A Maximum Likelihood Method for Analyzing Pseudogene Evolution: Implications for Silent Site Evolution in Humans and Rodents

Carlos D. Bustamante, Rasmus Nielsen,† and Daniel L. Hartl**

*Department of Organismic and Evolutionary Biology, Harvard University; and †Department of Biometrics, Cornell University

We present a new likelihood method for detecting constrained evolution at synonymous sites and other forms of nonneutral evolution in putative pseudogenes. The model is applicable whenever the DNA sequence is available from a protein-coding functional gene, a pseudogene derived from the protein-coding gene, and an orthologous functional copy of the gene. Two nested likelihood ratio tests are developed to test the hypotheses that (1) the putative pseudogene has equal rates of silent and replacement substitutions; and (2) the rate of synonymous substitution in the functional gene equals the rate of substitution in the pseudogene. The method is applied to a data set containing 74 human processed-pseudogene loci, 25 mouse processed-pseudogene loci, and 22 rat processed-pseudogene loci. Using the informatics resources of the Human Genome Project, we localized 67 of the human-pseudogene pairs in the genome and estimated the GC content of a large surrounding genomic region for each. We find that, for pseudogenes deposited in GC regions similar to those of their paralogs, the assumption of equal rates of silent and replacement site evolution in the pseudogene is upheld; in these cases, the rate of silent site evolution in the functional genes is ~70% the rate of evolution in the pseudogene. On the other hand, for pseudogenes located in genomic regions of much lower GC than their functional gene, we see a sharp increase in the rate of silent site substitutions, leading to a large rate of rejection for the pseudogene equality likelihood ratio test.

Introduction

It has long been held that pseudogenes provide the molecular evolutionist with an important tool for studying the rate and pattern of neutral evolution (Li, Gojobori, and Nei 1981). To the extent that pseudogenes evolve without selective constraint, the rate and pattern of substitutions in pseudogenes will faithfully reflect the underlying mutational process. Pseudogenes have, therefore, been used to infer the mutational process for nucleotide changes within species (Gojobori, Li, and Graur 1982; Li, Wu, and Luo 1984) and to compare this process between species (Petrov and Hartl 1999), as well as to study deletion rates among taxa (Graur, Shuali, and Li, 1989; Petrov, Lozovskaya, and Hartl 1996; Ophir and Graur 1997) and the effect that rates of DNA loss have on genome size (Petrov et al. 2000; Bensasson et al. 2001).

The question of whether silent sites are evolving at the neutral mutation rate is far from resolved. The analysis of polymorphism data and codon frequencies in *Escherichia coli* and *Salmonella enterica* suggest that there is considerable weak selection operating on silent sites (Sharp and Li 1986; Andersson and Kurland 1990; Hartl, Moriyama, and Sawyer 1994; Hartl et al. 2000). Recent work in the comparative analysis of *Drosophila* genes has also revealed evidence for constraint on synonymous site evolution, presumably because of codon bias (Akashi 1996, 1997; Akashi and Schaeffer 1997; McVean and Vieira 2001). In rodent genes, however, there seems to be no relationship between the rate of synonymous substitution and codon bias (Smith and Hurst 1999). It has been suggested that a balance between mutation and selection may account for genomic variation in GC content; this would affect both the apparent codon bias and rate of substitution at silent sites (for a review see Bernardi 2000).

Pseudogenes provide the molecular evolutionist with a direct opportunity to infer the strength of selection on changes at synonymous sites. Ophir et al. (1999) employed a distance method to analyze a set of 12 human and murid (rat and mouse) pseudogene gene trees; they found that, on average, murid and human third-position sites evolve at 40% the rate of pseudogene third-position sites. Because not all changes at third-position sites are synonymous, it is difficult to extrapolate from this result as to how much selection is acting on synonymous changes. The issue of selection on synonymous sites is of considerable practical importance in the study of molecular evolution because the ratio (ω) of the number of replacement substitutions per replacement site (dn) to the number of synonymous substitutions per synonymous site (ds) is widely used to detect adaptive evolution at the protein level.

The methods we present in this paper to study pseudogene evolution exploit recent statistical developments in molecular phylogenetics. In particular, we develop codon-based models for gene trees that contain a (processed) pseudogene, the functional gene from which the pseudogene was derived, and the ortholog of the functional gene from a closely related species (fig. 1). The models are implemented in a maximum likelihood framework and lead to two likelihood ratio tests of neutral evolution assuming a shared mutational process for all lineages—one to test the equality of silent and replacement substitution rates in a putative pseudogene, and another to test if the synonymous sites in the functional gene are evolving at the same rate as the pseu-

Abbreviations: BLRT, binomial likelihood ratio test; PELRT, pseudogene equality likelihood ratio test; SSELRT, silent site equality likelihood ratio test; WSRT, Wilcoxon signed rank test.

Key words: pseudogene evolution, synonymous site evolution, codon-based models, maximum likelihood, neutral evolution.

Address for correspondence and reprints: Daniel L. Hartl, 16 Divinity Avenue, Cambridge, Massachusetts 02138.
E-mail: dhartl@oeb.harvard.edu.

Mol. Biol. Evol. 19(1):110–117. 2002

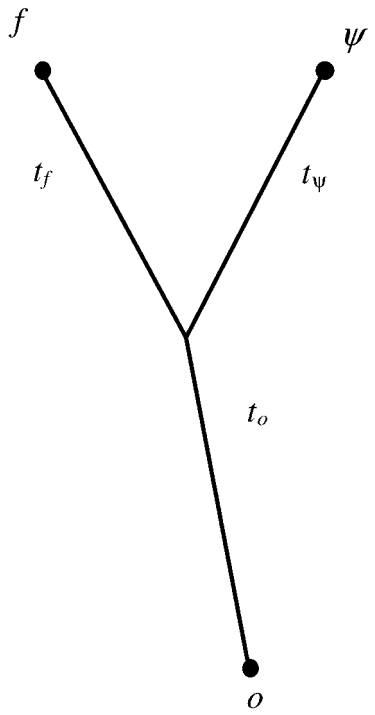

FIG. 1.—A graphical illustration of the model assumed in this paper. *f* is the codon sequence in the functional gene, ψ is the codon sequence in the pseudogene, *o* is the codon sequence in the outgroup and t_f, t_ψ, *and* t_o are the respective branch lengths of the lineages leading to these sequences.

dogene. We apply this method to a data set consisting of 121 processed pseudogenes (74 from human, 22 from rat, and 25 from mouse [Ophir and Graur 1997]) to test the assumption that processed pseudogenes evolve neutrally and to estimate the strength of selection on synonymous sites in the functional paralogs of pseudogenes.

Using the informatics resources of the Human Genome Project, we localized 67 of the gene-pseudogene pairs and estimated the GC content of the surrounding genomic areas for both. We used this data as well as the GC content of fourfold redundant sites (GC4) in the functional genes to assess whether rejection of pseudogene or silent site neutrality is correlated with the GC content.

It is important to note that the data sets presented here contain only retropseudogenes (i.e., pseudogenes generated by reverse transcription of mRNAs and insertion in the genome). These genes lack a promoter for expression and can, therefore, safely be termed ''dead on arrival.'' Furthermore, they generally insert far from their functional paralog and thus are expected to avoid the concerted evolution frequently observed for duplicated genes arranged in tandem.

Statistical Methods

The Model

The method we use to analyze pseudogene evolution is based on the likelihood models developed by Goldman and Yang (1994), Muse and Gaut (1994), Nielsen and Yang (1998), Yang (1998), and Yang and Nielsen (1998). In these models, the DNA sequence is treated as a sequence of triplets of nucleotides (codons). We will assume that the substitution processes in each codon site are independent and can be described by a continuous time Markov chain with state space on the 61 codons of the standard genetic code (excluding the stop codons). Furthermore, we assume that the process can be parameterized in terms of the transition-transversion rate ratio (κ), the *dn*/*ds* ratio (ω), and the stationary frequencies of each codon (π_i). The transition rate matrix $Q = \{q_{ij}\}$ is then defined as

$$q_{ij} = \begin{cases} 0, & \text{if codons } i \text{ and } j \text{ differ at more than one codon position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases} \quad (1)$$

The transition probabilities of the process can be calculated by exponentiating the rate matrix using standard numerical methods (e.g., numerical diagonalization of *Q*). The codon frequencies are estimated from the data using the method of moments to reduce the number of parameters in the model and to save computational time. The estimates are obtained from the observed base frequencies in the three codon positions (see Yang and Nielsen 1998 for details).

For the purpose of analyzing the pseudogenes, we assume a phylogenetic tree in which for each pseudogene there is a functional paralog from the same species and an ortholog of the functional gene from a related species (outgroup) (fig. 1). In the most general model, we will assume that the values of ω and the branch length of the three branches (t_f, t_o, and t_ψ) are independent parameters, and that the codon frequencies and κ do not vary among branches. The set of parameters that will be estimated by maximum likelihood is then $\Theta = \{t_f, t_o, t_\psi, \omega_f, \omega_o, \omega_\psi, \kappa\}$. The likelihood function to be optimized is proportional to

$$\Pr(S \mid \Theta) = \prod_{j=1}^{n} \sum_{i=1}^{61} \prod_{k=1}^{3} \pi_i P_\Theta(c_i \rightarrow S_{kj}), \quad (2)$$

where *n* is the number of codons in each sequence, *S* is the $3 \times n$ matrix of sequence data containing the codon in sequence *k* at position *j* in entry S_{kj}, and $P_\Theta(c_i \rightarrow S_{kj})$ is the transition probability along the appropriate branch of the phylogeny from codon c_i ($i = 1, 2, \ldots, 61$) in the internal node to codon S_{kj}. Parameter estimates are obtained by maximizing the logarithm of equation (2) with respect to Θ. In this model, branch lengths are not scaled by the expected number of substitutions. Hence, for a particular lineage in the phylogeny, the maximum likelihood estimate of the number of synonymous substitutions per synonymous site (*ds*) and the number of replacement substitutions per replacement site (*dn*) can be calculated from the maximum likelihood estimates of *t*, κ, and ω, using the methods described in Yang and Nielsen (1998).

Likelihood Ratio Tests

The first hypothesis we will test is whether the putative pseudogene has been an untranscribed pseudogene or a neutrally evolving gene in the entire evolution of the pseudogene lineage. If a putative pseudogene is truly a pseudogene, then ω_ψ ought to equal 1; that is, the rate of synonymous substitution should equal the rate of replacement substitution in the pseudogene. To do so, we compare the log maximum likelihood of the general model with seven parameters (t_f, t_o, t_ψ, ω_f, ω_o, ω_ψ, κ) to the log maximum likelihood of the nested six-parameter model (t_f, t_o, t_ψ, ω_f, ω_o, κ), which assumes that $\omega_\psi = 1$. Appealing to the usual large sample results for nested hypotheses (e.g., Stuart, Ord, and Arnold 1987, p. 246), two times the logarithm of the maximum likelihood ratio of the two hypotheses (LRT[6,7]) is asymptotically distributed as a χ^2 random variable with one degree of freedom (df). In particular, if the maximum log likelihood ratio under the seven-parameter model is more than 1.92 (3.84/2) log likelihood units larger than the maximum likelihood value under the six-parameter model, we reject the null hypothesis ($\omega_\psi = 1$) at the 5% significance level. Throughout the rest of the paper, we will refer to LRT[6,7] as the pseudogene equality likelihood ratio test (PELRT).

Once we have established that the rates of silent and replacement substitutions are statistically similar in the pseudogene, suggesting that the pseudogene rate of substitution reflects the underlying neutral mutation rate, we will wish to test whether silent sites in the functional gene evolve at the same rate as the pseudogene rate. To do so, we will exploit the genealogical relationship between the pseudogene, the functional gene, and the functional ortholog. If the pseudogene arose after the two species diverged, the pseudogene sequence and the functional ortholog will be more closely related to each other than either of them would be to the outgroup sequence. Therefore, if the rate of evolution of the pseudogene equals the rate of synonymous evolution in the functional gene, then $t_f = t_\psi$. To test this hypothesis, we will compare the maximum likelihood under the constrained model with six parameters $\{t_f, t_o, t_\psi, \omega_f, \omega_o, \kappa\}$ to the maximum likelihood under the nested model with five parameters $\{t_f = t_\psi, t_o, \omega_f, \omega_o, \kappa\}$. Again, significance is tested by comparing twice the log likelihood ratio, LRT[5,6], to a χ^2_1-distribution. Throughout the rest of this paper, we will refer to LRT[5,6] as the silent site equality likelihood ratio test (SSELRT).

Both these tests can be sensitive to the codon frequencies. Biased estimates of ω will be obtained if codon frequency biases are not properly accounted for (e.g., Yang and Nielsen 2000). This should cause little problem in the current context because the codon frequencies are similar. However, if the genomic context differs widely between pseudogenes and their functional paralog, the codon frequencies might be affected. In such cases, the result could potentially be an excess of significant results of the PELRT.

Pooled Likelihood Ratio Test

One issue of interest is whether we can reject the null hypotheses when combining the data across loci. If we assume that the genes are independent of one another, we can sum the log likelihoods under each model and perform the SSELRTs and the PELRTs on the pooled data. SSELRT applied to a combined data set would test whether all pseudogenes in the data set conform to $\omega_\psi = 1$. PELRT applied to the whole data set would test whether the average rate of silent site evolution in each gene is the same rate as in the respective pseudogene for all genes. As models 6 and 7 differ by one df, the PELRT for n genes will be distributed as χ^2_n. Likewise, the combined SSELRT statistic will be χ^2_{n-r} distributed, where r is the number of genes that reject the PELRT; the PELRTs are relevant because the SSELRT on a gene is justified only if the gene fails to reject the PELRT.

Binomial Likelihood Ratio Test

A separate (but not independent) analysis is based on the fraction of genes that reject a particular test. We perform this test to corroborate that rejection of the pooled test is not because of a handful of genes that have very large test statistics. If the null hypothesis, H_0, for n independent tests of the same hypothesis performed at the α significance level (e.g., 0.05) is true, the number of tests that reject the null hypothesis, k, is binomially distributed with parameters n and α. To test the hypothesis that the observed proportion of tests that reject the null hypothesis, k/n, is equal to α, we use a binomial likelihood ratio test (BLRT) which has a test statistic

$$\Lambda = -2\left[k \log\frac{\alpha}{p} + (n - k)\log\frac{1 - \alpha}{1 - p}\right] \qquad (3)$$

BLRT is approximately χ^2_1 if H_0 is true for all tests. A confidence interval for the true proportion of tests that reject H_0 is defined as the range of values of p for which one would not reject the BLRT.

Logistic Regression

To explore whether the GC content correlates with the rate of rejection of either the PELRT or the SSELRT, we used the logistic regression framework. This is an appropriate analysis to employ because our response variable is dichotomous (whether a particular test rejects the null hypothesis or not), and our predictor variable is continuous (GC content). We ran four separate logistic regressions using the genomic GC content or the GC content of fourfold redundant sites as explanatory variables, and whether or not the SSELRT or the PELRT rejects the null hypothesis as the outcome variables. We use the notation $\beta_{Genomic\ GC}$ and β_{GC4} to refer to the maximum likelihood estimates of the slope parameter for regressions using the genomic GC content and the GC4 content as predictor variables, respectively, and α to refer to the intercept parameter for each regression. The

P value reported for each regression is for a likelihood ratio test comparing the fit of the model with the maximum likelihood estimate of the slope parameter to a model with slope equal to 0 (Christensen 1997). The 95% confidence intervals for the predicted rate of rejection of each test as a function of GC content were found using nonparametric bootstrap sampling.

Bootstrap Test

To test whether the average number of replacement substitutions per replacement site in pseudogenes, $\overline{dn_\psi}$, equaled the average number of silent substitutions per silent site in pseudogenes, $\overline{ds_\psi}$, we used a paired bootstrapping procedure. Namely, we sampled (dn_ψ, ds_ψ) pairs as estimated from our data with replacement and calculated $\overline{dn_\psi}$ and $\overline{ds_\psi}$ for each sample of pseudogenes 100,000 times. The *P* value reported is twice the proportion of times $\overline{dn_\psi}$ was less than $\overline{ds_\psi}$ (because we were performing a two-sided test). Again, this test is not independent of the pooled likelihood ratio test.

Estimating Effective Level of Selection on Synonymous Sites

The difference in substitution rates at silent sites in functional genes and pseudogenes can be used to estimate *S*, the effective level of selection on mutations at silent sites; that is, the level of selection that would produce the observed reduction in substitution rate, assuming that all mutations at silent sites in functional genes have the same selective effect and that the mutation rates are the same for silent sites in functional genes and in pseudogenes. It can be shown (Kimura 1962, result [11]) that if mutations at silent sites have a selective effect *s* and that pseudogene sites evolve neutrally, the ratio of the rates of substitution at silent sites in functional genes, $\overline{ds_f}$, to the rate of substitution at pseudogene sites, $\overline{ds_\psi}$, is given by

$$\frac{\overline{ds_f}}{\overline{ds_\psi}} \approx \frac{S}{1 - e^{-S}} \qquad (4)$$

where $S = 4N_e s$ and N_e is the effective population size.

To generate confidence intervals for $\overline{ds_f}/\overline{ds_\psi}$, we used 100,000 nonparametric bootstrap samples generated by sampling with replacement (ds_f, ds_ψ) pairs estimated from our data. Using equation (4), we can transform the distribution of bootstrap estimates of $\overline{ds_f}/\overline{ds_\psi}$ into a distribution for estimates of *S*.

Materials and Methods

Data

For each pseudogene reported in Ophir and Graur (1997), we searched the nonredundant database (NR) of the NCBI server (http://www.ncbi.nlm.nih.gov/BLAST/) using the program (BLASTN) (Altschul et al. 1990) to find the closest available paralogous sequence from the same species. We then ran BLASTN on the paralog to find the single closest murid or human ortholog for the paralog-pseudogene pair (GenBank accession numbers available via the MBE website). Genes for which no functional ortholog could be found—as evidenced by the intercalation of phylogenetically incongruent sequences in the search results—were omitted (e.g., if a search using a functional human gene revealed a higher homology between the human gene and a crocodile gene rather than the human gene and the mouse gene, the pseudogene-ortholog pair was omitted).

Sequences were obtained, edited, and aligned by CLUSTALW using various versions (3.5–4.0.30) of the program DAMBE (Xia 2000). All stop codons and codons containing nucleotides with gaps in the alignment were removed and the reading frame was set to the reading frame of the functional gene. Pseudogenes that were identical to their functional paralog, except for gaps, were omitted from the analysis. Likewise, genes for which the estimated pseudogene, *t*, was longer than the outgroup, *t*, in model 7 were omitted because such a phylogeny indicated that the assumption of pseudogene emergence after the primate-rodent split was not met. We used BLASTN to identify the genomic position of 67 of the human pseudogene-functional gene pairs using the NCBI-curated sequences of the Human Genome Project. All genomic GC content measures are for the NCBI GenBank entry for the BAC (>100,000 bp) containing the functional gene.

Results and Discussion

In the Additional Information, which can be accessed via the *Molecular Biology and Evolution* web site, we summarize the maximum likelihood parameter estimates and the results of the individual likelihood ratio tests for each gene. The estimated number of silent substitutions per silent site between human and murid genes, as estimated from the full model, ranged from 0.143 to 2.427. Likewise, the estimated number of replacement substitutions per replacement site ranged from 0 to 0.447. The mean and standard deviation of these distances ($\overline{ds_f + ds_o} = 0.596 \pm 0.295$; $\overline{dn_f + dn_o} = 0.052 \pm 0.064$) are comparable to those previously cited for the human-mouse comparison (Matassi, Sharp, and Gautier 1999). The estimated distance between a pseudogene and its respective functional gene under the full model ranged from 0 to 0.684 silent substitutions per silent site ($\overline{ds_f + dn_\psi} = 0.128 \pm 0.106$) with 0–0.257 replacement substitutions per replacement site ($\overline{dn_f + dn_\psi} = 0.0645 \pm 0.0513$).

One key conclusion from our analysis is that, overall, the majority of pseudogene-gene pairs (>70%) do not reject either test. *This suggests that, as a first approximation, the most constrained model (i.e., equal codon frequencies and mutation rates in all lineages accompanied by neutral evolution in both pseudogenes and silent sites in functional genes) seems to fit the majority of the data.* We did find that the PELRT is significant at the 5% significance level for 13 of the 121 genes (11%). Among those genes for which the PELRT is not significant ($n = 108$), 25 genes (23%) have significant SSELRTs. Given the large number of tests, we would expect approximately 5% of the genes to reject either null hypothesis if the deviations were simply the

Table 1
Proportion of Genes that Reject Null Hypothesis for Pooled Data and Results of Pooled Likelihood Ratio Tests

Data Set	Likelihood Ratio Test[a]	Proportion of Tests that Reject H_0 (95% CI[b])	Pooled LRT (P value)
Human low–genomic GC.....	PELRT	1/34 = 0.0294 (0.0017, 0.1231)	33.414 (0.4962)
	SSELRT	5/33 = 0.152 (0.0571, 0.2978)	**57.092 (0.0057)**
Human high–genomic GC....	PELRT	7/33 = 0.2121 (0.0973, 0.3700)	**71.754 (0.0001)**
	SSELRT	7/26 = 0.2692 (0.1258, 0.4560)	**72.818 (2.6×10^{-6})**
Human low-GC4.....	PELRT	1/39 = 0.02564 (0.0015, 0.1081)	39.371 (0.4533)
	SSELRT	9/38 = 0.2368 (0.1216, 0.3864)	**83.888 (2.61×10^{-5})**
Human high-GC4.....	PELRT	8/35 = 0.2286 (0.1117, 0.3838)	**76.648 (6.06×10^{-5})**
	SSELRT	6/27 = 0.2222 (0.0949, 0.3999)	**75.338 (1.88×10^{-6})**
Rodent low-GC4.....	PELRT	1/22 = 0.0455 (0.0027, 0.1852)	20.286 (0.5651)
	SSELRT	3/21 = 0.1429 (0.0376, 0.3299)	**56.798 (3.85×10^{-5})**
Rodent high-GC4.....	PELRT	3/25 = 0.1200 (0.03129, 0.2836)	**48.602 (0.0032)**
	SSELRT	7/22 = 0.3182 (0.1513, 0.5252)	**53.630 (0.0002)**

[a] SSELRT = silent site equality likelihood ratio test; PELRT = pseudogene equality likelihood ratio test.
[b] CI = confidence intervals.

result of chance. However, for both tests the proportion of genes that reject the null hypothesis is significantly greater than 5% by the BLRT (PELRT: $P < 0.012$; SSELRT: $P \ll 0.0000001$). The source of these discrepancies is discussed in the following sections.

Implications of the Results for Pseudogene Evolution

To learn if the GC content correlates with an increase in the rejection rate of either test, we divided the data set into two equal groups based on the GC content of fourfold redundant sites in the functional paralogs of the pseudogenes (GC4). The rationale for using GC4 rather than overall GC content is that replacement sites are presumably constrained by purifying selection on amino acid changes. For 67 of the human genes, we were also able to estimate the GC content of the surrounding genomic region for the functional gene using the NCBI sequences of the Human Genome Project. Not surprisingly, GC4 was highly correlated with the surrounding genomic GC content ($r = 0.72$), and the conclusions are the same if one analyzes the flanking GC content of the functional gene rather than GC4.

In table 1, we report the rejection rates, 95% confidence intervals for the rejection rates, and the results of the pooled likelihood ratio tests for gene pairs grouped on the basis of the GC content of the functional gene for both the PELRT and the SSELRT. We find, by applying Fisher's Exact Test of homogeneity, that regardless of whether we classify humans genes on the basis of the regional GC content or GC4 of the functional gene, the processed pseudogenes derived from GC-rich genes reject the PELRT more frequently than do pseudogenes derived from GC-poor genes (Genomic GC, $P < 0.027$; GC4, $P < 0.011$). Likewise, a pooled likelihood ratio test implies that pseudogenes derived from GC-rich human genes, in general, have unequal rates of substitution at silent and replacement sites ($P < 0.0001$). This result is unexpected, but the same phenomenon is also observed in the rodent pseudogenes ($P < 0.0032$).

The inequality in substitution rates between silent and replacement sites is not observed in pseudogenes derived from GC-poor regions. From table 1, we see that the rejection rate of the PELRT is not significantly different from that expected by chance for human genes as classified by genomic GC (1/33; $P \approx 0.5766$), for human genes classified by GC4 (1/39; $P \approx 0.4430$), or for rodent genes (1/22; $P \approx 0.9209$). Furthermore, we do not reject the pooled PELRT for GC-poor genes in either humans or rodents ($P > 0.45$ for all; see table 1 for details). This tentatively suggests that in the cases where pseudogenes land in genomic regions similar to those in which their respective functional genes have evolved, the equal codon frequencies model and pseudogene neutrality is supported.

To investigate whether these results are because of our dichotomous classification of GC groups based on GC4, we used a logistic regression with GC4 in the functional gene as a continuous predictor variable and whether the pseudogene rejected or failed to reject

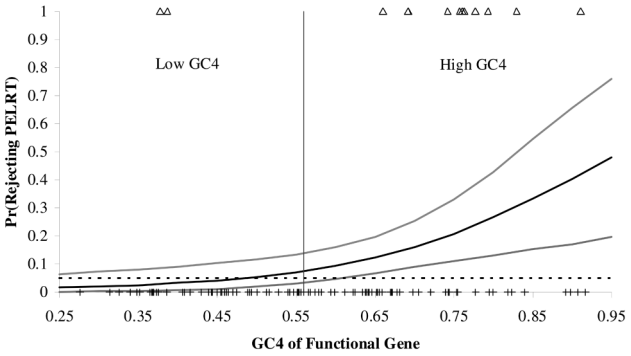

FIG. 2.—Probability of rejecting the PELRT as a function of GC4 content of the functional gene. The solid line is the predicted rate of rejection from the logistic regression analysis and the gray lines are the 95% confidence intervals. The locations of the genes that reject the test are denoted by open triangles, and the locations of the genes that do not reject the test are denoted by hatched marks. The dotted line traces out the expected 5% rejection rate.

PELRT as the outcome variable. As illustrated in figure 2, we found a strong positive association between GC4 in the functional gene and the probability of rejecting the PELRT ($\beta_{GC4} = 6.374$, $\alpha = -6.156$; $P < 0.0016$). However, for pseudogenes derived from genes with GC4 < 0.60, the 95% confidence intervals for the probability of rejecting the PELRT contain 5%. A logistic regression for human pseudogenes based on the local GC content yielded comparable results ($\beta_{Genomic\ GC} = 21.118$, $\alpha = -11.864$; $P < 0.0058$).

Why should pseudogenes derived from high-GC functional genes tend to have unequal rates of silent and replacement substitutions, whereas pseudogenes derived from low-GC functional genes conform to the expectations of neutral theory? One logical hypothesis is that selection for regional GC content (or differential mutation rates away from G or C) accounts for the difference in substitution rates at silent and replacement sites in processed pseudogenes derived from high-GC functional genes. Because mammalian functional genes tend to be nonrandomly distributed with respect to the GC content, such that a majority of genes are found in GC-rich regions, although a majority of the genome is composed of AT-rich regions (Mouchiroud et al. 1991; Saccone et al. 1997), if processed pseudogenes are randomly incorporated into the genome, the majority of pseudogenes will move to areas of lower GC content. Consequently, pseudogenes derived from genes in relatively GC-poor coding regions will usually land in genomic regions that are similar to the region in which their functional gene has evolved, whereas pseudogenes derived from GC-rich regions will usually land in regions of much lower GC content.

A prediction of this hypothesis is that, in the pseudogene lineage, the substitution rate per silent site should be elevated relative to the substitution rate per replacement site. In other words, the silent sites in the pseudogenes that move from regions of high GC content to regions of low GC content should be changing *faster* than the neutral rate. The reason that silent sites in pseudogenes are disproportionately affected is that it is the silent sites in *functional genes* that are most affected by local GC content because replacement sites are predominantly under purifying selection. Therefore, processed pseudogenes that move from an area of high GC to one of low GC have an excess of G or C at their silent sites, relative to their new genomic neighborhoods. This will cause silent sites to be more strongly affected by either mutation or selection pressure for GC content, leading to an excess of silent substitutions vis à vis replacement substitutions. The same phenomenon would hold for genes that move from low- to high-GC regions, but given the high abundance of AT-rich regions and the fact that most genes are in GC-rich regions, these events would occur infrequently.

Our data afford strong support for this hypothesis. For pseudogenes derived from high-GC genes, the average number of replacement substitutions per replacement site, $\overline{dn}_\psi$, relative to the average number of silent substitutions per silent site, $\overline{ds}_\psi$, is 0.0537/0.0935 = 0.5743; this is significantly different from unity by the bootstrap test ($P < 0.00001$). For pseudogenes derived from low-GC genes, $\overline{dn}_\psi = 0.0624$ and $\overline{ds}_\psi = 0.0619$, yielding a ratio of 1.008, which is extremely close to the expected ratio of 1.0 ($P \approx 0.9900$). Furthermore, a higher rate of silent substitution in pseudogenes derived from high-GC functional genes is evident from application of an unpaired *t*-test. The average replacement substitution rate does not vary between pseudogenes derived from high- and low-GC groups ($t = -1.027$; $P \approx 0.3062$), whereas the average silent substitution rate for pseudogenes derived from high-GC genes is significantly higher than the average rate for pseudogenes derived from low-GC genes ($t = 2.1808$; $P < 0.0312$). Together, these results suggest that silent and replacement sites in pseudogenes derived from low-GC genes, as well as replacement sites in genes derived from high-GC genes, are evolving at or near the neutral mutation rate, whereas silent sites in pseudogenes derived from high-GC genes are evolving at a rate faster than the neutral mutation rate.

In addition, pseudogenes derived from genes found in high-GC regions tend to decrease in GC4 content (Wilcoxon signed rank test [WSRT]: $Z = 3.561$, $P < 0.0005$) relative to their respective functional genes, whereas pseudogenes derived from low-GC regions tend to maintain the same level of GC4 content (WSRT: $Z = 0.950$, $P \approx 0.3422$). These findings corroborate the hypothesis that positive selection for GC content, or large differences in mutation rates among genomic regions, drives the observed difference in substitution rate between silent and replacement sites in a significant number of processed pseudogenes.

Implications of Results for Silent Site Evolution

We find that, for functional genes derived from both GC-rich and GC-poor regions, a significant number of pseudogenes evolve faster than silent sites in the genes from which they were derived. This comparison is statistically significant whether one considers the proportion of genes that reject the SSELRT or whether one

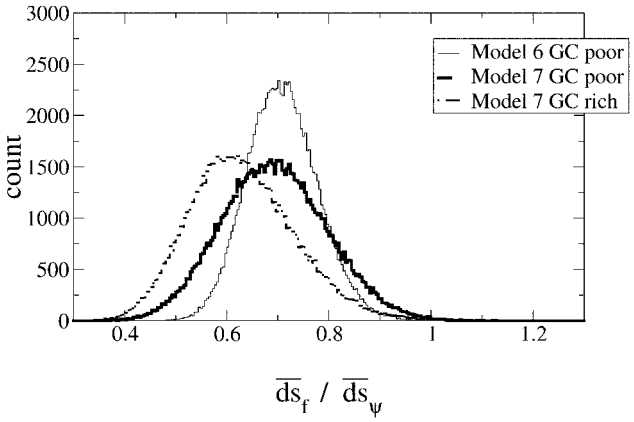

FIG. 3.—Distributions of $\overline{ds_f}/\overline{ds_\psi}$ estimated from 100,000 nonparametric bootstrap samples of estimates from model 6 (solid line) and model 7 (dotted line) from GC-poor gene-pseudogene pairs, and from model 7 (thick solid line) from the GC-rich pairs.

considers the pooled SSELRT across genes (table 1). For functional genes in GC-rich regions, the high rate of rejection is, in part, attributable to the excess of silent site fixation events along the pseudogene branch in the gene tree because of selection or mutation bias in the novel genomic region. This explanation does not apply to relatively GC-poor genes because we have shown that pseudogenes derived from these genes do not reject the PELRT more than expected by chance. In this class of genes, the substitution rates of silent and replacement sites in the pseudogenes are roughly equal. Furthermore, a logistic regression with GC4 of the functional gene as a predictor variable and whether the gene rejects the SSELRT as the outcome variable is not significant ($\beta = 1.471$, $\alpha = -2.040$; $P \approx 0.3434$), suggesting that GC4 alone cannot account for the overall level ($\sim$25%) of rejection for the SSELRT.

How does the detection of non-neutral evolution at silent sites in functional genes relate to the rate of substitution? To estimate the overall difference in the average rates of substitution between silent sites and pseudogene sites, we compare the ratio of the average rate of substitution per synonymous site in functional paralogs to the average rate of substitution per site in the pseudogenes ($\overline{ds_f}/\overline{ds_\psi}$). Figure 3 illustrates the distributions of 100,000 nonparametric bootstrap estimates of $\overline{ds_f}/\overline{ds_\psi}$ from model 6 (pseudogene neutrality model) and model 7 (free model) for gene-pseudogene pairs derived from GC-poor regions, and from model 7 only for gene-pseudogene pairs derived from GC-rich regions. We did not perform the bootstrap analysis for parameter estimates based on model 6 for gene-pseudogene pairs derived from GC-rich regions because this class of genes rejects model 6 as a group based on the pooled likelihood ratio test. Table 2 reports the mean and the 2.5% and 97.5% quantiles of these distributions. There is approximately a 30% reduction in the average substitution rate of silent sites in functional genes relative to silent sites in pseudogenes. We also note that the mean of the distribution of $\overline{ds_f}/\overline{ds_\psi}$ from GC-rich regions for model 7 (0.62) differs slightly from the mean of the distribution for models 6 and 7 for GC-poor regions (0.69–0.70), which reflects the accelerated substitution rate at silent sites in the pseudogene noted previously. The means of these distributions are all significantly different from unity because the interval between the 2.5% and 97.5% quantiles of these distributions does not contain 1 (table 2).

Using Kimura's result presented in the *Statistical Methods* section, one can transform the distributions of $\overline{ds_f}/\overline{ds_\psi}$ in figure 3 into a distribution for *S*. From table 2, we see that the means of these distributions are in the interval [−0.6753, −0.8814]. This suggests that the average strength of selection at synonymous sites is in the range where genetic drift and weak negative selection are important in determining the rate of substitution. However, if there is considerable variation within genes in selection intensity at silent sites, or if the mutation rates at silent sites differ considerably between pseudogenes and functional genes, then the above result underestimates the effect of selection. To estimate the effect of selection in these more complex scenarios, comparative data within and between species will be needed.

Acknowledgments

The authors thank H. Akashi and D. Petrov as well as two anonymous reviewers for comments on earlier drafts of this manuscript. This work was supported by a Howard Hughes Medical Institute award to C.D.B.

LITERATURE CITED

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

ANDERSSON, S. G. E., and C. G. KURLAND. 1990. Codon preferences in free-living microorganisms. Microbiol. Rev. **54**: 198–210.

AKASHI, H. 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster

Table 2
Average Ratio of Substitution Rates at Synonymous Sites in Functional Genes and in Pseudogenes and Corresponding values for 4*Ns*

Data Set (*n*)	$\overline{ds_f}/\overline{ds_\psi}$ (95% CI[a])	**4***Ns* (95% CI[a])
Model 6, GC-poor genes (61)	0.7071 (0.5801, 0.8523)	−0.6753 (−1.0055, −0.3115)
Model 7, GC-poor genes (61)	0.6908 (0.4985, 0.9010)	−0.699 (−1.2614, −0.2072)
Model 7, GC-rich genes (60).	0.6232 (0.4518, 0.8482)	−0.88144 (−1.4230, −0.3207)

[a] CI = confidence intervals.

rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics **144**:1297–1307.

———. 1997. Codon bias evolution in Drosophila. Population genetics of mutation-selection drift. Gene **205**:269–278.

AKASHI, H., and S. W. SCHAEFFER. 1997. Natural selection and the frequency distributions of ''silent'' DNA polymorphism in Drosophila. Genetics **146**:295–307.

BENSASSON, D., D. A. PETROV, D. ZHANG, D. L. HARTL, and G. M. HEWITT. 2001. Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol. Biol. Evol. **18**:246–253.

BERNARDI, B. 2000. The compositional evolution of vertebrate genomes. Gene **259**:31–43.

CHRISTENSEN, R. 1997. Log-linear model and logistic regression. Springer, New York.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

GOJOBORI, T., W. H. LI, and D. GRAUR. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18**:360–369.

GRAUR, D., Y. SHUALI, and W.-H. LI. 1989. Deletions in processed pseudogenes accumulate faster in murids than in humans. J. Mol. Evol. **28**:279–285.

HARTL, D. L., M. MORIYAMA, and S. A. SAWYER. 1994. Selection intensity for codon bias. Genetics **138**:227–234.

HARTL, D. L., E. F. BOYD, C. D. BUSTAMANTE, and S. A. SAWYER. 2000. The glean machine: what can we learn from DNA sequence polymorphism. Symp. Genomics Proteomics **4**:37–49.

KIMURA, M. 1962. On the probability of fixation of mutant genes in a population. Genetics **47**:713–719.

LI, W. H., T. GOJOBORI, and M. NEI. 1981. Pseudogenes as a paradigm of neutral evolution. Nature **292**:237–239.

LI, W. H., C. I. WU, and C. C. LUO. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21**:58–71.

MATASSI, G., P. M. SHARP, and C. GAUTIER. 1999. Chromosomal location effects on gene sequence evolution in mammals. Curr. Biol. **9**:786–791.

MCVEAN, G. A., and J. VIEIRA. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157**:245–257.

MOUCHIROUD, D., G. D'ONOFRIO, B. AISSANI, G. MACAYA, C. GAUTIER, and G. BERNARDI. 1991. The distribution of genes in the human genome. Gene **100**:181–187.

MUSE, S. V., and B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11**:715–724.

NIELSEN, R., and Z. YANG. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.

OPHIR, R., and D. GRAUR. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene **205**:191–202.

OPHIR, R., T. ITOH, D. GRAUR, and T. GOJOBORI. 1999. A simple method for estimating the intensity of purifying selection in protein-coding genes. Mol. Biol. Evol. **16**:49–53.

PETROV, D. A., E. LOZOVSKAYA, and D. L. HARTL. 1996. High intrinsic rate of DNA loss in Drosophila. Nature **384**:346–349.

PETROV, D. A., and D. L. HARTL. 1999. Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proc. Natl. Acad. Sci. USA **96**:1475–1479.

PETROV, D. A., T. A. SANGSTER, J. S. JOHNSTON, D. L. HARTL, and K. L. SHAW. 2000. Evidence for DNA loss as a determinant of genome size. Nature **287**:1060–1062.

SACCONE, S., S. CACCIO, P. PERANI, L. ANDREOZZI, A. RAPISARDA, S. MOTTA, and G. BERNARDI. 1997. Compositional mapping of mouse chromosomes and identification of the gene-rich regions. Chromosome Res. **5**:293–300.

SHARP, P. M., and W.-H. LI. 1986. An evolutionary perspective on synonymous codon usage in unicellular organims. J. Mol. Evol. **24**:28–38.

SMITH, N. G., and L. D. HURST. 1999. The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? Genetics **152**:661–673.

STUART, A., J. K. ORD, and S. ARNOLD. 1987. Kendall's advanced theory of statistics Vol. 2A: classical inference and the linear model. Sixth edition. Oxford University Press, New York.

XIA, X. 2000. DAMBE: data analysis in molecular biology and evolution. Department of Ecology and Biodiversity, University of Hong Kong.

YANG, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

YANG, Z., and R. NIELSEN. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46**:409–418.

———. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17**:32–43.

ADAM EYRE-WALKER, reviewing editor

Accepted September 27, 2001