.75	0.40 ± 0.05
*Ndh*6	68.8[a]	0.95 ± 0.18	4.38	1.58	0.60 ± 0.21

Δl is the difference in log likelihoods for observing the data under general reversible models with and without gamma distribution of rates among sites (REV+G and REV, respectively). α is the shape parameter of gamma distribution of rates as obtained using the likelihood method under REV+G model. Maximum parsimony estimates of gamma parameter are: α_{MP} by method-of-moments and α'_{MP} by YANG and KUMAR's (1996) method. α for "Mammals" was estimated using only placental mammal sequences.

[a] Rate constancy rejected at the 1% level; $2\Delta l > 6.6$.

1994b). This model is denoted REV +G and has 10 parameters (Table 2).

In the maximum likelihood analysis with the REV +G model, substitution rates (a–f), the gamma parameter (α), and the branch lengths and log likelihood (l) for the given tree were estimated. Hat (ˆ) was omitted from the symbols used for estimates of parameters because only the estimates are considered in this paper. Because only the relative values of substitution rates are important, they were scaled such that $a + b + c + d + e + f = 1$. Given the average nucleotide frequencies observed (π_T, π_C, π_A, π_G) and the substitution rates (a–f), the transition-transversion rate bias averaged over nucleotide frequencies (R) is computed by the following equation (GOLDMAN 1993):

$$R = (a\pi_T\pi_C + f\pi_A\pi_G)/(b\pi_T\pi_A + c\pi_T\pi_G + d\pi_C\pi_A + e\pi_G\pi_C). \quad (1)$$

The total number of substitutions per site that have occurred in the evolutionary history of the gene is the arithmetic sum of the maximum likelihood estimates of the branch lengths of the tree (S; overall amount of evolution per site).

The REV +G model assumes the reversibility of the evolutionary process (Table 1). Use of such a general model ensures that the effects of unequal nucleotide frequencies, biases in substitution rates between four nucleotides, and the heterogeneity of rates among sites are taken into account simultaneously in the estimation of evolutionary parameters. Thus, the maximum likelihood estimates obtained can be considered to be the most reliable. Comparisons of these estimates with those obtained by simpler models allows us to evaluate the effect of various restrictions on the model of nucleotide substitution in the estimation of evolutionary parameters.

An *unrestricted* model, which does not require the reversibility assumption, could also be used for estimating the pattern of nucleotide substitution. However, as noted by YANG (1994a), the use of the unrestricted model generally results in only marginal improvements of fit at the cost of adding three more parameters. This was indeed the case for the data sets analyzed here (results not shown).

Test of variability of substitution rates among sites: A likelihood ratio test of uniformity of substitution rates among sites (*i.e.*, $\alpha = \infty$) was conducted. In this test, log likelihoods of observing the data were obtained under the REV +G model (l_1) and the REV model (l_2), and a likelihood ratio test was constructed in which the statistic $2\Delta l$ ($\Delta l = |l_2 - l_1|$) follows a chi-square distribution with 1 degree of freedom.

Comparison of parametric models for a given data set: Several commonly used models of nucleotide substitution are special cases of the REV +G model (Table 2). For instance, the Tamura-Nei model (TN93+G) imposes the restriction that four types of transversional substitution rates are equal ($b = c = d = e$); thus, it involves three less parameters. We evaluated the effect of these and other restrictions on the fit of a model by the likelihood ratio test, where $2\Delta l$ (two times the difference in log likelihoods of observing the data under the models compared) follows a chi-square distribution with degrees of freedom equal to the difference in the number of free parameters between the two *nested* models compared. For example, in the comparison of TN93+G and REV +G, this amounts to 3 degrees of freedom.

GOLDMAN (1993) has suggested the use of parametric bootstrapping, instead of a chi-square approximation, for comparing models. However, the chi-square approximation for the likelihood ratio test may be reliable because the likelihoods under the two models compared may be affected in the same way when the "true tree" is given (YANG *et al.* 1995). Since the true tree is known in the present work, the chi-square test was employed for likelihood ratio tests. RZHETSKY and NEI (1995) tests for model selection were also used to evaluate the fit of different models to the data.

The estimates of R, α, and S obtained with simple models (involving two or three parameters) were compared with the corresponding estimates by REV +G. The statistical significance of the difference between the two estimates was determined by the normal-deviate tests, for which the standard errors were computed by the curvature method in the likelihood analysis (Table 7). Pairwise distances and their standard errors were computed by the MEGA program package (KUMAR *et al.* 1994). The relationships of the pairwise distances estimated by different models were examined graphically because of the lack of statistical tests for this purpose.

RESULTS

Nucleotide compositions: Table 3 shows the average nucleotide frequencies across species in different genes. In the 1st+2nd position data of genes encoded on the H strand, nucleotide G occurs with the smallest frequency, the frequency of nucleotide T is usually the largest, and $G+C$ content is in a narrow range of 40–45%. The observed nucleotide frequencies in different species were within 2–3% of their averages across species. However, the homogeneity (equality) of nucleotide compositions among species was rejected at 1% significance level in nine genes: *Atp*6, *Cox*1, *Cox*3, *Cyt*b, *Ndh*1, *Ndh*2 and *Ndh*4–6 data sets (RZHETSKY and NEI 1995). In placental mammals, the frequency of G was lower than that for other vertebrates and the frequency of A was slightly higher. In these five species, the homo-

TABLE 5

Substitution rates for 1st+2nd codon positions data under REV+G model

Gene	Transitions (a) C↔T	Transitions (f) A↔G	Transversions (b) A↔T	Transversions (c) G↔T	Transversions (d) A↔C	Transversions (e) G↔C	R	S	κ
*Atp*6	0.289	0.331	0.076	0.057	0.122	0.126	1.68	1.55	3.25
*Atp*8	0.239	0.428	0.050	0.030	0.120	0.134	1.56	4.24	3.62
*Cox*1	0.401	0.293	0.063	0.054	0.136	0.053	2.47	0.43	4.63[a]
*Cox*2	0.396	0.291	0.095	0.077	0.089	0.052	2.17	0.83	4.22
*Cox*3	0.431	0.252	0.140	0.037	0.103	0.038	2.26	0.70	4.24[a]
*Cyt*b	0.360	0.317	0.086	0.051	0.129	0.058	2.16	1.02	4.20
*Ndh*1	0.343	0.308	0.138	0.049	0.104	0.059	1.88	1.26	3.69
*Ndh*2	0.325	0.266	0.095	0.111	0.140	0.063	1.42	2.62	2.91
*Ndh*3	0.368	0.354	0.098	0.047	0.076	0.056	2.69	2.13	5.43
*Ndh*4	0.348	0.235	0.122	0.052	0.176	0.068	1.37	1.82	2.85[a]
*Ndh*4L	0.236	0.340	0.115	0.066	0.128	0.115	1.36	2.35	2.63
*Ndh*5	0.365	0.199	0.128	0.050	0.202	0.056	1.14	1.82	2.64[a]
*Ndh*6	0.256	0.304	0.118	0.072	0.121	0.128	1.15	2.26	2.79

R is transition/transversion rate bias averaged over nucleotide frequencies under REV+G model (Equation 1). *S* is sum of the maximum likelihood estimates of the branch lengths of the tree topology given in Figure 1. κ is the transition/transversion rate ratio under HKY85+G model.

[a] Data sets for which HKY85+G was rejected in comparison with REV+G.

geneity of base compositions across species was not rejected for any of the genes.

In the 3rd codon positions, the average frequency of nucleotide *G* is the lowest and that of *A* is the highest. However, the nucleotide compositions vary considerably among species (WOLSTENHOLME 1992). As an example, the base composition in the 3rd positions of chicken genes is shown (Table 3). Within mammals, nucleotide frequencies in *Atp*6, *Cox*1–2, *Cyt*b, and *Ndh*1–5 in opossum were quite different from those in the placental mammals. In particular, the frequency of *C* was lower in opossum (this decrease was apparently compensated by an increase in the frequency of *T*). This was another reason for limiting the analysis of 3rd position data only to placental mammal sequences. In this subset, the equality of nucleotide frequencies over species was rejected for *Atp*6, *Cox*2, *Cox*3, *Ndh*1, *Ndh*3, *Ndh*4L, and *Ndh*6 genes.

The rejection of uniformity of nucleotide frequencies among species indicates that the assumption of the stationarity of the evolutionary process has been violated. Such violations often occur in the analysis of distantly related sequences and for large data sets (RZHETSKY and NEI 1995; YANG 1995b). For example, much greater variation in base compositions among species was found in the 1st+2nd position data of *Atp*8 (or *Ndh*4L) as compared with *Atp*6 or *Ndh*5, but the stationarity was not rejected in *Atp*8 because of its short sequence length. In any case, the violation of the stationarity assumption indicates that the results should be interpreted with caution.

Differences in the substitution rates among sites: In the 1st+2nd position data, the likelihood estimates of α are <1 and the uniformity of substitution rates among sites is rejected for every gene (Table 4). The traditional parsimony-based estimates of α were up to seven times larger than the likelihood estimates for these data. In placental mammals, greater variation in rates among sites was inferred (as α values are smaller), but the difference between the estimates of α were generally not significant. Separate analysis of 1st and 2nd position data also demonstrated considerable rate variation among sites. The estimates of α

TABLE 6

Statistical test of fit of simple models to the 1st+2nd codon position data

	Likelihood ratio tests[a]			
	REV+G and TN93+G[c]	TN93+G and HKY85+G	HKY85+G and K80+G	Rzhetsky-Nei tests[b]
*Atp*6	+	+		TN93
*Atp*8	+	+		TN93
*Cox*1		+		TN93
*Cox*2	+	+	+	K80, TN93
*Cox*3		+		REV
*Cyt*b	+	+		REV
*Ndh*1	+	+		REV
*Ndh*2	+	+		REV
*Ndh*3	+	+		TN93
*Ndh*4		+		REV
*Ndh*4L	+	+		TN93
*Ndh*5		+		REV
*Ndh*6	+	+		REV

[a] + indicates that the difference in the fits of the models compared was statistically not significant at the 1% level.

[b] Models that were not rejected are listed. Presence of REV shows that the fits of all simpler models were significantly worse.

[c] Identical result obtained in the comparison of REV+G and HKY85+G.

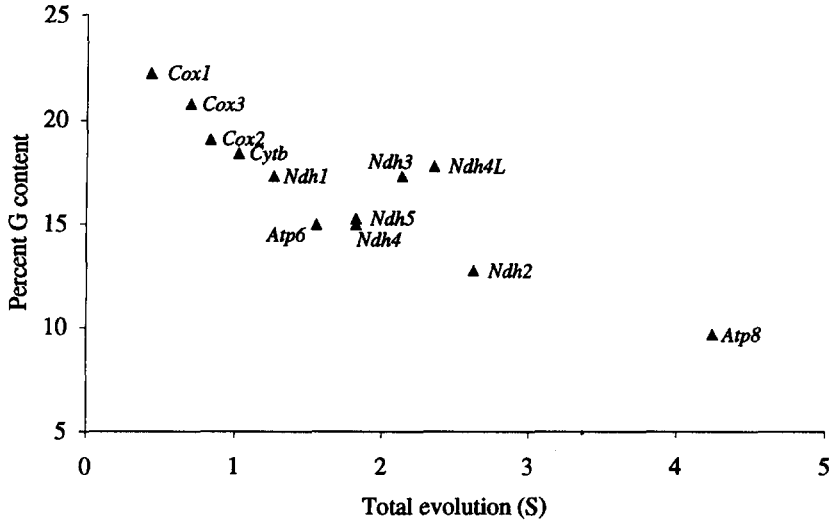

FIGURE 2.—The relation between the frequency of nucleotide *G* and the overall amount of evolution (*S*) in the 1st+2nd codon position data of mitochondrial genes encoded on the H strand. The correlation coefficient is −0.89.

were (1st and 2nd position): *Atp*6 (0.66, 0.39), *Atp*8 (1.06, 2.19), *Cox*1 (0.22, 0.08), *Cox*2 (0.54, 0.44), *Cox*3 (0.31, 0.37), *Cyt*b (0.27, 0.28), *Ndh*1 (0.44, 0.17), *Ndh*2 (0.84, 0.68), *Ndh*3 (0.84, 0.31), *Ndh*4 (0.62, 0.39), *Ndh*4L (0.98, 1.02), *Ndh*5 (0.53, 0.45), *Ndh*6 (0.83, 1.08).

In the 3rd position data of placentals, the uniformity of substitution rates was not rejected at the 1% significance level for *Atp*8, *Cox*1, *Cyt*b, *Ndh*1–6; for these genes α was >1. Therefore, the constancy of rates among sites was assumed in the analysis of these genes. For *Atp*6, *Cox*2, *Cox*3, and *Ndh*4L, estimates of α (±SE) were 0.51 ± 0.19, 0.57 ± 0.20, 0.84 ± 0.35, and 0.60 ± 0.39, respectively, and the uniformity of rates was rejected at the 1% level.

Substitution biases among four nucleotides: The estimates of two transitional (*a* and *f*) and four transversional (*b*–*e*) substitution rates for 1st+2nd position data are given in Table 5. The transversional rates that involve nucleotide *A* (*b*: *A*↔*T*; *d*: *A*↔*C*) are usually larger than others. Since the frequencies of *T* and *C* are higher than that of *A*, the rates of *A*→*T* and *A*→*C* changes are larger than the rates of *T*→*A* and *C*→*A* changes, respectively (see Table 1). In general, the transitional rates are up to five times larger than the transversional rates and the estimates of *R* are up to five times larger than those expected if transitional and transversional rates were equal. A similar trend was observed in the 1st+2nd position data of placental mammals (results not shown).

In the 3rd positions of placental mammals, *a* and *f* were up to two orders of magnitudes higher than *b*, *c*, *d*, or *e*, and these rates were estimated with very large variances. Similar results were obtained in substitution pattern analyses of the two closely related whale sequences. Furthermore, analysis of the 3rd position data of whales without two- or fourfold sites produced expected results; larger transition-transversion rate biases were observed when all fourfold sites were removed and lower transition / transversion rate biases were observed when all twofold sites were removed from the data analyzed. Comparable results were obtained when cow sequence was included in these analyses. However, these rate estimates are not reliable because of the saturation of transitional substitutions and the small sequence lengths of the 3rd positions (which led to very large variances) and because of the problem of correctly identifying the two- and fourfold sites in distantly related sequences. For these reasons, these estimates are not presented here.

LYNCH and JARRELL (1993) estimated the rate of evolution (number of amino acid substitutions per site per billion years, *r*) in animal mitochondrial genes by a Poisson model with constant substitution rates among sites. Following their generalized least squares approach, we could also estimate the rates of evolution at the nucleotide sequence level. Clearly, the correct species divergence times are required for estimating the rates of evolution reliably. Although it is well known that the fossil records underestimate the species divergence times (MARTIN 1993), it is now becoming clear that the extent of such underestimation is substantial and disproportionate for different lineages. For instance, an analysis of a large number of nuclear genes evolving at constant rates indicates that the divergence times of major mammalian orders are 50–90% larger than the fossil-based estimates (S. B. HEDGES, P. H. PARKER, C. G. SIBLEY, and S. KUMAR, unpublished data). Similarly, the divergence of mouse and rat may be up to three times larger than that known from the fossil record (FRYE and HEDGES 1995). Therefore, the use of fossil divergence times for the species analyzed in this work is likely to result in larger and biased estimates of rates of evolution. For this reason, evolutionary rate estimates are not reported here. Instead, the estimates of overall amount of evolution were computed to examine relative rates of evolution in different genes. The estimates of *S* for the 1st+2nd position data show high correlation (0.84) with *r* reported by LYNCH and JARRELL (1993). Thus, the differences in *S* roughly reflect

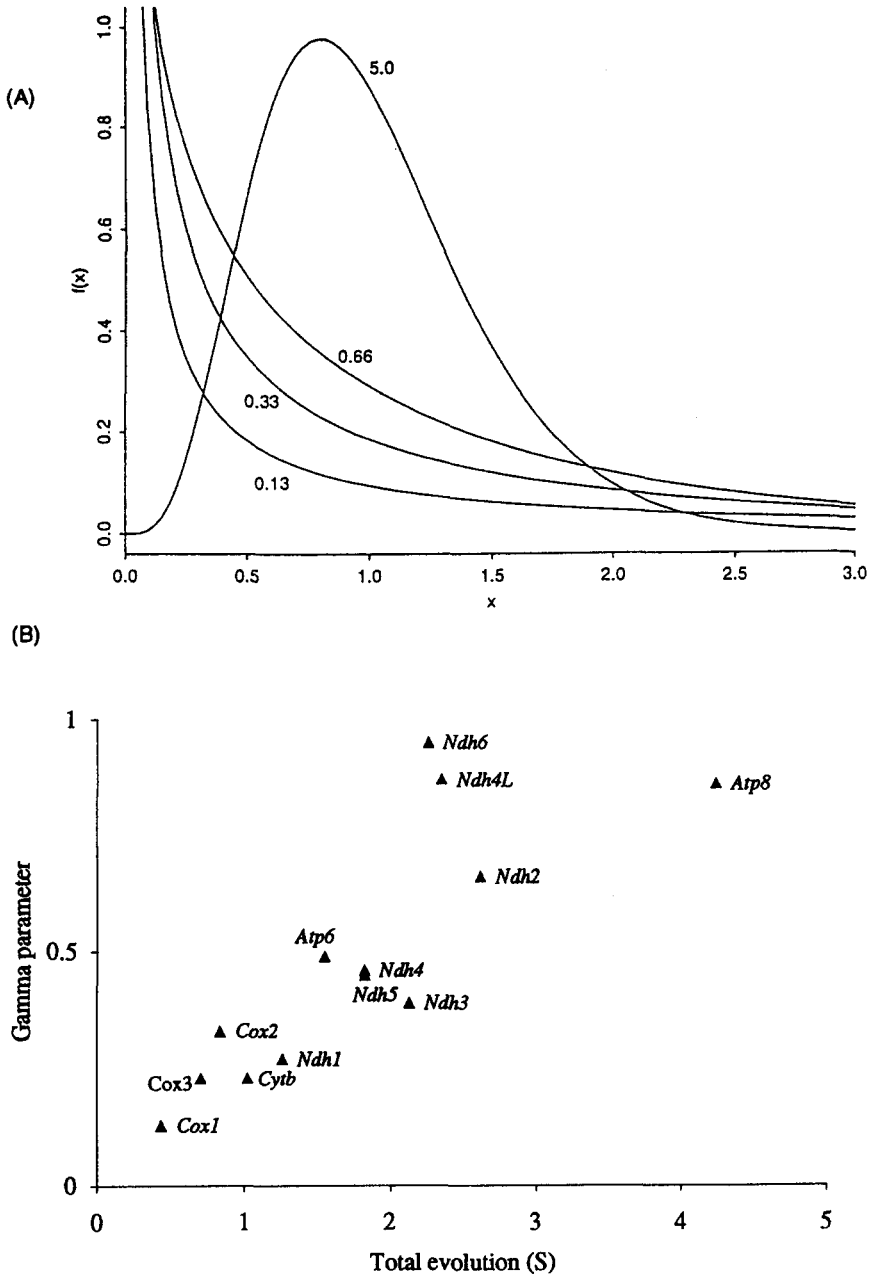

FIGURE 3.—(A) The density function of the gamma distribution with different values of α. The mean and variance of the distribution are 1 and $1/\alpha$, respectively. $f(x) = \alpha^{\alpha}\,\Gamma(\alpha)^{-1}\,e^{-\alpha x}\,x^{\alpha-1}$, $\alpha > 0$, $x > 0$. (B) The relationship between α and the overall amount of evolution (S) for the 1st+2nd codon position data of different genes. The correlation coefficient is 0.82.

differences in the rates of evolution. On the basis of the estimates of S, the mitochondrial protein-coding genes were arranged into three categories for convenience: genes evolving with low rates—*Cox*1–3, *Cyt*b, and *Ndh*1, moderate rates—*Atp*6 and *Ndh*3–6, and fast rates—*Atp*8 and *Ndh*2.

Comparison of models: In the likelihood ratio tests, we used a 1% significance level for rejecting the null hypothesis that a simple model describes the data as well as a more general one. In the 1st+2nd position data, the general reversible model with uniform rates (REV) fit the data significantly worse than the Tamura-Nei (1993) or Hasegawa-Kishino-Yano (1985) models with gamma-distributed rates (TN93+G and HKY85+G, respectively). These two models fit the *Atp*6, *Atp*8, *Cox*2, *Cyt*b, *Ndh*1–3, *Ndh*4L, and *Ndh*6 data well (Table 6), and all other models were rejected in the likelihood ratio tests. RZHETSKY-NEI's (1995) test also produced similar results (Table 6).

All simple models of substitution were rejected for *Cox*1, *Cox*3, *Ndh*4, and *Ndh*5 data. This is partly due to larger transversional rates b and d ($A \leftrightarrow T$, C) as compared with c and e ($G \leftrightarrow T$, C) and partly because of longer sequence lengths of *Cox*1, *Ndh*4, and *Ndh*5. By contrast, K80+G is not rejected for *Cox*2 because this gene is relatively short and because the transversional rates are quite similar and considerably smaller than the transitional rates.

In the 1st+2nd position data of placental mammals, difference in the fits of TN93+G and REV+G models was not significant for all but three genes: *cyt*b, *Ndh*4, *Ndh*5. Both TN93+G and HKY85+G fit these data

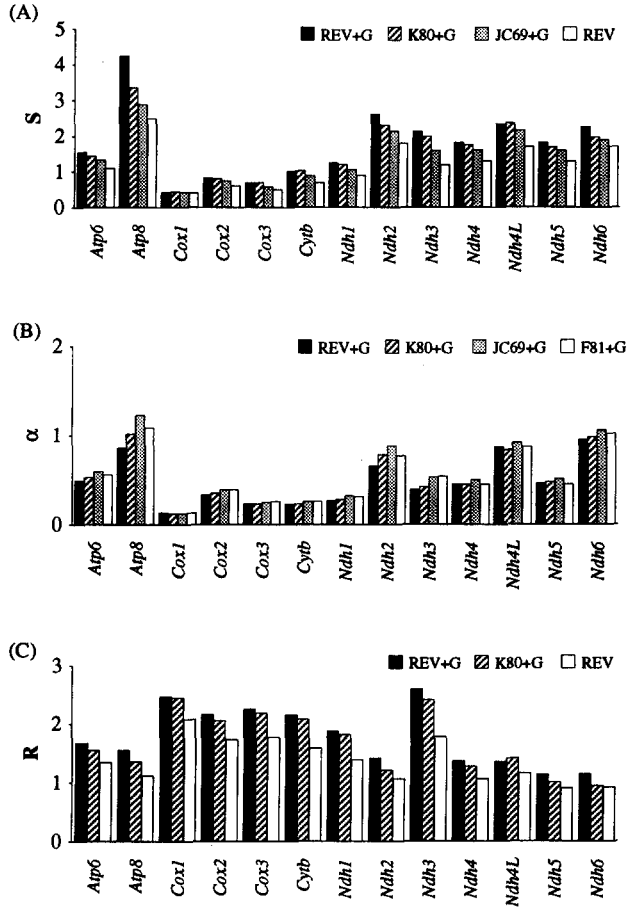

FIGURE 4.—Estimates of evolutionary quantities under different models of nucleotide substitution. (A) overall amount of evolution, *S*. (B) shape parameter of the gamma distributed substitution-rates among sites, α. (C) transition-transversion rate bias averaged over nucleotide frequencies, *R*.

equally well. In the 3rd position data of these mammals, TN93 and HKY85 were not rejected in the likelihood ratio tests (results not shown).

DISCUSSION

Nucleotide frequencies: In the 1st+2nd position data, the frequency of *G* is negatively correlated with overall amount of evolution for individual genes (Figure 2); no such trend was detected for frequencies of nucleotides *A*, *T*, or *C*. Furthermore, the frequency of *G* declines gradually from fishes > Xenopus > chicken > mammals; this tendency was more noticeable in the faster evolving genes. The fact that a change from *G* to *A*, *T*, or *C* in the first two codon positions causes an amino acid substitution and that the rates of amino acid substitution are lower in the basal lineages of the vertebrate phylogeny (ADACHI *et al.* 1993) may partly explain observed cline in the frequency of *G* in different lineages. However, reasons for the high substitution rates from *G* to *A*, *T*, or *C* are not yet clear. The susceptibility of nucleotide *G* to change is also highlighted in the 3rd codon positions where fast rates of sequence evolution have resulted in very low frequency of *G*.

Variability of substitution rates: The gamma parameter is inversely related to the extent of rate variation among sites. An $\alpha \leq 1$ indicates that most sites evolve with low rates, while some sites change with high rates ("hot spots"). If all sites are equally mutable or only small differences in substitution rates exist among sites, the value of α will be >1 (Figure 3A). When we combine the 1st and 2nd codon positions, substitution rates among sites are expected to be different because purifying selection operates with different intensities in the three positions. However, differences in rates of evolution between 1st and 2nd position cannot solely explain these observations since significant rate variation among sites was also found within 1st positions ($0.22 \leq \alpha \leq 1.06$) and within 2nd positions ($0.08 \leq \alpha \leq 2.19$). Moreover, the observed rate variation in the 1st position data cannot be attributed to the faster rates of evolution in the twofold redundant 1st positions in the Leucine codons. Leucine make up ~10–20% of all mitochondrial protein coding genes. However, only a few sites are twofold redundant in the Leucine codons because of low *G*-content in the third position, which results in a skewed relative synonymous codon usage. The estimates of gamma parameters for data sets with and without 1st positions of Leucine clearly showed that the Leucine codons are not contributing substantially to the rate heterogeneity among sites in the 1st position data.

Different regions of mitochondrial genes are under different functional constraints because of their transmembrane location, positions within the multimeric enzyme complexes of the electron transport chain, and co-evolution with some nuclear genes (see GILLHAM 1992 for review). The composite effect of these factors is reflected in rather small estimates of α when 1st and 2nd positions are analyzed together (Table 4) or individually (see results; $\alpha < 1$ for most genes). The estimates of α for *Cox*1, *Cox*3, *Cyt*b, and *Cox*2 (0.13, 0.23, 0.23, 0.33) show a correspondence with the number of putative transmembrane domains, 12, 7, 7–8, and 2, respectively (ESPOSTI *et al.* 1993; GILLHAM 1994). At any rate, of seven subunits of NADH dehydrogenase, the highest rate variation among sites is seen in *Ndh*1, which is thought to be functionally important in the activity of the mitochondrial NADH-ubiquinone reductase complex of the electron transport chain (RAGAN 1987). *Atp*8 codes for the smallest peptides in the mitochondria and is the least conserved of all protein-coding genes. It appears to be under relaxed selective constraints and has not been found in the mtDNA of nematodes.

Figure 3B shows that the estimates of α show significant correlation with the total amount of evolution inferred: highly conserved genes showing greater variation in rates among sites than the faster-evolving genes. It appears that in slowly evolving genes only a small proportion of sites are free to change and that these sites evolve rapidly. Smaller estimates of α for most of the genes of placental mammal sequences indicate greater variations

TABLE 7

Performance of simple models in the 1st+2nd position data analysis

Gene	Maximum likelihood estimation			Pairwise distances[d]
	Branch lengths[a]	R^b	α^c	
*Atp*6	HKY85+G	K80+G	K80+G	K80+G
*Atp*8	HKY85+G	K80+G	K80+G	TN93+G
*Cox*1	REV+G	K80+G	K80+G/F81+G	K80+G
*Cox*2	K80+G	K80+G	K80+G	K80+G
*Cox*3	REV+G	K80+G	K80+G	K80+G
*Cyt*b	HKY85+G	K80+G	K80+G	K80+G
*Ndh*1	HKY85+G	K80+G	K80+G	K80+G
*Ndh*2	HKY85+G	TN93+G	K80+G	K80+G/TN93+G
*Ndh*3	HKY85+G	K80+G	K80+G	TN93+G
*Ndh*4	REV+G	K80+G	K80+G	TN93+G
*Ndh*4L	HKY85+G	K80+G	K80+G/F81+G	TN93+G
*Ndh*5	REV+G	K80+G	K80+G/F81+G	K80+G
*Ndh*6	HKY85+G	K80+G	K80+G	TN93+G

[a] Simplest model not rejected when compared with REV+G (from Table 6).

[b] $z = (R_{TN93+G} - R_{K80+G})/SE(R_{K80+G})$, where SE is the standard error estimated by the curvature method in the PAML program. If $z > 1.96$, we rejected K80+G model at 5% level. The standard error of the estimate of R_{TN93+G} was not considered, which makes this test liberal in rejecting the use of simpler model for analysis.

[c] Same as above, but with values of α; both F81+G and K80+G were examined.

[d] Models were chosen based on Figure 5. The TN93+G model was selected whenever the choice was not obvious, because the variances of pairwise distances estimated by K80+G and TN93+G were quite similar.

in rates among sites (Table 4). It appears that the faster evolutionary rates in mammalian genomes have lead to greater acceleration of rates in the faster evolving sites as compared with other sites, which has resulted in larger differences in rates among sites.

Patterns of nucleotide substitution: In the 1st and 2nd positions, the rates of transitional and transversional substitutions are substantially different and the magnitudes of a and f and b, c, d, and e are disparate. The fact that the fit of TN93+G is not rejected for nine out of 13 genes indicate that the differences in the transversional rates (b, c, d, and e) are not significant. The two transitional substitution rates (a and f) are also not significantly different because HKY85+G and TN93+G fit these genes equally well. However, the equality of four nucleotide frequencies is rejected for all but one gene (*Cox*2) as is evident from the comparison of fits of HKY85+G and K80+G models. Similarly, the significantly worse fit of JC69+G as compared with HKY85+G (and other models) clearly shows significant transition-transversion rate bias.

Thus, the HKY85+G model, which accounts for differences in the transitional ($a = f$) and transversional ($b = c = d = e$) rates and the nucleotide composition bias, may adequately model the process of nucleotide substitution at 1st+2nd codon positions in a majority of mitochondrial protein-coding genes. In this model, a single transition/transversion rate ratio ($\kappa = a/b$) describes the substitution rate biases among four nucleotides, which are given in Table 5. From equation 1, it is clear that κ will be equal to $2R$ if the four nucleotide occur with equal frequencies and if $a = f$ and $b = c = d = e$. In fact, the estimates of κ are generally close to $2R$ (Table 5), because HKY85+G fits the 1st+2nd position data well and because the nucleotide compositions are not highly skewed.

It was argued previously that in slow evolving genes only a few sites are free to change and that these sites evolve very rapidly in slow evolving genes. It is clear from Figure 4 that the extent of transition/transversion rate bias (R or κ) is negatively correlated with S as well as with α, which suggest that the fast changing sites in highly conserved genes evolve with larger transitional-transversional rate bias than those in fast evolving genes.

Choosing models for analysis: In general, the TN93+G and HKY93+G models fit the 1st+2nd and the 3rd position data as well as the REV+G model (Table 6). However, it is desirable to use simpler models of nucleotide substitutions because they give, among others, distance estimates with smaller variances. Therefore, we studied the effects of assumptions made in the substitution model on estimates of α, R, S, and pairwise distances when the 1st+2nd position data are used for evolutionary analyses.

The sum of branch lengths is generally underestimated when simple and unrealistic models are employed (Figure 4A). The extent of underestimation of S primarily depends on the magnitude of rate variation among sites and the rate of evolution in the gene considered (Figure 4A), and the consideration of rate differences among sites is more important than the consideration of substitution rate biases among nucleotides (compare JC69+G and REV in Figure 4A). Thus, in genes evolving with slow or moderate rates, the estimates of S by K80+G are quite close to those by REV+G. How-

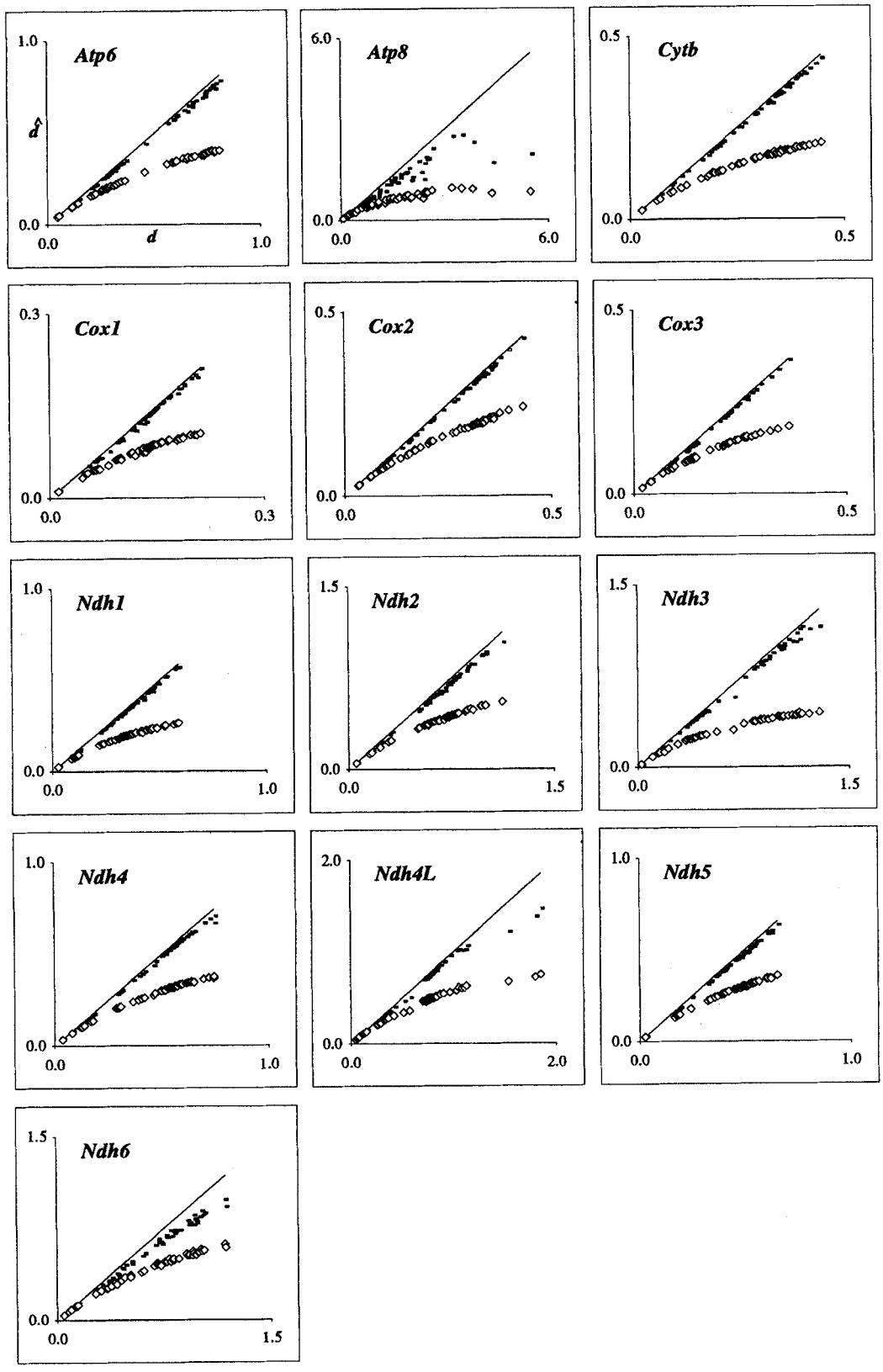

FIGURE 5.—Underestimation of pairwise distances by simple models when using 1st+2nd position data. The pairwise distances computed by Tamura-Nei (1993) model with values of α from Table 4 are on the abscissa (d) and those computed by Kimura's model with variable rates among sites (■; α from Table 4) and uniform rates among sites (◇) are on the ordinate ($\hat{d}$). Extent of underestimation of distances by simple models is reflected in the discrepancy of the curve from the straight line.

ever, the underestimation of S was not caused by the underestimation of each branch length. Instead, there was a tendency for simple models to overestimate short branches and underestimate long branches (results not shown). The underestimation of long branches was much more serious than the underestimation (or overestimation) of short branches (TATENO *et al.* 1994; YANG 1995a). Therefore, complex models found adequate in statistical tests should be employed for estimation of branch lengths and, thus, the sum of branch lengths and the rates of evolution (Tables 6 and 7).

The K80+G, F81+G and JC69+G models were almost always rejected in the likelihood ratio tests, and α was overestimated by these simple models (Figure 4B). However, differences in the estimates of α by K80+G (and F81+G) and REV+G were not statistically significant (Table 7). In general, neglecting transition/transversion rate bias resulted in larger overestimation of α than ignoring the nucleotide composition bias (compare K80+G and F81+G in Figure 4B). Of course, disregarding the nucleotide composition as well as the transition/transversion rate biases (JC69+G) leads to greater overestimates of α (WAKELEY 1994). However, these likelihood estimates were still remarkably less biased than those obtained from the parsimony-based analyses for genes evolving with moderate or fast rates (WAKELEY 1993). One reason for the biased estimates of α in the parsimony analysis is that the number of substitutions at a site inferred by parsimony are in fact the number of *differences* between the ancestral and the descendant sequences, which does not follow the negative-binomial distribution when rates are gamma distributed (YANG and KUMAR 1996). When a correct distribution of number of differences is derived, the bias in the computation of α becomes considerably smaller (compare α'_{MP} and α_{MP} to α in Table 4). In fact, for slowly evolving genes α'_{MP} overestimates α only slightly. In general, parsimony methods underestimate the extent of rate variation because the number of changes inferred at the fast evolving sites are severely underestimation due to long and unequal branches in the tree and the nucleotide frequency and transition rate biases (WAKELEY 1993), especially for genes evolving at moderate or fast rates (Table 4).

In the estimation of R, ignoring rate variation among sites resulted in more severe underestimation than ignoring the nucleotide frequency bias and/or the differences in substitution rates between nucleotides; R_{REV} R_{K80+G} R_{REV+G} (Figure 4C). This is because large variation in rates among sites in mtDNA genes causes more severe underestimation of the number of transitional substitutions than the number of transversional substitutions at the rapidly evolving sites (WAKELEY 1994). Thus, estimates of R by K80+G and REV show the smallest differences for genes with largest values of α (*i.e.*, least variation of rates among sites; Table 7).

To examine the underestimation of pairwise distances by simple models, we assumed that TN93+G model provides the most reliable estimates of pairwise distances (d). TN93+G was chosen because it is the most complex (and adequate) model for which analytical expressions for computing pairwise distances are available. The pairwise distances (and their variances) estimated by K80+G are quite close to the TN93+G estimates for *Atp*6, *Cyt*b, *Cox*1–3, *Ndh*1, and *Ndh*5 data (Figure 5, Table 7). Assumption of rate uniformity among sites resulted in 40–80% underestimation in pairwise distances (compare K80+G and K80), and, as expected, this underestimation was greater for distantly related species and in faster evolving genes (Figure 5). Pairwise distances estimated by K80 and TN93 were almost identical, and the estimates of pairwise distances by JC69+G model were larger than those by K80 (and TN93) but lower than those by K80+G and TN93+G (results not shown). Therefore, the consideration of rate variation among sites along with the transition/transversion rate bias appear to be of foremost importance in the estimation of pairwise distances from the 1st+2nd position data.

I thank Z. YANG, M. NEI, C. A. M. RUSSO, S. B. HEDGES, and J. ZHANG for helpful discussions and critically reading an earlier draft of this paper. This work was supported by National Science Foundation and National Institute of Health grants to M. NEI.

LITERATURE CITED

ADACHI, J., Y. CAO and M. HASEGAWA, 1993 Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid sequence level: rapid evolution in warm-blooded vertebrates. J. Mol. Evol. **36:** 270–281.

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

ARNASON, U., and A. GULLBERG, 1993 Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. J. Mol. Evol. **37:** 312–322.

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.

AVISE, J. C., 1994 *Molecular Markers, Natural History, and Evolution*. Chapman and Hall, New York.

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

BROWN, W. M., 1985 The mitochondrial genome of animals, pp. 95–130 in *Molecular Evolutionary Genetics*, edited by R. J. MACINTYRE. Plenum, New York.

BROWN, W. M., E. M. PRAGER, A. WANG and A. C. WILSON, 1982 Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18:** 225–239.

CARROL, R. L., 1988 *Vertebrate Paleontology and Evolution*. Freeman, New York.

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

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

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. **212:** 599–634.

ESPOSTI, M. D., S. DEVRIES, M. CRIMI, A. GHELLI, T. PATARNELLO *et al.*, 1993 Mitochondirial cytochrome *b*: evolution and structure of the protein. Biochim. Biophys. Acta **1143:** 243–271.

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

FRYE, M. S., and S. B. HEDGES, 1995 Monophyly of the order Rodentia inferred from mitochondrial DNA sequences of the genes for 12S rRNA, 16S rRNA, and tRNA-Valine. Mol. Biol. Evol. **12:** 168–176.

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

GILLHAM, N. W., 1994 *Organelle Genes and Genomes*. Oxford University Press, Oxford.

GINGERICH, P. D., B. H. SMITH and E. L. SIMONS, 1990 Hind limbs of Eocene Basilosaurus: evidence of feet in whales. Science **249:** 154–157.

GINGERICH, P. D., S. M. RAZA, M. ARIF, M. ANWAR and X. ZHOU, 1994 New whale from the Eocene of Pakistan and the origin of cetacean swimming. Nature **368:** 844–847.

GOLDING, G. B., 1983 Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. **1:** 125–142.

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

HASEGAWA, M., and H. KISHINO, 1989 Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. Jpn. J. Genet. **64:** 243–258.

HASEGAWA, M., H. KISHINO and T. YANO, 1985 Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

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

HOLMQUIST, R., M. GOODMAN, T. CONRY and J. CZELUSNIAK, 1983 The spatial distribution of fixed mutations within genes coding for proteins. J. Mol. Evol. **19:** 137–148.

HONEYCUTT, R. L., M. A. NEDBAL, R. M. ADKINS and L. L. JANECEK, 1995 Mammalian mitochondrial DNA evolution: a comparison of the cytochrome *b* and cytochrome *c* oxidase II genes. J. Mol. Evol. **40:** 260–272.

HORAI, S., K. HAYASAKA, R. KONDO, K. TSUGANE and N. TAKAHATA, 1995 Recent african origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. USA **92:** 532–536.

IRWIN, D. M., 1991 Evolution of the cytochrome *b* gene of mammals. J. Mol. Evol. **32:** 128–144.

JANKE, A., G. FELDMAIER-FUCHS, W. K. THOMAS, A. VON-HAESELER and S. PAABO, 1994 The marsupial mitochondrial genome and the evolution of placental mammals. Genetics **137:** 243–256.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. 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.

KISHINO, H., and M. HASEGAWA, 1990 Converting distance to time: an application to human evolution. Methods Enzymol. **183:** 550–570.

KOCHER, T. D., and A. C. WILSON, 1991 Sequence evolution of mitochondrial DNA in human and chimpanzees: control region and a protein-coding region, pp. 391–413 in *Evolution of Life*, edited by S. OSAWA and T. HONJO. Springer-Verlag, New York.

KUMAR, S., K. TAMURA and M. NEI, 1994 MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput. Appl. Biosci. **10:** 189–191.

LANAVE, C., G. PREPARATA, C. SACCONE and G. SERIO, 1984 A new method for calculating evolutionary substitution rates. J. Mol. Evol. **20:** 86–93.

LYNCH, M., and P. E. JARRELL, 1993 A method for calibrating molecular clocks and its application to animal mitochondrial DNA. Genetics **135:** 1197–1208.

MARTIN, R. D., 1993 Primate origins: plugging the gaps. Nature **363:** 223–234.

NAVIDI, W. C., G. A. CHURCHIL and A. VON HAESELER, 1991 Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariant. Mol. Biol. Evol. **8:** 128–143.

NEI, M., 1987 *Molecular Evolutionary Genetics*. Cambridge University Press, New York.

NEI, M., 1991 Relative efficiencies of different tree making methods for molecular data, pp. 133–147 in *Recent Advances in Phylogenetic Studies of DNA Sequences*, edited by M. M. MIYAMOTO and J. L. CRACRAFT. Oxford University Press, Oxford.

NEI, M., 1992 Age of the common ancestor of human mitochondrial DNA. Mol. Biol. Evol. **9:** 1176–1178.

PERNA, N. T., and T. D. KOCHER, 1995 Unequal base frequencies and the estimation of substitution rates. Mol. Biol. Evol. **12:** 359–361.

RAGAN, C. I., 1987 Structure of NADH-ubiquinone reductase (Complex I). Curr. Top. Bioenerg. **15:** 17–35.

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

RITLAND, K., and M. T. CLEGG, 1987 Evolutionary analysis of plant DNA sequences. Am. Nat. **130:** S74-S100.

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.

RUSSO, C. A. M., N. TAKEZAKI and M. NEI, 1996 Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. (In press).

RZHETSKY, A., and M. NEI, 1995 Tests of applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. **12:** 131–151.

SACCONE, C., C. LANAVE, G. PESOLE and G. PREPARATA, 1990 Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol. **183:** 570–583.

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

STORER, T. I., R. L. USINGER, R. C. STEBBINS and J. W. NYBAKKEN, 1971 *General Zoology*. McGraw Hill, New York.

TAMURA, K., 1994 Model selection in the estimation of the number of nucleotide substitutions. Mol. Biol. Evol. **11:** 154–157.

TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:** 512–526.

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

TAVARE, S., 1986 Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. **17:** 57–86.

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 variation among vertebrates. Nucleic Acids Res. **20:** 4853–4858.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES and A. C. WILSON, 1991 African populations and the evolution of human mitochondrial DNA. Science **253:** 1503–1507.

WAKELEY, J., 1993 Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

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

WOLSTENHOLME, D. R., 1992 Animal mitochondiral DNA: structure and evolution. Int. Rev. Cytol. **141:** 173–216.

YANG, Z., 1994a Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39:** 105–111.

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

YANG, Z., 1995a Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. J. Mol. Evol. **40:** 689–697.

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

YANG, Z., 1995c *PAML: Phylogenetic Analysis by Maximum Likelihood*. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park.

YANG, Z., and S. KUMAR, 1996 Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. (in press).

YANG, Z., N. GOLDMAN and A. E. FRIDAY, 1995 Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44:** 384–399.

Communicating editor: M. LYNCH