Mutation Patterns of Mitochondrial H- and L-Strand DNA in Closely Related Cyprinid Fishes

Joseph P. Bielawski*,†,[1] and John R. Gold*

**Center for Biosystematics and Biodiversity, Texas A&M University, College Station, Texas 77843-2258 and*
†Department of Biology, University College London, London WC1E 6BT, United Kingdom

Manuscript received January 14, 2002
Accepted for publication April 19, 2002

ABSTRACT

Mitochondrial genome replication is asymmetric. Replication starts from the origin of heavy (H)-strand replication, displacing the parental H-strand as it proceeds along the molecule. The H-strand remains single stranded until light (L)-strand replication is initiated from a second origin of replication. It has been suggested that single-stranded H-strand DNA is more sensitive to mutational damage, giving rise to substitutional rate differences between the two strands and among genes in mammalian mitochondrial DNA. In this study, we analyzed sequences of the cytochrome *b*, ND4, ND4L, and COI genes of cyprinid fishes to investigate rates and patterns of nucleotide substitution in the mitochondrial genome. To test for strand-asymmetric mutation pressure, a likelihood-ratio test was developed and applied to the cyprinid sequences. Patterns of substitution and levels of strand-asymmetric mutation pressure were largely consistent with a mutation gradient between the H- and L-strand origins of replication. Significant strand bias was observed among rates of transitional substitution. However, biological interpretation of the direction and strength of strand asymmetry for specific classes of substitutions is problematic. The problem occurs because the rate of any single class of substitution inferred from one strand is actually a sum of rates on two strands. The validity of the likelihood-ratio test is not affected by this problem.

INDIVIDUAL strands of double-stranded mitochondrial (mt)DNA molecules are distinguished by their buoyant density in a cesium chloride gradient as heavy (H-strand) *vs.* light (L-strand). This difference is a function of uneven nucleotide content of the two strands; the H-strand is guanine rich, whereas the L-strand is guanine poor. The strongest strand-specific biases are found at fourfold degenerate sites (Perna and Kocher 1995), where patterns of variation most likely result from base substitution processes that are unaffected by natural selection. This led to the hypothesis that mutation pressure acting on mtDNA is strand specific (Jermiin *et al.* 1995; Perna and Kocher 1995; Reyes *et al.* 1998).

The mode of mitochondrial replication has been hypothesized to be responsible for the high mutation rate of mtDNA relative to that of the nuclear genome (Brown and Simpson 1982; Richter *et al.* 1988; Lindahl 1993) and to produce a strand-specific DNA mutation process (Tanaka and Ozawa 1994). Replication of the mitochondrial genome is an asymmetric process initiated at two different times from two different origins of replication (Clayton 1982). H-strand replication is initiated first from an origin within the mitochondrial control region (Ori_H), and as replication of a daughter H-strand proceeds, the parental H-strand is displaced. L-strand replication is initiated at its origin (Ori_L) only after the H-strand replication complex passes through Ori_L, ~11 kb downstream from Ori_H. A presumed consequence of this mode of replication is a period of time in which a portion of the parental H-strand is single stranded. The process of mtDNA replication is slow, taking as much as 2 hr to replicate the genome completely (Clayton 1992). During this time, H-strand DNA could be exposed to greater mutational damage via one or more of the following: (i) hydrolytic deamination of cytosine, (ii) hydrolytic deamination of adenine, and (iii) oxidation of guanine (Tanaka and Ozawa 1994). It is important to note that coupled leading- and lagging-strand DNA synthesis (unidirectional) also can be initiated at or near Ori_H under certain conditions; however, it is not yet clear how prominent this mode of replication might be for cells *in vivo* (Holt *et al.* 2000).

If mutational damage to single-stranded H-strand DNA is significant, substitution rates should differ between H-strand and L-strand mtDNA (Tanaka and Ozawa 1994). Furthermore, a mutational gradient might exist between sequences that remain in the single-stranded state for longer periods of time as compared to those that exist in the single-stranded state for shorter periods (Tanaka and Ozawa 1994). These hypotheses lead to two predictions: (i) substitution frequencies associated with hydrolytic deamination of either cytosine or adenine and/or oxidation of guanine should be greater on H-strand as compared to the L-strand mtDNA, and

[1]*Corresponding author:* Department of Biology, University College London, Darwin Bldg., Gower St., London WC1E 6BT, United Kingdom. E-mail: j.bielawski@ucl.ac.uk

(ii) rates of base pair substitution, levels of strand-specific mutation bias, and/or nucleotide composition bias should be dependent on the position of a sequence relative to the origins of H- and L-strand replication. The second prediction is based on the assumption that the position of a sequence relative to the two origins of replication serves as an indirect measure of the time that the sequence exists in the single-stranded state.

Tests of these predictions in various animal taxa have yielded conflicting results. Tanaka and Ozawa (1994) found evidence for both strand-specific substitution rates and compositional gradients in human mtDNA, and Bielawski and Gold (1996) found a pattern consistent with a gradient in rates of synonymous substitution in mtDNA of fishes of the cyprinid genus Notropis. Nedbal and Flynn (1998), alternatively, examined fourfold degenerate sites among closely related mammalian mitochondrial genomes but were unable to demonstrate a mutational gradient between Ori_H and Ori_L. In a study of 25 mtDNA genomes representing 10 mammalian orders, Reyes *et al.* (1998) found a significant relationship between location of a DNA sequence in the mitochondrial genome and its levels of strand-specific nucleotide bias and variability. Reyes *et al.* (1998) hypothesized that replication-related, hydrolytic deamination of cytosine and adenine could explain the origin of observed asymmetries in nucleotide composition. A study of more closely related pairs of mammalian taxa, however, found a relatively uniform rate of synonymous substitution over the mitochondrial genome (Pesole *et al.* 1999).

The objective of this study was to utilize a phylogenetic framework to investigate the hypothesis of Tanaka and Ozawa (1994) that the mode of mtDNA replication influences mutational processes of mtDNA sequences. This investigation focused on cyprinid fishes of the genus Notropis because previous studies (Bielawski and Gold 1996) indicated that synonymous substitution rates might differ among mtDNA sequences of these taxa and because the thermal habit of fishes (poikilothermy) is perceived to influence the evolution of mtDNA differently from the thermal habit of mammals (homeothermy; Rand 1994). Sequences were chosen to represent three locations within the mitochondrial genome: (i) close to Ori_H [the cytochrome *b* (cyt *b*) gene, ~1 kb from Ori_H]; (ii) intermediate between Ori_H and Ori_L (the ND4 gene and ND4L gene, ~5.5 kb from Ori_H); and (iii) close to Ori_L [the cytochrome oxidase I gene (COI), ~10 kb from Ori_H].

DNA sequences were used to test the following null hypotheses: (i) substitution rate of a gene is independent of distance from Ori_H, and (ii) substitution rates are homogeneous between H- and L-strands of mtDNA. Tests of the first null hypothesis were based on maximum-likelihood estimates of synonymous rates and rates at fourfold degenerate sites. To test the second null hypothesis, a likelihood-ratio test for strand asymmetry was developed and applied to three different classes of substitutions. Results are consistent with the hypothesis that the mode of mitochondrial genome replication influences the mutational process of cyprinid DNA, with the pattern of substitution differing among genes and among H- and L-strands of the genome.

MATERIALS AND METHODS

Taxon and character sampling: Taxa examined in this study included seven species of the North American cyprinid genus Notropis: six of the species (*Notropis amabilis, N. atherinoides, N. photogenis, N. scepticus, N. stilbius,* and *N. suttkusi*) are members of the subgenus Notropis, while one species (*N. potteri*) belongs to the subgenus Alburnops (Bielawski and Gold 2001). Details of specimen procurement and tissue storage are given in Bielawski and Gold (2001). Sequences included the entire cyt *b* gene (1140 bp), a 915-bp fragment (ND4-ND4L sequence) that included the entire ND4L gene (297 bp) and an adjacent 625 bp of the ND4 gene, and an 873-bp fragment of the COI gene. Note that individual sequences of ND4 and ND4L do not sum to 915 because the reading frames of these two genes overlap by 7 bp. The overlapping codons were excluded. Published cyt *b* sequences (Bielawski and Gold 2001) were obtained from GenBank (AF352266, AF352269, AF352272, AF352280, AF352283, AF352285, and AF352287). COI, ND4, and ND4L were sequenced as part of this study. All sequences are encoded on the L-strand. Note that chi-square tests of nucleotide frequencies in each gene indicated that they are at compositional equilibrium (data not shown).

DNA extraction, amplification, and sequencing: Whole fish were ground in liquid nitrogen, and DNA was obtained by phenol:chloroform extraction and ethanol precipitation (Sambrook *et al.* 1989). PCR amplification of cyt *b* sequences used primers described in Schmidt *et al.* (1998). PCR amplifications of the ND4-ND4L sequence used the universal primer NAP2 (Hogan *et al.* 1997) and primers ARGBL (CAAGACCC TTGATTTCGGCTCA), ND4LB (CAAAACCTTAATCTYCTA CAATGCT), and LEUAH (CAAGAGTTTCAGGCTCCTAAG AACA). PCR amplifications of COI used primers COIHA (CCTGAGAATAAGGGGAATCAG), COIHB (GGTTATGTGG CTGGCTTGAAA), COILA (CCTGCAGGAGGAGGAGACCC), and COILB (GCATTCCCACGAATAAATA). PCR thermal profile consisted of 35 to 45 cycles of 95° denaturation for 1 min, 48–50° annealing for 1 min, and 72° extension for 45 sec. Excess primers, nucleotides, and polymerase were removed from DNA amplification products using the Prep-A-Gene DNA purification system (Bio-Rad, Richmond, CA). Double-stranded DNA amplification products were sequenced directly with ABI PRISM (Perkin-Elmer, Norwalk, CT) dye-terminator cycle-sequencing kits and an Applied Biosystems (Perkin-Elmer) automated DNA sequencer. The sequence of a single individual per taxon was determined for all three mtDNA sequences from a minimum of two independent sequencing reactions for each primer. ND4, ND4L, and COI sequences were deposited in GenBank under accession nos. AY116183–AY116203. Sequence alignment was trivial; sequences were aligned by eye and required no indels.

Testing the mutational gradient between Ori_H and Ori_L: Data analysis focused on either synonymous sites or third positions of fourfold degenerate codons (*i.e.*, fourfold degenerate sites) because mutations at those sites are assumed to be unaffected by natural selection acting on amino acid sequence. The phylogenetic hypothesis derived from previous

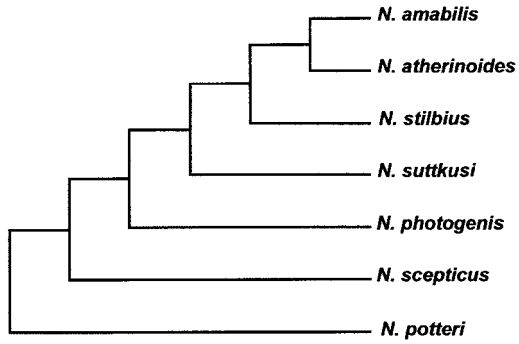

FIGURE 1.—Phylogenetic hypothesis for seven species of Notropis (BIELAWSKI and GOLD 2001). The tree topology (not including branch lengths) was used as a tree model for maximum-likelihood-based analysis of cyt *b*, ND4-ND4L, and COI sequence evolution.

analyses of mtDNA sequence data (Figure 1; BIELAWSKI and GOLD 2001) served as the topology assumed for maximum-likelihood analysis of cyt *b*, ND4-ND4L, and COI sequence evolution. Statistical tests were based on the following null hypotheses: (i) synonymous substitution rates are homogeneous among the three sequences, and (ii) substitution rates at fourfold degenerate sites are homogenous among the three sequences.

The mean number of synonymous substitutions per synonymous site was estimated by using the topology-dependent maximum-likelihood method of GOLDMAN and YANG (1994), as implemented in the computer program codeml of PAML (YANG 1997). The method employs a codon-based model for the evolution of protein-coding DNA sequences (GOLDMAN and YANG 1994). Synonymous substitution rates were estimated for each branch of the assumed topology by using the optimal model of codon evolution (see APPENDIX). The null hypothesis of homogenous synonymous rates among genes was tested by performing a repeated-measures ANOVA on the lengths of the individual branches of the topology. The absolute difference in the length of a particular branch was not expected to be equal among all branches but rather was expected to depend on the length of time represented by that branch. Therefore, all branch lengths were transformed using logarithm base 10 ($\log_{10}$) prior to performing the ANOVA.

The mean number of substitutions per fourfold degenerate site also was estimated for each branch of the assumed topology by using an optimal nucleotide-substitution model (see APPENDIX). Maximum-likelihood analyses were carried out using the baseml program of PAML (YANG 1997). The null hypothesis of homogenous rates at fourfold degenerate sites among genes was tested by performing a repeated-measures ANOVA on the lengths of the individual branches of the topology. All branch lengths were transformed using $\log_{10}$ prior to performing the ANOVA.

Repeated measures ANOVA, along with the posttest for linear trend, was performed by using GraphPad InStat v 3.01 (GraphPad Software, San Diego). In this case, the posttest determines whether the substitution rates increase (or decrease) systematically as the columns go from left to right. Analyses were conducted on both fourfold rates and synonymous rates to evaluate the robustness of results to these different approaches to estimating the silent rate.

DNA substitution models and testing asymmetry of the mutation process: Time-reversible models of DNA evolution place a restriction on the structure of the rate matrix, *Q*, such that all substitutions are "reversible"; *e.g.*, the expected number of $C \Rightarrow T$ substitutions is assumed to be equal to the expected number of $T \Rightarrow C$ substitutions on the same strand (*i.e.*, $\pi_T \times q_{TC} = \pi_C \times q_{CT}$). The most general form of the reversible models, the general time reversible (GTR) model, was introduced by YANG (1994a) and has the following form:

$$Q = \begin{bmatrix} — & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & — & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & — & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & — \end{bmatrix}.$$

Here, *a*–*f* are "rate parameters," the π_i's are the nucleotide frequency parameters, and the nucleotides are ordered T, C, A, G. For a more detailed discussion see YANG (1994a). It is easy to see that this model is reversible, as $\pi_i q_{ij} = \pi_j q_{ji}$. Although flexible, the GTR forces strand-asymmetric rates.

YANG (1994a) also introduced the most general form of a nucleotide substitution model without the restriction of reversibility. Because the restriction of reversibility leads to beneficial mathematical properties, studies of nucleotide evolution have been based predominantly on special cases of the GTR model rather than on the unrestricted model. Furthermore, the GTR model has 8 free parameters, whereas the unrestricted model has 11. However, the general unrestricted model provides the only framework for testing hypotheses of strand-asymmetric mutation pressure. Here we employ a special case of YANG's (1994a) unrestricted model that assumes symmetric substitution rates between the two strands of a DNA sequence. The rate matrix for this model has the following form:

$$Q = \begin{bmatrix} — & \beta & \chi & \varepsilon \\ \alpha & — & \delta & \phi \\ \chi & \varepsilon & — & \beta \\ \delta & \phi & \alpha & — \end{bmatrix}.$$

This model has five free parameters. Because this model is not necessarily reversible, likelihood calculations should be conducted on a rooted tree. The notion of a strand-symmetric model of evolution was first introduced by SUEOKA (1995).

The strand-symmetric model serves as the null model (M0) for likelihood-ratio tests (LRTs) of the hypothesis that substitution rates differ between the two strands of DNA. Let us take one set of substitutions, C to T on the H- and L-strands, as an example. First, note that a C to T transition on the H-strand ($C_H \Rightarrow T_H$) corresponds to a G to A transition on the L-strand ($G_L \Rightarrow A_L$), and a C to T transition on the L-strand ($C_L \Rightarrow T_L$) corresponds to a G to A transition on the H-strand ($G_H \Rightarrow A_H$). These substitutions are called "complementary" substitutions. If the DNA sequences are L-strand, complementary substitutions are modeled with the same parameter, α, under M0. This is true even if the rate of $C \Rightarrow T \neq G \Rightarrow A$. This is because the rate of, say, $C_L \Rightarrow T_L$ is an average over all sites, some of which had $C_L \Rightarrow T_L$ changes, and others had $G_H \Rightarrow A_H$ changes. Because of this averaging, rates of $C \Rightarrow T$ and $G \Rightarrow A$ estimated from any one strand will be equal if each type of substitution occurs at the same rate in both strands of DNA. Only when substitution rates are strand asymmetric will the rates of complementary substitutions differ. Hence, the alternative hypothesis can be modeled by specifying a separate rate parameter for $G_L \Rightarrow A_L$ and $C_L \Rightarrow T_L$. A LRT of M0 against such an alternative model is compared to a χ^2 distribution with 1 d.f. If significant, substitution rates must differ between the two DNA strands.

Strand asymmetry consistent with deamination of cytosine was investigated in the following way. The deamination product of cytosine is uracil (LINDAHL 1993), and because uracil base pairs with adenine instead of guanine, replication of a

mutated site (cytosine to uracil) in H-strand DNA yields a pyrimidine transition on the H-strand ($C_H \Rightarrow T_H$) and a purine transition on the L-strand ($G_L \Rightarrow A_L$). Complementary substitutions are $G_H \Rightarrow A_H$ and $C_L \Rightarrow T_L$, respectively. Hence, if the rate of G ⇒ A is strand symmetric, but the rate of C ⇒ T is greater on the H-strand due to an increased rate of cytosine deamination (Tanaka and Ozawa 1994), complementary substitution rates will differ. Strand asymmetry was tested by a LRT comparing M0 with a model permitting independent rates for $C_L \Rightarrow T_L$ and $G_L \Rightarrow A_L$ (M1).

In a similar way, strand asymmetry consistent with deamination of adenine also was investigated. The deamination product of adenine is hypoxanthine (Lindahl 1993). Because hypoxanthine base pairs with cytosine rather than thymine, replication of a mutated site in H-strand DNA yields a purine transition on the H-strand ($A_H \Rightarrow G_H$) and a pyrimidine transition on the L-strand ($T_L \Rightarrow C_L$). Complementary substitutions are $T_H \Rightarrow C_H$ and $A_L \Rightarrow G_L$, respectively. Strand asymmetry was tested for this set of complementary substitutions by a LRT of M0 against a model permitting independent rates for $A_L \Rightarrow G_L$ and $T_L \Rightarrow C_L$ (M2).

The primary mutagenic interaction of oxygen radicals with DNA is formation of a hydroxyl radical adduct of guanine, 8-hydroxyguanine (Croteau and Bohr 1997). Because 8-hydroxyguanine preferentially base pairs with adenine rather than cytosine, replication of a mutated site in H-strand DNA will yield a G to T transversion on the H-strand ($G_H \Rightarrow T_H$) and a C to A transversion on the L-strand ($C_L \Rightarrow A_L$). Complementary substitutions are $C_H \Rightarrow A_H$ and $G_L \Rightarrow T_L$, respectively. Strand asymmetry was tested by a LRT of M0 against a model permitting independent rates for $G_L \Rightarrow T_L$ and $C_L \Rightarrow A_L$ (M3).

The above models were applied to the fourfold degenerate sites of cyt *b*, ND4-ND4L, and COI of the species of Notropis. For comparison, the same models also were applied to the primate ψη-globin pseudogene dataset of Miyamoto *et al.* (1987). The test statistics of the LRTs provided a measure of the magnitude of strand asymmetry. It is important to note that the test statistic is influenced by the length of the sequence. Comparisons among different LRTs performed on a single dataset are straightforward. However, comparisons among LRTs performed on different datasets can be problematic; generally it is best to restrict comparisons to datasets of similar size.

RESULTS

Testing the mutational gradient between Ori_H and Ori_L: Two different measures of silent substitution (synonymous rates and rates at fourfold degenerate sites) were used to test the expectation of a mutational gradient between Ori_H and Ori_L. The pattern expected to arise if the H-strand is preferentially subjected to decay while in the displaced single-strand state (cyt *b* > ND4-ND4L > COI) was observed in both measures (Figure 2). Repeated-measures ANOVA also revealed significant heterogeneity among genes in both synonymous rate ($F = 4.63$, $P = 0.022$) and the rate at fourfold degenerate sites ($F = 4.45$, $P = 0.025$). Finally, posttests for a linear trend were significant for both measures of substitution rate (synonymous rates, slope = −0.144, $P = 0.008$; fourfold rates, slope = −0.224, $P = 0.007$), suggesting an inverse relationship between distance from Ori_H and substitution rate.

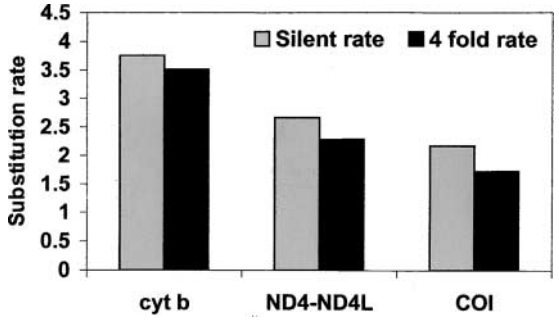

Figure 2.—Maximum-likelihood estimates of substitution rate at synonymous sites and at fourfold degenerate sites of cyt *b*, ND4-ND4L, and COI. The silent substitution rate is the mean number of synonymous substitutions per codon (d_S) estimated over the entire phylogeny presented in Figure 1. The fourfold substitution rate is the mean number of substitutions per fourfold site over the entire phylogeny (Figure 1). Rates were estimated by maximum likelihood under the optimal substitution models (see appendix). Both rates were estimated without clock constraints.

Testing strand asymmetry of the mutation process: L-strand sequences of cyt *b*, ND4-ND4L, and COI were analyzed under the null model of strand-symmetric substitution rates, M0, and three alternative models (M1, M2, and M3; Table 1). M1 permitted estimates of strand-asymmetric rates for complementary substitutions $G_L \Rightarrow A_L$ and $C_L \Rightarrow T_L$. LRTs of M0 against M1 indicated significant strand asymmetry in all three sets of sequences, *i.e.*, cyt *b*, ND4-DN4L, and COI (Table 1). M2 permitted estimates of strand-asymmetric rates for complementary substitutions $A_L \Rightarrow G_L$ and $T_L \Rightarrow C_L$. LRTs of M0 against M2 also were significant for all three sets of sequences, but evidence was not strong for COI as $P = 0.05$ (Table 1). M3 permitted estimates of strand-asymmetric rates for the complementary transversions $G_L \Rightarrow T_L$ and $C_L \Rightarrow A_L$. LRTs of M0 against M3 revealed a different pattern from LRTs of M0 against either M1 or M2 (Table 1). Strand symmetry was strongly rejected for ND4-ND4L, but not for cyt *b* and COI. This result suggests a unique pattern of evolution in ND4-ND4L.

Because all three sequences exhibited significant strand asymmetry for the complementary pairs $G_L \Rightarrow A_L/C_L \Rightarrow T_L$ and $A_L \Rightarrow G_L/T_L \Rightarrow C_L$, we estimated the ratio of their expected substitution rates across the assumed phylogeny (Table 2). Estimates were based on the unrestricted model of Yang (1994a). However, the expected number of substitutions can be high because a rate is high or because the frequency of the original nucleotide is high. Alternatively, the ratios of rate parameters (q_{ij}) are not dependent on the nucleotide frequencies. The ratio of either measure of complementary substitution rates will deviate from one when the substitution process is strand asymmetric. Consistent with the results of the LRTs, ratios of complementary substitution rates differed greatly from one (Table 2).

We also examined strand asymmetry in the ψη-globin

TABLE 1

Likelihood-ratio test statistics (2δ) for comparing models of strand-asymmetric substitution rates (M1, M2, and M3) with the strand-symmetric model (M0)

Models	NP	Cyt *b*	ND4-ND4L	COI	ψη-Globin pseudogene
M0: strand-symmetric model	5	($\ell = -1159.34$)	($\ell = -800.74$)	($\ell = -800.49$)	($\ell = -13805.23$)
M1: $G_L \Rightarrow A_L \neq C_L \Rightarrow T_L$	6	$\underline{2\delta = 71.0, P < 0.001}$	$\underline{2\delta = 57.5, P < 0.001}$	$\underline{2\delta = 34.2, P < 0.001}$	$2\delta = 0.05, P = 0.82$
M2: $T_L \Rightarrow C_L \neq A_L \Rightarrow G_L$	6	$\underline{2\delta = 29.5, P < 0.001}$	$\underline{2\delta = 19.8, P < 0.001}$	$\underline{2\delta = 3.91, P = 0.05}$	$2\delta = 0.61, P = 0.43$
M3: $C_L \Rightarrow A_L \neq G_L \Rightarrow T_L$	6	$2\delta = 0.47, P = 0.49$	$\underline{2\delta = 10.1, P = 0.001}$	$2\delta = 0.76, P = 0.38$	$2\delta = 1.66, P = 0.20$

NP is the number of parameters. Significant likelihood-ratio tests are underlined.

pseudogene. In contrast to the mitochondrial genes, there was no significant deviation for strand-symmetric substitution rates (Table 1). Moreover, complementary substitution rates estimated from the unrestricted model of Yang (1994a) were very similar, with ratios close to 1 (Table 2). Clearly, substitution patterns associated with these nuclear DNA sequences of mammals differ in a fundamental way from those found in mtDNA of closely related cyprinid fishes.

In addition to overall rates, the pattern of substitution at fourfold degenerate sites appeared to be related to distance from Ori_H. Both the magnitude of transitional strand asymmetry, as measured by 2δ (Table 1), and the amount of among-sites rate variation, as measured by the shape parameter (α) of the Gamma distribution (cyt *b*, $\alpha = 1.8$; ND4-ND4L, $\alpha = 2.3$; COI, $\alpha = \infty$), varied among the mitochondrial genes. The amount of transitional asymmetry appeared to decrease with increasing distance from Ori_H, while rates at sites appeared to become more homogenous with distance from Ori_H.

DISCUSSION

Rates measured at either synonymous sites or fourfold sites showed a statistically significant relationship with distance from Ori_H. Although very similar, estimates at fourfold sites were slightly lower than those at synonymous sites. This difference might be due to differences in modeling the processes of nucleotide and codon substitution. Specifically, nucleotide sites within a codon cannot evolve independently of one another, and nucleotide models assume independence among nucleotide sites, whereas codon models consider the probability of change from codon to codon. Regardless of method, patterns were consistent with the mutation-rate gradient hypothesis (Tanaka and Ozawa 1994), suggesting that sequence variation in mtDNA of cyprinids might be influenced by the mode of mitochondrial genome replication. However, our sample consisted of only a small portion of the coding sequences located between Ori_H and Ori_L, and differences in rates among the sampled sequences were not large. It is possible that rate variation among genes does not fit the mutation gradient hypothesis and that we have sampled three genes that fit the predicted pattern by chance. Additional analysis of complete mitochondrial genomes will be required to confirm a relationship between substitution rate and genome location. Future studies will need to focus on very closely related taxa, as more distantly related taxa often exhibit heterogeneity in equilibrium nucleotide frequencies and this can negatively effect estimates of substitution rates (*e.g.*, Kumar and Subramanian 2002).

In addition to overall rate, the pattern of DNA substitution was generally consistent with a gradient between Ori_H and Ori_L. In particular, levels of transitional asymmetry appeared to decrease as distance from Ori_H increases. Even if hypotheses about the specific mutational

TABLE 2

Ratio of the expected numbers of substitutions (d_{ij}) and rate parameters (q_{ij}) for complementary pairs of transitions

	d_{ij}[a]		q_{ij}[b]	
Genes	$G_L \Rightarrow A_L/C_L \Rightarrow T_L$	$T_L \Rightarrow C_L/A_L \Rightarrow G_L$	$G_L \Rightarrow A_L/C_L \Rightarrow T_L$	$T_L \Rightarrow C_L/A_L \Rightarrow G_L$
Cyt *b*	3.12	0.36	7.01	0.63
ND4-ND4L	2.32	0.43	8.98	0.51
COI	1.72	0.46	3.70	0.58
ψη-Globin pseudogene	1.12	1.01	1.06	0.96

[a] The expected numbers of substitutions (d_{ij}) were computed using the maximum-likelihood estimates of q_{ij}, π_i, and t under the unrestricted model of Yang (1994a).
[b] Rate parameters (q_{ij}) are from the unrestricted model of Yang (1994a).

mechanisms are incorrect, the pattern observed suggests a relationship between the presumed duration that the H-strand remains in a single stranded state and the process of mtDNA evolution in these cyprinids. However, sampling of the genome in this study was limited to only three regions, and an analysis of complete mitochondrial genomes will be needed to further investigate the apparent relationship between asymmetric substitution and genome location. Nevertheless, our current results, taken together with those of Tanaka and Ozawa (1994), suggest that direct damage to the mtDNA molecule, in addition to replication errors, might yield a substantial fraction of naturally occurring mtDNA mutations.

On the basis of laboratory assays (Frederico *et al.* 1990), the rate of hydrolytic deamination of cytosine to uracil may be as much as 200-fold higher in single-stranded DNA as compared with double-stranded DNA. Hence, hydrolytic deamination of cytosine is expected to produce higher rates of C ⇒ T transition on the H-strand as compared to the L-strand. Our analysis of complementary substitutions C ⇒ T and G ⇒ A indicated significant asymmetry in substitution rates on the mitochondrial H- and L-strands. Furthermore, the direction of strand bias is consistent with an increased sensitivity of H-strand DNA to spontaneous decay via hydrolytic deamination of cytosine, as the estimated rate of $G_L \Rightarrow A_L$ ($C_H \Rightarrow T_H$) was consistently higher than that of $C_L \Rightarrow T_L$. Similar inferences for mammalian mtDNA were made by Tanaka and Ozawa (1994) from an analysis of intraspecific patterns of nucleotide variation in humans and by Reyes *et al.* (1998) from an analysis of patterns of nucleotide composition in a wide variety of mammalian mitochondrial genomes.

Previous observations in mammalian mtDNA sequences (Tanaka and Ozawa 1994; Reyes *et al.* 1998) of decreases in guanine content and increases in adenine content of the H-strand relative to increasing distance from the origin of H-strand replication led to the hypothesis that displacement of the H-strand during mitochondrial genome replication also increased its sensitivity to spontaneous decay via hydrolytic deamination of adenine. The deamination product of adenine is hypoxanthine, and hypoxanthine base pairs with cytosine rather than thymine. Hence, hydrolytic deamination of adenine was predicted to produce higher rates of A ⇒ G transition on the H-strand. Although our analysis of the complementary substitutions A ⇒ G and T ⇒ C indicated significant asymmetry in substitution rates, the direction of the bias was not consistent with an elevated rate of A ⇒ G on the H-strand, as the estimated rate of $T_L \Rightarrow C_L$ ($A_H \Rightarrow G_H$) was consistently lower than that of $A_L \Rightarrow G_L$. Interestingly, it has been shown in laboratory assays that the rate of adenine deamination on single-stranded DNA is orders of magnitude lower than the rate of cytosine deamination (Lindahl 1993).

The mitochondrial electron transport chain is arguably the largest intracellular source of reactive oxygen species (Rand 1994). While there are hundreds of different oxidative lesions of DNA, the primary lesion is 8-hydroxyguanine, an oxidation product of guanine (Croteau and Bohr 1997). Because 8-hydroxyguanine preferentially base pairs with adenine rather than cytosine, replication of a mutated site yields a G ⇒ T transversion. If the H-strand has increased sensitivity to oxidation while in the single-stranded state, the rate of $C_L \Rightarrow A_L$ ($G_H \Rightarrow T_H$) should be higher than the rate of $G_L \Rightarrow T_L$. Our findings do not support the prediction of a consistently higher rate of G ⇒ T transversion on the mitochondrial H-strand of cyprinids. Although the genes and proteins involved in mtDNA repair are not yet well characterized, it is apparent from a diversity of studies that mitochondria have the capability to repair oxidative DNA damage (Croteau *et al.* 1999). One possible explanation of our findings is that the H-strand is more sensitive to oxidation, but repair mechanisms prevent or reduce strong rate asymmetry. It will be interesting to evaluate strand-asymmetric rates in mtDNA of mammals, where the importance of oxidative damage to guanine is suggested by observations of severalfold higher levels of 8-hydroyguanine in mtDNA as compared to nuclear DNA (Richter *et al.* 1988).

It is important to point out that for each of the examined sets of complementary substitutions, biological in-

terpretation of the ratio of complementary rates must be treated with caution. For example, the finding of strong asymmetries opposite of those expected under hydrolytic deamination of adenine should not be taken as strong evidence against its contribution to mitochondrial DNA evolution. Each rate in the ratio is actually the sum of two rates. For instance, the rate q_{TC} estimated from the L-strand is the sum of the following two rates: (i) q_{TC} on the L-strand ($q_{TC}^{[L]}$), and (ii) q_{AG} on the H-strand ($q_{AG}^{[H]}$). In the case of hydrolytic deamination of adenine, what we want to measure is $q_{AG}^{[H]}/q_{AG}^{[L]}$, but the ratio of complementary substitutions is actually $(q_{TC}^{[L]} + q_{AG}^{[H]})/(q_{AG}^{[L]} + q_{TC}^{[H]})$. The behavior of the ratio of complementary substitutions will be difficult to predict because different classes of substitution are likely to be affected differently by the processes of mutation and repair. Interpreting the LRT is not a problem, as it is based on the null hypothesis that $q_{TC}^{[L]} = q_{TC}^{[H]}$ and $q_{AG}^{[L]} = q_{AG}^{[H]}$. However, the LRTs do not provide a measure of the direction of strand asymmetry.

Because the approach used in this article is new, it was necessary to further evaluate its performance. For comparison, we analyzed sequences evolving under an independent evolutionary process, the primate ψη-globin pseudogene of the β-globin complex. Patterns of evolution in these sequences exhibited an interesting contrast to cyprinid mtDNA. Both LRTs and ratios of complementary substitution rates indicated a strand-symmetric substitution process. Interestingly, previous investigations of asymmetries in intergenic sequences of the β-globin complex generally support the notion that there are no mutational strand biases in those sequences (Bulmer 1991; Francino and Ochman 2000; but see Wu and Maeda 1987). This analysis illustrates that the present approach is not necessarily biased toward strand-asymmetric patterns of substitution.

Valuable discussions were contributed by Katherine A. Dunn, John Rice, Thomas F. Turner, and Ziheng Yang. We thank James N. Derr, Rodney L. Honeycutt, and Kirk O. Weinmiller for constructive comments on an early draft of this manuscript. We especially thank Diane Rowe for valuable assistance in a variety of areas. We are very grateful to Ziheng Yang for modifying the code of PAML to accept user-defined substitution matrices and for comments that substantially improved this manuscript. This article also was improved by the suggestions of two anonymous reviewers. Research was supported in part by a National Science Foundation doctoral dissertation improvement grant (DEB-9700717), in part by a Thomas Slick research fellowship (Texas A&M University), and in part by the Texas Agricultural Experimental Station under Project H-6703. J.P.B. was partially supported by a Biotechnology and Biological Sciences Research Council (United Kingdom) research grant (31/G10434). This article represents contribution no. 100 of the Center for Biosystematics and Biodiversity at Texas A&M University.

LITERATURE CITED

Bielawski, J. P., and J. R. Gold, 1996 Unequal synonymous substitution rates within and between two protein coding mitochondrial genes. Mol. Biol. Evol. **13:** 889–892.

Bielawski, J. P., and J. R. Gold, 2001 Phylogenetic relationships of cyprinid fishes in subgenus *Notropis* inferred from nucleotide sequences of the mitochondrially encoded cytochrome *b* gene. Copeia **2001:** 656–667.

Brown, G. G., and M. V. Simpson, 1982 Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. Proc. Natl. Acad. Sci. USA **79:** 3246–3250.

Bulmer, M., 1991 Strand symmetry of mutation rates in the beta-globin region. J. Mol. Evol. **33:** 305–310.

Clayton, D. A., 1982 Replication of animal mitochondrial DNA. Cell **28:** 693–705.

Croteau, D. L., and V. A. Bohr, 1997 Repair of oxidative damage to nuclear and mitochondrial DNA in mammalian cells. J. Biol. Chem. **272:** 25409–25412.

Croteau, D. L., R. H. Stierum and V. A. Bohr, 1999 Mitochondrial DNA repair pathways. Mutat. Res. **434:** 137–148.

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

Francino, M. P., and H. Ochman, 2000 Strand symmetry around the β-globin origin of replication in primates. Mol. Biol. Evol. **17:** 416–422.

Frederico, L. A., T. A. Kunkle and B. R. Shaw, 1990 A sensitive genetic assay for the detection of cytosine deamination: determination of rate constant and the activation energy. Biochemistry **29:** 2532–2537.

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

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 726–736.

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating the human-ape splitting by a molecular using clock mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Hogan, K. M., S. K. Davis and I. F. Greenbaum, 1997 Mitochondrial DNA analysis of the systematic relationships within the *Peromyscus maniculatus* species group. J. Mammal. **78:** 733–743.

Holt, I. J., H. E. Lorimer and H. T. Jacobs, 2000 Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. Cell **100:** 515–524.

Jermiin, L. S., D. Gaur and R. H. Crozier, 1995 Evidence from analyses of intergenic regions for strand specific directional mutation pressure in metazoan mitochondrial DNA. Mol. Biol. Evol. **12:** 558–563.

Kumar, S., and S. Subramanian, 2002 Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. USA **99:** 803–808.

Lindahl, T., 1993 Instability and decay of the primary structure of DNA. Nature **362:** 709–715.

Miyamoto, M. M., J. L. Slighton and M. Goodman, 1987 Phylogenetic relationships of humans and African apes from DNA sequences in the ψη-globin region. Science **238:** 369–373.

Nedbal, M. A., and J. J. Flynn, 1998 Do the combined effects of the asymmetric process of replication and DNA damage from oxygen radicals produce a mutation-rate signature in the mitochondrial genome? Mol. Biol. Evol. **15:** 219–223.

Perna, N. T., and T. D. Kocher, 1995 Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J. Mol. Evol. **41:** 353–358.

Pesole, G., C. Gissi, A. De Chirico and C. Saccone, 1999 Nucleotide substitution rate of mammalian mitochondrial genomes. J. Mol. Evol. **48:** 427–434.

Rand, D. M., 1994 Thermal habit, metabolic rate and the evolution of mitochondrial DNA. Trends Ecol. Evol. **9:** 125–131.

Reyes, A., C. Gissi, G. Pesole and C. Saccone, 1998 Asymmetric directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. **15:** 957–966.

Richter, C., J.-W. Park and B. N. Ames, 1988 Normal oxidative damage to mitochondrial and nuclear DNA is extensive. Proc. Natl. Acad. Sci. USA **85:** 6465–6467.

Sambrook, J., E. F. Fritsch and T. Maniatis, 1989 *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schmidt, T. R., J. P. Bielawski and J. R. Gold, 1998 Molecular phylogenetics and evolution of the cytochrome *b* gene in the cyprinid genus *Lythrurus* (Actinopterygii: Cypriniformes). Copeia **998:** 14–22.

Sueoka, N., 1995 Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. **40:** 318–325.

Tanaka, M., and T. Ozawa, 1994 Strand asymmetry in human mitochondrial mutations. Genomics **22:** 327–335.
Wu, C. I., and N. Maeda, 1987 Inequality in mutation rates of the 2 strands of DNA. Nature **327:** 169–170.
Yang, Z., 1994a Estimating the patterns of nucleotide substitution. J. Mol. Evol. **10:** 1396–1401.
Yang, Z., 1994b Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 306–314.
Yang, Z., 1997 PAML: a program package for phylogenetic analyses by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.
Yang, Z., N. Goldman and A. E. Friday, 1995 Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44:** 384–399.

Communicating editor: H. Ochman

APPENDIX

Testing models of DNA substitution: Below we evaluate codon models and nucleotide models (at fourfold sites) for the purpose of identifying the optimal sets of model parameters for estimating and comparing silent substitution rates in cyt *b*, ND4-ND4L, and COI of cyprinids. To accomplish this goal, the likelihood-ratio test was utilized to evaluate the contribution of parameters that differed between two nested models (Goldman 1993). The test statistic for the likelihood-ratio test (2δ) is a measure of the increase in likelihood provided by the more complex of the two models. The chi-square distribution was used to evaluate significance of the gain in likelihood associated with use of a more complex model (Yang *et al.* 1995). Note that testing in this fashion evaluates only the difference between a pair of parametric models; *i.e.*, inadequacies shared by any pair of models remained undetected.

Models of codon substitution: A hierarchy of codon substitution models was constructed on the basis of parameters accounting for transition/transversion bias, codon frequencies, and among-sites rate variation (Goldman and Yang 1994). Models permitting different rates of transition and transversion among codons employed the rate ratio κ. Models of codon frequencies were (i) equal codon frequencies (1/61) and (ii) empirical codon frequencies (F61). Models of among-sites rate variation were (i) equal rates among sites and (ii) variable rates among sites, modeled using a discrete approximation to the Gamma distribution (Gamma model; Yang 1994b).

LRTs (Table A1) indicated that a significant improvement in likelihood was obtained by accounting for transition bias. LRTs also indicated that codon frequencies were significantly biased. Interestingly, modeling codon usage by treating nucleotide frequencies at each codon position as evolving independently (F3 × 4; Goldman and Yang 1994) was not significantly better than assuming equal codon frequencies and was significantly worse than the F61 model (data not shown). Nucleotide frequencies in these sequences are not evolving independent of the context of the codon. Finally, permitting among-sites rate variation provides a significant improvement in the fit of the model to each gene (Table A1).

Nucleotide substitution at fourfold degenerate sites: Third codon positions of fourfold degenerate codons were sampled from each sequence (cyt *b*, 194 nt; ND4-ND4L, 147 nt; COI, 165 nt). A hierarchy of candidate DNA substitution models was constructed on the basis of three substitution matrices and two models of among-sites rate variation. Substitution matrices included the F81 matrix (Felsenstein 1981), the HKY85 matrix (Hasegawa *et al.* 1985), and the GTR matrix (Yang 1994a). Models of among-sites rate variation were (i) equal rates among sites and (ii) variable rates among sites, modeled using a discrete approximation to the Gamma distribution (Gamma model; Yang 1994b).

Likelihood-ratio tests (Table A2) revealed that the GTR substitution matrix (Yang 1994a) was optimal. Testing different models of among-sites rate variation (Table A2) indicated that the Gamma model was optimal for both cyt *b* and ND4-ND4L sequences, whereas the model of equal rates among sites was optimal for the COI sequences. Note that the equal-rates model is a special case of the Gamma model; the gamma model permits equal, or nearly equal, rates among sites when the value of the α parameter is large (Yang 1994b). Because the Gamma model can allow equal rates among sites for COI and because using the same model facilitates direct comparison of rates, we used the GTR + Gamma model to analyze fourfold degenerate sites of all three sets of sequences.

TABLE A1

Log-likelihood scores (ℓ) and test statistics (2δ) for likelihood-ratio tests of different models of codon substitution

Model				LRT with nested model		
κ	π_i's	ASRV	ℓ	2δ	d.f.	*P* value
			Cyt *b*			
1	1/61	None	−3372.81	None	None	None
ML est	1/61	None	−3250.11	245.4	1	<0.0001
ML est	Empirical	None	−3164.98	170.26	59	<0.0001
ML est	Empirical	Gamma	−3120.34	89.28	1	<0.0001
			ND4-ND4L			
1	1/61	None	−3016.35	None	None	None
ML est	1/61	None	−2848.94	334.82	1	<0.0001
ML est	Empirical	None	−2786.89	124.1	59	<0.0001
ML est	Empirical	Gamma	−2762.48	48.82	1	<0.0001
			COI			
1	1/61	None	−2356.77	None	None	None
ML est	1/61	None	−2224.63	264.28	1	<0.0001
ML est	Empirical	None	−2102.22	244.82	59	<0.0001
ML est	Empirical	Gamma	−2097.22	10	1	0.002

Note κ is the transition to transversion rate ratio and π_i's are the equilibrium codon frequencies. ASRV, among-sites rate variation.

TABLE A2

Log-likelihood scores (ℓ) and test statistics (2δ) for likelihood-ratio test of different models of nucleotide substitution

Model				LRT with nested model		
Substitution matrix	ASRV	NP	ℓ	2δ	d.f.	*P* value
			Cyt *b*			
F81	None	3	−1269.50	None	None	None
HKY85	None	4	−1148.54	241.92	1	<0.0001
GTR	None	8	−1127.66	42.76	4	<0.0001
GTR	Gamma	9	−1120.34	14.64	1	0.0001
			ND4-ND4L			
F81	None	3	−884.22	None	None	None
HKY85	None	4	−798.96	170.52	1	<0.0001
GTR	None	8	−770.00	57.92	4	<0.0001
GTR	Gamma	9	−768.01	3.98	1	0.0460
			COI			
F81	None	3	−911.68	None	None	None
HKY85	None	4	−800.08	223.2	1	<0.0001
GTR	None	8	−782.49	35.18	4	<0.0001
GTR	Gamma	9	−782.49	0	1	1

F81, the substitution matrix of Felsenstein (1981); HKY85, the substitution matrix of Hasegawa *et al.* (1985); GTR, the general time reversible model of Yang (1994a); ASRV, among-sites rate variation; NP, the number of parameters. Models were applied to the fourfold degenerate sites of cyt *b*, ND4-ND4L, and COI sequences.