The Rate Heterogeneity of Nonsynonymous Substitutions in Mammalian Mitochondrial Genes

Xuhua Xia

Department of Ecology and Biodiversity, The University of Hong Kong

Substitution rates at the three codon positions (r_1, r_2, and r_3) of mammalian mitochondrial genes are in the order of $r_3 > r_1 > r_2$, and the rate heterogeneity at the three positions, as measured by the shape parameter of the gamma distribution (α_1, α_2, and α_3), is in the order of $\alpha_3 > \alpha_1 > \alpha_2$. The causes for the rate heterogeneity at the three codon positions remain unclear and, in particular, there has been no satisfactory explanation for the observation of $\alpha_1 > \alpha_2$. I attempted to dissect the causes of rate heterogeneity by studying the pattern of nonsynonymous substitutions with respect to codon positions in 10 mitochondrial genes from 19 mammalian species. Nonsynonymous substitutions involve more different amino acid replacements at the second than at the first codon position, which results in $r_1 > r_2$. The difference between r_1 and r_2 increases with the intensity of purifying selection, and so does the rate heterogeneity in nonsynonymous substitutions among sites at the same codon position. All mitochondrial genes appear to have functionally important and unimportant codons, with the latter having all three codon positions prone to nonsynonymous substitutions. Within the functionally important codons, the second codon position is much more conservative than the codon position. This explains why $\alpha_1 > \alpha_2$. The result suggests that overweighting of the second codon position in phylogenetic analysis may be a misguided practice.

Introduction

Substitution rates at the three codon positions (r_1, r_2, and r_3) of mammalian mitochondrial genes are in the order of $r_3 > r_1 > r_2$ (Kimura 1983, p. 95; Nei 1987, p. 72; Irwin, Kocher, and Wilson 1991; Yang 1996*a*, 1996*b*). The standard explanation for this is that any nucleotide substitution at the second codon position is invariably nonsynonymous and should be under strong purifying selection, whereas most nucleotide substitutions at the third codon position and at least some nucleotide substitutions at the first codon position are synonymous and should evolve faster because of relatively weak purifying selection against synonymous substitutions (Kimura 1977, 1983, pp. 94–96; Nei 1987, p. 73; Li 1997, pp. 179–182).

This explanation leads one to expect higher heterogeneity of the substitution rate at the first than at the second codon position, because second codon positions are all nondegenerate sites, whereas first codon positions are more heterogeneous, consisting of nondegenerate sites (presumably with a low substitution rate) and twofold-degenerate sites (presumably with a relatively high substitution rate). This prediction, however, turned out to be wrong.

Available evidence shows that heterogeneity of substitution rate is the most dramatic at the second codon position and the least dramatic at the third codon position (Yang 1996*b*). Rate heterogeneity is commonly measured by fitting the gamma distribution of substitution rates, with the resulting shape parameter α inversely correlated with increasing rate heterogeneity. For four mitochondrial genes from six hominoid species, the α parameter equals 0.18, 0.08, and 1.58, respectively, for the first, second, and third codon positions (Yang 1996*b*), contrary to our expectation that the rate of nucleotide substitutions should be more homogeneous at the second codon position than at the first codon position.

The causes of rate heterogeneity at the three codon positions may be quite different. For mitochondrial protein genes, there are three sources of rate variation among sites: between genes, between codon positions within each gene (referred to hereafter as between-CP), and among sites at the same codon position (referred to hereafter as within-CP). Between-gene variation could be caused either by the difference in purifying selection or by the difference in amino acid composition (i.e., some genes having more conservative amino acids than others, see Graur 1985). Between-CP variation is measured by the difference among r_1, r_2, and r_3, and within-CP variation in substitution rate is measured by α.

Some of the causes of the differences in α values among the three codon positions are known. For example, the rate heterogeneity at the third codon position is presumably caused by the difference between the twofold- and fourfold-degenerate sites, whereas that at the first position may be caused by the difference in substitution rates between nondegenerate and twofold-degenerate sites. The nucleotide sites at the second codon position are all nondegenerate and presumably should have a homogeneously low rate of substitution. It is therefore not obvious why we should have $\alpha_1 > \alpha_2$.

Two hypotheses can be proposed to account for the unexpectedly high rate heterogeneity at the second codon position. The first hypothesis invokes differential purifying selection against different nonsynonymous substitutions. Different types of nonsynonymous substitutions are known to occur at very different rates within each gene (Zuckerkandl and Pauling 1965; Sneath 1966; Epstein 1967; Grantham 1974; Miyata, Miyazawa, and Yasunaga 1979; Kimura 1983, p. 152). For example, the substitution of a leucine codon by an isoleucine codon is much more frequent than that by an arginine codon. This is conventionally explained by differential purify-

Key words: nonsynonymous substitution, codon, mitochondrial gene, rate heterogeneity, mammal.

Address for correspondence and reprints: Xuhua Xia, Department of Ecology and Biodiversity, The University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: xxia@hkusub.hku.hk.

Mol. Biol. Evol. 15(3):336–344. 1998

ing selection that tolerates amino acid replacements involving similar amino acids and disallows amino acid replacements involving very different amino acids (Zuckerkandl and Pauling 1965; Sneath 1966; Epstein 1967; Grantham 1974; Miyata, Miyazawa, and Yasunaga 1979; Kimura 1983, p. 152). Nonsynonymous substitutions at the second codon position may involve both very similar and very different amino acid replacements, leading to dramatic rate heterogeneity. In contrast, nonsynonymous substitutions at the first codon position may involve amino acid replacements that are all similar to each other and have substitution rates similar to that of synonymous substitutions. This would explain the observation of both $r_1 > r_2$ and $\alpha_1 > \alpha_2$.

The second hypothesis assumes that a nucleotide change at the second codon position typically involves replacement of very different amino acids (Haig and Hurst 1991) and should generally be very rare. However, some codons may code for amino acids not located in functional domains and therefore are unimportant for the normal function of the protein. Such codons should be highly variable, with the substitution rate approaching that of neutral mutations. This implies that the second codon position will then have two groups of sites, one located in important codons and highly conservative, and the other located in unimportant codons and highly variable. These two groups of sites could give rise to the dramatic rate heterogeneity at the second codon position.

The second hypothesis, however, does not really explain $\alpha_1 > \alpha_2$, because given the functionally important and unimportant codons, we should expect the first codon position also to contain two heterogeneous groups of sites, one located in the important codons and having a low substitution rate, and the other located in the unimportant codons and having a high substitution rate. In short, the second hypothesis predicts that the substitution rate will be highly heterogeneous for both the first and the second codon positions, but falls short of explaining why the substitution rate should be more heterogeneous at the second than at the first codon position.

For the second hypothesis to explain why $\alpha_1 > \alpha_2$, we need the additional condition that nonsynonymous substitutions at the second codon position involve more different amino acid replacements than do those at the first codon position. Thus, for codons specifying unimportant protein segments, the substitution rate will be high for both the first and the second codon positions. For codons specifying important protein domains, the substitution rate will be much lower at the second than at the first codon position. Only when this additional condition is met can the second hypothesis explain both $r_1 > r_2$ and $\alpha_1 > \alpha_2$.

In this paper, I tested the validity of the two hypotheses outlined above by quantifying the empirical pattern of nonsynonymous substitutions in mammalian mitochondrial genes. The empirical data favor the second hypothesis.

Materials and Methods

The data consist of complete mitochondrial DNA sequences from 19 mammalian species: hedgehog (GenBank accession number X88898), mouse (J01420), rat (X14848), cat (U20753), gray seal (X72004), harbor seal (X63726), horse (X79547), donkey (X97337), rhinoceros (X97336), cow (V00645), fin whale (X61145), blue whale (X72204), gibbon (X99256), Sumatran orangutan (X97707), Bornean orangutan (D38115), gorilla (X93347), pygmy chimpanzee (D38116), chimpanzee (D38113), and human (X93334). Of the 13 protein-coding genes, only 10 were used in this study, with the three shortest genes (ATPase 8, ND3, and ND4L) excluded.

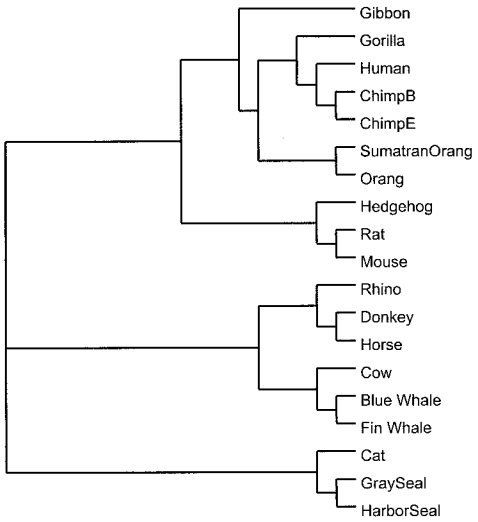


FIG. 1.—The unrooted phylogenetic tree for the 19 mammalian species.

There are now at least 23 completely sequenced mammalian mitochondrial genomes, several of which (e.g., opossum, wallaroo, and duckbill platypus) were not used for this paper because of difficulty in sequence alignment. Although sequences from diverse taxa such as mammalian, avian, and amphibian species have previously been aligned (Cummings, Otto, and Wakeley 1995; Otto, Cummings, and Wakeley 1996; Janke, Xu, and Arnason 1997), I am not certain of the accuracy of such alignments. This study needs reconstruction of ancestral sequences, and inaccuracy in sequence alignment may introduce systematic biases.

The unrooted phylogenetic tree (fig. 1) for the 19 mammalian species receives the strongest support from both traditional and molecular phylogenetics and represents our existing knowledge of mammalian evolution (Novacek, Wyss, and McKenna 1988; Cao et al. 1994; Cummings, Otto, and Wakeley 1995; Janke, Xu, and Arnason 1997). The only difference between the tree in figure 1 and that in Janke, Xu, and Arnason (1997) is that Janke, Xu, and Arnason grouped perissodactyls and carnivores together as sister taxa, whereas perissodactyls and artiodactyls are grouped as sister taxa in figure 1. The slight difference in topology between the two trees has only a negligible effect on the result in this paper. I used the tree in figure 1 for the reconstruction of ancestral states by using the BASEML program in the PAML package (Yang 1996*c*), which implements the likelihood-based method detailed in Yang, Kumar, and Nei

Table 1
Basic Information on Codon Differences from 35 Pairwise Comparisons Between Neighboring Nodes on the Phylogenetic Tree (fig. 1) for Each of the 10 Mitochondrial Genes Studied

	N_{codon}	N_{same}	N_{syn}	N_1	$N_{>1}$	R
COI	513	15,376	2,364	161	54	0.419
COIII	261	7,814	1,105	151	65	0.828
COII	227	6,773	983	121	68	0.833
Cyt-b	379	11,298	1,485	352	130	1.272
ND1	315	9,408	1,204	311	101	1.308
ATPase 6	226	6,615	937	251	107	1.584
ND4	459	13,407	1,882	532	241	1.684
ND5	600	17,332	2,311	955	393	2.247
ND6	172	5,009	600	312	96	2.372
ND2	343	9,842	1,255	656	252	2.647

NOTE.—N_{same}—number of identical codon pairs; N_{syn}—number of synonymous codon pairs; N_1, $N_{>1}$—numbers of nonsynonymous codon pairs differing at one and more than one positions, respectively. $R = (N_1 + N_{>1})/N_{codon}$, and is an overall measure of amino acid conservativeness. Sorted by R.

(1995). The reconstruction is generally satisfactory, with the overall accuracy being around 0.90.

Nonsynonymous codon substitutions were counted by comparing the DNA sequences between two neighboring nodes. For example, for the unrooted tree in figure 1, there are 35 pairwise comparisons, 16 between neighboring internal nodes and 19 between the terminal nodes and their neighboring internal nodes. This counting procedure differs from some other studies that counted substitutions from all possible pairwise comparisons, many of which are nonindependent and would introduce biases (Felsenstein 1992; Nee et al. 1996; Xia, Hafner, and Sudman 1996).

The 35 pairwise codon-by-codon comparisons result in four categories of codon pairs: identical codon pairs, different but synonymous codon pairs, nonsynonymous codon pairs differing at one codon position, and nonsynonymous codon pairs differing at more than one codon position. These basic data are summarized in table 1, together with relevant sequence information. Note that the nonsynonymous substitution rate (R in table 1) differs greatly among genes. All analyses in this paper are done separately for each gene.

Amino acid dissimilarity has been quantified in various ways (Sneath 1966; Grantham 1974; Miyata, Miyazawa, and Yasunaga 1979), but all measures are highly correlated with each other. I used only Grantham's and Miyata's distances in this study, and the results are similar. Only results from Grantham's distance (Grantham 1974) were presented. Grantham's distance will be referred to as D_G hereafter. If the first hypothesis is correct, then we should expect the mean and variance of D_G to be larger for nonsynonymous substitutions at the second than at the first codon position. If the second hypothesis is correct, we should expect mean D_G to be larger at the second than at the first codon position.

The second hypothesis also postulates that mitochondrial protein genes are structured into functionally important and unimportant codons or DNA segments. This implies that nonsynonymous substitutions will be clumped in the unimportant codons and absent or rare in important codons. This can be examined in two ways. First, all three codon positions should be prone to nonsynonymous substitutions in an unimportant codon, but should all be highly conservative in an important codon. Thus, the number of nonsynonymous substitutions occurring at one codon position should be correlated with the number of nonsynonymous substitutions occurring at other codon positions. This correlation is quantified both by Pearson's and Spearman's correlation coefficients and by the χ^2 test.

Second, if we examine one particular codon position (e.g., the first codon position) of all codons, we should find some first codon positions, if they are located in unimportant codons, to experience nonsynonymous substitutions repeatedly during mammalian evolution. In contrast, those first codon positions located in important codons should have no or few nonsynonymous substitutions along the mammalian lineages. The substitution rate will therefore differ much among these first codon positions. This can be revealed by fitting a negative binomial distribution to the observed number of nonsynonymous substitutions. I used the maximum-likelihood estimator in Johnson, Kotz, and Kemp (1992, p. 216) to estimate k through computer iteration. The iteration stops when the difference between the two sides of the equation is smaller than 0.00001.

Results and Discussion

The Expected Rate of Nonsynonymous Substitutions with Respect to Codon Position and Transition/Transversion Ratio

Of the 60 mitochondrial codons, there are 190 possible nonsynonymous codon pairs in which one codon can mutate into the other through a single nucleotide substitution, e.g., ACU-GCU. (Reciprocal codon pairs, e.g., ACU-GCU and GCU-ACU, were treated as the same type of nonsynonymous codon substitutions; otherwise, there would have been 380 possible nonsynonymous codon pairs differing at one codon position.) These 190 nonsynonymous codon pairs are grouped into five categories according to whether the nonsynonymous substitution occurs at the first, second, or third codon position and whether it is a transition or transversion. The result (table 2) shows that when we compare two DNA sequences and count nonsynonymous codon pairs that differ at one codon position, we should expect, assuming equal codon usage and equal probability of nonsynonymous substitutions, 43.2% (=82/190) of the nonsynonymous codon pairs to differ at the first codon position, 44.2% at the second codon position, and only 12.6% at the third codon position. Similarly, we should expect 28.4% of nonsynonymous codon pairs to differ by a transition, and 71.6% to differ by a transversion (table 2).

Our first hypothesis postulates that some nonsynonymous substitutions at the second codon position might involve very different amino acid replacements (having low substitution rates), whereas others might involve very similar amino acid replacements (having high substitution rates). Similar amino acid replace-

Table 2
Distribution of the 190 Possible Nonsynonymous Codon Pairs According to Codon Position and Transitions/Transversions

	Codon Position				
	1	2	3	Subtotal	Prop.
s	26	28	0	54	0.284
Mean D_G . .	63.92	92.64	Null	78.81	
Var D_G	1,758.95	1,955.5	Null	2,035.7	
v	56	56	24	136	0.716
Mean D_G . .	71.21	104.46	73.67	85.34	
Var D_G	2,568.68	1,541.13	4,879.54	2,764.4	
Sum	82	84	24	190	
Prop.	0.432	0.442	0.126	1	
Mean D_G . .	68.9	100.52	73.67		
Var D_G	2,298.71	1,688.78	4,879.54		

Note.—s—transition; v—transversion; D_G—Grantham's distance; Prop.—proportions; Var—variance. Transitions at the third codon position are all synonymous.

ments have small D_G values, and different amino acid replacements have large ones (Grantham 1974). If nonsynonymous substitutions at the second codon position involve both very different and very similar amino acid replacements, then D_G should have a large variance for nonsynonymous substitutions at the second codon position. This is not true (table 2). The variance of D_G is, in fact, smaller for nonsynonymous substitutions at the second codon position (1,688.78) than for those at the other two positions (2,298.71 and 4,879.54 for codon positions 1 and 3, respectively; table 2).

The expected values in table 2, however, should be adjusted for unequal codon usage because the assumption of equal codon usage is never true for real sequences. Let N_{jz} be the number of all possible nonsynonymous codon pairs that differ at codon position j ($j = 1, 2, 3$) with a substitution type z (z stands for either a transitional change or a transversional change), let n_{ijz} be the number of nonsynonymous codons into which codon i ($i = 1, 2, \cdots, 60$) can mutate through a change of type z at the jth codon position. With equal codon usage, we already have

$$N_{jz} = \sum_{i=1}^{60} n_{ijz}. \quad (1)$$

For real sequences with unequal codon usage, let F_i ($i = 1, 2, \ldots, 60$) be the empirical codon frequency of the 60 mitochondrial codons. Now we have

$$N_{jz} = \sum_{i=1}^{60} F_i n_{ijz}. \quad (2)$$

The distribution of nonsynonymous codon pairs adjusted for codon frequency in real sequences (table 3) differs only slightly from the unadjusted distribution shown in table 2. For example, the proportion of nonsynonymous codon substitutions being transversions at the first codon position is 29.5% (table 2), unadjusted for codon frequency, and the equivalent value adjusted for codon frequency for each of the 10 genes varied from 29.3% to 30.4% (table 3). Thus, the adjustment seems unnecessary.

Table 3
Distribution of All Possible Nonsynonymous Substitutions According to Codon Position and Transitions/Transversions (s/v), Adjusted for Codon Frequencies in Each of the 10 Genes

		Frequency: Codon Position			Proportion: Codon Position		
Gene	s/v	1	2	3	1	2	3
ATPase 6	s	6,509	7,855		0.13	0.15	
	v	15,788	15,069	6,738	0.30	0.29	0.13
COI	s	15,717	17,545		0.13	0.15	
	v	35,491	34,309	15,338	0.30	0.29	0.13
COII	s	6,918	7,832		0.13	0.15	
	v	15,389	15,064	7,320	0.29	0.29	0.14
COIII	s	7,898	8,854		0.13	0.15	
	v	17,863	17,578	7,926	0.30	0.29	0.13
Cyt-b	s	11,435	12,884		0.13	0.15	
	v	26,131	25,566	11,648	0.30	0.29	0.13
ND1	s	9,326	10,765		0.13	0.15	
	v	21,491	21,023	8,772	0.30	0.29	0.12
ND2	s	10,239	11,522		0.13	0.15	
	v	23,626	22,372	10,632	0.30	0.29	0.14
ND4	s	13,590	15,633		0.13	0.15	
	v	31,638	30,302	13,592	0.30	0.29	0.13
ND5	s	18,166	20,351		0.13	0.15	
	v	41,439	40,089	19,524	0.30	0.29	0.14
ND6	s	4,890	5,942		0.12	0.15	
	v	11,770	11,244	5,442	0.30	0.29	0.14

Note.—The proportion of nonsynonymous substitutions falling into each category is similar to the unadjusted value in table 2.

The variance of D_G is again the smallest for nonsynonymous substitutions at the second codon position (table 4). I conclude that, in comparison with nonsynonymous substitutions at the first and third codon positions, nonsynonymous substitutions at the second codon position tend to be more homogeneous in their effect. This contradicts the prediction of the first hypothesis, which is consequently rejected.

One might argue that this rejection is unjustified, because the effect of transition bias was not taken into account. Take data in table 2, for example. If no transitional mutations are repaired and all transversional mutations are repaired, then the mean and variance of D_G are 92.64 and 1,955.5, respectively, at the second codon position (table 2). These values are larger than the corresponding values at the first codon position (63.92 and 1,758.95, respectively; table 2). This supports the first hypothesis.

To accommodate the transition bias, we can use the following equations to calculate the mean and variance of D_G for each codon position:

$$\bar{D}_G = \frac{\kappa \sum_{i=1}^{N_s} D_{G_i} + \sum_{i=1}^{N_v} D_{G_i}}{(\kappa N_s + N_v) - 1}, \quad (3)$$

$$s^2_{\bar{D}_G} = \frac{\kappa \sum_{i=1}^{N_s} (D_{G_i} - \bar{D}_G)^2 + \sum_{i=1}^{N_v} (D_{G_i} - \bar{D}_G)^2}{(\kappa N_s + N_v) - 1}, \quad (4)$$

where κ designates transition bias, and N_s and N_v des-

Table 4
Mean and Variance of D_G Values for All Possible Nonsynonymous Substitutions, Grouped According to Codon Positions (represented by the numbers 1, 2, and 3)

	Mean D_G			Variance of D_G		
Gene	1	2	3	1	2	3
ATPase 6	49.05	99.82	39.81	1,440.34	1,373.08	1,911.59
COI	58.6	101.54	49.54	2,069.23	1,724.56	3,198.92
COII	56.61	103.11	50.95	1,911.35	1,481.26	2,746.35
COIII	60.94	101.57	57.28	2,016.63	1,879.77	4,037.77
Cyt-b	57.44	103.98	52.94	2,037.76	1,691.55	3,509.06
ND1	53.55	102.3	51.32	1,744.41	1,536.13	3,252.18
ND2	50.88	101.67	48.77	1,634.14	1,398.77	3,133.59
ND4	53.38	101.72	50.89	1,788.62	1,471.67	3,367.43
ND5	53.89	102.35	49.83	1,775.49	1,657.55	2,801.61
ND6	62.95	101.92	46.24	2,583.85	1,771.44	3,141.12

Note.—The variance of D_G is the smallest at the second codon position.

ignate the numbers of possible nonsynonymous transitions and transversions (e.g., 26 and 56, respectively, for the first codon position; table 2).

We do not know what value κ should take. However, Xia, Hafner, and Sudman (1996), working with mitochondrial genes from more closely related species, found that almost all observed transition bias arose at the twofold-degenerate sites (located mostly at third codon positions) where transitions are synonymous and transversions are nonsynonymous. The transition bias due to mutation is moderate, on the order of 2–4. In this study, the transition/transversion (s/v) ratios for observed nonsynonymous substitutions at the first and second codon positions of our mitochondrial genes ranged from 1 to 4, with more conserved genes (e.g., COI) having larger s/v ratios than less conserved genes (e.g., ND2). Even if we have $\kappa = 10$, the variance of D_G is still smaller for the second codon position than for the first codon position (1,848.56 and 1,852.27, respectively). Thus, the first hypothesis is not supported.

Table 5
The Observed Proportion (Obs) of Nonsynonymous Substitutions Is Invariably Higher than the Expected Value (Exp) at the First Codon Position, and Is Invariably Lower than the Expected Value at the Second Codon Position

		Codon Position					
Gene		1	2	3	N_{NS}	χ^2	φ
ATPase 6	Exp	0.429	0.441	0.13			
	Obs	0.606	0.295	0.1	251	32.20	0.358
COI	Exp	0.433	0.438	0.13			
	Obs	0.696	0.193	0.112	161	48.18	0.547
COII	Exp	0.425	0.436	0.139			
	Obs	0.612	0.24	0.149	121	20.70	0.414
COIII	Exp	0.429	0.44	0.132			
	Obs	0.603	0.325	0.073	151	19.18	0.356
Cyt-b	Exp	0.429	0.439	0.133			
	Obs	0.645	0.259	0.097	352	67.69	0.439
ND1	Exp	0.432	0.445	0.123			
	Obs	0.614	0.289	0.096	311	42.70	0.371
ND2	Exp	0.432	0.432	0.136			
	Obs	0.547	0.287	0.166	656	56.35	0.293
ND4	Exp	0.432	0.438	0.13			
	Obs	0.57	0.293	0.137	532	49.19	0.304
ND5	Exp	0.427	0.433	0.14			
	Obs	0.537	0.331	0.132	955	50.45	0.230
ND6	Exp	0.424	0.437	0.139			
	Obs	0.561	0.292	0.147	312	28.97	0.305

Note.—The difference between the observed and the expected values is statistically significant with χ^2 tests ($P < 0.001$). N_{NS} is the total number of observed nonsynonymous codon substitutions for each gene from 35 pairwise comparisons between neighboring nodes, and φ ($= \sqrt{\chi^2/n}$) is a sample-size-independent measure of deviation of the observed value from the expected value.

Nonsynonymous Substitutions at the Second Codon Position Involve More Different Amino Acid Replacements than Those at the First and Third Codon Positions

One prominent feature in tables 2 and 4 is that mean D_G for nonsynonymous codon substitutions at the second codon position is much greater than for those at the first and third codon positions. Because similar amino acids tend to replace each other more frequently than do different ones (Zuckerkandl and Pauling 1965; Sneath 1966; Epstein 1967; Clarke 1970; Grantham 1974; Miyata, Miyazawa, and Yasunaga 1979), we should expect nonsynonymous substitutions to be rarer at the second codon position than at the first or third codon position.

I compared the observed number of nonsynonymous substitutions among the three codon positions, with the expected values based on the assumption of equal probability of nonsynonymous substitutions. The observed proportion of nonsynonymous substitutions is invariably higher than the expected value at the first codon position, and is invariably lower than the expected value at the second codon position ($P < 0.001$ for all 10 genes; table 5). This is indicative of purifying selection operating on these genes.

If the deviation of the observed value from the expected value is truly caused by purifying selection, then we should expect greater deviation to be correlated with stronger purifying selection. I used φ ($=\sqrt{\chi^2/n}$) as a sample-size-independent measure of the deviation of the observed value from the expected value (table 5), and R (nonsynonymous substitution rate) in table 1 as a measure of the intensity of purifying selection (with larger R values corresponding to weaker selection). We should expect φ to be negatively correlated with R. The cor-

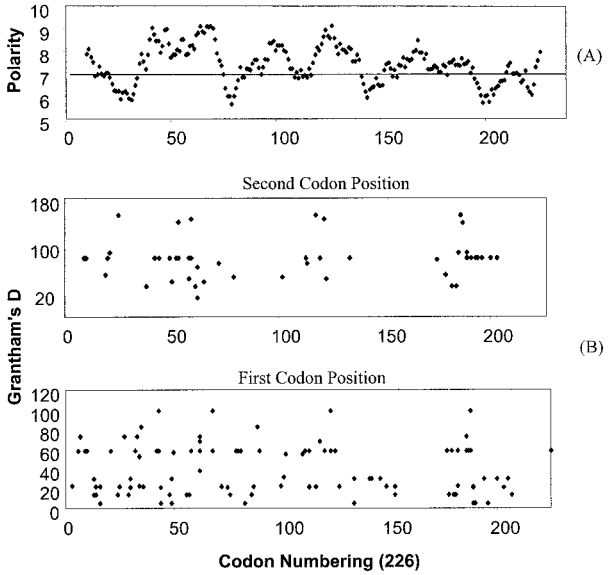


FIG. 2.—Protein structure and nonrandom distribution of nonsynonymous substitutions. *A*, Polarity plot for ATPase 6. Each point represents a moving average of polarity values for 10 neighboring amino acids. Polarity values are from Grantham (1974). *B*, Distribution of the observed nonsynonymous substitutions falling at the first and second codon positions along the ATPase gene. Each point is a Grantham's distance between two amino acids involved in a nonsynonymous codon substitution.

relation coefficient is -0.83 ($P = 0.003$) between the two.

The conventional hypothesis for the observation of $r_3 > r_1 > r_2$ is that any nucleotide substitution at the second codon position is invariably nonsynonymous and should be under strong purifying selection, whereas most nucleotide substitutions at the third codon position and at least some nucleotide substitutions at the first codon position are synonymous and should evolve faster because of relatively weak purifying selection against synonymous substitutions (Kimura 1977, 1983, pp. 94–96; Nei 1987, p. 73; Xia, Hafner, and Sudman 1996). This explanation ignores the rate heterogeneity of nonsynonymous substitutions among the three codon positions. Table 5 shows that, even if we consider nonsynonymous substitutions only, the second codon position is still much more conservative than the first and third codon positions.

Recall that our second hypothesis postulates that the mean D_G should be greater for nonsynonymous substitutions at the second than at the first codon position. The finding above is therefore consistent with the second hypothesis. However, to substantiate the second hypothesis, we also need to provide evidence to show the presence of functionally important and unimportant codons with strong and weak purifying selection, respectively.

Nonsynonymous Substitution Rate and the Functional Domains of Proteins

Most mitochondrial proteins are transmembrane proteins made of hydrophobic and hydrophilic domains associated with different nonsynonymous substitution rates (Kyte and Doolittle 1982; Irwin, Kocher, and Wilson 1991). The recognition of hydrophobic or hydrophilic segments is aided by a polarity plot, which I have done for ATPase 6 (fig. 2). The figure reveals some structural heterogeneity of the protein molecule. The distribution of nonsynonymous substitutions along the DNA sequence also exhibits apparent discontinuity, especially at the second codon position (fig. 2), where long stretches of the DNA sequence harbor no nonsynonymous substitutions at all for all 35 pairwise comparisons. For example, there is no nonsynonymous substitution at the second codon position between (and excluding) codon sites 79 and 102, 134 and 175, and 203 and 226. Such long stretches of DNA sequence devoid of nonsynonymous substitutions are indicative of functionally constrained protein domains.

The distribution of nonsynonymous substitutions over sites was fitted with the negative binomial distribution. All first codon positions were concatenated into one sequence, and all second codon positions into another (fig. 3). A window size of 1, 3, or 5 was then used as a sampling unit along the DNA sequence (fig. 3). I did not fit the negative binomial distribution to the nonsynonymous substitutions occurring at the third codon position, because a large number of sites at the third codon position simply cannot have nonsynonymous substitutions (e.g., all fourfold-degenerate sites). The *k* value will necessarily be small for nonsynonymous substitutions at the third codon position, but it is perhaps not biologically meaningful. Also, the Poisson distribution fits the data poorly.

The estimated *k* values (table 6) reveal two patterns. First, the *k* value is generally small, even for window sizes larger than one, confirming the suspected heterogeneous distribution of nonsynonymous substitutions along the DNA sequence. Second, the *k* value is correlated with gene conservativeness, with more conservative genes having smaller *k* values. The Pearson correlation coefficient between *k* and *R* (table 1) varies from

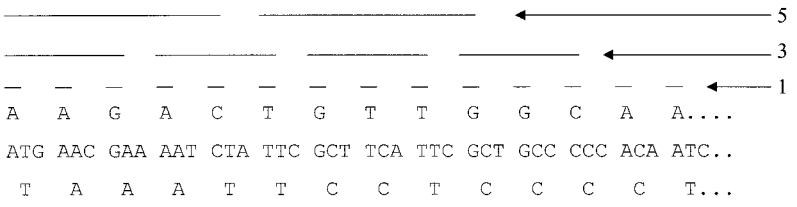
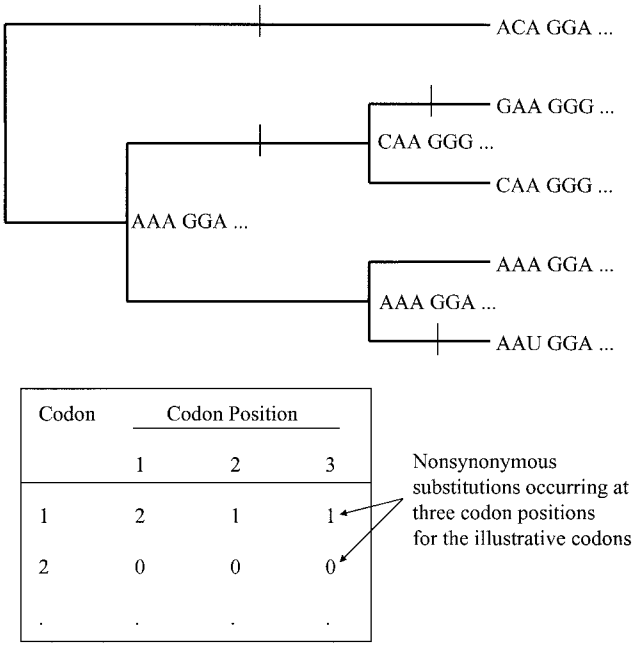


FIG. 3.—Illustration of sampling units for fitting the negative binomial distribution to observed nonsynonymous substitutions. Two new sequences were derived from the original protein gene, with the one above consisting of all first codon positions and the one below consisting of all second codon positions. The numbers 1, 3, and 5 designate window sizes (sampling units).

Table 6
Fitting of the Negative Binomial Distribution to the Number of Nonsynonymous Substitutions Along the DNA Sequence, Done Separately for First and Second Codon Positions (represented by numbers 1 and 2, respectively, in the top row)

	1			2		
GENE	1	3	5	1	3	5
COI	0.188	0.366	0.442	0.198	0.242	0.410
COIII	0.309	0.886	1.625	0.155	0.398	0.595
COII	0.770	1.535	2.073	0.283	0.646	0.750
Cyt-b	0.376	0.877	1.376	0.151	0.413	0.697
ND1	0.672	1.256	1.510	0.211	0.398	0.521
ATPase 6	0.733	1.062	1.341	0.216	0.448	0.487
ND4	0.651	1.469	2.266	0.325	0.609	0.980
ND5	0.865	1.353	1.505	0.459	0.756	0.720
ND6	1.496	2.413	2.354	0.844	1.130	1.432
ND2	1.883	3.313	4.305	0.807	1.526	1.616

NOTE.—Nonoverlapping window sizes of 1, 3, and 5 were used as sampling units (fig. 3). The numbers are values of the parameter *k* of the negative binomial distribution.



Codon	Codon Position		
	1	2	3
1	2	1	1
2	0	0	0
.	.	.	.

Nonsynonymous substitutions occurring at three codon positions for the illustrative codons

FIG. 4.—Illustration of conserved and highly variable codons. The table shows the type of data used in calculating correlation coefficients between the numbers of nonsynonymous substitutions occurring at different codon positions.

0.71 to 0.86 for different genes and is statistically significant ($P \leq 0.0222$). Thus, for protein genes subject to weak purifying selection (e.g., ND2), nonsynonymous substitutions tend to fall on codons in a relatively random fashion and with high frequencies. For proteins subject to strong purifying selection (e.g., COI), nonsynonymous substitutions can occur frequently only at unconstrained (unimportant) codon positions.

If a protein gene contains functionally important and unimportant codons, then we should expect different codon positions to covary. For example, if a codon is important, then purifying selection will prevent nonsynonymous substitutions from occurring at any of the three codon positions (e.g., the second codon in fig. 4). If a codon is unimportant, then nonsynonymous substitutions will fall on all three codon positions in different lineages (e.g., the first codon in fig. 4). This implies that the three columns of data headed by codon positions 1, 2, and 3 in figure 4 should be correlated with each other. Because the observed number of nonsynonymous substitutions at the third codon position is small, I have added the third column to the second column and calculated the correlation between this pooled column and the first column. This is done only for ND5 (the longest mitochondrial gene). The resulting correlation is 0.38 (Pearson) and 0.42 (Spearman), and both are highly significant ($P = 0.0001$).

An alternative way of quantifying the correlation is to count the number of codons that had (1) nonsynonymous substitutions at the first codon position and at one of the two remaining codon positions, (2) nonsynonymous substitutions at the first but not at the second or third codon position, (3) nonsynonymous substitutions at the second or third codon position but not at the first codon position, and (4) no nonsynonymous substitutions at any codon position. The corresponding numbers are 145, 112, 65, and 278 for ND2, with the resulting likelihood ratio chi-square value equal to 91 ($P = 0.0001$).

Even highly conserved genes (which presumably are under strong purifying selection) could have codons that seem to be weakly constrained. For example, COI is a highly conserved gene (table 1), yet quite a large number of nonsynonymous substitutions with a D_G value near 100 have escaped purifying selection (fig. 5). The distribution shown in figure 5 cannot be nicely fitted with selection models assuming uniform purifying selection across all codons. One has to assume that a certain fraction of codon sites are not constrained or are weakly constrained in order to account for the distribution. In other words, for all protein genes studied in this paper, no matter how conserved they may be, some codons must be under weak selection and be highly variable (i.e., having a substitution rate close to the neutral rate). This accords with the substitution pattern shown in figure 2, where the second codon position is shown graphically to be either highly conservative, with long stretches of unvaried sites, or highly variable, with a number of nonsynonymous substitutions having quite large D_G values.

The finding that the second codon position is either highly conservative or highly variable offers us a satisfactory explanation of the extreme rate heterogeneity at the second codon position. This seemingly trivial finding has significant implications for phylogenetic analysis. It is customary for phylogeneticists to assign more weight to substitutions at the second codon position (too many examples to cite), and popular computer programs for phylogenetic analysis, such as PHYLIP (Felsenstein 1993) and PAUP (Swofford 1993), have special provisions to facilitate the practice of overweighting the second codon position. However, this practice seems misguided given the finding that the sec-

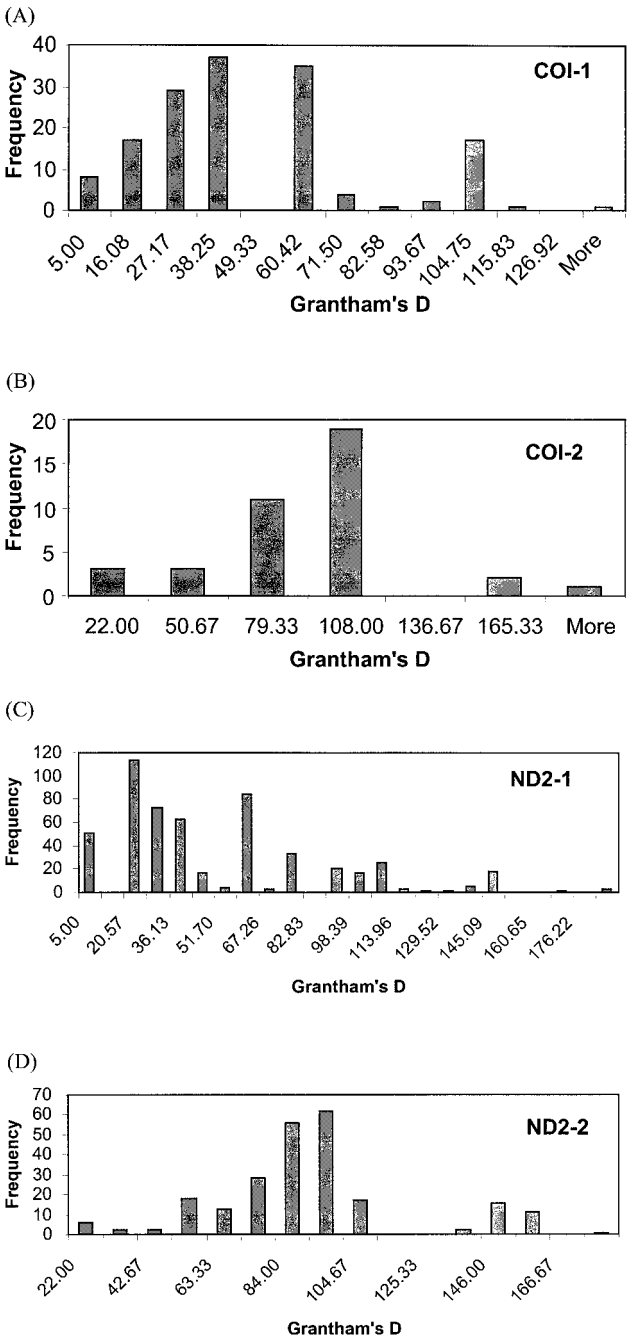


FIG. 5.—Frequency distribution of nonsynonymous substitutions along the range of Grantham's distances (*A* and *B*) for first and second codon positions, respectively, of COI (which is the most conservative mitochondrial gene) and (*C* and *D*) for first and second codon positions, respectively, of ND2 (which is the least conservative mitochondrial gene).

ond codon position appears either highly conservative or highly variable. Overweighting highly conservative sites has little effect on the outcome of the phylogenetic analysis, whereas overweighting highly variable sites is equivalent to overweighting potential homoplasies and may lead to serious bias in estimating the tree topology.

Let me illustrate this point with a simple analogy. Suppose we are to find the difference in body height between adult men and women. At one sampling point with a dense population, we measure 10 men and 10 women and find that the mean height for men is 1 cm greater than that for women. In another sampling point with a sparse population, we are able to measure only one man and one woman, and the woman happens to be 1.5 cm taller than the man. If we now give a weight of 10 to the two measurements from the sparse population and a weight of 1 to the 20 measurements from the dense population, we would reach a wrong conclusion that women were taller than men. The dense population is equivalent to the third codon position where a large number of substitutions were recorded. The sparse population is equivalent to the second codon position where only a few substitutions were observed. Overweighting a small and consequently unreliable sample and down-weighting a large and reliable sample seems undesirable. We have shown that the few observed substitutions at the second codon position do not have any magical phylogenetic resolving power hidden within—they are just substitutions on less constrained codons and represent a small sample that may mislead us to believe that women are taller than they really are. It is worth noting that substitutions at the third codon position, which are frequently down-weighted in phylogenetic analyses, have led to better parameter estimates (including the tree topology) than substitutions at the first and second codon positions (Yang 1996*b*).

Acknowledgments

Supported by an RGC grant from the Hong Kong government and a CRCG grant from the University of Hong Kong to X.X. I thank L. Choy for assistance, and J. Felsenstein, Y. Fu, H. Deng, D. M. Weinreich, B. T. Foley, I. Jakobsen, and members of the HKU Ecology and Evolutionary Genetics Group for discussion and comments. M. Dickman checked the grammar. N. Takahata and two anonymous reviewers offered helpful suggestions and corrected errors.

LITERATURE CITED

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

CLARKE, B. 1970. Selective constraints on amino-acid substitutions during the evolution of proteins. Nature **228**:159–160.

CUMMINGS, M., S. OTTO, and J. WAKELEY. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. **12**:814–822.

EPSTEIN, C. J. 1967. Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature **215**:355-359.

FELSENSTEIN, J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59**:139–147.

———. 1993. PHYLIP (phylogeny inference package). Version 3.5. Distributed by the author, Department of Genetics, University of Washington, Seattle.

GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. Science **185**:862–864.

GRAUR, D. 1985. Amino acid composition and the evolutionary rates of protein-coding genes. J. Mol. Evol. **22**:53–62.

HAIG, D., and L. D. HURST. 1991. A quantitative measure of error minimization in the genetic code. J. Mol. Evol. **33**: 412–417.

IRWIN, D. M., T. D. KOCHER, and A. C. WILSON. 1991. Evolution of the cytochrome b gene of mammals. J. Mol. Evol. **32**:128–144.

JANKE, A., X. XU, and U. ARNASON. 1997. The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. Proc. Natl. Acad. Sci. USA **94**: 1276–1281.

JOHNSON, N. L., S. KOTZ, and A. W. KEMP. 1992. Univariate discrete distributions. John Wiley & Sons, New York.

KIMURA, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**:275–276.

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

KYTE, J., and R. F. DOOLITTLE. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

MIYATA, T., S. MIYAZAWA, and T. YASUNAGA. 1979. Two types of amino acid substitution in protein evolution. J. Mol. Evol. **12**:219–236.

NEE, S., E. C. HOLMES, A. RAMBAUT, and P. H. HARVEY. 1996. Inferring population history from molecular phylogenies. Pp. 66–80 *in* P. H. HARVEY, A. J. L. BROWN, J. MAYNARD SMITH, and S. NEE, eds. New uses for new phylogenies. Oxford University Press, Oxford.

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

NOVACEK, M. J., A. R. WYSS, and M. MCKENNA. 1988. The major groups of eutherian mammals. Pp. 31–71 *in* M. J. BENTON, ed. The phylogeny and classification of the tetrapods. Vol. 2. Clarendon Press, Oxford.

OTTO, S. P., M. P. CUMMINGS, and J. WAKELEY. 1996. Inferring phylogenies from DNA sequence data: the effects of sampling. Pp. 103–115 *in* P. H. HARVEY, A. J. L. BROWN, J. MAYNARD SMITH, and S. NEE, eds. New uses for new phylogenies. Oxford University Press, Oxford.

SNEATH, P. H. A. 1966. Relations between chemical structure and biological activity. J. Theor. Biol. **12**:157–195.

SWOFFORD, D. L. 1993. Phylogenetic analysis using parsimony (PAUP). University of Illinois, Champaign.

XIA, X., M. S. HAFNER, and P. D. SUDMAN. 1996. On transition bias in mitochondrial genes of pocket gophers. J. Mol. Evol. **43**:32–40.

YANG, Z. 1996*a*. Among-site rate variation and its impact on phylogenetic analysis. TREE **11**:367–372.

———. 1996*b*. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. **42**:587–596.

———. 1996*c*. Phylogenetic analysis by maximum likelihood (PAML). Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park.

YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

ZUCKERKANDL, E., and L. PAULING. 1965. Evolutionary divergence and convergence in proteins. Pp. 97–166 *in* V. BRYSON and H. L. VOGEL, eds. Evolving genes and proteins. Academic Press, New York.

NAOYUKI TAKAHATA, reviewing editor

Accepted November 14, 1997