Rates of Nucleotide Substitution and Mammalian Nuclear Gene Evolution: Approximate and Maximum-Likelihood Methods Lead to Different Conclusions

Joseph P. Bielawski, Katherine A. Dunn and Ziheng Yang

Department of Biology, University College London, London NW1 2HE, United Kingdom

Manuscript received December 9, 1999
Accepted for publication July 20, 2000

ABSTRACT

Rates and patterns of synonymous and nonsynonymous substitutions have important implications for the origin and maintenance of mammalian isochores and the effectiveness of selection at synonymous sites. Previous studies of mammalian nuclear genes largely employed approximate methods to estimate rates of nonsynonymous and synonymous substitutions. Because these methods did not account for major features of DNA sequence evolution such as transition/transversion rate bias and unequal codon usage, they might not have produced reliable results. To evaluate the impact of the estimation method, we analyzed a sample of 82 nuclear genes from the mammalian orders Artiodactyla, Primates, and Rodentia using both approximate and maximum-likelihood methods. Maximum-likelihood analysis indicated that synonymous substitution rates were positively correlated with GC content at the third codon positions, but independent of nonsynonymous substitution rates. Approximate methods, however, indicated that synonymous substitution rates were independent of GC content at the third codon positions, but were positively correlated with nonsynonymous rates. Failure to properly account for transition/transversion rate bias and unequal codon usage appears to have caused substantial biases in approximate estimates of substitution rates.

IT is well known that synonymous substitution rates vary among mammalian nuclear genes (Bernardi *et al.* 1993; Wolfe and Sharp 1993; Mouchiroud *et al.* 1995). Investigations of this variation, however, are complicated by nonuniform patterns of base composition among different regions of the mammalian genome. Mammalian genomes are structured into large regions (>300 kb) of distinct and homogeneous nucleotide composition known as isochores (Bernardi 1993). Both natural selection (Bernardi and Bernardi 1986; Gautier and Mouchiroud 1998; Eyre-Walker 1999) and mutation pressure (Filipski 1988; Wolfe and Sharp 1993; Francino and Ochman 1999) have been hypothesized as important mechanisms for the origin and maintenance of isochores. Consequently, the relationship between synonymous rate and nucleotide composition has been the subject of debate (*e.g.*, Bernardi *et al.* 1993).

Most studies report that genes with high GC content have lower silent substitution rates than genes with intermediate GC content (Filipski 1988; Ticher and Graur 1989; Wolfe *et al.* 1989; Wolfe and Sharp 1993; Eyre-Walker 1994). However, others (Miyata *et al.* 1982; Bernardi *et al.* 1993; Matassi *et al.* 1999) concluded that synonymous substitution rates are independent of nucleotide composition. Recently, Smith and Hurst (1999) analyzed a large sample of mouse and rat genes and found a significant positive correlation when maximum likelihood (ML) was used and no correlation when approximate methods were used. Smith and Hurst (1999) suggested that this methodological bias hindered further investigation of the relationship between synonymous rate variation and GC content.

A number of authors have reported that synonymous and nonsynonymous rates are positively correlated in mammalian genes (Graur 1985; Li *et al.* 1985; Wolfe and Sharp 1993; Mouchiroud *et al.* 1995; Ohta and Ina 1995; Makalowski and Boguski 1998; Smith and Hurst 1999). This observation, taken together with patterns of within-gene rate variation, recently led Alvarez-Valin *et al.* (1998) to hypothesize that selection is acting to enhance translational accuracy in mammals. However, this interpretation of the correlation between synonymous and nonsynonymous substitution rates also is controversial (Eyre-Walker 1991; Smith and Hurst 1999). Smith and Hurst (1999) hypothesized that selection for RNA structure and tandem substitutions, rather than translational accuracy, dominates the evolution of silent sites of rodent genes. Further investigations of selection at synonymous sites will require more reliable estimates of substitution rates.

To date, most studies of mammalian genes have employed approximate methods of estimating substitution rates. Although such studies intended to examine the effect of nucleotide content, their estimation proce-

Corresponding author: Joseph P. Bielawski, Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom. E-mail: j.bielawski@ucl.ac.uk

dures ignored unequal nucleotide frequencies. Most approximate methods also ignored the transition/transversion rate bias. Recent studies suggest that ignoring the transition/transversion rate bias or codon usage bias could lead to systematically biased estimates of substitution rates (Ina 1995; Yang and Nielsen 1998, 2000). A method that accounts for those features of DNA sequence evolution is ML. By employing a codon model of substitution, the ML method also uses probability theory to correct for multiple hits and weight evolutionary pathways between codons (Goldman and Yang 1994; Muse and Gaut 1994).

The objective of this study was to evaluate differences between ML and approximate methods and to evaluate their impacts on hypothesis testing. We compiled a sample of 82 homologous genes from three mammalian orders and estimated the rates of synonymous and nonsynonymous substitution for each gene using the ML method and two popular approximate methods (Nei and Gojobori 1986; Ina 1995). These data were used to evaluate the sensitivity of testing the following two null hypotheses: (i) the rate of synonymous substitution is independent of nucleotide composition, and (ii) the rate of synonymous substitution is independent of the rate of nonsynonymous substitution. ML analysis indicated that synonymous substitution rates were positively correlated with GC content at third codon positions but were independent of the nonsynonymous rate. Approximate methods, however, indicated opposite relationships, *i.e.*, synonymous substitution rates were independent of GC content at third codon positions but were positively correlated with the nonsynonymous rate. The differences were found to be due to the failure of approximate methods to properly account for the transition/transversion rate bias and unequal codon frequencies.

MATERIALS AND METHODS

Sequence data: We analyzed the aligned sequences of 82 nuclear genes from the mammalian orders Artiodactyla, Primates, and Rodentia. The data set is a composite of 49 genes analyzed by Ohta (1995) and 48 genes analyzed by Alvarez-Valin *et al.* (1998). The total number of genes in our analysis is 82 because 7 genes were used by both studies and because 8 genes were excluded due to regions of ambiguous alignment. Small differences between studies in number of codons analyzed are due to removal of initiation codons and minor adjustments to alignments.

Nucleotide composition and synonymous codon usage: G + C content at third codon positions (GC3) and codon usage bias, measured using the effective number of codons (ENC; Wright 1990), were calculated for each gene. ENC ranges from 20 to 61 with a smaller value indicating a greater bias. GC3 and ENC were computed using the program Codon W written by John Penden. Tests of compositional homogeneity among mammalian orders were conducted for each gene using chi-square tests of contingency tables of nucleotide counts.

Estimation of the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site: ML analysis was performed using the PAML package (Yang 1999). The models account for transition/transversion rate bias (κ) and codon usage bias (see Yang and Nielsen 1998 for details). We used two models to determine equilibrium codon frequencies. The first model used the nucleotide frequencies at the three positions of the codon and had $3 \times (4 - 1) = 9$ parameters. The second model used empirical estimates of 61 codon frequencies and had 60 parameters. Likelihood ratio tests comparing those two models (d.f. $= 60 - 9 = 51$) were significant for 81 of the 82 genes (data not shown). Analyses of substitution rates using both models were similar and hence only results obtained using empirical estimates of codon frequencies are presented.

Likelihood ratio tests of the assumption that the nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$) is homogeneous for all three mammalian lineages were performed by comparing two models of d_N/d_S ratios (Yang and Nielsen 1998). Model 0 assumed the same ratio (ω_0) for all three branches of the artiodactyl, primate, and rodent tree, whereas model 1 allowed independent d_N/d_S ratios (ω_A, ω_P, ω_R) for the three branches. Twice the log-likelihood difference under these two models was compared to a χ^2 distribution with d.f. = 2. This constitutes a likelihood ratio test of the strict neutral hypothesis. Model 1 also was used to obtain lineage-specific estimates of d_S and d_N for each gene.

Estimates of d_S and d_N also were computed pairwise between sequences using the approximate methods of Nei and Gojobori (1986) and Ina (1995). The PAML package (Yang 1999) was used to implement the method of Nei and Gojobori (1986) and Ina's program (dists1, available at ftp.nig.ac.jp) was used to implement method 1 of Ina (1995). To facilitate comparison of approximate and ML methods, we also estimated d_S and d_N in a pairwise fashion between the three orders of mammals using ML (Goldman and Yang 1994).

ML estimation can be performed under different model assumptions. We thus changed the models to investigate the effects of nucleotide (codon) frequencies and transition/transversion rate bias on estimation of d_S and d_N. If one compares a model in which κ is fixed to 1 (the rate of transition is set equal to the rate of transversion) to a model without such a constraint, the difference in d_S and d_N indicates the bias that arises from failure to account for the transition/transversion ratio. Likewise, if one compares a model in which codon frequencies are assumed to be equal (1/61) to a model where codon frequencies are free parameters, the difference in d_S and d_N indicates the bias that arises from failure to account for unequal codon usage.

RESULTS

Nucleotide (codon) usage bias and transition/transversion bias are common features of mammalian DNA sequence evolution: GC content at third codon positions (GC3) varied greatly among genes, ranging from 29 to 96%. Consistent with the suggestion that most mammalian genes are located in GC-rich isochores (Bernardi 1993), we observed that the majority of genes (60%) were GC rich (GC3 $>$ 60%) at third codon positions. Only a small proportion of genes (5%) were AT rich (AT3 $>$ 60%) at third codon positions. Mean values of GC3 were 65, 62, and 62% for artiodactyl, primate, and rodent genes, respectively.

Consistent with patterns of nucleotide bias, codon usage also varied greatly among genes, with ENC rang-

ing from small values indicating highly biased codon usage (*e.g.*, primate neurophysin 1 = 30.8) to large values indicating unbiased codon usage (*e.g.*, rodent transforming growth factor β1 = 60.4). Mean values of ENC were 46.8, 47.6, and 49.6 in artiodactyls, primates, and rodents, respectively. ML estimates of the transition/transversion rate ratio, κ, indicated that a transition bias was also present in all the sampled genes (Table 1). Collectively, these data show that transition/transversion bias and biased nucleotide (codon) frequencies are common features of DNA sequence evolution in mammalian genes.

Lineage-specific estimation of substitution rates by maximum likelihood: Results of ML analyses using model 0 (one d_N/d_S ratio) and model 1 (lineage-specific d_N/d_S ratios) are presented in Table 1. Using a likelihood ratio test, homogeneity of d_N/d_S ratio was rejected for 33 (40%) of the sampled genes (Table 1). Furthermore, there were 6 genes in the primate lineage (CD3 ε antigen, growth hormone receptor, insulin-like growth factor 1, interleukin 6 receptor, interleukin 7, osteopontin) and one gene in the artiodactyl lineage (interleukin 2) for which d_N/d_S ratios were >1.0. Because positive selection could adversely affect our investigation (Makalowski and Boguski 1998), gene and lineage combinations for which the d_N/d_S ratio was >1 were excluded from further analysis.

Values of d_N and d_S were estimated separately for the artiodactyl, primate, and rodent lineages using model 1 (Table 1). Estimates of d_S for these lineages were positively correlated (artiodactyl *vs.* primate, $r^2 = 0.1343$, $P = 0.0013$; artiodactyl *vs.* rodent, $r^2 = 0.2993$, $P < 0.0001$; primate *vs.* rodent, $r^2 = 0.2632$, $P < 0.0001$). Similarly, estimates of d_N were correlated between lineages (artiodactyl *vs.* primate, $r^2 = 0.5758$, $P < 0.0001$; artiodactyl *vs.* rodent, $r^2 = 0.6401$, $P < 0.0001$; primate *vs.* rodent, $r^2 = 0.5763$, $P < 0.0001$). These findings are consistent with previous reports that substitution rates were variable among genes, and genes with higher substitution rates in one lineage tended to have higher rates in other lineages as well (Bulmer *et al.* 1991; Mouchiroud *et al.* 1995).

Hypothesis testing using maximum-likelihood estimates of substitution rates: The null hypothesis that the rate of synonymous substitution is independent of nucleotide composition was evaluated by linear regression of lineage-specific estimates of d_S and GC3. There was a significant positive correlation between d_S and GC3, with $r^2 = 0.45$, 0.27, and 0.26 in artiodactyls, primates, and rodents, respectively. Because results were similar for all three lineages, only results for artiodactyl genes are presented in Figure 1.

Because nonstationary genes could have negative impacts on analyses of substitution rates (Lanave *et al.* 1984; Saccone *et al.* 1989; Mouchiroud and Gautier 1990), each gene was tested for homogeneity of nucleotide frequencies. Chi-square tests at third positions of the codon indicated significant heterogeneity among lineages in 27 (33%) of the genes (Table 1). Reanalysis of the subset of genes defined by homogeneity of nucleotide frequencies also yielded a significant positive relationship between d_S and GC3 (artiodactyls, $r^2 = 0.5053$, $P < 0.0001$; primates, $r^2 = 0.2351$, $P = 0.0004$; rodents, $r^2 = 0.4225$, $P < 0.0001$). This finding indicated that a positive correlation between d_S and GC3 was not a consequence of including genes that were nonstationary for nucleotide frequencies.

The null hypothesis that synonymous and nonsynonymous substitution rates are independent was evaluated by linear regression of lineage-specific estimates of d_S and d_N. In the artiodactyl and rodent lineages, the correlation between d_S and d_N did not differ significantly from zero (Figure 2, a and b). Primate genes, however, exhibited a significant positive correlation between d_S and d_N (Figure 2c). This plot has an outlier gene (growth hormone), and Makalowski and Boguski (1998) demonstrated that outliers could have adverse effects on linear regression of d_S and d_N. When growth hormone was removed, the correlation between d_S and d_N did not differ significantly from zero (Figure 2d). Reanalysis of artiodactyl and rodent lineages to the exclusion of other outlier genes had no effect on the inferred relationship between d_S and d_N (data not shown). Given that previous analyses of growth hormone indicated episodes of positive selection (Ohta 1993; Wallis 1996), we excluded it from further analyses.

The null hypothesis that synonymous and nonsynonymous substitution rates are independent was retested by using d_S and d_N estimated from the subset of genes defined by homogeneous d_N/d_S ratios. None of the comparisons exhibited a significant correlation (artiodactyls, $r^2 = 0.0297$, $P = 0.2367$; primates, $r^2 = 0.0304$, $P = 0.2413$; rodents, $r^2 = 0.0025$, $P = 0.7318$). Similar results also were obtained from reanalysis of the subset of genes defined by stationary nucleotide frequencies (artiodactyls, $r^2 = 0.0003$, $P = 0.9074$; primates, $r^2 = 0.0284$, $P = 0.2525$; rodents, $r^2 = 0.0013$, $P = 0.7919$). These results indicate that lack of a correlation between d_S and d_N was not a consequence of including genes with nonstationary nucleotide frequencies or with variable d_N/d_S ratios among lineages.

Hypothesis testing using approximate estimates of substitution rates: The two null hypotheses were tested using two approximate methods (Nei and Gojobori 1986; Ina 1995). Consistent with some previous analyses that used approximate methods (Miyata *et al.* 1982; Bernardi *et al.* 1993; Matassi *et al.* 1999; Smith and Hurst 1999), the correlation between d_S estimated between a pair of lineages and the average GC3 between the same pair of lineages did not differ significantly from zero. Also consistent with previous analyses based on approximate methods (Graur 1985; Li *et al.* 1985; Wolfe and Sharp 1993; Mouchiroud *et al.* 1995; Ohta and Ina 1995; Makalowski and Boguski 1998; Smith

TABLE 1

Maximum-likelihood estimates of synonymous and nonsynonymous rates

				Model 0		Model 1						
Gene	L_c	GC3[a]	χ^2	κ_0	ω_0	d_S(A)	d_S(P)	d_S(R)	d_N(A)	d_N(P)	d_N(R)	$\ell_1-\ell_0$
A1 Adenosine receptor	325	82	10.7	5.40	0.029	0.487	0.242	0.885	0.020	0.006	0.020	0.88
Acetylcholine receptor α	456	68	4.1	2.97	0.045	0.182	0.156	0.460	0.006	0.012	0.017	1.39
Acetylcholine receptor β	500	66	4.8	2.91	0.095	0.175	0.225	0.437	0.020	0.023	0.036	0.38
Acid phosphatase type 5	322	73	8.3	4.01	0.087	0.294	0.398	0.770	0.050	0.028	0.048	2.80
Alkaline phosphatase intestine	495	72	26.8***	2.15	0.158	0.364	0.378	0.578	0.089	0.041	0.082	3.18*
Alkaline phosphatase liver	514	77	8.2	2.43	0.059	0.395	0.264	0.749	0.025	0.026	0.030	3.13*
Apolipoprotein A1	262	81	26.2***	1.76	0.163	0.406	0.286	0.924	0.056	0.050	0.168	0.16
Apolipoprotein E	292	85	29.3***	3.48	0.119	0.794	0.095	1.472	0.090	0.062	0.104	3.93*
Apolipoprotein H	344	48	28.5***	2.63	0.190	0.541	0.176	0.520	0.069	0.049	0.115	2.04
Aspartate aminotransferase cytosolic	412	61	2.8	2.50	0.084	0.245	0.165	0.383	0.023	0.013	0.011	4.13*
Aspartate aminotransferase mitochondrial	429	63	6.4	2.24	0.059	0.253	0.158	0.416	0.026	0.016	0.028	0.81
ATP synthase α	543	47	34.8***	2.91	0.025	0.184	0.128	0.445	0.006	0.007	0.006	2.65
ATP synthase β	357	48	3.2	1.40	0.013	0.196	0.104	0.443	0.005	0.000	0.005	1.63
β-1,4-Galactosyl transferase	396	64	3.5	2.67	0.202	0.257	0.100	0.362	0.067	0.025	0.052	2.20
Ca-ATPase	1125	39	83.5***	3.15	0.010	0.186	0.110	0.446	0.004	0.001	0.004	1.99
Carboxypeptidase	432	55	8.9	1.42	0.042	0.259	0.156	0.571	0.019	0.004	0.017	2.75
CD3 ε antigen	184	56	5.6	1.60	0.515	0.196	0.116	0.539	0.120	0.119	0.173	3.10*
Connexin	381	61	15.4*	1.51	0.017	0.120	0.285	0.552	0.005	0.006	0.006	1.52
Corticotropin-releasing factor	182	78	7.5	3.70	0.185	0.469	0.123	0.584	0.113	0.018	0.087	0.75
D-Amino acid oxidase	344	62	3.4	3.50	0.211	0.189	0.118	0.553	0.065	0.029	0.079	3.77*
Dipeptidase	410	79	62.9***	3.12	0.113	0.643	0.349	0.969	0.044	0.067	0.111	2.71
Dopamine receptor D2	442	79	25.8***	3.76	0.023	0.281	0.275	0.450	0.005	0.009	0.008	0.63
Endothelin	200	58	6.5	2.15	0.294	0.165	0.202	0.477	0.070	0.089	0.080	2.94
Erythropoietin	191	67	6.1	4.00	0.192	0.201	0.176	0.793	0.069	0.058	0.073	5.34**
Fibrinogen α	433	43	37.1***	2.75	0.198	0.187	0.132	0.648	0.072	0.032	0.073	8.96***
Flavin-containing monooxygenase	530	51	4.2	3.10	0.206	0.159	0.122	0.386	0.031	0.036	0.067	1.31
Glucose transporter protein I	491	77	6.9	5.52	0.019	0.371	0.116	0.726	0.008	0.008	0.007	4.22*
Glutathione peroxidase	197	78	9.9	3.47	0.063	0.738	0.377	0.650	0.029	0.027	0.056	1.14
GMP-phosphodiesterase β	847	76	130.2***	4.02	0.024	0.577	0.550	0.951	0.018	0.020	0.012	4.54*
Growth hormone	214	76	10.4	4.81	0.123	0.501	0.593	0.609	0.041	0.160	0.047	3.80*
Growth hormone receptor	636	42	3.3	2.99	0.388	0.162	0.056	0.352	0.030	0.058	0.133	9.05***
H,K, ATPase β subunit	287	82	24.1***	2.59	0.052	1.252	0.248	0.903	0.032	0.033	0.053	4.38*
Hexokinas I	915	66	6.7	1.71	0.068	0.233	0.176	0.608	0.029	0.018	0.020	16.01***
Heat-shock 108-kD protein	397	44	4.5	1.50	0.043	0.243	0.123	0.494	0.009	0.007	0.022	0.17
Insulin-like growth factor binding protein 1	258	66	9.1	2.73	0.179	0.506	0.360	0.722	0.083	0.113	0.088	2.24
Insulin-like growth factor binding protein 3	287	78	5.0	3.36	0.109	0.802	0.051	0.702	0.054	0.044	0.061	5.82*
Insulin-like growth factor 1	114	71	4.4	3.30	0.031	0.368	0.000	0.689	0.008	0.004	0.020	0.56
Insulin-like growth factor 2	149	78	6.7	3.19	0.097	0.523	0.231	0.430	0.047	0.025	0.042	0.03
Interleukin 1A	261	51	3.1	2.26	0.400	0.185	0.214	0.462	0.085	0.090	0.165	0.22
Interleukin 1B	252	61	1.9	2.59	0.469	0.270	0.213	0.348	0.180	0.089	0.127	1.29
Interleukin 2	152	44	10.0	3.53	0.732	0.117	0.058	0.735	0.240	0.046	0.251	7.71***

(*continued*)

TABLE 1

(Continued)

Gene	L_c	GC3[a]	χ^2	Model 0		Model 1						
				κ_0	ω_0	d_S(A)	d_S(P)	d_S(R)	d_N(A)	d_N(P)	d_N(R)	ℓ_1–ℓ_0
Interleukin 2 receptor	219	58	7.8	3.08	0.406	0.416	0.383	0.328	0.189	0.095	0.184	1.36
Interleukin 6 receptor	198	54	6.6	3.07	0.942	0.278	0.070	0.565	0.185	0.188	0.426	1.60
Interleukin 7	153	33	2.5	2.18	0.631	0.102	0.059	0.296	0.069	0.102	0.102	3.46*
Lactate dehydrogenase A	331	50	34.9***	2.55	0.064	0.145	0.103	0.626	0.017	0.020	0.015	9.86***
Lactoferrin	662	62	15.6*	2.51	0.308	0.422	0.171	0.500	0.134	0.069	0.129	1.13
Link protein	352	49	3.0	4.12	0.043	0.141	0.098	0.390	0.010	0.008	0.009	2.23
Luteinizing hormone receptor	685	50	11.4	3.42	0.200	0.139	0.117	0.388	0.030	0.042	0.052	6.51**
Macrophage scavenger receptor	441	37	2.7	2.87	0.318	0.338	0.158	0.491	0.119	0.045	0.151	0.20
Myelin proteolipid protein	148	56	0.8	1.74	0.081	0.077	0.036	0.114	0.009	0.009	0.001	3.49*
Na Glucose transporter	601	65	7.3	3.10	0.108	0.244	0.160	0.565	0.040	0.029	0.031	9.71***
Na-H exchange protein	807	81	9.0	2.91	0.043	0.321	0.183	0.427	0.009	0.009	0.024	3.47*
Na-K ATPase β-1 subunit	302	57	13.7*	1.94	0.107	0.113	0.173	0.328	0.034	0.009	0.022	5.97**
Neuroleukin	557	74	2.0	1.80	0.078	0.209	0.300	0.482	0.018	0.016	0.044	1.13
Neurophysin 1	124	90	10.0	5.15	0.048	0.550	0.343	0.879	0.027	0.033	0.024	0.99
Neurophysin 2	162	91	14.0*	4.93	0.056	0.223	0.405	1.582	0.025	0.024	0.061	1.03
Ornithine decarboxylase	460	48	1.6	2.42	0.084	0.235	0.270	0.333	0.017	0.016	0.038	1.54
Osteopontin	238	48	10.5	2.31	0.382	0.161	0.236	0.594	0.150	0.460	0.176	4.21*
Phagocytic glycoprotein I	324	57	5.8	2.67	0.014	0.279	0.105	0.522	0.035	0.023	0.066	0.55
Plasminogen activator inhibitor	386	73	2.7	2.91	0.109	0.349	0.278	0.802	0.040	0.033	0.081	0.11
Polymeric Ig receptor	723	68	42.4***	2.49	0.303	0.512	0.128	0.519	0.130	0.072	0.139	2.99
Prolactin receptor	550	49	12.1	3.36	0.468	0.270	0.108	0.314	0.114	0.086	0.116	2.68
Prolyl-4-hydroxylase β	502	70	25.6***	1.90	0.030	0.460	0.187	0.789	0.013	0.009	0.019	0.88
Proopiommelanocortin	211	56	4.7	3.87	0.043	0.647	0.387	1.044	0.020	0.015	0.058	0.58
Prostaglandin E receptor	328	82	2.7	3.01	0.077	0.312	0.258	0.938	0.043	0.017	0.051	2.90
Protein disulphide isomerase	505	70	25.3***	1.80	0.041	0.461	0.234	0.691	0.013	0.017	0.026	1.97
Retinol	192	75	13.8*	2.48	0.067	0.372	0.165	0.846	0.017	0.017	0.059	0.42
Selectin	451	50	1.9	2.96	0.374	0.162	0.164	0.517	0.115	0.051	0.135	5.92**
Serum albumin	604	42	29.2***	1.76	0.257	0.192	0.263	0.636	0.092	0.057	0.121	5.41**
Stem cell factor/Kit ligand	272	39	0.3	2.41	0.358	0.151	0.066	0.173	0.041	0.029	0.070	0.52
Terminal transferase	506	50	12.7*	2.39	0.205	0.102	0.154	0.545	0.034	0.041	0.080	4.40*
Thrombomodulin	341	77	86.2***	3.30	0.132	0.632	0.478	1.712	0.122	0.097	0.115	4.77**
Tissue factor	272	53	1.0	2.32	0.395	0.303	0.190	0.527	0.075	0.093	0.253	1.74
Transforming growth factor β1	315	77	9.2	2.72	0.053	0.363	0.344	0.796	0.016	0.014	0.053	0.73
Transforming growth factor β2	413	61	12.8*	2.67	0.031	0.179	0.122	0.432	0.001	0.003	0.019	2.90
Transforming growth factor β3	408	73	6.6	3.19	0.064	0.277	0.103	0.395	0.040	0.002	0.009	11.60***
Transforming growth factor β3 receptor	843	61	47.4***	2.80	0.155	0.396	0.142	0.436	0.051	0.038	0.059	3.28*
tRNA ligase	466	69	24.5***	2.25	0.063	0.444	0.274	0.414	0.021	0.011	0.042	3.48*
Tumor necrosis factor α	229	76	4.0	3.15	0.162	0.405	0.231	0.551	0.083	0.024	0.090	0.64
Tumor necrosis factor β	200	73	9.6	4.49	0.228	0.365	0.122	0.466	0.044	0.087	0.077	4.14*
Urate oxidase	298	54	13.6*	2.33	0.092	0.170	0.122	0.601	0.025	0.025	0.026	6.79**
Urokinase-plasminogen activator	403	60	8.1	2.08	0.348	0.208	0.205	0.362	0.065	0.079	0.126	0.12
Mean		63	—	2.88	0.172	0.327	0.201	0.599	0.054	0.045	0.073	—

L_c is the length of a gene in codons. χ^2 is the chi-square value for a contingency test of equal nucleotide frequencies at third codon positions. κ_0 and ω_0 are the transition/transversion rate ratio and nonsynonymous/synonymous rate ratios estimated using model 0. (A), (P), and (R) indicate the artiodactyl, primate, and rodent lineages. ℓ_1–ℓ_0 is the log-likelihood difference between model 0 and model 1. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

[a] Indicates the average of GC3 in artiodactyl, primate, and rodent genes.

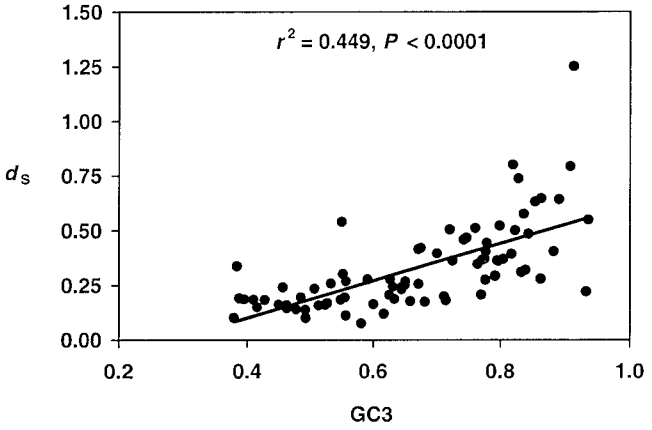

FIGURE 1.—The relationship between ML estimates of d_S and GC3 in artiodactyl genes.

and HURST 1999), there was a significant positive correlation between d_S and d_N. Because results were similar for all three comparisons, only comparisons between artiodactyl and primate genes are presented in Figure 3. These findings indicate that approximate and ML methods led to exactly opposite conclusions.

Pairwise estimation of d_S and d_N using maximum likelihood is consistent with lineage-specific estimation of d_S and d_N: Approximate methods are applicable only to pairwise sequence comparisons, whereas ML results discussed above were obtained from joint analysis of all sequences on a phylogeny. To facilitate direct comparison of approximate and ML methods, d_S and d_N were re-estimated in a pairwise fashion between the sampled lineages using ML (GOLDMAN and YANG 1994). In all three pairwise comparisons, estimation of substitution rates via ML yielded results similar to those obtained by using lineage-specific estimates of substitution rates; *i.e.*, a significant positive correlation was observed between d_S and GC3, and a nonsignificant correlation was observed between d_N and d_S (Figure 3, c and f). These findings indicate that comparisons could be made between approximate and ML methods by utilizing ML to estimate d_N and d_S in a pairwise fashion between lineages.

Reconciling differences between methods: We have shown that transition/transversion bias is a common feature of DNA sequence evolution in these genes. The approximate method of NEI and GOJOBORI (1986) ignores the transition/transversion bias by assuming rate equality. We changed the parameters of the codon model to investigate the effects of this assumption on the estimation of d_S and d_N (see MATERIALS AND METHODS). The effect of ignoring the transition/transversion rate bias was consistent underestimation of the numbers of synonymous sites (S; Figure 4a). Because transitions at third codon positions are more likely to be synonymous than transversions, ignoring the transition/transversion bias leads to underestimation of S and overestimation of d_S (LI *et al.* 1985; PAMILO and BIANCHI 1993; INA 1995; YANG and NIELSEN 1998).

We also have shown that biased nucleotide (codon) frequencies were characteristic of the sampled genes. Both the methods of NEI and GOJOBORI (1986) and INA (1995) ignore this feature of DNA sequence evolution. We changed the parameters of the codon model to investigate the effect of this assumption on estimation of d_S and d_N (see MATERIALS AND METHODS). Ignoring codon bias had the opposite effect to ignoring the transition/transversion bias, in that $S$was consistently overestimated (Figure 4b). These results indicate that the number of synonymous sites (S) available to mutation was restricted to varying degrees by biased codon usage. Because approximate methods (NEI and GOJOBORI 1986; INA 1995) assume unbiased codon usage, counts of the number of synonymous substitutions will be mea-

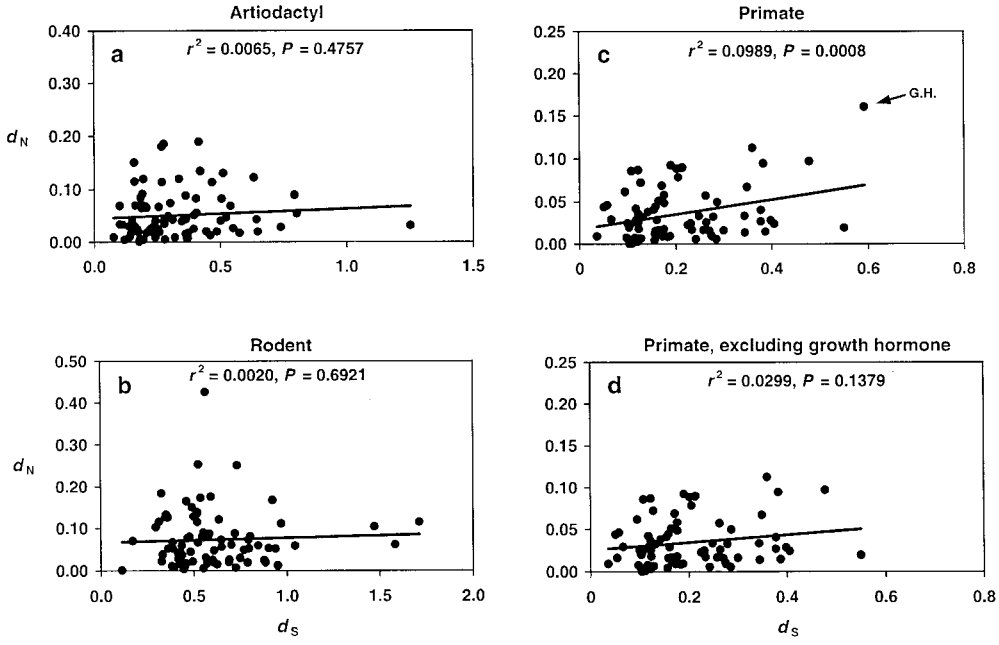

FIGURE 2.—The relationship between ML estimates of d_S and d_N in artiodactyl (a), rodent (b), and primate (c and d), genes. G.H. indicates the growth hormone gene.

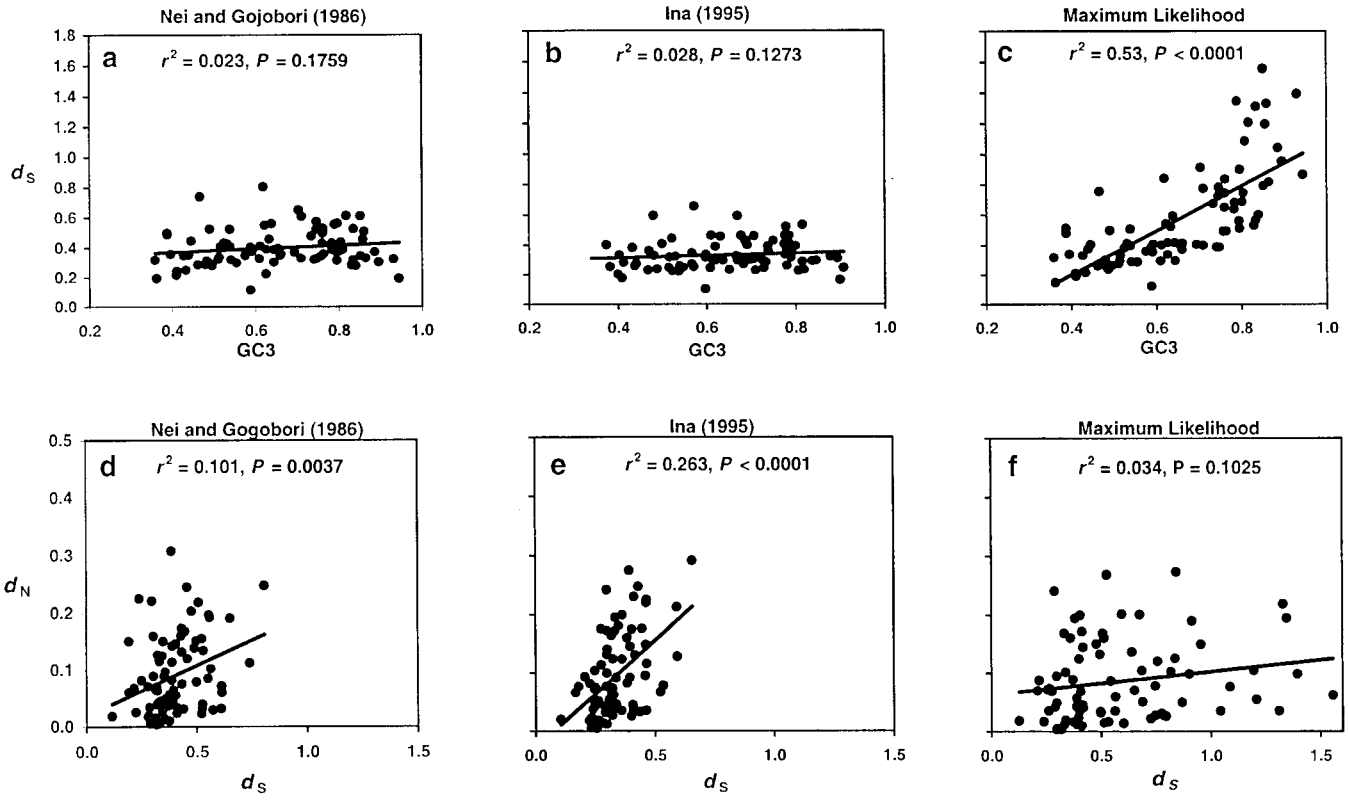

FIGURE 3.—The relationship between pairwise estimates of d_S and mean GC3 (a–c) and the relationship between pairwise estimates of d_S and d_N (d–f). All plots represent a pairwise comparison between artiodactyl and primate genes. Pairwise estimates of substitution rates were computed by using the approximate methods of NEI and GOJOBORI (1986) and INA (1995) and also by using ML (GOLDMAN and YANG 1994).

sured against too large a number of synonymous sites, and therefore d_S will be underestimated. Because the total number of sites is fixed in a gene, the bias in estimation of nonsynonymous sites (N) is opposite to that of S.

To understand why different methods produced different results concerning the correlation of d_S with GC3 or d_N, we examined the following two summary statistics: (i) the ratio of the approximate estimate of d_S to the ML estimate of d_S (d_S ratio) and (ii) the ratio of the approximate estimate of d_N to the ML estimate of d_N (d_N ratio). Plots of the d_S ratio and d_N ratio against GC3 illustrate the complexity of the biases involved in approximate estimation of d_S and d_N (Figure 5). For genes with highly biased nucleotide (codon) usage (GC3 > 60%), both approximate methods were consistent with our earlier analysis of codon models that ignored nucleotide (codon) frequencies (Figure 4b) in that d_S was underestimated and d_N was overestimated (Figure 5). However, when nucleotide (codon) bias was weak (GC3 < 60%), the two approximate methods differed in the direction of bias, with the method of NEI and GOJOBORI (1986) overestimating d_S and underestimating d_N (Figure 5a) and the method of INA (1995) underestimating d_S and overestimating d_N (Figure 5b).

Estimates of d_S and d_N by the method of NEI and GOJOBORI (1986) were affected differently in genes with high and low codon bias (Figure 5a) because this method ignores both the transition/transversion rate bias and codon usage bias, and these two features of DNA sequence evolution have opposite effects on estimation of d_S and d_N (Figure 4). The method of INA (1995) overestimated d_S and underestimated d_N in genes with both weak as well as strong codon usage bias because this method overcorrects for the transition/transversion rate bias (YANG and NIELSEN 1998), thereby producing bias in the same direction as when codon usage is highly biased. For both methods, codon usage bias had the largest effect on approximate estimation of d_S and d_N (Figure 5).

To understand the difference between methods concerning the d_S and d_N correlation, we examined the relationship between d_S ratios and ML estimates of d_N and the relationship between d_N ratios and ML estimates of d_S. Although approximate methods produced highly biased estimates of d_S (Figure 5), there was no significant correlation between this bias (d_S ratio) and d_N (*e.g.*, artiodactyl *vs.* primate: NEI and GOJOBORI 1986, r^2 = 0.025, P = 0.1539; INA 1995, r^2 = 0.017, P = 0.2505). However, there was a significant positive correlation between the d_N ratio and d_S (*e.g.*, artiodactyl *vs.* primate: NEI and GOJOBORI 1986, r^2 = 0.227, $P < 0.0001$; INA 1995, r^2 = 0.258, $P < 0.0001$). These findings suggest that approximate estimation of d_N could interpose a

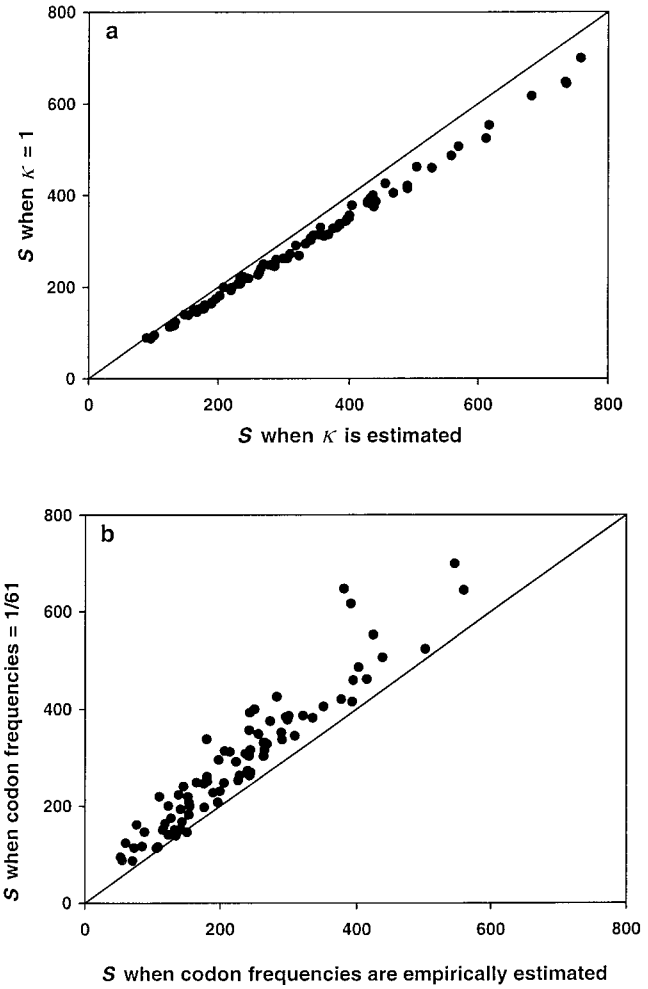

FIGURE 4.—Bias in the estimated number of synonymous sites (S) when (a) transition/transversion ratio (κ) is ignored and (b) when unequal codon frequencies are ignored. Data presented in (a) were estimated using two models with equal codon frequencies (1/61), and in one model κ was a free parameter and in the other model $\kappa = 1$ (transition and transversion rates assumed to be equal). Data presented in (b) were estimated using two models with $\kappa = 1$, where one model used empirical codon frequencies and the other model assumed equal codon frequencies (1/61).

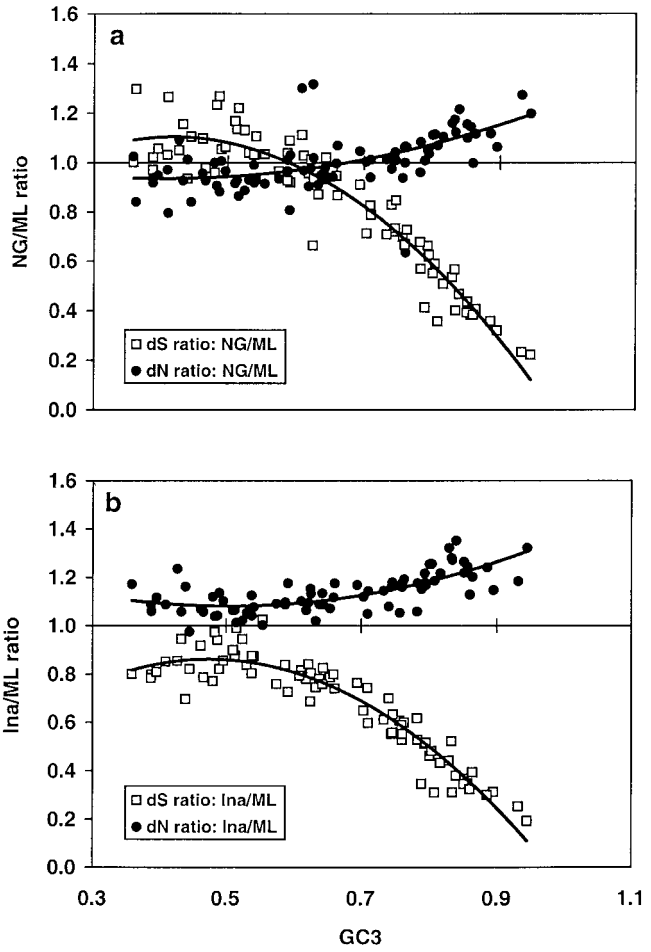

FIGURE 5.—Bias in (a) the method of NEI and GOJOBORI (1986) and (b) the method of INA (1995) as compared to ML. Data represent pairwise comparisons between artiodactyl and primate genes. Bias was measured using the ratio of the approximate estimate of d_S to the ML estimate of d_S (d_S ratio) and the ratio of the approximate estimate of d_N to the ML estimate of d_N (d_N ratio). "NG" indicates the method of NEI and GOJOBORI (1986), "Ina" indicates method 1 of INA (1995), and "ML" indicates the maximum-likelihood method of GOLDMAN and YANG (1994).

positive correlation between estimates of nonsynonymous and synonymous substitution rates.

The preceding analyses suggested that failure of the approximate methods to properly account for the transition/transversion rate bias and unequal codon usage has resulted in seriously biased estimates of substitution rates. These biases appear to be the source of conflict between the methods. To test this prediction, we retested the two null hypotheses using substitution rates estimated from a codon model that was modified to ignore biased nucleotide (codon) frequencies and transition/transversion ratio. Linear regression of substitution rates estimated using this codon model yielded results that fit the prediction, *i.e.*, there was no significant correlation between d_S and GC3 (*e.g.*, artiodactyl *vs.* primate: $r^2 = 0.027$, $P = 0.137$), and there was a significant positive correlation between d_S and d_N (*e.g.*, artiodactyl *vs.* primate: $r^2 = 0.124$; $P = 0.001$).

DISCUSSION

Synonymous substitution rate is positively correlated with nucleotide composition: Mammalian genomes exhibit a degree of structure in the form of long (>300 kb) compositionally homogenous regions of DNA known as isochores (BERNARDI 1993). The well-known correlation between GC content at third codon positions of a gene and GC content of the isochore in which that gene resides, permits us to study substitution rates at the level of the isochore (MOUCHIROUD *et al.* 1991; BERNARDI 1995; CLAY *et al.* 1996). Our results indicate that synony-

mous substitution rates differ among isochores and therefore among different regions of the mammalian genome. Furthermore, the most GC-rich isochores appear to have the highest synonymous substitution rate. These results are significant because arguments against a mutation-based hypothesis for the origin and maintenance of isochores have relied, in part, upon the assumption that synonymous substitution rates do not differ among regions of the mammalian genome (Bernardi *et al.* 1993; Mouchiroud *et al.* 1995).

The hypothesis that synonymous substitution rates vary among different isochores was originally proposed by Wolfe *et al.* (1989). Moreover, Wolfe *et al.* (1989) found remarkably similar rates of silent substitution in six physically linked genes in mouse and rat. Support for the hypothesis of Wolfe *et al.* (1989) can be found in other studies. Matassi *et al.* (1999) investigated synonymous substitution rates among genes lying within one centimorgan of each other in mouse and human. Synonymous substitution rates among these neighboring genes were more similar than among genes that were farther apart on the chromosome (Matassi *et al.* 1999). The results of our study, taken together with those of Wolfe *et al.* (1989) and Matassi *et al.* (1999), suggest that the perceived gene specificity of synonymous substitution rate reflects, at least in part, region-specific effects on the rate of synonymous substitution.

Matassi *et al.* (1999) also investigated GC3 content of genes within one centimorgan of each other and found that the same sets of neighboring genes were more similar to each other in GC content than genes found farther apart on the chromosome. However, in contrast to our study, Matassi *et al.* (1999) did not find a significant correlation between d_S and GC3 and hypothesized that regional similarities in both synonymous substitution rates and nucleotide composition were evolving independently of each other. Values of d_S used in their correlation analysis were estimated using the approximate method of Li (1993). Because this method is similar to the method of Ina (1995) in that it does not account for biased nucleotide (codon) frequencies, their estimates might be biased.

Our results have important implications for the hypothesis of Alvarez-Valin *et al.* (1998) that selection is acting to enhance translational accuracy in mammals. If selection is acting to enhance translational accuracy, then we should observe a negative correlation between nucleotide (codon) bias and synonymous substitution rate (Akashi 1994). Our finding of a positive correlation between d_S and GC3 suggests that synonymous codon usage in mammals is not subject to this type of selective constraint. In support of Akashi (1994), a negative correlation between synonymous substitution rate and codon bias has been observed in Drosophila, bacteria, and yeast (Sharp and Li 1987, 1989; Shields *et al.* 1988; Moriyama and Gojobori 1992; Powell and Moriyama 1997), and in these taxa codon usage also matches tRNA abundance.

The results of this study do not preclude a role for selection in the maintenance of mammalian isochores. It has been suggested that selection might be acting regionally to elevate GC content (Bernardi *et al.* 1985, 1988). In this hypothesis, selection acts to elevate GC content in regions of the genomes of warm-blooded vertebrates as a means of protecting DNA from heat degradation (Bernardi *et al.* 1985, 1988). In support of this hypothesis, Eyre-Walker (1999) reported that patterns of silent site variation in major histocompatibility genes of mammals were not consistent with neutral expectations, but were consistent with the influence of selection on nucleotide composition. However, Francino and Ochman (1999) recently reported that interspecific variation in two globin pseudogenes that reside in different isochores was consistent with the effect of differential GC mutation pressure. Although data presented here are not sufficient to resolve this long-standing controversy, our conclusion that synonymous substitution rates vary among different isochores, taken together with the recent findings of Francino and Ochman (1999), suggest at least a partial role for mutation in the maintenance of mammalian isochores.

Synonymous substitution rate is independent of nonsynonymous substitution rate: Smith and Hurst (1999) estimated substitution rates between pairs of rat and mouse genes and found that the correlation between d_S and d_N obtained from ML was less than neutral expectations (Ohta and Ina 1995), whereas the correlation obtained from approximate methods was greater than neutral expectations. In this regard the results of their study are compatible with ours. However, the findings of Smith and Hurst (1999) differ from ours in that a positive correlation between ML estimates of d_S and d_N, although less than neutral expectations, was significant. The reason for this difference is unclear.

A potential source of correlation between d_S and d_N is variation among loci in codon usage and base frequencies. The significant correlation between d_S and d_N indicated by the approximate methods, which ignore codon usage bias, disappeared after we corrected for codon usage and base frequencies. Results of simulation studies (Yang and Nielsen 1998, 2000) support the view that the differences among methods observed in the present study may be attributed to biases in estimation.

What is clear from both this study and the study of Smith and Hurst (1999) is the sensitivity of such analyses to the estimation method and to assumptions concerning the transition/transversion rate bias and nonrandom codon usage. Unbiased estimation of substitution rates is a critical aspect of reliably measuring the effectiveness of selection at synonymous sites.

This study was supported by a Biotechnology and Biological Sciences Research Council grant (31/G10434) to Z.Y.

LITERATURE CITED

Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136:** 927–935.

Alvarez-Valin, F., K. Jabbari and G. Bernardi, 1998 Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. J. Mol. Evol. **46:** 37–44.

Bernardi, G., 1993 The vertebrate genome: isochores and evolution. Mol. Biol. Evol. **10:** 186–204.

Bernardi, G., 1995 The human genome: organization and evolutionary history. Annu. Rev. Genet. **29:** 445–476.

Bernardi, G., and G. Bernardi, 1986 Compositional constraints and genome evolution. J. Mol. Evol. **24:** 1–11.

Bernardi, G., B. Oloffson, J. Filipski, M. Zerial, J. Salinas *et al.*, 1985 The mosaic genome of warm-blooded vertebrates. Science **228:** 953–958.

Bernardi, G., D. Mouchiroud, C. Gautier and G. Bernardi, 1988 Compositional patterns in vertebrate genomes: conservation and change in evolution. J. Mol. Evol. **28:** 7–18.

Bernardi, G., D. Mouchiroud and C. Gautier, 1993 Silent substitutions in mammalian genomes and their evolutionary implications. J. Mol. Evol. **37:** 583–589.

Bulmer, M., K. H. Wolfe and P. M. Sharp, 1991 Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationships of mammalian orders. Proc. Natl. Acad. Sci. USA **88:** 5974–5978.

Clay, O., S. Caccio, S. Zoubak, D. Mouchiroud and G. Bernardi, 1996 Human coding and noncoding DNA: compositional correlations. Mol. Phylogenet. Evol. **5:** 2–12.

Eyre-Walker, A., 1991 An analysis of codon usage in mammals: selection or mutation bias. J. Mol. Evol. **33:** 442–449.

Eyre-Walker, A., 1994 DNA mismatch repair and synonymous codon evolution in mammals. Mol. Biol. Evol. **11:** 88–98.

Eyre-Walker, A., 1999 Evidence for selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152:** 657–683.

Filipski, J., 1988 Why the rate of silent codon substitution is variable within a vertebrate's genome. J. Theor. Biol. **134:** 159–164.

Francino, M. P., and H. Ochman, 1999 Isochores result from mutation not selection. Nature **400:** 30–31.

Gautier, N., and D. Mouchiroud, 1998 Isochore evolution in mammals: a human-like ancestral sequence. Genetics **150:** 1577–1584.

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

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

Ina, Y., 1995 New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40:** 190–226.

Lanave, C., G. Preparata, C. Saccone and G. Serio, 1984 A new method for calculating evolutionary substitution rates. J. Mol. Evol. **20:** 86–93.

Li, W.-H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96–99.

Li, W.-H., C.-I. Wu and C.-C. Luo, 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.

Makalowski, W., and M. S. Boguski, 1998 Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. J. Mol. Evol. **47:** 119–121.

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

Miyata, T., H. Hayashida, R. Kikuno, M. Hasagawa, M. Kobayashi *et al.*, 1982 Molecular clock of silent substitution: at least six fold preponderance of silent changes in mitochondrial genes over those of nuclear genes. J. Mol. Evol. **19:** 28–35.

Moriyama, E. N., and T. Gojobori, 1992 Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics **143:** 847–858.

Mouchiroud, D., and C. Gautier, 1990 Codon usage changes and sequence dissimilarity between human and rat. J. Mol. Evol. **31:** 81–91.

Mouchiroud, D., G. D'Onofrio, G. Aissani, G. Macaya, C. Gautier *et al.*, 1991 The distribution of genes in the human genome. Gene **100:** 181–187.

Mouchiroud, D., C. Gautier and G. Bernardi, 1995 Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. **40:** 107–113.

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.

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

Ohta, T., 1993 Pattern of nucleotide substitution in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. Genetics **134:** 1271–1276.

Ohta, T., 1995 Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. **40:** 56–63.

Ohta, T., and Y. Ina, 1995 Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. J. Mol. Evol. **41:** 717–720.

Pamilo, P., and N. O. Bianchi, 1993 Evolution of the *Zfx* and *Zfy* genes—rates and interdependence between genes. Mol. Biol. Evol. **10:** 271–281.

Powell, J. R., and E. N. Moriyama, 1997 Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94:** 7784–7790.

Saccone, C., G. Pesole and G. Preparata, 1989 DNA microenvironments and the molecular clock. J. Mol. Evol. **29:** 407–411.

Sharp, P. M., and W.-H. Li, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

Sharp, P. M., and W.-H. Li, 1989 On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28:** 3398–3402.

Shields, D. C., P. M. Sharp, D. G. Higgins and F. Wright, 1988 "Silent" sites in Drosophila genes are not neutral: evidence for selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

Smith, N. G. C., and L. D. Hurst, 1999 The effect of tandem substitutions on the correlation of synonymous and nonsynonymous rates in rodents. Genetics **153:** 1395–1402.

Ticher, A., and D. Graur, 1989 Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. J. Mol. Evol. **28:** 286–298.

Wallis, M., 1996 The molecular evolution of vertebrate growth hormones: a pattern of near-stasis interrupted by sustained bursts of rapid change. J. Mol. Evol. **43:** 93–100.

Wolfe, K. H., and P. M. Sharp, 1993 Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. **37:** 441–456.

Wolfe, K. H., P. M. Sharp and W.-H. Li, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

Yang, Z., 1999 *Phylogenetic Analysis by Maximum Likelihood* (*PAML*), *Version 2.* University College London, England.

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

Yang, Z., and R. Nielsen, 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17:** 32–43.

Communicating editor: M. K. Uyenoyama