Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage

Ziheng Yang and Rasmus Nielsen†*

*Department of Biology, Galton Laboratory, University College London, London, United Kingdom; and †Department of Biology, University of Copenhagen, Copenhagen, Denmark

Current models of codon substitution are formulated at the levels of nucleotide substitution and do not explicitly consider the separate effects of mutation and selection. They are thus incapable of inferring whether mutation or selection is responsible for evolution at silent sites. Here we implement a few population genetics models of codon substitution that explicitly consider mutation bias and natural selection at the DNA level. Selection on codon usage is modeled by introducing codon-fitness parameters, which together with mutation-bias parameters, predict optimal codon frequencies for the gene. The selective pressure may be for translational efficiency and accuracy or for fine-tuning translational kinetics to produce correct protein folding. We apply the models to compare mitochondrial and nuclear genes from several mammalian species. Model assumptions concerning codon usage are found to affect the estimation of sequence distances (such as the synonymous rate d_S, the nonsynonymous rate d_N, and the rate at the 4-fold degenerate sites d_4), as found in previous studies, but the new models produced very similar estimates to some old ones. We also develop a likelihood ratio test to examine the null hypothesis that codon usage is due to mutation bias alone, not influenced by natural selection. Application of the test to the mammalian data led to rejection of the null hypothesis in most genes, suggesting that natural selection may be a driving force in the evolution of synonymous codon usage in mammals. Estimates of selection coefficients nevertheless suggest that selection on codon usage is weak and most mutations are nearly neutral. The sensitivity of the analysis on the assumed mutation model is discussed.

Introduction

In protein-coding genes, synonymous codons that code for the same amino acid do not appear at the same frequency (Ikemura 1981, 1985). Whether the origin and maintenance of such codon usage bias is due to biases in the mutation process or to natural selection has been a matter of much controversy (see, e.g., Duret 2002 for review). Mutation bias must play a role, but the significance of selection in driving the evolution of codon usage is less certain and may depend on the species. In fast-growing organisms with large population sizes, such as *Escherichia coli*, *Saccharomyces cerevisiae*, and yeast, codon usage is generally thought to be under selective pressure, as supported by several lines of evidence. First, codon frequencies are correlated with the cellular cognate tRNA concentrations (Ikemura 1981, 1985; Bennetzen and Hall 1982; Bulmer 1987; Sharp and Li 1987; Moriyama and Powell 1997). Preferential use of so-called major codons to match the most abundant tRNAs may enhance translational speed and improve translational accuracy (for reviews, see Akashi 1995; Sharp et al. 1995; Duret 2002). In addition, major codons may reduce the energetic cost of translation by reducing the chances of amino acid misincorporations and ribosomal drop-offs (Kurland 1992) and by freeing up the protein synthesis machinery through faster ribosomal elongation. Second, in both *Drosophila* and *Caenorhabditis elegans*, codon usage is correlated with gene expression, with highly expressed genes having strongly biased codon usage, presumably because of stronger selective pressure (Duret and Mouchiroud 1999; Castillo-Davis and Hartl 2002). Third, silent substitution rate (measured by the sequence distances d_S or d_4 at the synonymous or 4-fold degenerate sites) is lower in genes with highly biased codon usage, implying stronger purifying selection on silent mutations in highly biased genes (e.g., Sharp and Li 1987). This correlation was nevertheless found to depend on the method used to estimate silent rates (Dunn et al. 2001; Bierne and Eyre-Walker 2003). Fourth, in *Drosophila*, codon usage is more biased for conserved amino acids than for nonconserved amino acids (Akashi 1994). This may be explained by selection for translational accuracy because highly conserved amino acids are expected to be functionally more important and less tolerant to misincorporations of wrong amino acids and are thus under stronger selective pressure.

In slowly growing organisms with small population sizes such as vertebrates, natural selection may be inefficient and indeed its effect on codon usage is controversial (see, e.g., Duret 2002 for a review). In contrast to results for bacteria, yeast, and *Drosophila*, strong evidence for selection on codon usage is lacking in vertebrates. For example, Kanaya et al. (2001) found a correspondence between codon bias and tRNA gene copy number (a proxy for tRNA concentration) in *Schizosaccharomyces pombe* and *C. elegans* but not in *Xenopus laevis* and *Homo sapiens*; in the later species, highly expressed genes such as ribosomal genes and histone genes do not have strong codon bias. Some studies (e.g., Musto et al. 2001) found a correlation between codon bias and putative expression levels (as measured by expressed sequence tag frequencies), but this correlation could be explained by transcription-coupled repair (Duret 2002).

Besides selection for translational efficiency and accuracy, recent experimental work suggests that the selective pressure on codon usage may also be due to the need for an optimal translation kinetics, to ensure correct protein folding. Protein folding is thought to be cotranslational, occurring at the same time the protein is translated from the mRNA (Frydman 2001). The use of preferred and unpreferred codons may affect the rate at which the protein is translated. The translation kinetics may be important in separating temporally folding events during protein synthesis

Key words: codon substitution model, codon usage, mutation, selection, synonymous substitution.

E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 25(3):568–579. 2008
doi:10.1093/molbev/msm284
Advance Access publication January 3, 2008

on the ribosome, thus ensuring "beneficial" interactions and avoiding "unwanted" interactions within the growing peptide, to achieve high yield of the correctly folded protein. Kimchi-Sarfaty et al. (2007) reported that certain synonymous mutations in the multidrug resistance 1 gene resulted in altered drug and inhibitor interactions. They found similar mRNA and protein levels but altered protein conformations between the "wild type" and mutant protein products and hypothesized that the incorporation of rare synonymous codons may have affected the timing of folding. This form of selection differs from translational selection in that preferred codons are not always advantageous if the optimal folding requires a slow translation. It is unclear how important such selection for protein folding is to the evolutionary process of protein-coding genes.

A number of authors have studied population genetics models in which the proportions of synonymous codons are modeled as the product of interactions between mutation bias, natural selection, and genetic drift (Kimura 1983; Li 1987; Bulmer 1991; McVean and Charlesworth 1999). McVean and Vieira (1999) applied maximum likelihood (ML) to fit such a model to counts of synonymous codons for 2-fold amino acids in protein-coding genes in several *Drosophila* species, to estimate parameters of mutation bias and selective pressure. The analysis does not consider the evolutionary relationships among species, which may provide useful information concerning relative mutation rates between nucleotides. This model was extended by McVean and Vieira (2001) to analyze synonymous differences between different species, with nonsynonymous differences ignored. Nielsen et al. (2007) implemented a codon-substitution model in which a mutation is favored or disfavored by natural selection depending on whether it changes an unpreferred codon into a preferred one or vice versa. The model was applied to *Drosophila* protein–coding genes to obtain ML estimates of parameters measuring the strength of selection. This method requires a priori partitioning of synonymous codons into preferred and unpreferred categories and also assumes only one selection coefficient to accommodate selection on codon usage.

In this paper, we implement a few new models of codon substitution that relax those assumptions. Our motivations for this study are 2-fold. First, we devise a likelihood ratio test (LRT) of neutral evolution of codon usage to infer possible effects of natural selection. Whereas many previous studies have performed correlation analysis to test the various predictions of the mutation and selection theory of codon usage bias (see above), the LRT addresses this problem directly. Our model also provides direct measurements of selection acting on silent sites. Second, we examine the effects of model assumptions about codon usage on estimation of sequence distances such as d_S, d_N, and their ratio $\omega = d_N/d_S$. There has been considerable interest in the use of the ω ratio to detect positive selection affecting protein evolution, and some concerns have been expressed as to whether this inference is affected by natural selection acting on silent sites (Kreitman and Akashi 1995; Yang and Bielawski 2000). We analyze 2 sets of data to address these issues, the first of the human and chimpanzee mitochondrial protein–coding genes and the second of 5,639 protein-coding genes from the 5 mammalian species: human, chimpanzee, macaque, mouse, and rat.

Theory

A Mutation-Selection Model of Codon Substitution

We construct a model of codon substitution by specifying the instantaneous rate of substitution from sense codons $I = i_1i_2i_3$ to $J = j_1j_2j_3$, where i_1 is the nucleotide at the first position in codon I, and so on. We assume that point mutations occur independently at nucleotide sites and thus the rate is zero if I and J differ at more than 2 or 3 codon positions (Goldman and Yang 1994). Thus, we focus on the rate between 2 codons that differ at only one position, say position k, with $i_k \neq j_k$. We explicitly model the process of one codon substituting another codon, that is, mutation, selection on the DNA (selection on codon usage), and selection on the protein.

Mutation Bias

Let the mutation rate from nucleotides i to j be μ_{ij} per generation. The mutation model applies to all 3 codon positions, although the base compositions at the 3 positions may differ. We use the general time reversible (GTR or REV) model (e.g., Yang 1994) to describe the mutation process so that $\mu_{ij}=a_{ij}\pi_j^*$, with $a_{ij} = a_{ji}$ for all $i \neq j$. Here π_j^* reflects mutation bias; if π_T^* is large, mutations are biased toward T. One of the mutation-bias parameters is redundant, and we scale them so that $\sum \pi_j^*=1$. If the HKY mutation model (Hasegawa et al. 1985) is used, $\mu_{ij}=\mu\kappa\pi_j^*$ if i and j differ by a transition and $\mu_{ij}=\mu\pi_j^*$ if i and j differ by a transversion, with κ to be the transition/transversion rate ratio. Our analysis below is based mostly on the HKY model, but GTR is used in some analyses to examine the robustness of the results.

Selection on Codon Usage

We model selection on codon usage by introducing a fitness parameter f_I for codon I. The selection coefficient for the mutation that changes the wild type codon I into a new mutant codon J is thus $s_{IJ} = f_J - f_I$. The probability of fixation of the mutation is $\frac{2s_{IJ}}{1-e^{-2Ns_{IJ}}}$, where N is the effective chromosomal population size (Fisher 1930; Wright 1931; Kimura 1957). Let $F_I = 2Nf_I$ be the scaled fitness of codon I, and $S_{IJ} = 2Ns_{IJ} = 2N(f_J - f_I) = F_J - F_I$ be the scaled selection coefficient. As the number of the $I \rightarrow J$ mutations in a generation is $N\mu_{i_kj_k}$, the substitution rate from codons I to J is given as

$$N\mu_{i_kj_k} \times \frac{2s_{IJ}}{1-e^{-2Ns_{IJ}}} = a_{i_kj_k}\pi_{j_k}^* \times \frac{S_{IJ}}{1-e^{-S_{IJ}}} = a_{i_kj_k}\pi_{j_k}^* \times h(S_{IJ}), \quad (1)$$

where $h(S_{IJ}) = S_{IJ}/(1-e^{-S_{IJ}})$ is the ratio of the fixation probability of the $I \rightarrow J$ mutation to the fixation probability of a neutral mutation, with $h(S_{IJ}) < 1$, $= 1$ and > 1 for deleterious mutations (with $S_{IJ} < 0$), neutral mutations ($S_{IJ} = 0$), and advantageous mutations ($S_{IJ} > 0$), respectively.

When the model is applied to sequence data from different species, we have in this study assumed that the effective population size N and the selection coefficients are the same among lineages. Those assumptions can be relaxed at the expense of including more parameters (McVean and Vieira 2001; Nielsen et al. 2007).

Selection on the Protein

To describe selection on the protein, we multiply the substitution rate by ω if and only if the mutation is nonsynonymous (Goldman and Yang 1994; Yang and Nielsen 1998). Thus, ω is the nonsynonymous/synonymous substitution rate ratio. The use of one single ω to describe selection on the protein is very simplistic. However, previous models that incorporate amino acid chemical properties to specify codon substitution rates achieved only moderate (although statistically significant) improvements to the model's fit to data, and furthermore, such models produced rather similar estimates of mutation parameters to the simple model of one ω ratio (Goldman and Yang 1994; Yang et al. 1998). Here our focus is on the effect of selection on synonymous codon usage. We also implement the site models that assume variable ω ratios among codons in the gene (Nielsen and Yang 1998; Yang et al. 2000).

To summarize, the substitution rate from codons I to J is specified as

$$q_{IJ} = \begin{cases} 0, \text{if the 2 codons differ at more than one position,} \\ a_{i_k j_k} \pi^*_{j_k} h(S_{IJ}), \text{for synonymous substitution,} \\ \omega a_{i_k j_k} \pi^*_{j_k} h(S_{IJ}), \text{for nonsynonymous substitution.} \end{cases} \quad (2)$$

The diagonals of the rate matrix $Q = \{q_{IJ}\}$ are determined by the requirement that each row in the matrix sums to zero. As only the difference $S_{IJ} = F_J - F_I$ enters the probability calculation under the model, we fix one of the 61 F_I's to zero and estimate 60 free parameters for the universal genetic code. The model thus includes the following parameters in the substitution rate matrix Q: 8 parameters in the GTR mutation model (or 4 parameters in HKY: κ, π^*_T, π^*_C, and π^*_A), 60 scaled fitness parameters, and ω. The sequence distance t or branch lengths on the tree are additional parameters to be estimated from the data.

After the Q matrix is constructed, the stationary distribution of the Markov chain, $\pi = \{\pi_1, \pi_2, \ldots, \pi_{61}\}$, is given by the system of linear equations $\pi Q = 0$, subject to the constraint that the π_j's sum to one. This distribution can also be calculated directly (see eq. 4 below). The matrix is then multiplied by a constant so that the "average" rate is one: $-\sum_I \pi_I q_{II} = 1$. The transition probability matrix $P(t) = e^{Qt}$ is calculated following standard theory. (Note that we have used π_J, where the subscript J is a codon to indicate the equilibrium frequency of codon J, and π^*_j, where the subscript j is a nucleotide to represent the mutation-bias parameter in the HKY or GTR mutation models.)

The Markov model of codon substitution specified by equation (2) is time reversible. To show this, it is sufficient to write the rate matrix as a product of a symmetrical matrix and a diagonal matrix (e.g., Yang 2006, p. 33–34). The rate q_{IJ} in equation (2) for a synonymous change can be rewritten as

$$\begin{aligned} q_{IJ} &= a_{i_k j_k} \pi^*_{j_k} \times \frac{F_J - F_I}{1 - e^{F_I - F_J}} \\ &= \left[a_{i_k j_k} \times \frac{1}{\prod_{k' \neq k} \pi^*_{j_{k'}}} \times \frac{F_J - F_I}{e^{F_J} - e^{F_I}} \right] \times \left(\pi^*_{j_1} \pi^*_{j_2} \pi^*_{j_3} e^{F_J} \right). \end{aligned} \quad (3)$$

Here $\prod_{k' \neq k} \pi^*_{j_{k'}}$ is the product of the mutation-bias parameters for the 2 unchanged nucleotides (i.e., $\pi^*_T \pi^*_C$ if I = TCA and J = TCG). The quantity in the square brackets, denoted A_{IJ}, satisfies $A_{IJ} = A_{JI}$ for all $I \neq J$, whereas the quantity in the parentheses is a function of J only. The rate q_{IJ} when the $I \rightarrow J$ substitution is nonsynonymous can be written in this form as well. Thus, the rate matrix $Q = \{q_{IJ}\}$ can be written as a product of a symmetrical matrix $\{A_{IJ}\}$ and a diagonal matrix so that the Markov process is time reversible, with the stationary frequency for codon J given as

$$\pi_J \propto \pi^*_{j_1} \pi^*_{j_2} \pi^*_{j_3} \times e^{F_J}. \quad (4)$$

For example, the equilibrium frequency of codon TCG is proportional to $\pi^*_T \pi^*_C \pi^*_G \times e^{F_{TCG}}$. This result makes it clear that the stationary codon frequencies are determined by both mutation bias (represented by $\pi^*_{j_1} \pi^*_{j_2} \pi^*_{j_3}$) and selection on codon usage (represented by e^{F_J}). The model is referred to below as the FMutSel model. It may also be noted that instead of the codon fitness parameters (F_J), one may use the codon frequencies (π_J) as parameters. The latter parametrization is convenient for an approximate implementation to be described below.

An LRT of Selection on Codon Usage

We implement a special case of the mutation-selection model of codon substitution (eq. 2), in which all synonymous codons (codons that encode the same amino acid) have the same fitness. Thus, instead of 60 (=61 − 1) codon fitness parameters for the universal genetic code, only 19 (=20 − 1) amino acid fitness parameters are used. The model assumes that the amino acid frequencies are determined by the functional requirements of the protein, but there is no fitness difference among the synonymous codons. From the theory above (eq. 4), the relative frequencies of synonymous codons are determined solely by the mutational-bias parameters. This model is referred to as FMutSel0.

An LRT can be constructed by comparing models FMutSel0 against FMutSel. Twice the log-likelihood difference between the 2 models is compared with the χ^2 distribution with degree of freedom = 60 − 19 = 41 for the universal code (or 40 for the vertebrate mitochondrial code). This constitutes a test of the null hypothesis that codon usage is due to mutation bias alone and not to selection acting at silent sites.

Measurements of Selection on Codon Usage

As our model explicitly separates mutation bias from selection affecting codon usage, we devise a few measures

of the strength of natural selection on codon usage. Imagine observing the Markov process of codon substitution at any site (any codon triplet) for an infinitely long time. In a proportion π_I of the time, the wild-type codon at the site in the population is codon I. The mutation (from codon I) to codon J, which changes the nucleotides i_k into j_k at codon position k and which has scaled fitness $S_{IJ} = F_J - F_I$, occurs at the rate $\mu_{i_k j_k}$. Averaged over time, the proportion of the $I \to J$ mutation among all mutations is

$$m_{IJ} = \frac{\pi_I \mu_{i_k j_k}}{\sum_{I \neq J} \pi_I \mu_{i_k j_k}} = \frac{\pi_I a_{i_k j_k} \pi^*_{j_k}}{\sum_{I \neq J} \pi_I a_{i_k j_k} \pi^*_{j_k}}, \quad (5)$$

where the sum in the denominator is over all pairs of codons I and J with $I \neq J$.

One may then calculate the proportion of advantageous mutations among all mutations as

$$P_+ = \sum_{I \neq J} m_{IJ} \mathbb{I}_{S_{IJ}>0}, \quad (6)$$

where the indicator function $\mathbb{I}_{S_{IJ}>0} = 1$ if $S_{IJ} > 0$ or $= 0$ if otherwise. Similarly, the proportion of deleterious mutations among all mutations is

$$P_- = \sum_{I \neq J} m_{IJ} \mathbb{I}_{S_{IJ}<0} = 1 - P_+. \quad (7)$$

The strength of positive selection on an average advantageous mutation may be measured by

$$\bar{S}_+ = \sum_{I \neq J} m^+_{IJ} S_{IJ} \mathbb{I}_{S_{IJ}>0}, \quad (8)$$

where

$$m^+_{IJ} = \frac{\pi_I \mu_{i_k j_k} \mathbb{I}_{S_{IJ}>0}}{\sum_{I \neq J} \pi_I \mu_{i_k j_k} \mathbb{I}_{S_{IJ}>0}}, \quad (9)$$

is the proportion of the $I \to J$ mutation among all advantageous mutations. Here m^+_{IJ} is defined only if the $I \to J$ mutation is advantageous, with $S_{IJ} > 0$. Similarly, the strength of negative selection may be measured by the average S_{IJ} among deleterious mutations with $S_{IJ} < 0$.

One may also calculate the proportion of advantageous mutations among all "substitutions," that is, among those mutations that have passed the filtering by natural selection. This can be calculated using equation (6), with the proportion m_{IJ} calculated using equation (5) but with $\pi_I \mu_{i_k j_k}$ replaced by $\pi_I \mu_{i_k j_k} h(S_{IJ})$ or $\pi_I q_{IJ}$ (eq. 2). Because the substitution process is reversible, the proportion of advantageous mutations among substitutions is exactly $\frac{1}{2}$.

An Approximate Implementation

In the FMutSel and FMutSel0 models, the codon fitness and amino acid fitness parameters are estimated by numerical optimization under ML. We also implement approximate versions of these models by fixing the predicted codon or amino acid frequencies to the observed frequencies in the sequence data. These are referred to as "FMutSel-F" and "FMutSel0-F," respectively. This strategy reduces the number of parameters to be estimated by numerical iteration by 60 under FMutSel-F for the universal genetic code and by 19 under FMutSel0-F. Early models concerning codon usage, such as F1 × 4, F3 × 4, and Fcodon, were all implemented using the observed base or codon frequencies as parameter estimates (Yang 1997). For fair comparison, they are now also implemented using proper numerical optimization of the frequency parameters. Models implemented using the approximation are referred to using the suffix "-F" (e.g., F1 × 4-F).

Analysis of Real Data

We analyze 2 sets of data. The first consists of the mitochondrial genes of the human (GenBank accession number D38112) and the chimpanzee (D38113) of Horai et al. (1995). The 12 protein-coding genes on the same strand of the genome are concatenated into one "supergene," with 3,569 codons in the alignment. The data were analyzed previously by Hasegawa et al. (1998). We fit both the new models implemented in this paper and many old models implemented in the CODEML program (Yang 1997). Several distances between the 2 sequences are calculated under different models, and our objective in this analysis is to examine the impact of model assumptions concerning codon usage on distance estimation.

The second set of data consists of the 5,639 human–chimpanzee–macaque–mouse–rat quintet alignments of orthologous genes from the macaque genome-sequencing project (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). Codons that had alignment gaps in at least one species are removed. The data were analyzed as the primate pair of human and macaque genes, the rodent pair of mouse and rat genes, as well as the quintet including all 5 species. Our objectives in those analyses are to conduct the LRT of neutral evolution at silent sites and to estimate the coefficients of selection acting on codon usage.

Effects of the Model of Codon Usage on Distance Estimation

The log-likelihood values and estimates of sequence distances are shown in table 1 for the human and chimpanzee mitochondrial data set. The assumed mutation model is HKY, but different models are used concerning codon usage. The F1 × 4, F3 × 4, and Fcodon models specify the codon-substitution rate to be proportional to the frequency of the target codon, with the codon frequencies calculated using the 4 nucleotide frequencies (F1 × 4), the nucleotide frequencies at the 3 codon positions (F3 × 4), or with all codon frequencies treated as free parameters (Fcodon) (Yang 1997). The F1 × 4MG model was proposed by Muse and Gaut (1994) and assumes that the codon-substitution rate is proportional to the frequency of the target nucleotide. F3 × 4MG is an extension of F1 × 4MG and uses different base frequencies at the 3 codon positions. F1 × 4MG and F1 × 4 predict the same equilibrium codon frequencies, as do F3 × 4MG and F3 × 4.

Table 1
Estimates of Parameters between the Human and Chimpanzee Mitochondrial Genes under Different Models

Model	p	p'	$\hat{t}$	$\hat{\omega}$	ℓ	$\hat{d}_N$	$\hat{d}_S$	$\hat{d}^*_N$	$\hat{d}^*_S$	$\hat{d}_{3B}$	$\hat{d}_4$
F1 × 4-F	6	3	0.488	0.063	−17,190.2	0.026	0.414	0.022	0.590	0.426	0.442
F1 × 4	6	6	0.482	0.062	−17,172.5	0.025	0.408	0.022	0.590	0.424	0.424
F1 × 4MG-F	6	3	0.431	0.075	−17,132.0	0.026	0.355	0.023	0.513	0.366	0.364
F1 × 4MG	6	6	0.431	0.072	−17,115.5	0.026	0.357	0.022	0.521	0.370	0.367
F3 × 4-F	12	3	0.534	0.039	−16,565.2	0.024	0.629	0.024	0.631	0.418	0.416
F3 × 4	12	12	0.491	0.041	−16,538.6	0.023	0.578	0.023	0.582	0.387	0.380
F3 × 4MG-F	12	3	0.460	0.051	−16,503.5	0.025	0.485	0.024	0.535	0.380	0.376
F3 × 4MG	12	12	0.449	0.051	−16,487.2	0.024	0.475	0.023	0.525	0.370	0.366
Fcodon-F	62	3	0.603	0.040	−16,186.6	0.025	0.613	0.023	0.718	0.477	0.441
Fcodon	62	62	0.553	0.042	−16,175.6	0.024	0.577	0.023	0.656	0.443	0.409
FMutSel0-F	25	6	0.457	0.070	−16,233.9	0.027	0.383	0.023	0.532	0.376	0.369
FMutSel0	25	25	0.457	0.070	−16,233.9	0.027	0.382	0.023	0.533	0.376	0.369
FMutSel-F	65	6	0.490	0.054	−16,093.3	0.025	0.467	0.023	0.570	0.424	0.405
FMutSel	65	65	0.490	0.054	−16,093.2	0.025	0.467	0.023	0.570	0.414	0.405

NOTE.—p is the number of parameters, whereas p' is the number of parameters estimated by ML iteration. The -F models use observed base or codon frequencies as parameter estimates. Estimates of κ range from 22 to 34 among models.

The new FMutSel model has a much higher log-likelihood value than all the old models, indicating better fit to the data. Note that except for F1 × 4MG, which is equivalent to FMutSel with all codons having the same fitness, none of the other old models are nested within FMutSel and the χ^2 distribution cannot be used to compare them. However, use of the Akaike information criterion (Akaike 1974) leads to clear preference of FMutSel over all old models (table 1). Besides the better fit, we emphasize the better explanatory power of the new model.

We are interested in whether model assumptions concerning codon usage affect estimation of the distances between 2 protein-coding genes. The familiar nonsynonymous and synonymous distances d_N and d_S are calculated according to Goldman and Yang (1994). Previous studies have found that those distances are sensitive to assumptions about codon usage (e.g., Yang and Nielsen 1998, 2000). Estimates of d_N are very similar among models, but estimates of d_S vary considerably. Estimates of the ω ratio differ by 2-folds among models. Nevertheless, the new FMutSel model produced estimates that are within the range of the old estimates. The estimates of ω under the commonly used F3 × 4 and Fcodon models are slightly smaller than that under FMutSel.

Distances $\hat{d}^*_N$ and $\hat{d}^*_S$ are the number of nonsynonymous substitutions per nonsynonymous site and the number of synonymous substitutions per synonymous site, respectively, based on the "physical site" definition of sites (Yang 2006: eq. 2.20). These distances are more stable across models, as noted previously. $\hat{d}_{3B}$ is the number of nucleotide substitutions per site at the third codon position before selection on the protein, whereas $\hat{d}_4$ is the number of nucleotide substitutions per 4-fold degenerate site, estimated from the codon model under ML (Yang 2006, p. 63–64). Distances d_{3B} and d_4 are very similar to each other and their estimates are also similar among different models of codon usage (table 1). See Yang (2006) and Bierne and Eyre-Walker (2003) for a discussion of those distances in analysis of codon usage bias.

Overall, estimates of sequence distances and ω ratio under the old models, especially models F3 × 4 and Fcodon, are similar to estimates under the new FMutSel model. We also note that FMutSel produced almost identical results to FMutSel-F, indicating that the approximation of fixing the equilibrium codon frequencies at their observed values worked well in the data set. FMutSel-F has a big computational advantage and may be useful in real data analysis.

Test of Selection on Synonymous Codon Usage

We applied the LRT of neutral evolution of codon usage to nuclear genes from the mammalian species. The FMutSel and FMutSel0 models are fitted to each of the 5,639 genes for the human–macaque pair, the mouse–rat pair, and the 5-species quintet. The histograms of the log-likelihood difference between the 2 models ($\Delta\ell$) are shown in figure 1. Table 3 lists the number and proportion of genes in which the LRT is significant. At the 5% level, the null hypothesis of neutral evolution is rejected in 87%, 90%, and 94% of genes for the primate pair, the rodent pair, and the quintet, respectively. The differences in the proportions appear to reflect the information content in the data sets rather than any real biological differences between primates and rodents. The mouse–rat pair is more divergent than the human–macaque pair so that the data are more informative and the test has higher power. Similarly, the quintet data are most informative so that the null hypothesis is rejected in the greatest number of genes. The analysis thus provides statistical evidence that synonymous codon usage in most genes is influenced by natural selection. Nevertheless, the LRT may be sensitive to the mutation model assumed in the FMutSel and FMutSel0 models, and we suggest caution should be exercised in interpreting those results (see Discussion).

We also conducted the LRT by comparing FMutSel0-F against FMutSel-F, using the approximation of fixing equilibrium codon frequencies at their observed values. This approximate test produced very similar results to those of figure 1. The test statistics ($\Delta\ell$) calculated using the 2 procedures are plotted against each other in figure 2 for the quintet data sets.

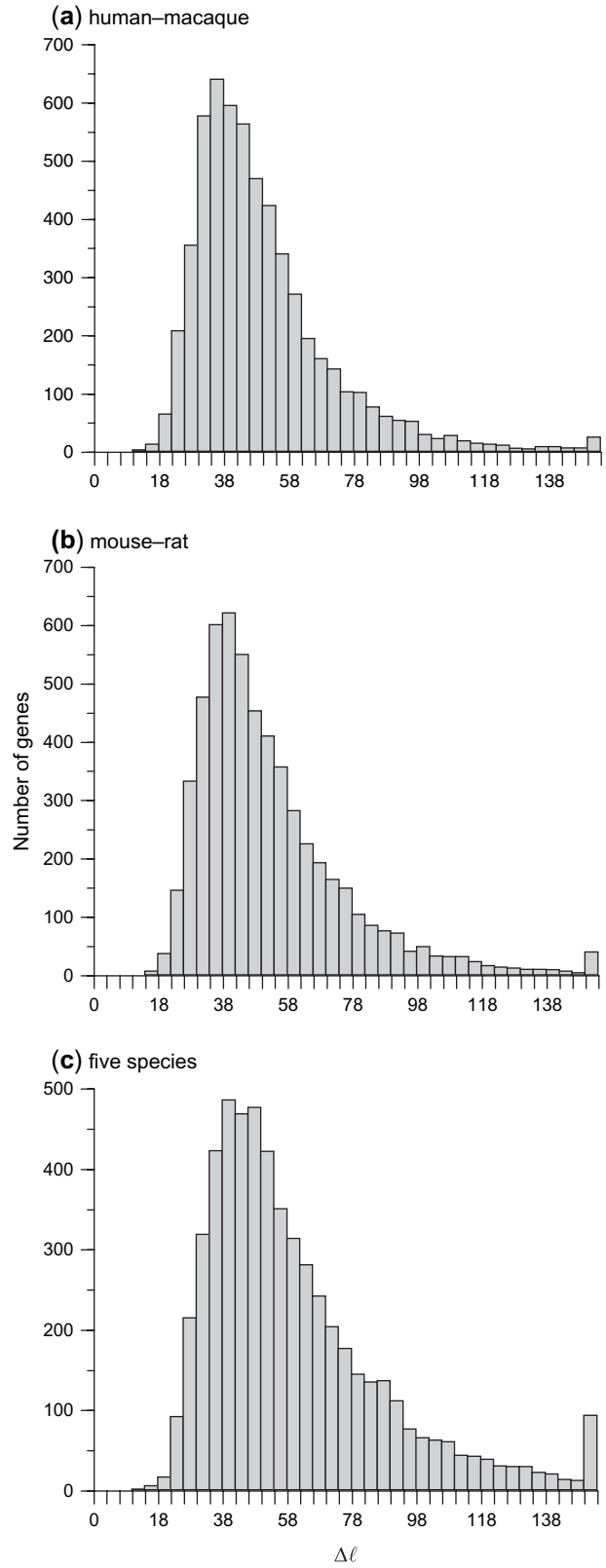

FIG. 1.—Histograms of the log-likelihood difference ($\Delta\ell$) for test of selection on codon usage for (*a*) the human–macaque genes, (*b*) the mouse–rat genes, and (*c*) the quintet of all 5 species. Values greater than 150 are grouped into the last bin. As $2\Delta\ell$ is asymptotically distributed as χ^2_{41} under the null model, the critical values for $\Delta\ell$ are 28.47 and 32.48 at the 5% and 1% levels, respectively.

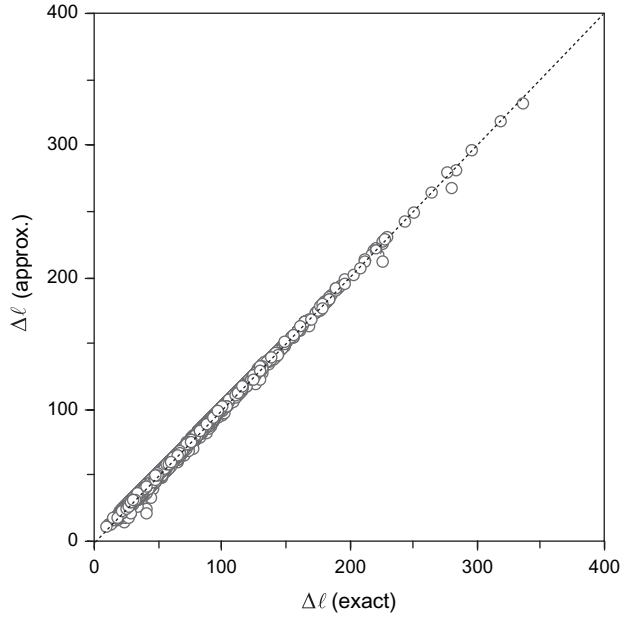

FIG. 2.—The log-likelihood difference ($\Delta\ell$) for test of selection on codon usage when the codon frequency parameters are estimated by ML iteration (exact) or by fixing them at the observed values. The 5-species mammalian genes are analyzed.

The Distribution of Selection Coefficients

We used the FMutSel model to calculate the proportions of mutations with different selective coefficients (S), generating an estimation of the distribution of S among new mutations. For this analysis, we use 4 large data sets: the concatenated mitochondrial genes from the human and chimpanzee and the concatenated nuclear genes for the human–macaque pair, the mouse–rat pair, and the quintet. We used both model M0 (1-ratio), which assumes the same ω ratio for all codons, and M3 (discrete), which assumes 2 site classes in proportions p_0 and p_1 with different ω ratios ω_0 and ω_1 (Yang et al. 2000). The results are shown in table 2. The log-likelihood values under models M0 (1-ratio) and M3 (discrete) are hugely different, indicating that the ω ratio is highly variable among codons. Nevertheless, estimates of the mutation bias parameters ($\pi^*_T, \pi^*_C, \pi^*_A$) and codon fitness parameters (not shown) are very similar between the 2 models in each of the 4 large data sets (table 2).

We used parameter estimates obtained under model M0 (1-ratio) to calculate the scaled selective coefficients (S) for mutations that involve 2 codons differing at exactly one position and thus have nonzero rates. Those are the possible mutations allowed by the model, and their probabilities of occurrences are given by equation (5). There are 526 and 508 such mutations (codon pairs) for the universal and mitochondrial codes, respectively. The S values for those mutations were binned into 21 bins to generate a histogram, with the mid value in each bin used as the representative for that bin and with the proportion for the bin calculated as the sum of proportions (m_{IJ} in eq. 5) of all mutations falling into that bin. The results are shown in figure 3*a*. The proportion (P_+) of advantageous mutations among all mutations is shown in table 2, as well as the average selective coefficients of advantageous and deleterious mutations ($\bar{S}_+$and $\bar{S}_-$).

Table 2
Parameter Estimates under the Mutation-Selection (FMutSel) Model in 4 Concatenated Data Sets

Data Sets	p	$\hat{b}$	$\hat{\kappa}$	$\hat{\omega}$	$\hat{\pi}^*_T$	$\hat{\pi}^*_C$	$\hat{\pi}^*_A$	$\hat{\pi}^*_G$	ℓ	P_+	$\bar{S}_+$	$\bar{S}_-$
Human–chimpanzee mitochondria												
M0 (1-ratio)	65	0.490	29.0	0.054	0.213	0.186	0.419	0.182	−16,093.2	0.319	0.612	−0.916
M3 (discrete)	67	0.501	28.7	$\hat{p}_0 = 0.848, \hat{p}_1 = 0.151$ $\hat{\omega}_0 = 0.006, \hat{\omega}_1 = 0.374$	0.222	0.181	0.363	0.234	−16,079.9	0.306	0.645	−1.017
Human–macaque												
M0 (1-ratio)	66	0.072	4.1	0.149	0.169	0.316	0.227	0.288	−10,095,235.6	0.370	0.435	−0.648
M3 (discrete)	68	0.076	4.2	$\hat{p}_0 = 0.997, \hat{p}_1 = 0.003$ $\hat{\omega}_0 = 0.125, \hat{\omega}_1 = 17.2$	0.167	0.319	0.225	0.290	−10,091,520.6	0.368	0.444	−0.662
Mouse–Rat												
M0 (1-ratio)	66	0.195	3.5	0.111	0.190	0.284	0.275	0.252	−10,903,772.7	0.385	0.390	−0.562
M3 (discrete)	68	0.199	3.6	$\hat{p}_0 = 0.982, \hat{p}_1 = 0.018$ $\hat{\omega}_0 = 0.079, \hat{\omega}_1 = 2.623$	0.189	0.284	0.274	0.253	−10,897,867.4	0.385	0.390	−0.563
5 species												
M0 (1-ratio)	72	0.657	3.2	0.119	0.183	0.287	0.262	0.268	−14,743,509.9	0.385	0.393	−0.565
M3 (discrete)	74	0.676	3.3	$\hat{p}_0 = 0.882, \hat{p}_1 = 0.118$ $\hat{\omega}_0 = 0.040, \hat{\omega}_1 = 0.851$	0.184	0.284	0.261	0.271	−14,683,601.6	0.386	0.388	−0.560

NOTE.—p is the number of parameters in the model. $\hat{b}$ is the distance between 2 sequences or the tree length for the 5-species data, measured by the expected number of nucleotide substitutions per codon. Estimates of the 60 (for the mitochondrial data) or 61 (for nuclear genes) codon fitness parameters are not shown. P_+ is the proportion of advantageous mutations. $\bar{S}_+$ and $\bar{S}_-$ are the average selection coefficients of advantageous and deleterious mutations, respectively.

Because preferred codons with higher fitness are more common and most mutations lead to unpreferred codons with lower fitness, the distribution of S among new mutations is skewed to the left, with the proportion $P_+ < \frac{1}{2}$. The proportion of advantageous mutations among substitutions is higher than P_+ because an advantageous mutation has a higher fixation probability and makes a greater contribution to substitutions than does a deleterious mutation. Indeed, the proportions of advantageous mutations among substitutions is $\frac{1}{2}$, due to the reversibility of the substitution model.

The estimates of $\bar{S}_+$ and $\bar{S}_-$ are greater and thus selection on silent sites is stronger in the mitochondrial genes than in the nuclear genes (table 2). In the former, ~31% of new mutations are advantageous, whereas in the latter, the proportion is 37–40%. The much lower ω ratios in the mitochondrial genes than in the nuclear genes indicate that the mitochondrial proteins are under much stronger selective constraint than the nuclear proteins. The difference is more striking when one considers the fact that the effective population size for mitochondrial genes is $\sim \frac{1}{4}$ that of the nuclear genes and that selection is less efficient in smaller populations. The higher efficiency of selection in mtDNA, with respect to both codon usage and protein evolution, may be due to the fact that the haploid mitochondrial genome makes it easy to remove recessive mutations, whereas they may remain hidden in the heterozygous state in nuclear genes. Another possible explanation is the hypothesis of selection for translational accuracy, which predicts stronger selection on codon usage on highly conserved proteins or on highly conserved amino acids in a protein because the fitness cost of translational misincorporation should depend on how the amino acid change affects protein function (Akashi 1994). If mitochondrial genes perform crucial biological functions and are more highly expressed than nuclear genes, this hypothesis may explain both the stronger selection on protein evolution and the stronger selection on codon usage.

It should be noted that in our model, all S values are nonzero, and P_+ in table 2 includes mutations with S only very slightly positive, the evolutionary dynamics of which may be indistinguishable from that of neutral mutations. For example, mutations with $|S| > 2$ are rare in all data sets. The estimated proportions of mutations with $S > 2$ and $S < -2$ are 0.2% and 1.7%, respectively, for the mitochondrial genes, 0.2% and 1.6% for the human–macaque pair, 0.1% and 1.0% for the mouse–rat pair, and 0.1% and 0.9% for the quintet. Thus, although the LRT rejects the null model of neutral evolution of silent sites, selection on codon usage is mostly weak, and most mutations appear to be nearly neutral with respect to selection on codon usage.

We are also interested in how natural selection on codon usage changes the fitness distribution of mutations, that is, how mutations of different fitness contribute to substitutions. A histogram of S after filtering by natural selection on codon usage can be generated using the same procedure as described above, except that the proportion m_{IJ} is calculated using equation (5), with $\pi_I \mu_{i_k j_k}$ replaced by $\pi_I \mu_{i_k j_k} h(S_{IJ})$. The resulting histograms (fig. 3*b*) show the proportion of mutations with scaled fitness S that has survived natural selection on codon usage. Similarly, If we replace $\pi_I \mu_{i_k j_k}$ by $\pi_I q_{IJ}$ in equation (5), the resulting histograms (fig. 3*c*) represent the proportion of mutations with fitness S among observed substitutions, that is, among mutations that have passed the filtering by selection both on codon usage bias and on amino acid replacements. Because of the detailed balance condition of the reversible Markov model of substitution, the distributions in figure 3*b* and *c* are all symmetrical. Note that here the distinction between selection on codon usage and selection on amino acid replacements is more conceptual than temporal, with no implication that one necessarily occurs before the other.

Discussion

Mechanistic Models of Codon Usage and Protein Evolution

A number of authors have studied the frequencies of synonymous codons for 2-fold degenerate amino acids as

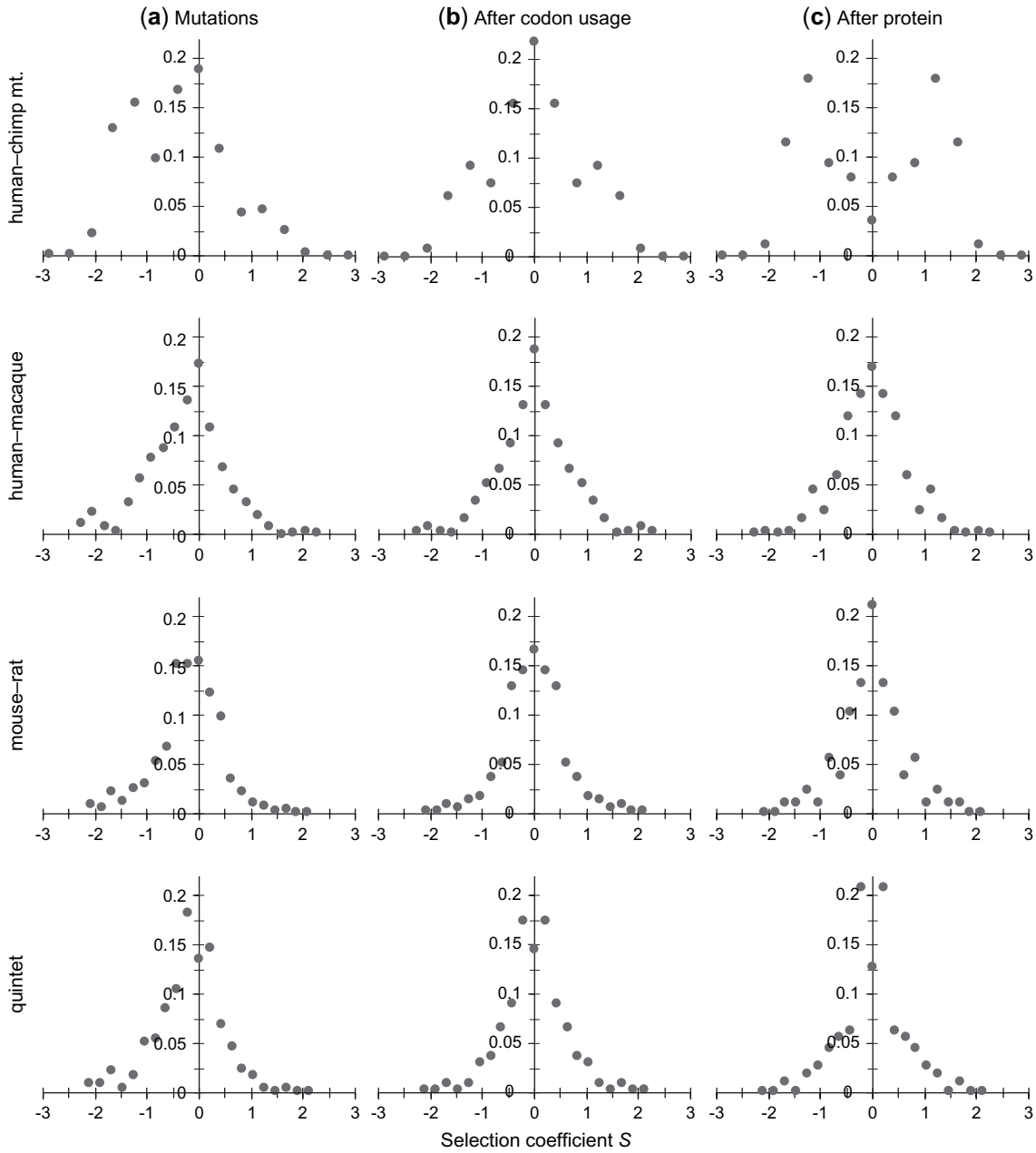

FIG. 3.—Estimated distributions of selection coefficient $S = 2Ns$ from 4 data sets: concatenated human–chimpanzee mitochondrial genes, concatenated human–macaque nuclear genes, concatenated mouse–rat nuclear genes, and concatenated data for all 5 mammalian species. The histograms show the proportion of mutations with scaled selection coefficient S (*a*) among all mutations, (*b*) after filtering by natural selection on codon usage, and (*c*) after filtering by selection on both codon usage and on amino acid replacements. Model M0 (1-ratio) is used, with the same ω ratio for all nonsynonymous changes. Parameter estimates are shown in table 2.

the result of interactions between mutation, genetic drift, and natural selection (Kimura 1983; Li 1987; Bulmer 1991; McVean and Charlesworth 1999). Let the 2 alleles be 1 (preferred codon) and 0 (unpreferred codon), with the mutation rate from 0 to 1 to be μ_1 and that in the reverse direction be μ_0. Suppose that the 2 alleles have fitness f_0 and f_1 so that the selection coefficient of the $0 \rightarrow 1$ mutation in the allele-0 population is $s = f_1 - f_0$ and that of the $1 \rightarrow 0$ mutation in the allele-1 population is $-s$. At mutation-selection-drift equilibrium, the probability density of the frequency p of allele 1 is given as

$$f(p) \propto e^{2Nsp} p^{2N\mu_1 - 1}(1-p)^{2N\mu_0 - 1}, \quad (10)$$

(Wright 1931). This theory can be used to analyze codon usage in a single species, under the assumption that one of the alleles is fixed. The probability that the population is fixed at the preferred codon can be obtained by integrating the density $f(p)$ from $1 - 1/N$ to 1 (e.g., Li 1987), as

$$\pi_i \propto \mu_i e^{2Nf_i}, i = 0, 1, \quad (11)$$

where the proportionality constant is determined to ensure that $\pi_0 + \pi_1 = 1$. If we assume that the same selective pressure applies to synonymous codons for all 2-fold degenerate amino acids in a gene, π_1 will be the proportion of preferred codons in the gene. The contributions of mutation and selection to the equilibrium frequencies of synonymous codons are apparent from equation (11). This may also be considered a special case of equation (4), which gives the equilibrium distribution of the codon-substitution process.

McVean and Vieira (1999) used equation (11) to analyze observed counts of preferred codons for 2-fold amino acids in several *Drosophila* species, fitting binomial models by ML. The analysis used information on codon usage but ignored differences between species. McVean and Vieira (2001) implemented a population genetics model that is very similar to equation (2) to describe substitutions between synonymous codons between species. The authors analyzed between-species synonymous differences to estimate the strength of natural selection on synonymous codon usage, with nonsynonymous differences ignored. The FMutSel models extend the work of McVean and Vieira to a full codon substitution model, which is suitable for comparative analysis of protein-coding genes from multiple species.

Previous models of codon substitution (Goldman and Yang 1994; Muse and Gaut 1994) aim to describe nucleotide substitutions and do not explicitly accommodate mutation bias and natural selection acting on the DNA level. The models may thus be ill suited for studying the forces and mechanisms of the evolutionary process at silent sites. The mutation-selection models implemented in this paper address this drawback, by introducing parameters that explicitly describe mutation bias and natural selection acting on codon usage. We suggest that such models, with the easy interpretation of the model parameters, may be very useful for studying the process of molecular sequence evolution.

There has been considerable interest in incorporating fitness effects of new mutations in constructing substitution models for phylogenetic analysis. Halpern and Bruno (1998) considered a codon-substitution model in which at every amino acid site in the protein, different amino acids have different fitness and thus different equilibrium frequencies. The model was developed for distance calculation but is not practical for real data analysis due to its use of too many parameters. Moses et al. (2003) adapted the theory to describe nucleotide substitutions and to estimate site-specific substitution rates in noncoding regulatory elements such as transcription factor–binding sites. Note that from equation (4), we have

$$\frac{\pi_I}{\pi_J} \times \frac{\mu_{i_k j_k}}{\mu_{j_k i_k}} = \frac{\pi^*_{i_k} e^{F_I}}{\pi^*_{j_k} e^{F_J}} \times \frac{a_{i_k j_k} \pi^*_{j_k}}{a_{j_k i_k} \pi^*_{i_k}} = e^{F_I - F_J}, \qquad (12)$$

from which equation (9) of Halpern and Bruno (1998) can be seen to equal $h(S_{IJ})$ in equation (1), with $h(S_{IJ}) = 1$ for $S_{IJ} = 0$. Thus, the underlying population genetics theory is the same although the applications are very different. Note that given a reversible mutation model such as HKY or GTR, reversibility of codon substitution is a natural property of the model and not an additional assumption, as made by Halpern and Bruno (1998) and Moses et al. (2003).

The FMutSel model also has similarities to the site-class models of amino acid replacement implemented by Koshi et al. (1999), which assume that different site classes have different amino acid frequencies and different substitution patterns and that in each site class, every amino acid J has its own "propensity" F_J. Koshi et al. (1999, eq. 4) applied a truncation on the substitution rate, equivalent to fixing $h(S_{IJ}) = 1$ whenever the difference in propensity $S_{IJ} = F_J - F_I > 0$. Like FMutSel, this model is also time reversible, with the same equilibrium distribution, where the frequency of amino acid J is proportional to e^{F_J}. Except for the truncation mentioned above, the model of Koshi et al. (1999) can be given a population genetics interpretation, with the propensity interpreted as the scaled fitness F_J. However, the truncation of rates means that the model assumes that an advantageous mutation is fixed at the same rate as a neutral mutation, which is unrealistic biologically. A similar criticism was made by Thorne et al. (2007).

More recent work by Yu and Thorne (2006), Thorne et al. (2007), and Choi et al. (2007) assigned a fitness to the sequence when they developed mutation-selection models to describe the evolution of RNA or protein sequences. An advantage of those models is that they allow dependence among sites due to RNA or protein structural constraints.

We note that there has been some debate in the literature concerning whether use of the ω ratio to detect natural selection acting on the protein (for reviews, see Yang and Bielawski 2000; Yang 2002) requires the assumption of neutral evolution at silent sites. Many authors take it for granted that this assumption is needed. A concern is that if selection acts on codon usage, codon models may be misled to produce an ω ratio greater than one because selection on silent sites has reduced d_S and not because positive selection has elevated d_N. From the mutation-selection models implemented in this paper, it is clear that the assumption is not necessary and it is possible to use the ω ratio to detect positive selection acting on the protein even if silent sites are under natural selection, as assumed in FMutSel. Comparison between d_S and d_N is a contrast between the rates before and after the action of selection on the protein (Yang 2006, eq. 2.19) so that the comparison is valid whether evolution at silent sites is driven by mutation or selection. In this regard, selection on silent sites may be more accurately described as selection on the DNA level as it affects both silent and replacement sites.

Sensitivity of the LRT to the Mutation Model

The mutation-selection model of codon substitution makes many simplistic assumptions about the evolutionary process. For our purpose of testing for selection acting on silent sites, the most worrying assumptions appear to be those concerning the mutation process as the mutation-bias and codon-fitness parameters are expected to be highly correlated in such an analysis. Indeed, the effects of the 2 would be virtually impossible to separate if we had used only information on codon frequencies (see eqs. 4 and 11).

To examine the impact of the assumed mutation model on the LRT of selection on codon usage, we implemented the GTR mutation model (e.g., Yang 1994). The codon frequency parameters are estimated using the observed

Table 3
Number and Percentage (in Parentheses) of Mammalian Genes for Which the Null Model of Neutral Evolution at Silent Sites Is Rejected

	Significance Level	
Data	5%	1%
HKY mutation model		
Human–macaque	4,909 (87%)	4,336 (77%)
Mouse–rat	5,073 (90%)	4,587 (81%)
5 species	5,282 (94%)	4,945 (88%)
GTR mutation model		
Human–macaque	4,815 (85%)	4,216 (75%)
Mouse–rat	4,988 (88%)	4,479 (79%)
5 species	5,240 (93%)	4,870 (86%)

NOTE.—A total of 5,639 genes are analyzed.

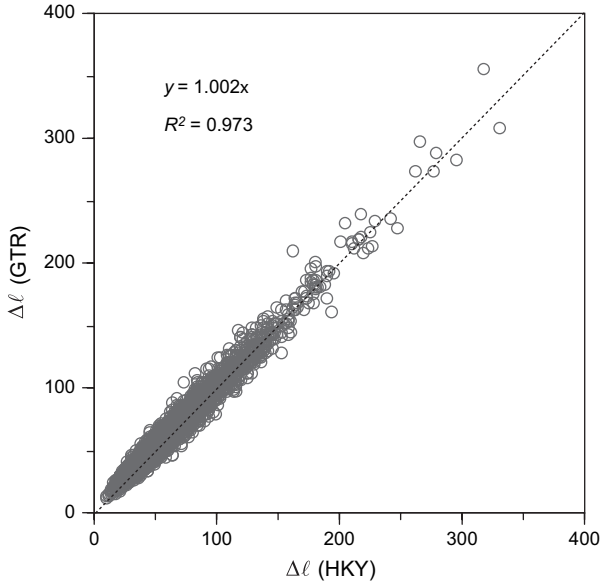

FIG. 4.—The log-likelihood difference ($\Delta\ell$) for test of selection on codon usage when the assumed mutation model is HKY or GTR. The 5-species mammalian genes are analyzed.

frequencies rather than by ML iteration. Application of the LRT under the GTR model to the mammalian data produced results very similar to those obtained under HKY. The proportions of genes for which the LRT is significant under GTR (table 3) are slightly lower (by 1–2%) than under HKY. Figure 4 plots the test statistic ($\Delta\ell$) for the 2 mutation models for the quintet data sets. The results suggest that the LRT may not be very sensitive to the assumed mutation model.

However, the estimates of codon-fitness parameters for the concatenated data under the 2 mutation models are very different (results not shown). This is the case even though both mutation models predicted very similar codon frequency parameters, which closely match the observed frequencies. Our estimates of the selection coefficients are affected by the mutation model. Thus, we found that the LRT is somewhat insensitive to the assumed mutation model but the estimates of codon fitness parameters are.

Both HKY and GTR assume independent mutations at nucleotide sites. There is considerable evidence suggesting that the mutation rate of a nucleotide may depend on neighboring nucleotides (e.g., Bulmer 1986; Hwang and Green 2004; Siepel and Haussler 2004). One well-known example of such context effects is the high mutation rate of CpG dinucleotides in mammalian genomes. As the cytosine in CpG is prone to methylation and deamination, CpG dinucleotides have a very high rate of mutating into TpG (Scarano et al. 1967). With such mutational context effects, both the null and alternative hypotheses (FMutSel0 and FMutSel) in the LRT are violated, but the 2 models may not be affected to the same extent, in which case the violation of assumptions may cause the test to generate excessive false positives. For example, FMutSel0 predicts that the relative frequencies of 4-fold degenerate codons encoding the same amino acid are given by the mutation-bias parameters (π_{j_3}), independent of the encoded amino acid. If the mutation rate and pattern at the third codon position depend on the nucleotides at the first and second positions, FMutSel0 may fit the data poorly, but FMutSel may still achieve a reasonable fit because of its use of a separate codon fitness parameter F_J for each target codon J. Although both FMutSel0 and FMutSel make use of information from nonsynonymous differences as well as synonymous differences, the test may nevertheless be sensitive to such mutational context effects. It has also been suggested that one mutation event may affect multiple nucleotides and the assumption of independent mutations may be unrealistic (e.g., Yang et al. 1998; Whelan and Goldman 2004). However, those studies typically analyze substitutions instead of mutations, and the apparent double or triple substitutions may reflect artifacts of the inadequate substitution model rather than true double or triple mutations. The models developed here concern the mutation process, and it would appear that double or triple mutations, if not rare, should affect the 2 models in similar ways. At any rate, the sensitivity of the LRT to violations of the assumed mutation model is not well understood and merits further research.

We consider several strategies that may alleviate the confounding effect of mutation and selection. The first is to make certain assumptions concerning either the mutation or the selection process. For example, the method of Nielsen et al. (2007) required prior knowledge of preferred and unpreferred codons and also assumed the same selective strength acting on all codons. The latter assumption may be unrealistic in some data sets. A second strategy is to analyze pseudogenes or noncoding DNA to estimate mutation parameters and then use them in the mutation-selection model of codon substitution to analyze coding genes. Similarly, one may analyze coding and neighboring noncoding regions jointly, with the same mutation-bias parameters applied to both regions and the selection parameters applied to the coding regions only. This requires that the same mutation process operates in both coding and noncoding regions, an assumption that may be violated due to translation-coupled repair (Duret 2002). A third strategy, suitable for joint analysis of many genes from the same set of species, is to assume that the mutation parameters are shared among genes or at least among genes with similar codon usage bias or GC content at the third codon positions, whereas the strengths of selection on codon usage may differ among

genes. In this paper, we analyzed the 5,639 mammalian genes separately, fitting 66 or more parameters to each gene, so that the model is rather parameter-rich. Finally, developing models that explicitly accommodate mutational context effects may also be very useful in improving the realism of the models implemented here. In this regard, our likelihood model provides a natural framework for such extensions.

Program Availability

The new FMutSel and FMutSel0 models developed in this paper are implemented independently by the 2 authors for error checking. All models described in this paper are implemented in the CODEML program in PAML 4 (Yang 2007).

Acknowledgments

We thank 3 referees for many useful comments. This study is supported by a grant from the Biotechnological and Biological Sciences Research Council to Z.Y. and grants from FNU (Danish Natural Science Research Council) and Danmarks Grundforskningsfond to R.N.

Literature Cited

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Automat Contr. 19:716–723.

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

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics. 139:1067–1076.

Bennetzen JL, Hall BD. 1982. Codon selection in yeast. J Biol Chem. 257:3026–3031.

Bierne N, Eyre-Walker A. 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics. 165:1587–1597.

Bulmer M. 1986. Neighboring base effects on substitution rates in pseudogenes. Mol Biol Evol. 3:322–329.

Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature. 325:728–730.

Bulmer MG. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics. 129:897–907.

Castillo-Davis CI, Hartl DL. 2002. Genome evolution and developmental constraint in *Caenorhabditis elegans*. Mol Biol Evol. 19:728–735.

Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL. 2007. Quantifying the impact of protein tertiary structure on molecular evolution. Mol Biol Evol. 24:1769–1782.

Dunn KA, Bielawski JP, Yang Z. 2001. Substitution rates in *Drosophila* nuclear genes: implications for translational selection. Genetics. 157:295–305.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12:640–649.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, Arabidopsis. Proc Natl Acad Sci USA. 96:4482–4487.

Fisher R. 1930. The distribution of gene ratios for rare mutations. Proc R Soc Edinb. 50:205–220.

Frydman J. 2001. Folding of newly translated proteins in vivo: the role of molecular chaperones. Annu Rev Biochem. 70:603–647.

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

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 15:910–917.

Hasegawa M, Cao Y, Yang Z. 1998. Preponderance of slightly deleterious polymorphism in mitochondrial DNA: replacement/synonymous rate ratio is much higher within species than between species. Mol Biol Evol. 15:1499–1505.

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

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA. 92:532–536.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA. 101:13994–14001.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol. 146:1–21.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2:13–34.

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol. 53:290–298.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science. 315:525–528.

Kimura M. 1957. Some problems of stochastic processes in genetics. Ann Math Stat. 28:882–901.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Koshi JM, Mindell DP, Goldstein RA. 1999. Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. Mol Biol Evol. 16:173–179.

Kreitman M, Akashi H. 1995. Molecular evidence for natural selection. Annu Rev Ecol Syst. 26:403–422.

Kurland CG. 1992. Translational accuracy and the fitness of bacteria. Annu Rev Genet. 26:29–50.

Li W-H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol. 24:337–345.

McVean GA, Charlesworth B. 1999. A population genetics model for the evolution of synonymous codon usage: patterns and predictions. Genet Res. 74:145–158.

McVean GA, Vieira J. 1999. The evolution of codon preferences in *Drosophila*: a maximum-likelihood approach to parameter estimation and hypothesis testing. J Mol Evol. 49:63–75.

McVean GA, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics. 157:245–257.

Moriyama EN, Powell JR. 1997. Codon usage bias and tRNA abundance in *Drosophila*. J Mol Evol. 45:514–523.

Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. 2003. Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol Biol. 3:19.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11:715–724.
Musto H, Cruveiller S, D'Onofrio G, Romero H, Bernardi G. 2001. Translational selection on codon usage in *Xenopus laevis*. Mol Biol Evol. 18:1703–1707.
Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. Mol Biol Evol. 24:228–235.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.
Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the Rhesus macaque genome. Science. 316:222–234.
Scarano E, Iaccarino M, Grippo P, Parisi E. 1967. The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. Proc Natl Acad Sci USA. 57:1394–1400.
Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. 1995. DNA sequence evolution: the sounds of silence. Philos Trans R Soc Lond B Biol Sci. 349:241–247.
Sharp PM, Li WH. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol. 4:222–230.
Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol. 21:468–488.
Thorne JL, Choi SC, Yu J, Higgs PG, Kishino H. 2007. Population genetics without intraspecific data. Mol Biol Evol. 24:1667–1677.
Whelan S, Goldman N. 2004. Estimating the frequency of events that cause multiple-nucleotide changes. Genetics. 167:2027–2043.
Wright S. 1931. Evolution in Mendelian populations. Genetics. 16:97–159.
Yang Z. 1994. Estimating the pattern of nucleotide substitution. J Mol Evol. 39:105–111.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13: 555–556.
Yang Z. 2002. Inference of selection from multiple species alignments. Curr Opin Genet Dev. 12:688–694.
Yang Z. 2006. Computational molecular evolution. Oxford (UK): Oxford University Press.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.
Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 46: 409–418.
Yang Z, Nielsen R. 2000. Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.
Yang Z, Nielsen R, Goldman N, Pedersen A-MK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.
Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol. 15:1600–1611.
Yu J, Thorne JL. 2006. Dependence among sites in RNA evolution. Mol Biol Evol. 23:1525–1537.

Jeffrey Thorne, Associate Editor

Accepted December 19, 2007