J Mol Evol (1996) 43:32–40

JOURNAL OF MOLECULAR EVOLUTION

On Transition Bias in Mitochondrial Genes of Pocket Gophers

Xuhua Xia,[1] Mark S. Hafner,[1,2] Philip D. Sudman[1,*]

[1] Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA
[2] Department of Zoology and Physiology, Louisiana State University, Baton Rouge, LA 70803, USA

Received: 29 November 1994 / Accepted: 5 January 1996

Abstract. The relative contribution of mutation and purifying selection to transition bias has not been quantitatively assessed in mitochondrial protein genes. The observed transition/transversion (*s/v*) ratio is $(\mu_s P_s)/(\mu_v P_v)$, where μ_s and μ_v denote mutation rate of transitions and transversions, respectively, and P_s and P_v denote fixation probabilities of transitions and transversions, respectively. Because selection against synonymous transitions can be assumed to be roughly equal to that against synonymous transversions, $P_s/P_v \approx 1$ at fourfold degenerate sites, so that the *s/v* ratio at fourfold degenerate sites is approximately μ_s/μ_v, which is a measure of mutational contribution to transition bias. Similarly, the *s/v* ratio at nondegenerate sites is also an estimate of μ_s/μ_v if we assume that selection against nonsynonymous transitions is roughly equal to that against nonsynonymous transversions. In two mitochondrial genes, cytochrome oxidase subunit I (COI) and cytochrome *b* (cyt-*b*) in pocket gophers, the *s/v* ratio is about two at nondegenerate and fourfold degenerate sites for both the COI and the cyt-*b* genes. This implies that mutation contribution to transition bias is relatively small. In contrast, the *s/v* ratio is much greater at twofold degenerate sites, being 48 for COI and 40 for cyt-*b*. Given that the μ_s/μ_v ratio is about 2, the P_s/P_v ratio at twofold degenerate sites must be on the order of 20 or greater. This suggests a great effect of purifying selection on transition bias in mitochondrial protein genes because transitions are synonymous and transversions are nonsynonymous at twofold degenerate sites in mammalian mitochondrial genes. We also found that nonsynonymous mutations at twofold degenerate sites are more neutral than nonsynonymous mutations at nondegenerate sites, and that the COI gene is subject to stronger purifying selection than is the cyt-*b* gene. A model is presented to integrate the effect of purifying selection, codon bias, DNA repair and GC content on *s/v* ratio of protein-coding genes.

Key words: Transition bias — Mitochondrial protein gene — Purifying selection — Molecular evolution

Introduction

Transition bias in nucleotide substitution is a ubiquitous phenomenon in animal mitochondrial DNA (mtDNA), having been reported in both vertebrate species (Brown and Simpson 1982; Brown et al. 1982; Aquadro and Greenberg 1983; Thomas and Bechkenbach 1989; Bechkenbach et al. 1990; Edwards and Wilson 1990; Irwin et al. 1991) and invertebrate species (DeSalle et al. 1987; Satta et al. 1987; Thomas et al. 1989; Thomas and Wilson 1991). Although the phenomenon of transition bias is poorly understood, two contributing factors have been suggested. One is that the spontaneous mutation rate involving a transitional change is much greater than that involving a transversional change (Brown et al. 1982; Li et al. 1984; DeSalle et al. 1987; Beckenbach et al. 1990). The second is purifying selection (Li et al. 1985), and is applicable only to protein-coding genes. Purifying selection can affect the transition bias because (1) purifying selection tolerates synonymous mutations and

* *Present address:* Department of Biology, University of South Dakota, Vermilion, SD 57069, USA
Correspondence to: X. Xia

eliminates nonsynonymous mutations and (2) transitional mutations are more likely to be synonymous than transversional mutations.

The relative contribution of these two factors to the transition/transversion (*s*/*v*) ratio has not been studied in a quantitative way. We here summarize the joint effect of the two factors on the *s*/*v* ratio as follows:

$$\frac{s}{v} = \frac{\mu_s \cdot P_s}{\mu_v \cdot P_v} \quad (1)$$

where μ_s and μ_v are the mutation rate of transitions and transversions, respectively, and P_s and P_v are the fixation probability of a transitional mutation and a transversional mutation, respectively. Thus, transition bias can arise either from differential mutation pressure favoring transitions (i.e., a large μ_s/μ_v ratio) or from differential purifying selection against transversions, which would decrease P_v and consequently increase the P_s/P_v ratio.

At fourfold degenerate sites, both transitions and transversions are synonymous and may be assumed to be nearly neutral, with $P_{s_4} \approx P_{v_4}$, where the subscript 4 denotes fourfold degenerate sites. This leads to

$$\frac{s_4}{v_4} = \frac{\mu_s \cdot P_{s_4}}{\mu_v \cdot P_{v_4}} \approx \frac{\mu_s}{\mu_v} \quad (2)$$

Similarly, if we assume that purifying selection acts roughly equally against nonsynonymous transitions and nonsynonymous transversions, then $P_{s_0} \approx P_{v_0}$ (where the subscript 0 denotes nondegenerate sites), so that

$$\frac{s_0}{v_0} = \frac{\mu_s \cdot P_{s_0}}{\mu_v \cdot P_{v_0}} \approx \frac{\mu_s}{\mu_v} \quad (3)$$

Equations (2) and (3) state that the s_4/v_4 ratio and the s_0/v_0 ratio offer two independent estimates of the μ_s/μ_v ratio, which measures mutational contribution to transition bias. Thus, the s_4/v_4 ratio and the s_0/v_0 ratio are expected to be similar because they both reflect the same μ_s/μ_v ratio. An *s*/*v* ratio close to 1 at fourfold degenerate sites and at nondegenerate sites would suggest little mutational contribution to transition bias.

At twofold degenerate sites,

$$\frac{s_2}{v_2} = \frac{\mu_s \cdot P_{s_2}}{\mu_v \cdot P_{v_2}} \quad (4)$$

where the subscript 2 denotes twofold degenerate sites. Because transitions are synonymous and transversions are nonsynonymous at twofold degenerate sites in animal mitochondrial genes, P_{s_2} is expected to be larger than P_{v_2} under neutral theory (Kimura 1983), so the P_{s_2}/P_{v_2} ratio should be larger than one for functional genes. This P_{s_2}/P_{v_2} ratio can serve as a measure of the contribution of purifying selection to transition bias. The *s*/*v* ratio at twofold degenerate sites is expected to increase with increasing intensity of purifying selection. An *s*/*v* ratio at twofold degenerate sites similar to that at nondegenerate and fourfold degenerate sites suggests the absence of purifying selection.

The intensity of purifying selection against nonsynonymous substitutions can be assessed by the following three ratios:

$$\frac{P_{s_2}}{P_{v_2}} = \frac{s_2\,\mu_v}{v_2\,\mu_s} \quad (5)$$

$$\frac{P_{s_4}}{P_{s_0}} = \frac{s_4}{s_0} \quad (6)$$

$$\frac{P_{v_4}}{P_{v_0}} = \frac{v_4}{v_0} \quad (7)$$

These three ratios are expected to be the same if we assume that $P_{s_0} = P_{v_0} = P_{v_2}$ (i.e., all nonsynonymous mutations are subject to equally intense purifying selection and have the same probability of fixation regardless of whether they occur at nondegenerate or twofold degenerate sites), and $P_{s_2} = P_{s_4} = P_{v_4}$ (i.e., all synonymous mutations are nearly neutral and have the same probability of fixation regardless of whether they occur at twofold degenerate or fourfold degenerate sites). These assumptions have never been critically examined, although they are generally accepted as true when calculating the rate of synonymous and nonsynonymous substitutions (Li et al. 1985; Nei and Gojobori 1986; Li 1993).

It is possible for the first assumption ($P_{s_0} = P_{v_0} = P_{v_2}$) to be violated because, when a nonsynonymous mutation occurs, the original and the replacement amino acids could be very similar to each other in physical and chemical properties, or they could be very different. If nonsynonymous mutations at twofold degenerate sites tend to involve amino acid pairs that are more similar to (or more different from) each other than do nonsynonymous mutations at nondegenerate sites, then P_{v_2} would be larger (or smaller) than either P_{0_s} or P_{0_v}, so the P_{s_2}/P_{v_2} ratio would be smaller (or larger) than the other two ratios. This could bias estimates of the rate of synonymous and nonsynonymous substitutions.

The three ratios in equations (5–7) can be used to study differential purifying selection acting on different genes in the same genome. The three ratios can all be considered as measures of the strength of purifying selection, with stronger purifying selection being correlated with larger ratios. Because previous studies have shown that the amino acid sequence of cytochrome oxidase subunit I (COI) is more conservative than the cytochrome *b* (cyt-*b*) gene (reviewed by Brown 1985), we are interested in testing whether the *s*/*v* ratio at twofold degenerate sites is larger for the COI gene than for the cyt-*b* gene.

There are at least two more reasons for a detailed study of the relative contribution of mutation and purifying selection to transition bias. First, if purifying selection is a dominant factor shaping the rate of nucleotide substitution, then about 72% of the nucleotide sites (i.e., the proportion of nonsynonymous sites) are constrained. Such a large proportion of constrained sites would bring into question the concept of the molecular clock, because the presence of such a clock would now depend largely on the constancy of purifying selection. At present, there is little evidence that purifying selection is constant over geological time.

An understanding of the relative contribution of mutation and purifying selection to transition bias would also help us to choose a phylogenetic method for systematic analysis. For example, certain computer programs such as DNAML and DNADIST in the PHYLIP package (Felsenstein 1993) include a correction for transition bias by allowing the user to input a single *s/v* ratio. Such implementation would be adequate of mutation is the dominant factor shaping the transition bias, but would be insufficient if there is strong purifying selection generating great heterogeneity in the *s/v* ratio at nondegenerate, twofold degenerate, and fourfold degenerate sites. For nuclear genes, this heterogeneity appears to be small, with *s/v* ratios equal to ~4 at twofold degenerate sites and ~2 at nondegenerate and fourfold degenerate sites (Li et al. 1985). How great the heterogeneity is in mitochondrial genes is unknown. Considering that the ratio of synonymous to nonsynonymous substitutions is much greater in mitochondrial genes than in nuclear genes (Thomas and Beckenbach 1989), we suspect that the effect of purifying selection is greater in mitochondrial genes than in nuclear genes, which would lead to greater heterogeneity in the *s/v* ratio at nondegenerate, twofold degenerate, and fourfold degenerate sites.

In this paper, we investigate the relative contribution of mutation and purifying selection to transition bias by using mtDNA sequence data obtained in our laboratory for the COI and cyt-*b* genes in 15 species of pocket gophers (Rodentia: Geomyidae). We ask the following questions: (1) Is the *s/v* ratio at nondegenerate sites similar to that at fourfold degenerate sites as expected? (2) Is the *s/v* ratio at twofold degenerate sites greater than that at fourfold degenerate and nondegenerate sites? (3) Are the three ratios in equations (5–7) similar to each other? (4) Are the three ratios larger for the COI gene than for the cyt-*b* gene? and (5) Has COI experienced stronger purifying selection than cyt-*b* in the evolution of the 15 species of pocket gophers?

Materials and Methods

We sequenced regions of mtDNA from two genes in 15 species of pocket gopher representing six genera. These sequences (379 bp of COI and 402 bp for cyt-*b*) have been deposited in GenBank with accession numbers of L32682–L32696 for COI, and L11900, L11902, L11906 (DeWalt et al. 1993), and L38465–L38476 for cyt-*b*. The method of amplifying and sequencing cyt-*b* and COI genes has been described in DeWalt et al. (1993) and Hafner et al. (1994), respectively. Voucher specimens are deposited in the Louisiana State University Museum of Natural Science (LSUMZ) or the New Mexico Museum of Natural History (NMMNH): *Orthogeomys underwoodi* (LSUMZ 29493), *O. hispidus* (LSUMZ 29231), *O. cavator* (LSUMZ 29253), *O. cherriei* (LSUMZ 29539), *O. heterodus* (LSUMZ 29501), *Geomys breviceps* (LSUMZ 33940), *G. personatus* (LSUMZ 29539), *G. bursarius halli* (LSUMZ 31463), *G. b. majusculus* (LSUMZ 31448), *Cratogeomys castanops* (LSUMZ 31455), *C. merriami* (LSUMZ 34343), *Pappogeomys bulleri* (LSUMZ 34338), *Zygogeomys trichopus* (LSUMZ 34340), *Thomomys bottae* (LSUMZ 29320 and 29569), and *T. talpoides* (NMMNH 1634 and 1637).

Published estimates of the *s/v* ratio are typically based on pair-wide comparisons. For example, with 15 DNA sequences from pocket gopher species, we could make 105 pair-wise comparisons and report the average. However, there are two major disadvantages with this type of pair-wise comparisons in studying transition bias. First, the estimates are not statistically independent. For example, if there is one species that has recently experienced a large number of transitions and few transversions, then all 14 pair-wise comparisons between this species and the other 14 species will each contribute one data point with a large transition bias. Second, one has to assume that the nondegenerate, twofold degenerate, and fourfold degenerate sites have remained nondegenerate, twofold degenerate, and fourfold degenerate throughout the entire evolutionary history of the species studied. This is a tenuous assumption because the three categories of sites could potentially change into each other (i.e., a nondegenerate site could become a twofold degenerate site, which in turn could become a fourfold degenerate site). One way to avoid these problems is to reconstruct ancestral states of DNA sequences and estimate the number of transitions and transversions between neighboring nodes on the phylogenetic tree.

For phylogeny reconstruction, we used PAUP (Swofford 1993) and DNAML in the PHYLIP package (Felsenstein 1993) to generate initial trees. For further refinement, we used the computer program CODEML in the package PAML (Yang 1995), which implements the codon-based maximum-likelihood method in Goldman and Yang (1994). CODEML is exceedingly slow and is used mainly for verifying subtree topology. The resulting tree (Fig. 1) has a maximum likelihood value greater than trees generated with PAUP and DNAML. CODEML was also used for estimating the parameter ω, which can be interpreted as a measure of the intensity of purifying selection when the number of substitutions due to fixation of favorable mutations accounts for a negligible fraction of total substitutions. The relationship between ω and the *V* parameter in Goldman and Yang (1994) is $\omega = 100/V$.

We used PAUP (Swofford 1993) to reconstruct ancestral states, using the tree topology in Fig. 1. Different options for character optimization (DELTRANS, ACCTRANS, and MINF) resulted in slightly different reconstructions of internal nodes. Recently, the maximum likelihood method has also been used in reconstruction of ancestral states, and the method has been implemented in BASEML in the PAML package (Yang 1995). The BASEML program evaluates posterior probability values for alternative parsimony reconstructions of ancestral states. For example, given the terminal states in Fig. 1, internal nodes 23 and 24 could both be T or C, both being parsimonious reconstructions. However, the reconstruction with both nodes being T has a higher posterior probability (0.319) than the reconstruction with both nodes being C (0.042). Only the reconstruction with the highest posterior probability is used for subsequent phylogeny-based analysis.

For each pair of neighboring nodes on the tree, the number of transitions and transversions were counted separately for nondegenerate, twofold degenerate, and fourfold degenerate sites according to the method in Li (1993). The maximum-likelihood estimate of the *s/v* ratio is

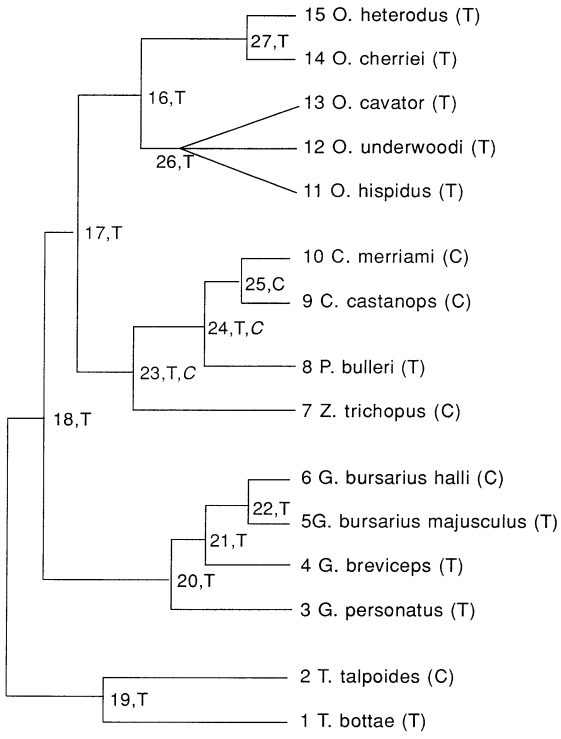

Fig. 1. Reconstructed phylogenetic consensus tree for the 15 species of pocket gophers based on COI and cyt-*b* sequences. The nucleotides (T and C) following each node illustrate two alternative reconstructions, one with internal nodes 23 and 24 being T, and the other with internal nodes 23 and 24 being *C* (*italics*). The first reconstruction has a higher posterior probability (0.319) than the second (0.042), and is preferred over the second reconstruction.

$$\frac{s}{v} = \frac{\sum_{i=1}^{N} s_i}{\sum_{i=1}^{N} v_i} \tag{8}$$

where N is the number of branches, and s_i and v_i are the estimated number of transitions and transversions, respectively, between two neighboring nodes of the ith branch. For example, the *s/v* ratio at twofold degenerate sites for the COI gene (data in Fig. 2) is

$$\frac{s}{v} = \frac{0.0000 + 0.0568 + 0.2413 + \ldots + 0.2063}{0.0000 + 0.0134 + 0.0213 + \ldots + 0.0000} = 48 \tag{9}$$

Results

Transition Bias at Nondegenerate, Twofold Degenerate, and Fourfold Degenerate Sites

The *s/v* ratio at nondegenerate sites is similar to that at fourfold degenerate sites for both COI and cyt-*b* genes (Table 1), which is consistent with expectations based on equations (2) and (3). That is, both are estimates of the same parameter, i.e., mutation bias (μ_s/μ_v). The *s/v* ratio at twofold degenerate sites is much greater (Table 1 and Fig. 2), suggesting that nucleotide substitution at the twofold degenerate sites is constrained by strong purifying selection against transversions, which are nonsynonymous at these sites.

Table 1. Calculation of *s/v* ratios for fourfold degenerate, twofold degenerate, and nondegenerate sites for the coding sequences of COI and cyt-*b* in pocket gopher mitochondrial DNA, based on equation (8)[a]

Genes	Site	Σs	Σv	*s/v*
COI	4-fold	2.627	2.045	1.285
	2-fold	3.670	0.076	48.101
	0-fold	0.042	0.019	2.230
cyt-*b*	4-fold	2.228	1.856	1.201
	2-fold	2.968	0.073	40.441
	0-fold	0.064	0.053	1.224

[a] Σs and Σv are the total number of transitions (s) and transversions (v) per site summed over the 26 branches (Fig. 1)

The high rate of nucleotide substitution at the fourfold degenerate sites for the COI and the cyt-*b* genes (Table 1) indicates substitutional saturation. Because substitutional saturation eventually leads to a reduction of available information for estimating the number of transitions and transversions, the *s/v* ratio for the fourfold degenerate sites based on all pair-wise comparisons between neighboring nodes may be a biased estimate. To obtain an unbiased estimate of the *s/v* ratio for the fourfold degenerate sites, we need data with negligible substitutional saturation (i.e., recently diverged taxa) and without mutual statistical dependence among data points. For this reason, we selected a subset of pair-wise comparisons involving more closely related species and their reconstructed ancestors (Table 2). From these we obtained an *s/v* ratio of 2.0 at fourfold degenerate sites for both COI and cyt-*b* genes (Table 2). These estimates are similar to those observed at the nondegenerate sites (Table 1), where little substitutional saturation should have occurred.

Our results show that the contribution of mutation (μ_s/μ_v) to the *s/v* ratio is relatively small, and clearly cannot explain the much larger *s/v* ratio observed at the twofold degenerate sites (Table 1 and Fig. 2). Given that the μ_s/μ_v ratio estimated from the fourfold degenerate and nondegenerate sites is about 2, the P_{s_2}/P_{v_2} ratio should be ~24 for COI and ~20 for cyt-*b* to account for the observed *s/v* ratio at the twofold degenerate sites (equation [4] and Table 1). Because transitions are synonymous and transversions nonsynonymous at twofold degenerate sites, the observed transition bias at twofold degenerate sites is attributable to purifying selection acting against amino acid substitutions.

Nonsynonymous Mutations at Twofold Degenerate Sites Are More Neutral Than Those at Nondegenerate Sites

The three ratios in equations (5–7) are estimates of the intensity of purifying selection. P_{s_2}/P_{v_2} has already been

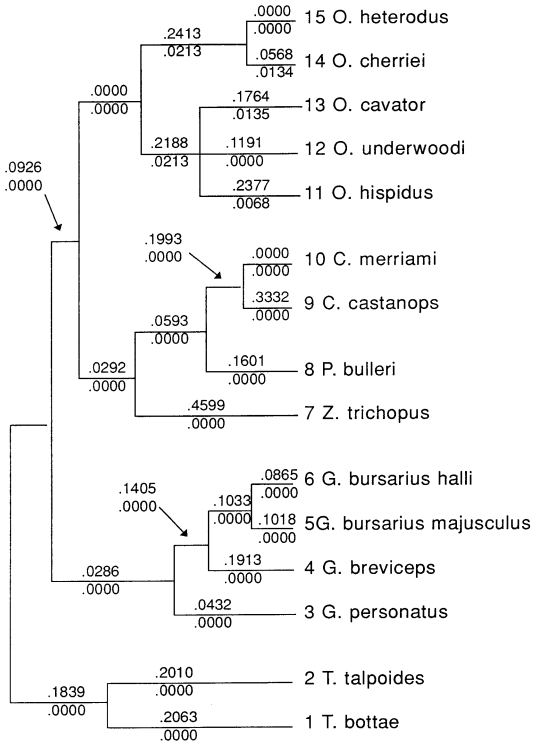

Fig. 2. The *numbers* above and below each branch are the estimated number of transitions and transversions, respectively, per site for twofold degenerate sites for the COI gene. Note that transitional substitutions have occurred in each branch, but transversions are rare. The equivalent diagram for the cyt-*b* gene shows a similar pattern, although the difference between the number of transitions and that of transversions is less dramatic.

Table 2. The number of transitional (*s*) and transversional (*v*) substitutions per site at fourfold degenerate sites, estimated with Kimura's two-parameter model, by using independent pair-wise comparisons between more closely related species and their reconstructed ancestors[a]

	COI		Cyt-*b*	
	s	*v*	*s*	*v*
G. b. majusculus—node 22	0.0526	0.0167	0.1772	0.0473
G. b. halli—node 22	0.0729	0.0513	0.0685	0.0484
C. castanops—node 25	0.2967	0.0518	0.0728	0.0640
C. merriami—node 25	0.0000	0.0167	0.0153	0.0842
O. underwoodi—node 26	0.0973	0.0903	0.1659	0.0500
O. hispidus—node 26	0.1410	0.1076	0.0909	0.0673
O. cavator—node 26	0.0708	0.0165	0.1356	0.0685
O. cherrieri—node 27	0.0895	0.0000	0.0500	0.0000
O. heterodus—node 27	0.0000	0.0691	0.0879	0.0000
Sum	0.8208	0.4200	0.8641	0.4297
s/*v* ratio	1.9543		2.0109	

[a] The *s*/*v* ratio = $\Sigma s/\Sigma v$, which is a maximum-likelihood estimate of the ratio

calculated as 24 for COI and 20 for cyt-*b*. The P_{s_4}/P_{s_0} and P_{v_4}/P_{v_0} ratios, however, are much larger (63 and 108 for COI and 35 and 35 for cyt-*b*). The average number of nondegenerate, twofold degenerate, and fourfold degenerate sites are 241, 74.5, and 62.5, respectively, for COI and 249, 85.5, and 64.5 for cyt-*b*. Thus, P_{s_2}/P_{v_2} can be shown to be significantly smaller ($P < 0.01$) than either P_{s_4}/P_{s_0} or P_{v_4}/P_{v_0} for both COI and cyt-*b* genes.

Note that the three ratios should be the same if $P_{s_0} = P_{v_0} = P_{v_2}$ and $P_{s_2} = P_{s_4} = P_{v_4}$. The fact the P_{s_2}/P_{v_2} is significantly smaller than either P_{s_4}/P_{s_0} or P_{v_4}/P_{v_0} suggests that at least one of these two assumptions must be wrong. Of the two assumptions, $P_{v_2} = P_{s_0} = P_{v_0}$ is obviously weaker. Whereas all synonymous mutations may be nearly neutral, nonsynonymous mutations can differ greatly in their effect on the function of the gene product. Some amino acids are very similar in physical and chemical properties, whereas others differ greatly from each other (Grantham 1974). Thus, nonsynonymous mutations involving similar amino acids are expected to be nearly neutral and to have a high fixation probability, whereas nonsynonymous mutations involving very different amino acids are expected to disrupt normal functioning of the protein and, therefore, to have a low fixation probability. If nonsynonymous substitutions tend to replace an amino acid with a similar one at twofold degenerate sites, but with a very different one at nondegenerate sites, then P_{v_2} would be greater than either P_{v_0} or P_{s_0}, resulting in a P_{s_2}/P_{v_2} ratio smaller than either P_{s_4}/P_{s_0} or P_{v_4}/P_{v_0}. We investigated this possibility in more detail.

Pair-wise amino acid dissimilarity can be measured by Grantham's distances (Grantham 1974), with amino acid dissimilarity increasing with distance. There are 54 nonsynonymous codon pairs that can mutate into each other through a single transition at a nondegenerate site (e.g., GAC-AAC), 104 nonsynonymous codon pairs that can mutate into each other through a single transversion at a nondegenerate site (e.g., GAC-GCC), and 32 nonsynonymous codon pairs that can mutate into each other through a single transversion at a twofold degenerate site (e.g., GAG-GAU). The mean Grantham distances for the three groups of codon pairs are 92.73, 79.04, and 57.00, respectively, which are significantly different ($P = 0.002$, Table 3). A least-significant-difference test (SAS Institute 1990, p. 222) showed that the mean Grantham's distance for nonsynonymous substitutions at twofold degenerate sites is significantly smaller than those for nonsynonymous transitions and transversions at nondegenerate sites (Table 3). We therefore conclude that nonsynonymous mutations at twofold degenerate sites tend to replace the original amino acid with a similar one, whereas nonsynonymous mutations at nondegenerate sites tend to replace the original amino acid with a relatively more different one. This suggests that P_{v_2} should be greater than either P_{v_0} or P_{s_0}, and explains why the observed P_{s_2}/P_{v_2} ratio is smaller than either P_{s_4}/P_{s_0} or P_{v_4}/P_{v_0}.

Transition Bias and the Strength of Purifying Selection

Previous studies have shown that the amino acid sequence of COI is more conserved evolutionarily than that

Table 3. One-way analysis of variance on Grantham's distances for nonsynonymous transversions at twofold degenerate sites (G-v_2, all transversions), and nonsynonymous transitions and transversions at nondegenerate sites (G-s_0 and G-v_0, respectively)[a]

(a) ANOVA Table

Source	DF	SS	MS	F	P
Between	2	3.232	1.616	6.52	0.0020
Within	187	46.370	0.248		
Total	189	49.61			

(b) LSD (T) pairwise comparisons of means:

Variable	Mean	Homogeneous groups	
G-v_0	92.73	I	
G-s_0	79.04	I	
G-v_2	57.00		I

[a] LSD is the least-significant-difference test for pair-wise comparison of means, with rejection level of 0.05. G-v_2 is significantly smaller than G-v_0 and G-s_0

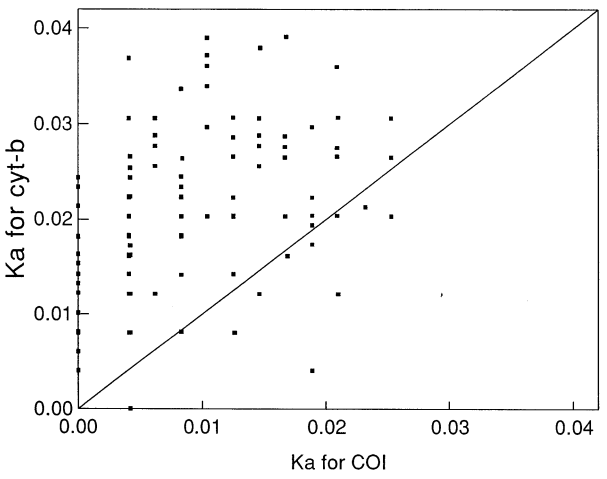

Fig. 3. Estimated number of nonsynonymous substitutions per site (*Ka*) between each pair of species for cyt-*b* plotted against that for COI. For 15 species, there are 105 pair-wise comparisons. *Ka* for COI and *Ka* for cyt-*b* are estimated for each pair. The *diagonal line* has a y-intercept of 0 and a slope of 1. Points are expected to distribute equally above and below this line if *Ka* for COI is equal to *Ka* for cyt-*b*.

of cyt-*b* (reviewed by Brown 1985). For example, the percent amino acid conservation for COI in pair-wise comparisons for human/mouse, human/cow, mouse/cow, and *Drosophila*/mouse is 90, 91, 93, and 75, respectively. The corresponding values for cyt-*b* are 78, 79, 84, and 67, respectively (Brown 1985). This indicates that purifying selection may be stronger for COI than for cyt-*b*. In our data, the transition bias at twofold degenerate sites is greater for COI than for cyt-*b* (48 vs 40 in Table 1). In addition, we have shown previously that P_{s_4}/P_{s_0} and P_{v_4}/P_{v_0} equal 63 and 108, respectively, for COI, but both ratios are much smaller for cyt-*b* (both are 35). This is consistent with the idea of stronger purifying selection acting on COI than on cyt-*b*.

The total number of nonsynonymous substitutions is also smaller for COI than for cyt-*b* (Fig. 3). If the two genes accumulate nonsynonymous substitutions at equal rates, then the points in Fig. 3 should scatter equally above and below the diagonal line, which has an intercept of 0 and a slope of 1. However, there are clearly more points above the diagonal (94 points) than below (11 points), reflecting more nonsynonymous substitutions for cyt-*b* than for COI.

One problem with the above argument is that the points in Fig. 3 are not statistically independent. For example, if there is just one species in which the rate of nonsynonymous substitutions at cyt-*b* is much higher than at COI, then all 14 pair-wise comparisons between this species and the other 14 species will each contribute one point above the diagonal. Thus, random fluctuations in substitution rate could be exaggerated and lead to rejection of a potentially correct null hypothesis of no difference in the rate of nonsynonymous substitution between COI and cyt-*b*.

Recently, a codon-based maximum-likelihood method has been proposed by Goldman and Yang (1994) that can be used to estimate the intensity of purifying selection if we accept the neutralist claim that positive selection has little effect on the rate of nucleotide substitution (Kimura 1983). The method is implemented in the program CODEML in the PAML package (Yang 1995), with the parameter ω measuring the intensity of purifying selection, assuming that the number of substitutions due to fixation of favorable mutations is negligible. Because the program is exceedingly slow when working with more than six terminal taxa, we grouped our 15 species into three subtrees and applied the method to each of the three subtrees separately. Subtree 1 contains the four *Geomys* species and subspecies plus *T. talpoides;* subtree 2 contains *T. bottae*, *P. bulleri*, *Z. trichopus,* and the two *Cratogeomys* species; and subtree 3 contains the five *Orthogeomys* species plus *T. talpoides.*

Table 4. Comparison of intensity of purifying selection between the COI gene and the cyt-*b* gene: intensity of purifying selection (ω) for COI is significantly greater than that for cyt-*b* (one-tailed test)

	COI		cyt-*b*			
Subtree	ω	SE	ω	SE	*z*	P
1	11.5525	0.3433	8.2828	1.0338	3.00	0.001
2	13.2895	1.5829	6.6553	0.7731	3.77	0.000
3	7.0554	0.8463	5.1910	0.6319	1.77	0.038

The estimated ω for COI (ω_{COI}) is significantly larger than that for cyt-*b* ($\omega_{cyt\text{-}b}$) in each of the three subtrees (Table 4). The difference between ω_{COI} and $\omega_{cyt\text{-}b}$ is tested by a method suggested by Z. Yang (pers. comm.). The CODEML program estimates ω and its SE, which is the square root of the variance of ω. The variance of parameters in CODEML is obtained by finding the inverse of the information matrix, whose elements are the expected values of the second derivatives of the log-

likelihood function with respect to the parameters. The variances estimated are, therefore, large-sample variances. To test the significance of the difference between ω_{COI} and $\omega_{cyt\text{-}b}$, we calculate

$$z = \frac{\omega_{COI} - \omega_{cyt-b}}{\sqrt{\text{var}_{\omega_{COI}} + \text{var}_{\omega_{cyt-b}}}} \quad (10)$$

and compare it with a normal critical value, such as 1.65 for a one-tailed test. In our study, all three z values are larger than 1.65 (Table 4).

It is unknown how large the sample size should be for the above test to be valid. However, we sampled randomly half of the codons and did repeated tests, and the results (not shown) suggest that the significance test above is robust. We conclude, therefore, that COI is subject to stronger purifying selection than cyt-*b* during the evolutionary history of the pocket gophers.

Discussion

Transition Bias

Our results suggest that transition bias in protein-coding sequences of mtDNA is mainly caused by strong purifying selection acting against transversions at twofold degenerate sites. However, differential mutation pressure also may have contributed to the transition bias. The *s/v* ratio at nondegenerate and fourfold degenerate sites is approximately 2 (Tables 1 and 2), which suggests a higher spontaneous rate of transitional mutations relative to transversional mutations in these species of pocket gophers. Whether our conclusions are generally applicable requires further empirical studies.

The *s/v* ratio at twofold degenerate sites in the two mitochondrial genes studied is much greater (48 for COI and 40 for cyt-*b*) than that reported for mammalian nuclear genes, where the *s/v* ratio at these sites is about 4 (Li et al. 1985). However, the *s/v* ratios at nondegenerate and fourfold degenerate sites for our mitochondrial genes (about 2) are comparable to those for nuclear genes (Li et al. 1985). This suggests that the mutational contribution to transition bias is similar between nuclear DNA and mtDNA and that the dramatic difference in the *s/v* ratio at the twofold degenerate sites between mitochondrial genes and nuclear genes is attributable to much stronger purifying selection against nonsynonymous mutations in mitochondrial DNA than in nuclear DNA. This is corroborated by our result—that the P_{s_4}/P_{s_0} and P_{v_4}/P_{v_0} ratios are also much greater for the mitochondrial genes (63 and 108 for COI, and 35 and 35 for cyt-*b*) than for the nuclear genes.

The dramatic heterogeneity in the *s/v* ratio among nondegenerate, twofold degenerate, and fourfold degenerate sites shown in both COI and cyt-*b* genes (Tables 1 and 2) argues strongly against the common practice of lumping all transitions and all transversions together to obtain an overall *s/v* ratio. Such a ratio is of little meaning, and it obscures important biological information. In fact, even the common practice of lumping synonymous and nonsynonymous substitutions at all sites to generate synonymous and nonsynonymous rates (e.g., Li et al., 1985; Nei and Gojobori 1986; Li 1993) is questionable, because nonsynonymous mutations at twofold degenerate sites tend to replace the original amino acid with a similar one, whereas nonsynonymous mutations at nondegenerate sites tend to replace the original amino acid with a relatively more different one (Table 3). For this reason, the rate of nonsynonymous substitution at twofold degenerate sites (all transversions) is expected, and documented (Table 1), to be higher than that at nondegenerate sites.

A substitutional process is characterized by the rate of transition and the rate of transversion, and it is evident that substitutional processes at nondegenerate, twofold degenerate, and fourfold degenerate sites differ. For example, the substitutional process at nondegenerate sites differs from that at twofold degenerate sites mainly in the rate of transitional substitution, and it differs from that at fourfold degenerate sites in the rate of both transitional and transversional substitutions. Finally, the process at twofold degenerate sites differs from that at fourfold degenerate sites in rate of transversional substitution. Thus, the overall process of nucleotide substitution in protein-coding sequences is much more complex than is assumed in current computer programs used for phylogenetic analysis. To reduce this heterogeneity in nucleotide substitution at nondegenerate, twofold degenerate, and fourfold degenerate sites, one should either use the amino-acid-based maximum likelihood method (e.g., Kishino et al. 1990) for phylogenetic reconstruction or use the codon-based maximum-likelihood method developed by Goldman and Yang (1994).

Other Factors That May Contribute to Transition Bias

Other factors, such as codon bias, DNA repair, and GC content, can also affect transition bias. Let us focus on the transition/transversion ratio at twofold degenerate sites, which is the source of most of the observed transition bias in our genes. Our explanation for the high transition bias in mtDNA at twofold degenerate sites is that transversional mutations at these sites are nonsynonymous and are reduced by purifying selection, whereas transitional mutations at these sites are synonymous and are not reduced by purifying selection. Thus, purifying selection operates differentially against transversional substitutions, resulting in a high transition bias.

How other factors affect *s/v* ratio can best be illustrated by a symbolic model. The number of transitional substitutions at twofold degenerate sites per generation is:

$$R_t = N(u_1 \cdot P_{t.1} + u_2 \cdot P_{t.2} \cdot 2 + \ . \ . \ . \ + u_n \cdot P_{t.n} \cdot n)$$
$$= N \sum_{i=1}^{n} u_i \cdot P_{t.i} \cdot i \qquad (11)$$

where N is effective population size, u_i is the mutation rate per generation of the sequence (gene) involving i transitions at different nucleotide sites of the gene, $P_{t.i}$ is the fixation probability of a mutation involving i transitions, and n is the number of twofold degenerate sites, which is the maximum number of transitional mutations a gene can accumulate in one generation.

The number of transversional substitutions at twofold degenerate sites per generation is:

$$R_v = N \sum_{i=1}^{n} v_i \cdot P_{v.i} \cdot i \qquad (12)$$

where v_i is the mutation rate per generation of the gene involving i transversions at different sites of the gene, and $P_{v\cdot i}$ is the fixation probability of a mutation involving i transversions. This gives an *s/v* ratio as:

$$\frac{R_t}{R_v} = \frac{\sum_{i=1}^{n} u_i \cdot P_{t.i} \cdot i}{\sum_{i=1}^{n} v_i \cdot P_{v.i} \cdot i} \qquad (13)$$

When synonymous mutations are neutral, transitional mutations at twofold degenerate sites (which are synonymous) will have $P_{t.i}$ equal to the fixation probability of a neutral mutation, whereas $P_{v.i}$ will decrease with increasing intensity of purifying selection. Thus, increasing purifying selection will result in decreasing $P_{v.i}$ *and* R_v, and consequently an increasing *s/v* ratio.

The effect of DNA repair on *s/v* ratio can be seen clearly through equations (11–13). We note that $P_{v.i}$ is expected to decrease with increasing i. For example, the effect of having many nonsynonymous transversions on the fitness of the mutant is expected to be more deleterious than that of having a single transversion. Consequently, $P_{v.i}$ in equations (11–13) should decrease with increasing i. However, $P_{t.i}$ will remain as the fixation probability of a neutral mutation regardless of what value i takes because transitions are synonymous at twofold degenerate sites. This implies that the ratio of $P_{t.i}/P_{v.i}$ will increase with increase i, leading to an increase in the R_t/R_v ratio. Thus, in the absence of DNA repair, i will be large, and the $P_{t.i}/P_{v.i}$ ratio, as well as the *s/v* ratio, is expected to be large. When there is DNA repair, i can take only small values, so that the ratio of $P_{t.i}/P_{v.i}$ decreases. As a consequence, R_t/R_v also decreases.

Understanding the effect of DNA repair on *s/v* ratio helps to answer a long-standing question concerning the difference in transition bias between mitochondrial genes and nuclear genes. In sharp contrast to the dramatic bias toward transitional substitutions in animal mtDNA, transitions are generally found only one-half to twice as often as transversions in interspecific comparisons of nuclear genes (Vogel and Kopun 1977; van Ooyen et al. 1979; Fitch 1980; Gojobori et al. 1982; Li et al. 1985). A major difference between mtDNA and nuclear DNA is that the latter has several enzymatic systems for DNA repair, whereas the former has none in itself, although mitochondria have a limited importation of repair enzymes from the cytoplasm (Myers et al. 1988; Satoh et al. 1988). For this reason, *s/v* ratio is expected to be higher in protein-coding genes of mtDNA than in those of nuclear genomes.

Codon bias can affect the *s/v* ratio because, with selection maintaining codon bias, synonymous substitutions are no longer neutral, and $P_{t.i}$ (and R_t) will decrease with the intensity of purifying selection for codon bias. This decreases the numerator in equation (13), and we should expect to see little transitional bias at twofold degenerate sites in genes experiencing strong purifying selection for codon bias. This suggests an alternative explanation for the difference in *s/v* ratio between nuclear DNA and mtDNA. One can explain the difference by supposing strong selection for codon bias in nuclear DNA and little selection for codon bias in mtDNA.

Selection for biased GC content can also affect *s/v* ratio. Suppose an extreme situation in which selection for GC pairs is so strong that the gene is made entirely of GC. Now if a G or C mutates into an A or T, the mutant will be eliminated by purifying selection. So the only visible and possible substitution is a G by a C or a C by a G, both being transversions. This implies that the *s/v* ratio will simply be 0 because transitions are selectively eliminated. Thus, genes with extreme bias in GC content are expected to exhibit a low *s/v* ratio.

In summary, our empirical study addresses only a small part of the theoretical framework on transition bias. However, many predictions stemming from our analysis are readily testable, and we hope that the tests of these predictions will eventually lead to a deeper understanding of the causal factors underlying the deceptively simple *s/v* ratio.

Acknowledgments. Part of the computation was done by using programs from W.H. Li and A. Zharkikh. We thank F.X. Villablanca, J.W. Demastes, T.A. Spradling, and R.D.M. Page for discussion and comments. Critical reviews from two referees and suggestions from E. Zuckerkandl greatly improved the content and clarify of the paper. This project is supported by NSF grant BSR-8817329 and NSF/LaSER (1992-96)-ADP-02.

References

Aquadro CF, Greenberg BD (1983) Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics 103:287–312

Bechkenbach AT, Thomas WK, Homayoun S (1990) Intraspecific sequence variation in the mitochondrial genome of rainbow trout (*Oncorhynchus mykiss*). Genome 33:13–15

Brown WM (1985) The mitochondrial genome of animals. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum Press, New York, p 95

Brown GG, Simpson MV (1982) Novel features of animal mtDNA evolution as shown by sequences of two rate cytochrome oxidase subunit II genes. Proc Natl Acad Sci USA 79:3246–3250

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: the tempo and mode of evolution. J Mol Evol 18:225–239

Curtis SE, Clegg MT (1984) Molecular evolution of chloroplast DNA sequences. Mol Biol Evol 1:291–301

DeSalle R Freedman T, Prager EM, Wilston AC (1987) Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila.* J Mol Evol 26:157–164

DeWalt TS, Sudman PD, Hafner MS, Davis SK (1993) Phylogenetic relationship of pocket gophers (*Cratogeomys* and *Pappogeomys*) based on mitochondrial DNA cytochrome b sequences. Mol Phylogen Evol 2:193–204

Edwards SV, Wilson AC (1990) Phylogenetically informative length polymorphisms and sequence variability in mitochondrial DNA of Australian songbirds (*Pomatostomus*). Genetics 126:695–711

Felsenstein J (1993) PHYLIP 3.5 (phylogeny inference package). Department of Genetics, University of Washington, Seattle

Fitch W (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin messenger RNAs. J Mol Evol 16:153–209

Gojobori T, Li WH, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Hafner MS, Sudman PD, Villablanca FX, Spradling TA, Demastes JW, Nadler SA (1994) Disparate rates of molecular evolution in cospeciating hosts and parasites. Science 265:1087–1090

Irwin DM, Kocher TD, Wilson AC (1991) Evolution of the cytochrome b gene of mammals. J Mol Evol 32:128–144

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31:151–160

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Li WH, Wu CI, Luo CH (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions and its evolutionary implications. J Mol Evol 21:58–71

Li WH, Wu CI, Luo CH (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Miyata T, Miyazawa S, Yasunaga T (1979) Two types of amino acid substitution in protein evolution. J Mol Evol 12:219–236

Myers KA, Raffhill R, O'Conner PJ (1988) Repair of alkylated purines in the hepatic DNA of mitochondria and nuclei in the rat. Carcinogenesis 9:285–292

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

SAS Institute (1994) SAS/STAT User's guide, version 6, 4th ed, vol 1. p 222

Satta Y, Ishiwa H, Chigusa SI (1987) Analysis of nucleotide substitutions of mitochondrial DNAs in *Drosophila melanogaster* and its sibling species. Mol Biol Evol 4:638–650

Satoh MS, Huh N, Rajewsky MF, Turoki T (1988) Enzymatic removal of O-ethylguanine from mitochondrial DNA in rat tissues exposed to N-ethyl-N-nitrosourea *in vivo.* J Biol Chem 263:6854–6856

Swofford DL (1993) Phylogenetic analysis using parsimony (PAUP), version 3.2. University of Illinois, Champaign

Thomas WK, Beckenbach AT (1989) Variation in Salmonid mitochondrial DNA: evolutionary constraints and mechanisms of substitution. J Mol Evol 29:233–245

Thomas WK, Wilson AC (1991) Mode and tempo of molecular evolution in the nematode *Caenorhabditis:* cytochrome oxidase II and calmodulin sequences. Genetics 128:269–279

Thomas WK, Maa J, Wilson AC (1989) Shifting constraints on tRNA genes during mitochondrial DNA evolution in animals. New Biol 1:93–100

van Ooyen A, van den Berg J, Mantel N, Weissmann C (1979) Comparison of total sequence of a cloned rabbit β-globin gene and its flanking regions with a homologous mouse sequence. Science 206: 337–344

Vogel F, Kopun M (1977) Higher frequencies of transitions among point mutations. J Mol Evol 9:159–180

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proc Natl Acad Sci USA 84:9054–9058

Yang Z (1994) Maximum likelihood phlogenetic estimation from DNA sequences with variable rates over sites: approximately methods. J Mol Evol 39:306–314

Yang Z (1995) Phylogenetic analysis by maximum likelihood (PAML), version 1.1. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University