Detecting Positive and Purifying Selection at Synonymous Sites in Yeast and Worm

Tong Zhou,[1,2] Wanjun Gu,[3] and Claus O. Wilke*[,1,2,4]

[1]Center for Computational Biology and Bioinformatics, University of Texas at Austin
[2]Section of Integrative Biology, University of Texas at Austin
[3]Key Laboratory of Child Development and Learning Science of Ministry of Education of China, Southeast University, Nanjing, Jiangsu, China
[4]Institute for Cell and Molecular Biology, University of Texas at Austin

***Corresponding author:** E-mail: cwilke@mail.utexas.edu.

Associate editor: Koichiro Tamura

Research article

Abstract

We present a new computational method to identify positive and purifying selection at synonymous sites in yeast and worm. We define synonymous substitutions that change codons from preferred to unpreferred or vice versa as nonconservative synonymous substitutions and all other substitutions as conservative. Using a maximum-likelihood framework, we then test whether conservative and nonconservative synonymous substitutions occur at equal rates. Our approach replaces the standard rate of synonymous substitutions per synonymous site, dS, with two new rates, the conservative synonymous substitution rate (dS_C) and the nonconservative synonymous substitution rate (dS_N). Based on the ratio dS_N/dS_C, we find that 0.05% of all yeast genes and none of worm genes show evidence of positive selection at synonymous sites ($dS_N/dS_C > 1$). On the other hand, 9.44% of all yeast genes and 5.12% of all worm genes show evidence of significant purifying selection on synonymous sites ($dS_N/dS_C < 1$). We also find that dS_N correlates strongly with gene expression level, whereas the correlation between expression level and dS_C is very weak. Thus, dS_N captures most of the signal of selection for translational accuracy and speed, whereas dS_C is not strongly influenced by this selection pressure. We suggest that the ratio dN/dS_C may be more appropriate than the ratio dN/dS to identify positive or purifying selection on amino acids.

Key words: codon usage bias, preferred codon, positive selection, synonymous substitution rate, translational selection.

Introduction

The rate of evolution in protein-coding genes is commonly assessed with the two quantities dN (rate of nonsynonymous substitutions per nonsynonymous site, also called K_a) and dS (rate of synonymous substitutions per synonymous site, also called K_s). If synonymous evolution is neutral, then the ratio of dN/dS identifies the type of selection pressure acting on a gene. $dN/dS \ll 1$ indicates strong purifying selection, $dN/dS \gg 1$ indicates positive selection, and $dN/dS \sim 1$ implies that amino acids evolve largely neutrally (Nielsen and Yang 1998; Suzuki and Gojobori 1999; Hurst 2002; Koonin and Rogozin 2003; Bustamante et al. 2005; Yang et al. 2005; Petersen et al. 2007). But synonymous substitutions are not neutral. Selection for translational efficiency and accuracy operates from bacteria to mammals and shapes codon usage bias (Ikemura 1981; Sharp et al. 1986; Akashi 1994; Stenico et al. 1994; Drummond et al. 2006; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008; Higgs and Ran 2008; Zhou et al. 2009). Other selective pressures on synonymous sites relate to splicing (Chamary and Hurst 2005a; Dewey et al. 2006; Parmley et al. 2006; Warnecke and Hurst 2007) and to DNA and messenger RNA (mRNA) secondary structure and stability (Vinogradov 2003; Chamary and Hurst 2005b; Hoede et al. 2006; Stoletzki 2008).

Because selection on synonymous sites is widespread, several authors have developed methods to infer the strength of selection on synonymous sites and/or to correct dN/dS ratios for selection on synonymous sites. McVean and Vieira (2001) proposed a maximum-likelihood method to infer the strength of selection on different codons within each codon family. More recently, Yang and Nielsen (2008) developed a similar but more general approach that can estimate selection pressures on both synonymous and nonsynonymous mutations at the same time. The downside to their approach is that a large number of parameters are being estimated from the data. Thus, reliable estimates can be obtained only for very large data sets. An alternative approach, proposed by Nielsen et al. (2007), takes into account prior knowledge on codon bias and estimates only the overall strength of selection against unpreferred codons. Other approaches include comparing dS with the substitution rate in introns, dI (Resch et al. 2007), regressing dS against codon bias (Hirsh et al. 2005) or using other corrections based on codon usage frequencies (Liberles 2001) or testing the asymmetry between high- and low-expression genes (Higgs et al. 2007).

Here, we develop a novel method to both assess the amount of selection on synonymous sites and to derive improved dN/dS estimates that are less affected by selection on synonymous sites. We introduce two novel evolutionary rates, the rate of conservative synonymous substitutions per conservative synonymous site, dS_C, and the rate of nonconservative synonymous substitutions per nonconservative synonymous site, dS_N. Conservative synonymous substitutions are synonymous substitutions that connect

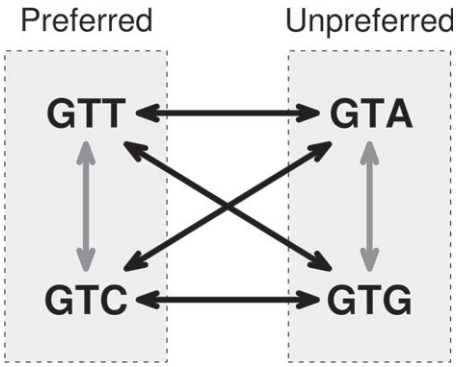

FIG. 1. Illustration of conservative and nonconservative synonymous substitutions for the amino acid valine in yeast. The black arrows represent all possible nonconservative synonymous substitutions and the grey arrows represent all possible conservative synonymous substitutions.

two preferred or two unpreferred codons. All other synonymous substitutions are nonconservative. Our approach provides two novel ratios, dS_N/dS_C as a measure of selection on synonymous sites and dN/dS_C as a measure of selection on nonsynonymous sites. We apply our model to two species known to experience strong selection for codon usage, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. We also consider *Drosophila melanogaster*, but because fly violates some of our model assumptions (see Discussion), all fly results are relegated to Supplementary Material online.

Our approach is related to the method of Nielsen et al. (2007), but there are three differences: 1) We treat selection on synonymous and nonsynonymous sites conceptually the same by calculating evolutionary rate ratios for both; Nielsen et al. (2007) calculated a selection coefficient for the former and an evolutionary rate ratio for the latter; 2) we assume that selection favors the preservation of codon status (either preferred or nonpreferred), whereas Nielsen et al. (2007) assumed that selection unconditionally favors specific codons over others; 3) the selection coefficient of Nielsen et al. (2007) for synonymous selection is affected by nonconservative nonsynonymous substitutions, whereas all nonsynonymous substitutions contribute only to nonsynonymous selection in our model. We discuss the implications of these differences below.

Materials and Methods

Model

We assume that all the codons for each amino acid can be subdivided into two groups, preferred codons and unpreferred codons. We further assume that there is some selection pressure that keeps codons in the preferred or unpreferred state. Under this assumption, a synonymous substitution leading from a preferred codon to another preferred codon or from an unpreferred codon to another unpreferred codon does not experience this selection pressure, whereas a synonymous substitution leading from a preferred codon to an unpreferred codon or vice versa will experience it. We refer to the former type of synonymous substitution as conservative and to the latter as nonconservative (fig. 1).

We use a codon-based continuous-time Markov model to compute the rates of conservative and nonconservative synonymous substitutions. Because there are 61 sense codons, the transition matrix of our model is 61×61. We define the instantaneous substitution rate from codon i to codon j ($i \neq j$) as

$$q_{ij} = \begin{cases} 0 & \text{the two codons differ at more than one position,} \\ \psi\alpha_{i_k j_k}\pi_j & \text{one nonconservative synonymous substitution between codons } i \text{ and } j \\ \alpha_{i_k j_k}\pi_j & \text{one conservative synonymous substitution between codons } i \text{ and } j, \\ \omega\alpha_{i_k j_k}\pi_j & \text{one nonsynonymous substitution between codons } i \text{ and } j. \end{cases} \quad (1)$$

Here, ψ and ω capture selection on synonymous and nonsynonymous substitutions, respectively, and the π_j and $\alpha_{i_k j_k}$ reflect the mutation process. As usual, $\omega < 1$ implies purifying selection at the peptide level and $\omega > 1$ suggests positive selection at the peptide level. Similarly, $\psi < 1$ implies purifying synonymous selection and $\psi > 1$ suggests positive selection at synonymous sites. The mutation process is fully described by the equilibrium frequency of codons, given by π_j, and the relative substitution rates between nucleotides, given by $\alpha_{i_k j_k}$. Here, i_k and j_k represent the nucleotides at position k in codons i and j ($i_k, j_k \in \{A, C, G, T\}$). We use the general time-reversible mutation model ($\alpha_{i_k j_k} = \alpha_{j_k i_k}$) and therefore have six free mutation rate parameters (α_{AC}, α_{AG}, α_{AT}, α_{CG}, α_{CT}, and α_{GT}).

The substitution process between two sequences separated by t time units is described by the matrix $P(t) = \exp(Qt)$, where $Q = \{q_{ij}\}$ as defined above. We normalize the Q matrix so that one substitution is expected to occur in one time unit ($\sum_{i \neq j} \pi_i q_{ij} = 1$) (Goldman and Yang 1994; Yang 2006). The numbers of nonsynonymous (N_d) and synonymous (S_d) substitutions per codon are as follows:

$$N_d = \sum_{i \to j \notin \mathcal{S}} \pi_i q_{ij} t, \quad (2)$$

$$S_d = \sum_{i \to j \in \mathcal{S}} \pi_i q_{ij} t. \quad (3)$$

Here, $\mathcal{S}$ is the set of all possible synonymous substitutions. The numbers of conservative (S_d^C) and nonconservative (S_d^N) synonymous substitutions per codon are

$$S_d^C = \sum_{i \to j \in \mathcal{C}} \pi_i q_{ij} t, \quad (4)$$

$$S_d^N = \sum_{i \to j \in \mathcal{N}} \pi_i q_{ij} t. \quad (5)$$

Here, $\mathcal{C}$ is the subset of $\mathcal{S}$ for which codon preference remains unchanged between i and j and $\mathcal{N}$ is the subset of $\mathcal{S}$

for which codon preference changes between i and j. We calculate evolutionary rates according to the physical-site definition. We count a nondegenerate site as one nonsynonymous site, a 2-fold degenerate site as one-third synonymous and two-third nonsynonymous site, a 3-fold site as two-third synonymous and one-third nonsynonymous site, and a 4-fold site as one synonymous site (Yang 2006). Similarly, synonymous sites are also divided into fractional conservative synonymous and nonconservative synonymous sites according to the proportion of possible changes at a site that are either conservative or nonconservative. For example, for codon GTT, the third nucleotide position is a synonymous site. One possible nucleotide substitution at this site leads to a conservative synonymous codon change (GTC), whereas the other two possible substitutions are nonconservative (GTA and GTG) (fig. 1). Thus, the third nucleotide position of GTT is counted as one-third of a conservative synonymous site and two-third of a nonconservative synonymous site. We average these counts over all codons in the sequence to obtain the average number of synonymous, nonsynonymous, conservative synonymous, and nonconservative synonymous sites per codon, represented by S, N, S^C, and S^N. We then define evolutionary rates by $dN = N_d/N$, $dS = S_d/S$, $dS_C = S_d^C/S^C$, and $dS_N = S_d^N/S^N$.

Model Fitting

We implemented our model in the software package HyPhy (Kosakovsky Pond et al. 2005). We estimated the parameters ψ, ω, and α_{rs} by maximum likelihood, but estimated codon frequencies π_j from the nucleotide frequency at the three codon positions (model F3×4). HyPhy scripts to carry out this analysis can be downloaded from http://openwetware.org/images/5/52/Zhou-Gu-Wilke-synonymous-selection.zip. For comparison, we also fitted our model with a fixed $\psi = 1$. Throughout this manuscript, all parameters obtained under this setting are indicated with a superscript 1, as in ω^1, dN^1, and dS^1.

Data Sources

We obtained genomic sequences and orthologs from the following sources: the *Saccharomyces* Genome Database (ftp://genome-ftp.stanford.edu/) for yeast (*S. cerevisiae* vs. *S. bayanus*) and the WormBase (http://www.wormbase.org) for worm (*C. elegans* vs. *C. briggsae*). For each pair of orthologs, we aligned the peptide sequences using MUSCLE (Edgar 2004) and then translated the peptide sequences back to restore the original nucleotide sequences. We only retained complementary DNAs with 80% of alignment to their orthologs and at least 100 codons. We also excluded all genes for which HyPhy could not fit the model with an optimization precision of 0.001 within 10^5 likelihood function evaluations. This procedure yielded 4,047 genes for yeast and 5,623 for worm.

We used previously published expression data for yeast (Holstege et al. 1998) and worm (Hill et al. 2000). Multiple signals for the same transcript were averaged. After combining the expression data with the genomic data, we ended up with 3,816 genes for yeast and 4,711 for worm.

Inferring Preferred Codons

We calculated the adjusted effective number of codons (ENC′) for each gene, according to the method developed by Novembre (2002), which corrects for nucleotide content. We then compared the codon usage pattern between the gene groups showing the lowest 5% and highest 5% ENC′ in each species. We defined codons as "preferred" if they showed a statistically significant increase in frequency in the lowest ENC' group, as determined by a chi-square test (supplementary table S1, Supplementary Material online). The codon usage patterns of the orthologous species are quite similar (supplementary fig. S1, Supplementary Material online) with several exceptions: For *S. bayanus*, the codon ATT for Ile is assigned as unpreferred, whereas the codon CGT for Arg is assigned as preferred. For *C. briggsae*, the codon ACT for Thr and the codon GTT for Val are assigned as unpreferred while the codon AGA is assigned as preferred for Arg. The codon preference for the 2-fold degenerate amino acid Glu is reversed between the two worm species. The preferred codons we identified corresponded to transfer RNAs (tRNAs) with increased gene copy number (supplementary fig. S2, Supplementary Material online).

Results

Conservative and Nonconservative Synonymous Evolutionary Rates

Our method makes the assumption that certain codons in a codon family are either preferred or unpreferred and that there is a selection pressure to keep codons in the preferred or unpreferred state. We make no a priori assumptions about why a specific codon is preferred or unpreferred, but in general, preferred codons will be the ones that are efficiently and rapidly translated because their cognate tRNAs are highly abundant. Codons with highly abundant cognate tRNAs have increased translation speed and/or accuracy, and these properties tend to confer a selective advantage to the gene in which they occur (Ikemura 1981; Sharp et al. 1986; Akashi 1994; Drummond and Wilke 2008; Zhou et al. 2009). At the same time, there is evidence that unpreferred codons are selected for at specific sites, presumably to aid in cotranslational protein folding (Thanaraj and Argos 1996; Komar et al. 1999; Cortazzo et al. 2002; Kimchi-Sarfaty et al. 2007; Widmann et al. 2008; Zhang et al. 2009).

To identify preferred and unpreferred codons, we selected the 5% of genes with the strongest codon usage bias (lowest ENC') and the 5% of genes with the weakest codon usage bias (highest ENC') and identified codons as preferred if they were significantly enriched in the low-ENC' group (see Materials and Methods). We then defined substitutions from a preferred codon to another preferred codon coding for the same amino acid or from a unpreferred codon to another unpreferred codon coding for the same amino acid as "conservative synonymous substitutions." We defined all other synonymous substitutions as "nonconservative."

We fitted our model to coding sequences of yeast and worm and calculated conservative and nonconservative

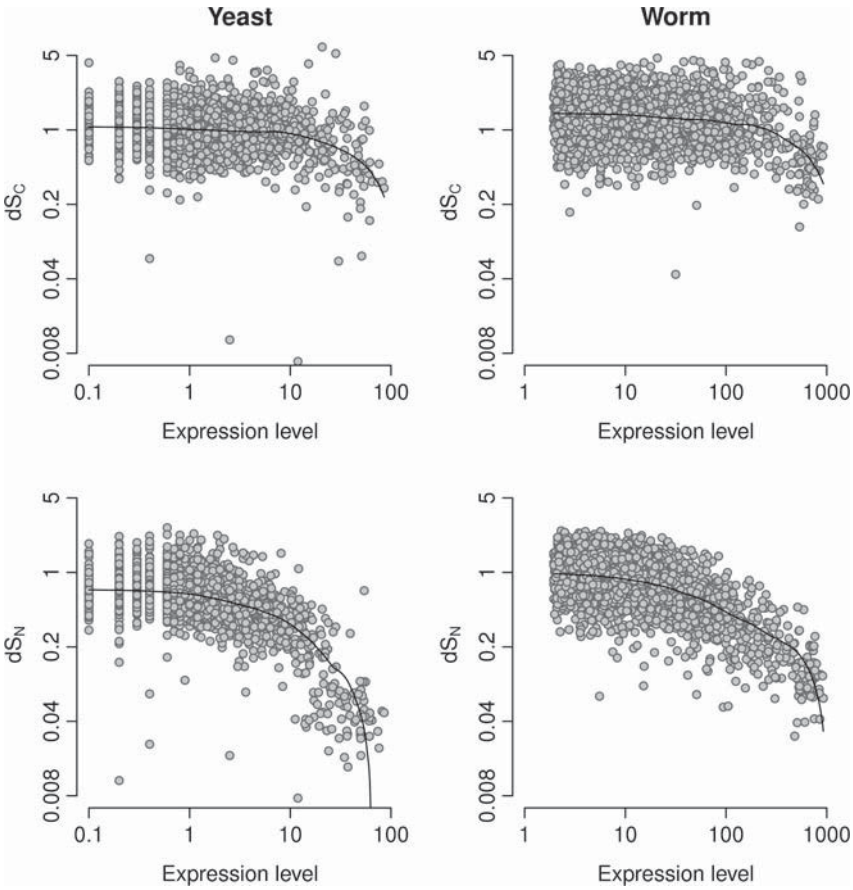

FIG. 2. Evolutionary rates dS_C and dS_N versus expression level for yeast (left) and worm (right). Solid lines show lowess smoothed data.

evolutionary rates dS_C and dS_N. Because translational selection is strong in these species (Drummond and Wilke 2008), we calculated the correlation of both dS_C and dS_N with gene expression level (fig. 2 and table 1). We found that dS_N correlates strongly with expression level in yeast and worm. By contrast, the correlation between dS_C and expression level, even though statistically significant, is very weak: Gene expression level accounts only for 1.8% of the variation in dS_C in yeast and 1.3% of the variation in dS_C in worm.

As is well known from prior work (Drummond and Wilke 2008), both dN and dS as calculated by traditional methods correlate strongly with expression level in yeast and worm. We found that the same is true for dN and dS calculated from our model (table 1). Moreover, the correlation coefficients of dS expression level and dS_N expression level are comparable, and the correlation between dS and dS_N (Spearman's $\rho = 0.812, P \ll 10^{-100}$ for yeast and $\rho = 0.889, P \ll 10^{-100}$ for worm) is stronger than that between dS and dS_C (Spearman's $\rho = 0.643, P \ll 10^{-100}$ for yeast and $\rho = 0.702, P \ll 10^{-100}$ for worm). The correlation between dS_N and dS_C is much weaker (Spearman's $\rho = 0.176, P \ll 10^{-100}$ for yeast and $\rho = 0.361, P \ll 10^{-100}$ for worm). These observations show that the variation in dS due to translational selection is largely captured in dS_N and removed from dS_C. Therefore, dS_C is a suitable neutral baseline to which we can compare both dN and dS_N.

To obtain an independent verification that our model works as expected, we also devised a counting model based on the method by Nei and Gojobori (1986) (see Supplementary Material online for details). We computed the proportion of conservative (PS_C) and nonconversitive (PS_N) synonymous differences for each gene and repeated the same analyses as with dS_C and dS_N. We obtained largely the same results as we did with the maximum-likelihood method (supplementary fig. S3, Supplementary Material online).

Hirsh et al. (2005) proposed an adjusted measure of dS, denoted dS', which takes the relationship between codon bias and synonymous divergence into account. We thus analyzed dS', dS_N, and dS_C in yeast (dS' data obtained from Hirsh et al. 2005). We found that the correlation between dS' and dS_N (Spearman's $\rho = 0.501, P \ll 10^{-100}$) does not differ greatly from the correlation between dS' and dS_C (Spearman's $\rho = 0.449, P \ll 10^{-100}$). The correlation between dS' and expression level (Spearman's $\rho = -0.168, P = 2.1 \times 10^{-19}$) is comparable but slightly stronger than the one between dS_C and expression level (table 1). Thus, our method seems to work at least as well in controlling for effects of translational selection as the method of Hirsh et al. (2005), with the added benefit that

Table 1. Spearman Correlations of Expression Level with dS_C, dS_N, dS, dS^1, dN, dN^1, dN/dS_C, and dS_N/dS_C.

Variable	Yeast		Worm	
	ρ	P	ρ	P
dS_C	−0.135	5.7×10^{-17}	−0.116	1.2×10^{-15}
dS_N	−0.441	2.6×10^{-181}	−0.400	1.2×10^{-176}
dS	−0.400	8.3×10^{-147}	−0.343	7.8×10^{-130}
dS^1	−0.434	8.9×10^{-175}	−0.364	1.5×10^{-147}
dN	−0.522	1.7×10^{-265}	−0.290	4.4×10^{-92}
dN^1	−0.523	3.8×10^{-267}	−0.280	7.1×10^{-86}
dN/dS_C	−0.447	1.1×10^{-186}	−0.232	1.2×10^{-58}
dS_N/dS_C	−0.293	1.8×10^{-76}	−0.301	6.1×10^{-99}

our model is based on mechanistic model of molecular evolution formulated in a coherent maximum-likelihood framework.

Selection Restricts Nonconservative Synonymous Substitutions

In the previous subsection, we found that the conservative synonymous substitution rate (dS_C) is a reasonably neutral baseline of evolutionary variation. Therefore, we can use the parameter ψ to estimate the amount and type of selection on synonymous sites. Under positive selection, we expect ψ to be larger than 1, whereas under purifying selection, ψ should be less than 1. The distributions of ψ in yeast and worm are very similar (fig. 3), and they are slightly shifted to the left of 1 (t-test: $P \ll 10^{-100}$ for yeast and $P \ll 10^{-100}$ for worm). Thus, there is some purifying selection at synonymous sites in both species. Interestingly, the distributions of ω are approximately an order of magnitude further to the left than the distributions of ψ (fig. 3). Averaged over all sites in a gene, nonsynonymous substitutions accumulate approximately an order of magnitude slower than nonconservative synonymous substitutions.

A comparison of the right and left panel of figure 3 shows that we found very similar results when considering the ratios dN/dS_C and dS_N/dS_C (based on physical sites) instead of the ratios ω and ψ (based on mutational opportunity). Throughout the remainder of the manuscript, we will focus on the physical site–based ratios for genome-wide correlation studies, as these ratios tend to perform more reliably for those kinds of studies (Bierne and Eyre-Walker 2003). We will, however, use ω and ψ for statistical tests of positive or purifying selection in individual genes because the likelihood-ratio test provides us with a straightforward means to test for the null hypotheses $\omega = 1$ and $\psi = 1$.

We next studied the correlation of both dN/dS_C and dS_N/dS_C with expression level. We found that both quantities decline as the gene expression level increases (fig. 4 and table 1). Because the correlation of dS_C with expression level is very weak and negative (table 1), this result shows that the amount of purifying selection on both synonymous and nonsynonymous sites increases with expression level in both species. Results were similar for ω and ψ, but the correlations were slightly weaker (supplementary fig. S4, Supplementary Material online). Interestingly, dN/dS_C starts increasing again for the genes with the highest expression levels in yeast. This effect may be caused by very strong selection on synonymous sites in those genes. Even though dS_C is largely independent of expression level, it does decrease for genes with very high expression level (fig. 2).

We also determined all genes with ψ significantly above or below 1 by testing for the null hypothesis $\psi = 1$ using a likelihood-ratio test. We found 18 genes in yeast and 41 in worm with $\psi > 1$ at $P < 0.05$. These numbers correspond to 0.44% and 0.73% of the genomes of yeast and worm, respectively. After applying a false discovery rate correction for multiple testing (Benjamini and Hochberg 1995) and allowing for a false discovery rate of 5%, only two genes (0.05%) remained significant in yeast and no gene remained significant in worm. On the other hand, we found 1,076 yeast genes (26.59% of the genome) and 1,164 worm genes (20.70% of the genome) with $\psi < 1$ at $P < 0.05$. But only 382 (9.44%) and 288 (5.12%) genes survived the correction for multiple testing.

Because selection on preferred codons is generally associated with the translation process and increases with expression level, we compared the number of genes with $\psi < 1$ and corrected $P < 0.05$ between the top 10% highest

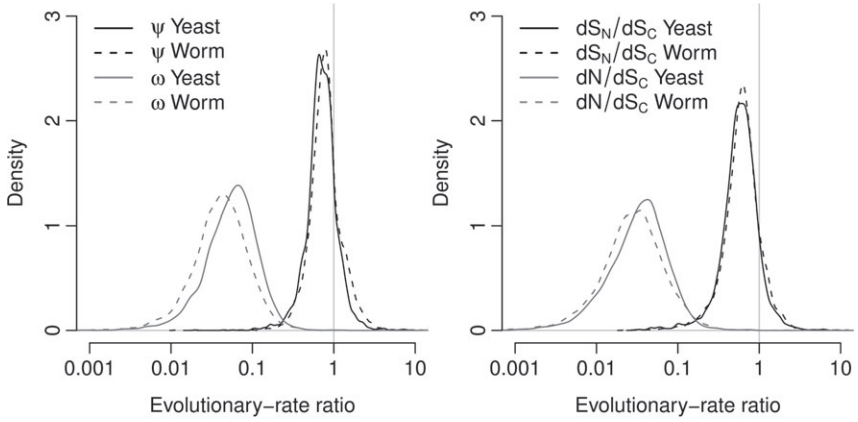

FIG. 3. Distribution of evolutionary rate ratios. Left panel: distribution of the ratios ω and ψ. Right panel: distribution of the ratios dN/dS_C and dS_N/dS_C. The y axes are scaled such that the area under the curves equals 1.

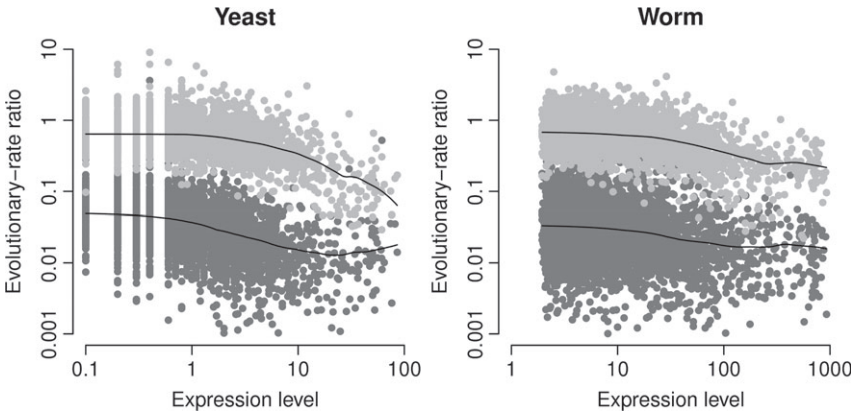

FIG. 4. Evolutionary rate ratios dN/dS_C (dark points) and dS_N/dS_C (light points) versus expression level for yeast (left) and worm (right). Both ratios decline with increasing expression levels. Solid lines show lowess smoothed data.

expressed and the bottom 10% lowest expressed genes. In yeast, we found 96 such genes of 374 in the high expression group but only 13 of 152 in the low expression group. These fractions are significantly different (Fisher's exact test, $P = 4.9 \times 10^{-6}$). The group sizes differ because numerous genes had identical expression levels, preventing us from choosing exactly equal-sized groups. In worm, we found 45 significant genes of 472 in the high expression group and 15 of 469 in the low expression group. These fractions are also significantly different (Fisher's exact test, $P = 8.1 \times 10^{-5}$). Thus, highly expressed genes are more likely to experience purifying synonymous selection.

In our model, the parameter ψ measures the extent of selection on synonymous sites. Nielsen et al. (2007) proposed a similar model that estimates the overall strength of selection (S) against unpreferred codons. To compare their model with our model, we implemented a variant of their model and fitted it to our data sets. The one modification we made to the model of Nielsen et al. (2007) is that we used the same general time-reversible mutation model we used for our models. We calculated the selection coefficient S for each gene in yeast and worm. Although S and ψ were correlated, the correlations were weak (Spearman's $\rho = -0.155, P = 2.6 \times 10^{-23}$ for yeast and $\rho = -0.130, P = 1.8 \times 10^{-22}$ for worm).

Application for Detecting Selection on Protein Sequence

A large dN/dS is usually interpreted as signal for positive selection at nonsynonymous sites. In our model, the traditional dN/dS ratio (without selection on synonymous sites) is reflected by ω^1 (ω estimated under a constant $\psi = 1$), whereas our ω value corresponds to the ratio between dN and dS_C. Under strong purifying synonymous selection, dS should be small, whereas dS_C is not necessarily small. In this case, ω^1 would overestimate the amount of nonsynonymous divergence. Similarly, for a gene under positive synonymous selection, dS might be inflated and ω^1 would underestimate the amount of nonsynonymous divergence. Table 2 lists the genes with either ω or ω^1 significantly larger than 1 ($P < 0.05$ under the null hypothesis $\omega = 1$ or $\omega^1 = 1$, respectively) in our data set. We found the strongest effect for yeast gene YDR133C for which we may greatly underestimate the dN/dS value without considering the effect caused by selection at synonymous sites ($\omega = 5.118$ and $\omega^1 = 0.648$).

We also correlated the ω values with ψ. We found that, in both species, there is a significant correlation between ω and ψ (Spearman's $\rho = 0.343, P \ll 10^{-100}$ for yeast and $\rho = 0.366, P \ll 10^{-100}$ for worm) (supplementary fig. S5, Supplementary Material online). This result most likely reflects the increasing strength of selection on both synonymous and nonsynonymous sites with increasing expression level, as seen by the correlation of both ω and ψ with expression level (supplementary fig. S4, Supplementary Material online).

An Alternative Definition for Selection on Synonymous Sites

All results we reported in the preceding subsections were obtained with a model in which synonymous selection happens only within codon families. We refer to this model also as the "main model." We also considered an alternative model in which the rate of nonsynonymous mutations that change codon preference also differs by a factor ψ from the rate of nonsynonymous mutations that do not change codon preference (see Supplementary Material online for details). By and large, the main model and the alternative model produced comparable results (fig. 5 and supplementary figs. S6–S9, supplementary table S2, and supplementary text, Supplementary Material online), and all results were consistent among yeast and worm. The largest deviations between the two models arise for ψ and dS_C. For ψ, the

Table 2. Comparison between ω and ω^1.

	Gene	ω	P	ω^1	P
Yeast	YDR133C	5.118	5.0×10^{-2}	0.648	3.0×10^{-1}
	YDR433W	26.253	4.2×10^{-4}	12.096	3.0×10^{-5}
	YJL009W	4.055	3.2×10^{-2}	3.048	1.3×10^{-2}

NOTE.—Only genes with ω or ω^1 significantly larger than 1 ($P < 0.05$) were listed.

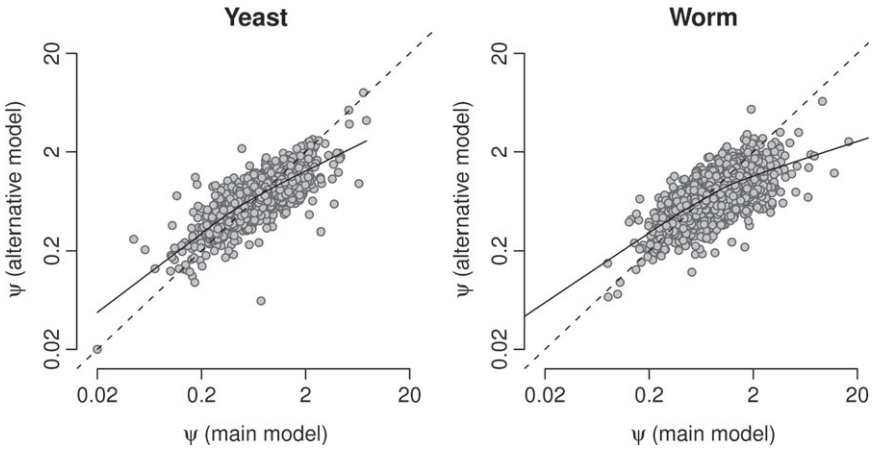

FIG. 5. Evolutionary rate ratio ψ for the model described in the main text and for the alternative model described in the Supplementary Material online for yeast (left) and worm (right). The Spearman correlations for the two data sets are 0.715 ($P \ll 10^{-100}$) for yeast and 0.694 ($P \ll 10^{-100}$) for worm. The solid lines show lowess smoothed data and the dashed lines indicate exact agreement between the two models.

values derived from the main model explain only 51%, and 48% of the variance in the values derived from the alternative model for yeast and worm, respectively (supplementary table S2, Supplementary Material online). For dS_C, the amount of variance explained is 73% and 76% in yeast and worm, respectively (supplementary table S2, Supplementary Material online). For comparison, for dS_N, the variance explained is above 90% in both species (supplementary table S2, Supplementary Material online).

It makes intuitive sense that ψ would be more strongly affected by the change in the model definition than ω. Most of the selection on nonsynonymous substitutions is likely due to amino acid–level constraints and not to selection on codon preference. Therefore, ω should be largely the same regardless of which of the two model definitions we use. By contrast, ψ measures a much weaker and more subtle effect, and thus, even a small selection pressure on codon preference among amino-acid families would have a noticeable effect on ψ.

In general, we found that ψ as estimated under the alternative model indicates weaker synonymous selection than ψ as estimated under the main model. The former ψ is consistently closer to 1; it tends to be smaller than the latter ψ when the latter ψ is large, and it tends to be larger than the latter ψ when the latter ψ is small (fig. 5). We interpret this result as follows: The selection pressure among codon families is dominated by amino acid–level effects; codon preference plays only a minor role when the amino acid is changed. However, when the amino acid remains unchanged, codon preference is important. The main model, by disregarding synonymous selection pressures among codon families, can fully measure the synonymous selection pressure within codon families. In the alternative model, on the other hand, ψ gets diluted by the weak selection pressure on codon preference among codon families.

Consistent with this interpretation, the correlation between dS_C and expression level is slightly stronger for the alternative model than for the main model in yeast and worm (Spearman's $\rho = -0.143, P = 3.5 \times 10^{-19}$ for yeast and $\rho = -0.133, P = 3.0 \times 10^{-24}$ for worm, see also supplementary fig. S7, Supplementary Material online, and table 1). Under the alternative model, dS_C is likely somewhat confounded by amino acid–level selection pressures mediated by expression level.

A Model with Four Synonymous Rates

As a generalization of our main model, we also developed a model in which each type of synonymous substitution (from either preferred or unpreferred to either preferred or unpreferred codon) can occur at an independent rate for a total of four different synonymous rates. We added two additional parameters (η and θ) into our main model (see Supplementary Material online for details). In this model with four rates, η measures the ratio between the substitution rates from unpreferred to preferred codons and from preferred to unpreferred codons, whereas θ measures the ratio between the substitution rates from preferred codons to preferred codons and from unpreferred codons to unpreferred codons. All other parameters in the model have exactly the same meaning as before. On the basis of ψ, η, and θ, we calculated the synonymous substitution rates for preferred to unpreferred codon change (dS_{PU}), unpreferred to preferred codon change (dS_{UP}), unpreferred to unpreferred codon change (dS_{UU}), and preferred to preferred codon change (dS_{PP}).

We tested for correlations between expression level and the four synonymous rates. We considered first the two substitution rates associated with conservative synonymous substitutions. For dS_{UU}, we found that it did not correlate significantly with expression level (Spearman's $\rho = -0.020, P = 0.280$ for yeast and $\rho = -0.033, P = 0.076$ for worm). For dS_{PP}, we found a moderate negative correlation in yeast ($\rho = -0.248, P = 3.0 \times 10^{-40}$) and none in worm ($\rho = -0.038, P = 0.038$). For nonconservative substitutions, we found that dS_{PU} was negatively correlated with expression level (Spearman's $\rho = -0.541, P \ll 10^{-100}$ for yeast and $\rho = -0.491, P \ll 10^{-100}$ for worm),

whereas the correlation between dS_{UP} and expression level was positive (Spearman's $\rho = 0.314, P \ll 10^{-100}$ for yeast and $\rho = 0.330, P \ll 10^{-100}$ for worm). We also correlated both η and θ with expression level. We found that η increased with gene expression level (Spearman's $\rho = 0.538, P \ll 10^{-100}$ for yeast and $\rho = 0.486, P \ll 10^{-100}$ for worm). The correlations between θ and expression level were much weaker (Spearman's $\rho = 0.034, P = 0.075$ for yeast and $\rho = 0.167, P \ll 10^{-100}$ for worm). Overall, these results support our approach of considering conservative synonymous mutations largely free of expression-related selection pressures, whereas nonconservative synonymous mutations are not.

To assess whether we were justified in combining the two conservative rates and the two nonconservative rates into a single rate each in the main model, we considered the distributions of η and θ. We found that they were very similar across species (supplementary fig. S10, Supplementary Material online) and nearly centered around 1. The distribution of η was shifted slightly to the right of 1 (t-test: $P \ll 10^{-100}$ for both species), whereas the distribution of θ was shifted slightly to the left of 1 (t-test: $P \ll 10^{-100}$ for both species). That both these distributions were nearly centered around 1 supports our approach of using only two synonymous evolutionary rates in the main model. At the same time, the small but statistically significant shifts to the right and left of 1 indicate that the four-rate model is not superfluous but can instead resolve subtle second-order effects that are not visible under the main model.

Discussion

We have developed a statistical method to identify positive and purifying selection at synonymous sites in yeast and worm. We tested whether synonymous substitutions from preferred to unpreferred codons or vice versa happen more or less frequently than expected by chance. If the rate of synonymous substitutions is independent of codon preference, then the conservative synonymous substitution rate (dS_C) should equal the nonconservative rate (dS_N). If synonymous substitutions tend to conserve codon preference, we expect $dS_N < dS_C$ ($\psi < 1$), whereas if they tend to change codon preference, we expect $dS_N > dS_C$ ($\psi > 1$). By testing for the null hypothesis $\psi = 1$, we found that 0.05% of the yeast genes and no worm genes were positively selected at synonymous sites (assuming a 5% false discovery rate). On the other hand, we found 9.44% of yeast genes and 5.12% of worm genes to undergo significant purifying synonymous selection. The percentage of positively selected genes we found is substantially lower than what Resch et al. (2007) found for mammals using a different method (comparing the synonymous rate to the rate of divergence in introns). They found that roughly 12% of the genes (without correction for multiple testing) have undergone positive synonymous selection in mouse–rat orthologs.

By correlating dS, dS_N, and dS_C with gene expression level, we found that much of the signal of translational selection commonly found in dS (Drummond et al. 2006; Drummond and Wilke 2008) is captured by dS_N, whereas dS_C is largely unaffected by expression level. The correlation between dS_C and expression level, although significant, is very weak. The amount of variance explained is on the order of 2% (yeast) or less (worm). For this reason, we propose that dS_C may be a better measure of neutral variation than dS and that the ratio dN/dS_C may be more appropriate to detect positive selection than the ratio dN/dS.

We do not claim, however, that dS_C is free from any selection pressure. Our model is fundamentally based on the concept of codon bias and of preferred and unpreferred codons. Any selection pressure that acts on the DNA or mRNA level, such as selection pressures related to transcription (Xia 1996), splicing (Chamary and Hurst 2005b; Dewey et al. 2006; Parmley et al. 2006; Warnecke and Hurst 2007), expression regulation (Parmley and Huynen 2009), protein structure (Xie and Ding 1998; Gu et al. 2004; Clarke and Clark 2008), DNA secondary structure (Vinogradov 2003; Hoede et al. 2006), or mRNA secondary structure and stability (Chamary and Hurst 2005a; Stoletzki 2008) will likely affect dS_C as much as it affects dS. The relative strength of such selection pressures compared with translational selection in organisms that experience strong translational selection, as is the case with yeast and worm, is not well understood at present and deserves future study.

A common use of the dN/dS method is to identify individual branches in a larger phylogeny that have experienced altered selection pressures. We here applied our model only to species pairs, but our maximum-likelihood approach allows us also to use our model in more complex settings and to test, for example, whether specific branches have experienced particularly strong purifying or positive synonymous selection. Such an analysis makes only sense, however, if the preferred codons remain largely unchanged throughout the phylogeny. We believe that this condition will often be satisfied for species that are not too distantly related. For example, beyond *S. cerevisiae* and *S. bayanus*, we found a very similar set of preferred codons in five further *Saccharomyces* species (*S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. castellii*, and *S. kluyveri*), as well as in *Kluyveromyces lactis*. Even the distantly related *Schizosaccharomyces pombe* had only minor differences in preferred codon usage. As the number of fully sequenced species is only going to increase in the future, we expect that there will be many situations where our approach may be useful.

Our model differs in three ways from previous work by Nielsen et al. (2007). First, Nielsen et al. (2007) used an explicit selection term for synonymous substitutions in their model and thus estimated a selection strength S rather than an evolutionary rate ratio. Although obtaining a direct estimate for the strength of selection on synonymous sites is desirable, we believe that there are advantages to our approach. Under our approach, ψ and ω are both evolutionary rate ratios measured in comparable units. We can directly compare ψ and ω to assess the relative strength of selection on synonymous and nonsynonymous substitutions. By contrast, it is not obvious how to compare the estimated S to the estimated ω in the model of Nielsen et al. (2007).

Second, the selection term of Nielsen et al. (2007) also assumes that preferred (or unpreferred, for $S < 0$) codons have systematically higher fitness than the other type of codon. This assumption is different from our assumption, which states that selection tends to preserve codon status, regardless of whether the codon status is preferred or unpreferred. Our assumption is based on the observations that unpreferred codons are selected for at specific sites (Thanaraj and Argos 1996; Komar et al. 1999; Cortazzo et al. 2002; Kimchi-Sarfaty et al. 2007; Widmann et al. 2008; Zhang et al. 2009) and that codon usage bias is highly regulated even in genes that are not encoded primarily by preferred codons (Dong et al. 1996). We believe that the difference in assumption about how selection acts on synonymous codons is the main reason why S and ψ correlate only weakly.

Third, in our model, ψ is purely a measure for the difference in conservative and nonconservative substitutions within codon families. In principle, a substitution from a preferred codon in one codon family to a unpreferred codon in another codon family could experience both a selective effect because the amino acid and the codon preference were changed. We absorbed the latter effect into ω in our model and thus counted it as a nonsynonymous selection pressure as well. We proceeded in this manner because a priori it is not clear that selection pressures on synonymous sites can be compared across codon families. For example, the translational efficiency of a unpreferred codon in one codon family could be higher than the translational efficiency of a preferred codon in another codon family simply because all codons of the first family have higher translational efficiency than all codons of the second family. Nielsen et al. (2007) made a different choice and included a term representing synonymous selection into all substitutions that connected preferred with unpreferred codons or vice versa. To determine to what extent our results were affected by this choice, we also fitted a model in which all substitutions that connected preferred with unpreferred codons or vice versa received a factor ψ in the transition matrix. We found that our main model and our alternative model gave by-and-large similar results. However, the alternative model tended to predict weaker selection on synonymous sites than the main model. This observation shows that selection for preferred or unpreferred codons is not a major force among codon families, and it justifies our approach of disregarding synonymous selection among codon families.

In our model with four synonymous rates, we categorized both conservative and nonconservative substitutions into two subgroups, respectively. The two rates dS_{UU} and dS_{PP} were not strongly correlated with expression level. This finding supports our strategy in the main model to combine these two rates into dS_C and use the latter as a baseline to estimate the pace of the neutral evolutionary process. Interestingly, the two rates dS_{PU} and dS_{UP} were both significantly correlated with expression level, but the correlation was negative for dS_{PU} and positive for dS_{UP}. This result suggests that highly expressed genes experience positive selection to increase their number of preferred codons. This result also implies that the effects of dS_{UP} and dS_{PU} may partly cancel each other when we combine these two rates into dS_N. Therefore, dS_N may actually underestimate the amount of selection on synonymous sites.

We fitted all models we developed to both yeast and worm, with largely identical results. However, when applied to fly, our models gave somewhat comparable results to yeast and worm but also produced some differences (see Supplementary Material online for details). These observations beg the question of how generally applicable our models are to other systems. Our models are valid if two conditions are met: First, codons need to separate clearly into preferred and nonpreferred ones. For organisms for which a clear distinction cannot be made, for example, because codon preference is better described on a continuous scale from preferred to nonpreferred and everything in between, our models would not be appropriate. Indeed, in fly, preferred and nonpreferred codons do not separate as cleanly as they do in yeast or worm. Second, the mutation process needs to be reversible. Although this assumption is commonly made when fitting evolutionary models to sequence data, the assumption is not always justified. In particular, the evolution of the *D. melanogaster* line relative to other *Drosophila* species did likely not follow a time-reversible process (Nielsen et al. 2007). In summary, there are likely many more systems than just yeast and worm for which our models may be useful, but one should not assume that any species with strong codon bias is a suitable candidate for our approach.

Supplementary Material

Supplementary text, tables S1 and S2, and figures S1–S9 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the National Institute of Health grant R01 AI065960 to C.O.W. and by a grant from the National Natural Science Foundation of China (No. 30900836) to W.G. We would like to thank David Liberles for helpful comments and suggestions on this work.

References

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

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series* B. 57:289–300.

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.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD,

et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

Chamary JV, Hurst LD. 2005a. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet.* 21:256–259.

Chamary JV, Hurst LD. 2005b. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6:R75.

Clarke TFIV, Clark PL. 2008. Rare codons cluster. *PLoS One* 3:e3412.

Cortazzo P, Cervenansky C, Marin M, Reiss C, Ehrlich R, Deana A. 2002. Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochem Biophys Res Commun.* 293:537–541.

Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* 7:311.

Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J Mol Biol.* 260:649–663.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol.* 23:327–337.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

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

Gu W, Zhou T, Ma J, Sun X, Lu Z. 2004. The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. *Biosystems* 73:89–97.

Higgs PG, Hao W, Golding GB. 2007. Identification of conflicting selective effects on highly expressed genes. *Evol Bioinform.* 3:1–13.

Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol Biol Evol.* 25:2279–2291.

Hill AA, Hunter CP, Tsung BT, Tucker-Kellogg G, Brown EL. 2000. Genomic analysis of gene expression in *C. elegans*. *Science* 290:809–812.

Hirsh AE, Fraser HB, Wall DP. 2005. Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol Biol Evol.* 22:174–177.

Hoede C, Denamur E, Tenaillon O. 2006. Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. *PLoS Genet.* 2:e176.

Holstege FCP, Jennings E, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95:717–728.

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18:486.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol.* 151:389–409.

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.

Komar AA, Lesnik T, Reiss C. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.* 462:387–391.

Koonin EV, Rogozin IB. 2003. Getting positive about selection. *Genome Biol.* 4:331.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Liberles DA. 2001. Evaluation of methods for determination of a reconstructed history of gene sequence evolution. *Mol Biol Evol.* 18:2040–2047.

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

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

Nielsen R, DuMont VLB, 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.

Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol.* 19:1390–1394.

Parmley J, Chamary J, Hurst L. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol.* 23:301–309.

Parmley JL, Huynen MA. 2009. Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation. *PLoS Genet.* 5:e1000548.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in *Escherichia coli*. *Genome Res.* 17:1336–1343.

Resch AM, Carmel L, Mariño-Ramírez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol.* 24:1821–1831.

Sharp PM, Tuohy T, Mosurski K. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* 14:5125–5143.

Stenico M, Lloyd AT, Sharp PM. 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res.* 22:2437–2446.

Stoletzki N. 2008. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. *BMC Evol Biol.* 8:224.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol.* 24:374–381.

Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol.* 16:1315–1328.

Thanaraj TA, Argos P. 1996. Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* 5:1594–1612.

Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Res.* 31:1838–1844.

Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol Biol Evol.* 24:2755–2762.

Widmann M, Clairo M, Dippon J, Pleiss J. 2008. Analysis of the distribution of functionally relevant rare codons. *BMC Genomics* 9:207.

Xia X. 1996. Maximizing transcription efficiency causes codon usage bias. *Genetics* 144:1309–1320.

Xie T, Ding D. 1998. The relationship between synonymous codon usage and protein structure. *FEBS Lett.* 434:93–96.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*. 25:568–579.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22:1107–1118.

Zhang G, Hubalewska M, Ignatova Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol*. 16:274–280.

Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol*. 26:1571–1580.