Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins

Tong Zhou,† Mason Weems,‡ and Claus O. Wilke*†‡*

*Center for Computational Biology and Bioinformatics, The University of Texas at Austin; †Section of Integrative Biology, The University of Texas at Austin; and ‡Institute for Cell and Molecular Biology, The University of Texas at Austin

The mistranslation-induced protein misfolding hypothesis predicts that selection should prefer high-fidelity codons at sites at which translation errors are structurally disruptive and lead to protein misfolding and aggregation. To test this hypothesis, we analyzed the relationship between codon usage bias and protein structure in the genomes of four model organisms, *Escherichia coli*, yeast, fly, and mouse. Using both the Mantel–Haenszel procedure, which applies to categorical data, and a newly developed association test for continuous variables, we find that translationally optimal codons associate with buried residues and also with residues at sites where mutations lead to large changes in free energy ($\Delta\Delta G$). In each species, only a subset of all amino acids show this signal, but most amino acids show the signal in at least one species. By repeating the analysis on a reduced data set that excludes interdomain linkers, we show that our results are not caused by an association of rare codons with solvent-accessible linker regions. Finally, we find that our results depend weakly on expression level; the association between optimal codons and buried sites exists at all expression levels, but increases in strength as expression level increases.

Introduction

Synonymous mutations are usually referred to as "silent," but increasing evidence shows that they experience significant selection pressures in a wide range of organisms. For example, selection on synonymous sites has been linked to transcription, splicing, DNA secondary structure, and messenger RNA secondary structure and stability (Xia 1996; Vinogradov 2003; Chamary and Hurst 2005a, 2005b; Hoede et al. 2006; Parmley et al. 2006; Warnecke and Hurst 2007; Stoletzki 2008). The strongest selection pressure on synonymous sites, at least in microbes, seems to be selection for translational efficiency. This pressure causes highly expressed genes to be encoded predominantly by codons corresponding to highly abundant transfer RNAs (tRNAs). It has been observed in bacteria, plants, yeast, fly, worm, and even mammals (Ikemura 1981, 1985; Sharp et al. 1986; Akashi and Eyre-Walker 1998; Duret 2002; Urrutia and Hurst 2003; Comeron 2004; Wright et al. 2004; Lavner and Kotlar 2005).

Selection for translational efficiency may reflect selection for rapid translation (speed selection), selection for translation with high fidelity (accuracy selection), or both. Akashi (1994) argued that selection for translational accuracy should lead to inhomogeneous codon usage within genes. More important sites (i.e., sites that are less robust to translation errors) should be more frequently encoded by codons with high fidelity than other sites. Akashi (1994) found such a signal in *Drosophila*. Later, others found similar results in *Escherichia coli*, yeast, worm, and mammals (Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008).

Akashi (1994) suggested to classify sites as important or not depending on whether they are conserved at the amino acid level in an orthologous sequence. But although there is good evidence that evolutionary conservation reflects functional (Lichtarge et al. 1996) and structural (Koshi and Goldstein 1995; Mirny and Shakhnovich 1999, 2001a, 2001b; Schueler-Furman and Baker 2003) constraints, multiple other factors influence evolutionary conservation as well. Among them are the divergence time between orthologous sequences, the background rate of amino acid substitutions (which can vary over several orders of magnitude among genes within the same organism), mutational biases, and random chance. Moreover, the relative frequency of mutations at sites with different biochemical properties (e.g., solvent accessibility) changes with sequence divergence (Sasidharan and Chothia 2007). Ultimately, instead of linking codon usage bias to conserved or variable sites, we would like to link codon usage bias to sites with specific biochemical properties. In line with this reasoning, Akashi (1994) carried out a second analysis in which he showed that preferred codon usage was increased in putative zink finger and homeodomain regions of transcription factors.

Motivated by Drummond and Wilke's (2008) hypothesis that translational accuracy selection minimizes the misfolding of mistranslated proteins, we test here whether translationally optimal codons are associated with structurally sensitive sites, that is, sites at which translation errors are particularly likely to cause misfolding. We focus on residues' solvent accessibility because substitutions in the solvent-shielded core of proteins tend to be particularly disruptive (Matthews 1993; Tokuriki et al. 2007). As a second measure of structural sensitivity, we use computationally predicted changes in free energy upon mutation ($\Delta\Delta G$ values).

We consider four model organisms, *E. coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Mus musculus*. We define translationally optimal codons as those that are overrepresented in highly expressed genes and address the following questions: 1) Are optimal codons more likely to encode residues in the core of proteins or on the surface? 2) Is such an association a general characteristic for all amino acids or does it depend on the type of amino acid encoded? 3) Are the results affected by

Key words: codon usage bias, optimal codon, protein structure, protein evolution, translational accuracy selection.

E-mail: cwilke@mail.utexas.edu.

Mol. Biol. Evol. 26(7):1571–1580. 2009
doi:10.1093/molbev/msp070
Advance Access publication April 6, 2009

gene expression level? 4) Are the results affected by an excess of rare codons in interdomain linkers? 5) Are optimal codons more likely to occur at sites for which computational modeling predicts that amino acid substitutions are particularly disruptive?

Materials and Methods

Structural Data

To match gene sequences to protein structures, we used the GTOP (Genomes TO Protein structures and functions) database (Kawabata et al. 2002). We saved a match in the database for further calculation if its region of similarity was longer than 80% of the protein length and its sequence identity was larger than 40% of the sequence in the Protein Data Bank (PDB). This process yielded 822, 403, 947, and 1464 matches in *E. coli*, *S. cerevisiae*, *D. melanogaster*, and *M. musculus*, respectively.

For each protein with a match, we obtained the corresponding 3D crystal structure from the PDB. For every match, we retained only the specific matching peptide chain. After aligning the gene sequence and the sequence from the crystal structure with MUSCLE (Edgar 2004), we calculated the percent solvent-accessible surface area for each aligned residue with the DSSP program (Kabsch and Sander 1983). We normalized these results by the reference surface areas of an extended Gly-X-Gly peptide (Creighton 1992). We considered residues with less than 25% relative solvent accessibility as buried. Because we discarded all but the matching peptide chain, residues involved in protein–protein interfaces were considered as exposed for the purpose of our study.

We used the Rosetta $\Delta\Delta G$ module (Kortemme and Baker 2002; Kortemme et al. 2004) to estimate the change in the free energy gap, $\Delta\Delta G$, for all 19 possible single-point amino acid substitutions at each site. Although most mutations are destabilizing ($\Delta\Delta G > 0$ kcal/mol) (Tokuriki et al. 2007), proteins are often quite mutationally robust (Markiewicz et al. 1994; Guo et al. 2004; Bloom et al. 2005; Bloom, Drummond, et al. 2006) and only approximately 20% of all mutations are significantly destabilizing ($\Delta\Delta G > 2.0$ kcal/mol) (Tokuriki et al. 2007). We considered mutations with $\Delta\Delta G > 3.0$ kcal/mol as strongly destabilizing mutations and calculated the "structural importance" of a site as its fraction of strongly destabilizing mutations.

Because the protein sequences in our data set are not always 100% identical to the sequences of the structure homologs in PDB (see fig. S1), we tested the effect of sequence dissimilarity on our structural data (solvent accessibility and structural importance). We collected 58 groups of structure homologs from the PDB-REPRDB database (Noguchi et al. 2001; Noguchi and Akiyama 2003), which is a database of representative protein chains selected from the PDB. Each group of structure homologs contained one representative protein and three homologs with sequence identity to the representative protein between 80% and 100%, between 60% and 80%, and between 40% and 60%, respectively. We used the following criteria to collect groups of structure homologs: We only included structures for which we found at least one homolog in each sequence-identity interval. If there was more than one homolog falling into one sequence-identity bin, we retained only the homolog with the lowest sequence identity to the representative protein. This choice made our test more conservative. For each group, we calculated the correlation between the solvent accessibility of the representative protein and its homologs and the correlation between the structural importance of the representative protein and its homologs, respectively. We found that both solvent accessibility and structural importance tended to correlate strongly among homologous proteins, but the correlation coefficient decreased with decreasing sequence identity (see fig. S2). For solvent accessibility, the median Spearman correlation coefficient ranged from 0.964 (80%–100% identity) to 0.871 (40%–60% identity), and for structural importance, it ranged from 0.797 (80%–100% identity) to 0.626 (40%–60% identity).

We obtained protein domain boundary information from the CATH domain structure database (http://www.cathdb.info/) (Orengo et al. 1997; Greene et al. 2007). We defined domain linkers to be regions that are centered at CATH domain boundaries and extend at least 10 residues in both directions. Loops (continuous peptides composed of residues with DSSP classes S, T, and "-"), which neighbor on or overlap this region were also considered as parts of domain linkers. Finally, we considered both termini of the protein as domain boundaries. These criteria are strict and yield a conservative analysis of nonlinker regions.

Genomic Data

We obtained genomic sequences from the following sources: the Comprehensive Microbial Resource (http://cmr.tigr.org/) for *E. coli*, the *Saccharomyces* Genome Database (ftp://genome-ftp.stanford.edu/) for *S. cerevisiae*, the Eisen Lab (http://rana.lbl.gov/drosophila/) for *D. melanogaster*, and Ensembl (http://www.ensembl.org/) for *M. musculus*.

We designated as evolutionarily conserved all sites at which the amino acid was unchanged compared with the orthologous gene in a closely related species. We used the following orthologs: For *E. coli*, we obtained orthologs between *E. coli* and *Salmonella typhimurium* from the Comprehensive Microbial Resource. For yeast, we obtained orthologs between *S. cerevisiae* and *Saccharomyces bayanus* from the *Saccharomyces* Genome Database. For fly, we obtained orthologs between *D. melanogaster* and *Drosophila yakuba* from the *Drosophila* 12-genome project AAAWiki at http://rana.lbl.gov/drosophila/. For mouse, we obtained orthologs between *M. musculus* and *Rattus norvegicus* from Biomart through the Ensembl Homology track. We aligned each pair of orthologs at the peptide level using MUSCLE (Edgar 2004).

Expression Data

We used previously published expression data for each species: For *E. coli*, we obtained gene expression levels measured in messenger RNAs per cell from Covert et al. (2004); for *S. cerevisiae*, we used expression data from Holstege et al. (1998); for *D. melanogaster*, we used as

Table 1
Example of a 2×2 Contingency Table for Codon ACT in One Particular Gene in *Escherichia coli*

Codon	Buried Sites	Exposed Sites
ACT	15	5
ACC, ACA, ACG	3	6

NOTE.—Codon ACT encodes amino acid Thr. The other three non-ACT codons encoding Thr are ACC, ACA, and ACG. The odds ratio of ACT usage between buried and exposed sites is $(15/5)/(3/6) = 6$ for this contingency table. Because there is one table of ACT per gene, we applied the Mantel–Haenszel procedure to calculate the joint odds ratio of use frequency between buried and exposed sites for all genes.

expression level the geometric mean of expression data from different tissues obtained by Stolc et al. (2004); and for *M. musculus*, we measured expression level as the breadth of expression among different tissues (Su et al. 2004). After combining the expression data with the structural data, we ended up with 698, 384, 123, and 569 genes for *E. coli*, *S. cerevisiae*, *D. melanogaster*, and *M.musculus*, respectively.

Inferring Optimal Codons

To identify which codons are translationally optimal in each species, we compared the codon usage pattern between the gene groups showing the lowest 5% and highest 5% expression level in each species. We defined codons as "optimal" if they showed a statistically significant increase in frequency in the highly expressed group, as determined by a chi-square test. We defined codon optimality (C_{opt}) as the odds ratio of codon usage between highly and lowly expressed groups, calculated separately for each codon:

$$C_{\text{opt}} = \frac{n_{\text{high}}/(N_{\text{high}} - n_{\text{high}})}{n_{\text{low}}/(N_{\text{low}} - n_{\text{low}})}. \qquad (1)$$

Here, n_{high} and n_{low} are the observed numbers of the codon in the highly and lowly expressed groups, respectively, and N_{high} and N_{low} are the observed numbers of the corresponding amino acid in the highly and lowly expressed groups, respectively.

Statistical Tests of Association

We used two different methods to test for association, one using categorical variables and one using continuous variables. For pairs of categorical variables (e.g., optimal vs. nonoptimal codons and buried vs. exposed sites), we stratified the data by gene and synonymous codon family within each gene and constructed a separate 2×2 contingency table for each stratum. We then combined either the tables for all genes and a given codon family or the tables for all genes and all codon families into an overall analysis using the Mantel–Haenszel procedure (Mantel and Haenszel 1959; Mantel 1963). The null hypothesis in this analysis assumes that the status of the site (e.g., buried or exposed) is independent of the codon type in any given stratum. Because the Mantel–Haenszel procedure yields undefined results on contingency tables whose sum of all four entries is less than 2 (i.e., 0 or 1), we excluded all such tables from the analyses.

For pairs of continuous variables (e.g., codon optimality and solvent accessibility), we also stratified the data by gene and synonymous codon family within each gene. Then, for each stratum, we separately calculated the Pearson correlation coefficient between the two variables. As test statistic, we used the mean of the correlation coefficients over all strata. We calculated the sampling distribution by randomly reshuffling, separately for each gene, synonymous codons among sites with identical amino acid, and recalculating all correlation coefficients. We generated 1,000 resampled sequences for each gene.

We carried out all statistical analyses using the software R (R Development Core Team 2008). In the analyses of individual amino acids, we corrected for multiple testing using the false discovery rate method of Benjamini and Hochberg (1995), as implemented in the R function `p.adjust()`.

Results

Optimal Codons are Preferred at Buried Sites

We first assessed whether there was any relationship between a codon's tendency to be preferentially used in highly expressed genes and the same codon's tendency to be preferentially used at buried sites. We calculated 2 odds ratios each for 59 codons (excluding ATG for Met, TGG for Trp, and three stop codons). The first odds ratio, which we refer to as codon optimality (C_{opt}, see Materials and Methods), measures whether the codon is preferred in highly expressed genes compared with all other codons encoding the same amino acid. The second odds ratio, which we denote by O_{buried}, measures whether the codon is preferred at buried sites compared with all other codons encoding the same amino acid. To control for confounding effects of differing amino acid usage among genes, we calculated O_{buried} by first constructing 2×2 contingency tables of codon usage within each gene (see table 1 for an example) and then using the Mantel–Haenszel procedure (Mantel and Haenszel 1959; Mantel 1963) to combine the odds ratios for each individual contingency table into an overall odds ratio O_{buried}. We list the values of C_{opt} and O_{buried} for each codon in table S1. In all species except fly, we found a significant positive correlation between C_{opt} and O_{buried} (Spearman's $\rho = 0.38, P = 0.003$ for *E. coli*; $\rho = 0.54, P < 0.001$ for *S. cerevisiae*; $\rho = 0.24, P = 0.069$ for *D. melanogaster*; and $\rho = 0.78, P \ll 0.001$ for *M. musculus*; see also fig. 1).

The correlation between C_{opt} and O_{buried} reveals that there is an association between codon usage and protein structure. To determine whether this correlation is consistent across all amino acids or if different amino acids have different trends, we carried out a similar statistical test on each amino acid separately. We inferred a set of optimal codons for each species (see Materials and Methods and table S2). For each gene, we then constructed separate 2×2 contingency tables for the 18 amino acids encoded by at least two codons (see table 2 for an example). For each of these 18 amino acids, we calculated a joint odds ratio (O_{joint}) of optimal codon usage between buried and exposed sites using the Mantel–Haenszel procedure. A value of O_{joint} greater than 1 signifies a preference for

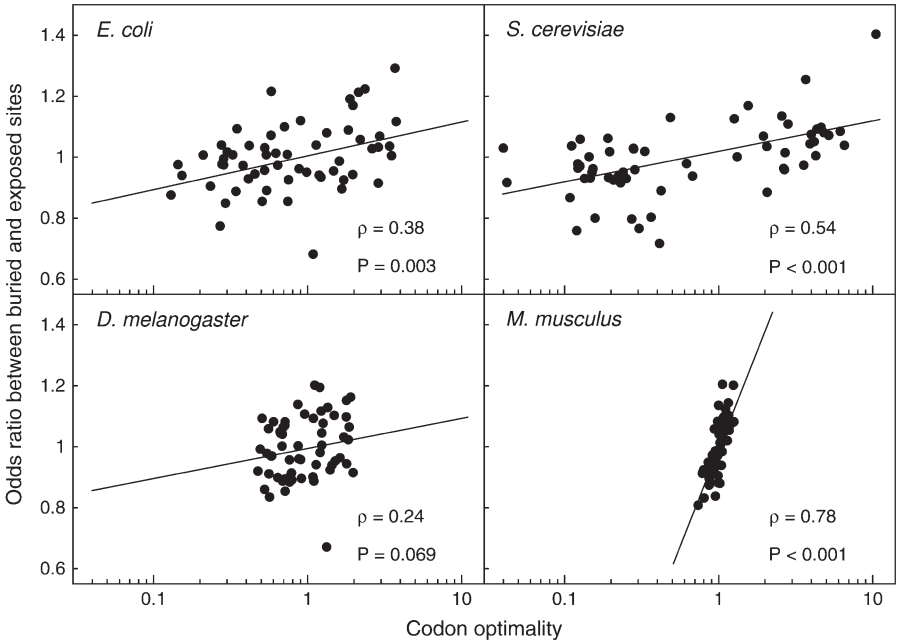

FIG. 1.—Odds ratio O_{buried} versus codon optimality C_{opt}. With the exception of fly, all organisms show a significant correlation between these two quantities.

optimal codons at buried sites (and nonoptimal codons at exposed sites).

We found that 13 of 18 amino acids showed, in at least one species, a significant preference for optimal codons at buried residues (table 3). Unexpectedly, in *E. coli*, 2 of these 13 amino acids (Ala and Val) showed a significant preference for optimal codons at exposed residues. The remaining codons (Cys, Glu, His, Pro, and Tyr) showed no significant preference for optimal codons to be buried or exposed in any of the four species tested. Of a total of 72 association tests, 23 showed a significant preference for buried optimal codons, whereas only 2 showed a significant preference for exposed optimal codons.

Why did some amino acids show a signal and others did not? We found no clear pattern related to amino acid biochemistry, such as polarity or volume. Most amino acids showed a signal in at least one species. Instead, we found that amino acid frequency was the best predictor of a significant association between codon optimality and solvent exposure. We observed a negative correlation between the significance level (*P* value after correction for multiple testing) from the Mantel–Haenszel test and the relative amino acid frequency (Spearman's $\rho = -0.33, P = 0.007$, data of all four species pooled; see also fig. 2). This finding suggests that the absence of a significant association for some amino acids is likely caused by lack of statistical power rather than by a specific biochemical mechanism.

For each species, we also used the Mantel–Haenszel procedure to combine all 2×2 contingency tables for all genes and all amino acids into a single overall odds ratio. This analysis corresponds to Drummond and Wilke (2008)'s analysis but using buried sites instead of evolutionary conserved sites. We found a statistically significant association between optimal codons and buried sites in all species (table 3). These results were not strongly dependent on solvent-accessibility cutoff used to identify buried residues (table S3).

To control for evolutionary conservation, we tested for an association between optimal codons and buried sites considering only evolutionarily conserved residues in each species. The results remained largely unchanged from the results including all residues (table S4).

Because many ribosomal proteins contain natively unstructured regions, which assume their structure only upon binding other parts of the ribosome, we also repeated the association test on a data set excluding all ribosomal proteins. Our full data set contained 40 ribosomal proteins in *E. coli*, 29 in yeast, 22 in fly, and 14 in mouse. The results remained largely unchanged after excluding these genes (table S5).

To determine if the association between optimal codons and buried sites was affected by expression level, we calculated the overall odds ratio separately for the highest (top 25%) and lowest (bottom 25%) expressed genes (see fig. 3). In all species except mouse, the overall odds

Table 2
Example of a 2×2 Contingency Table for Amino Acid Thr in One Particular Gene in *Escherichia coli*

	Codon	Buried Sites	Exposed Sites
Optimal	ACT, ACC	16	6
Nonoptimal	ACA, ACG	2	5

NOTE.—Codons ACT and ACC are optimal codons for amino acid Thr in *E. coli* (see table S2). The odds ratio of optimal codon usage between buried and exposed sites is $(16/6)/(2/5) = 6.67$ for this contingency table. Because there is one table of Thr per one gene, we applied the Mantel–Haenszel procedure to calculate the joint odds ratio for all tables of Thr across all genes.

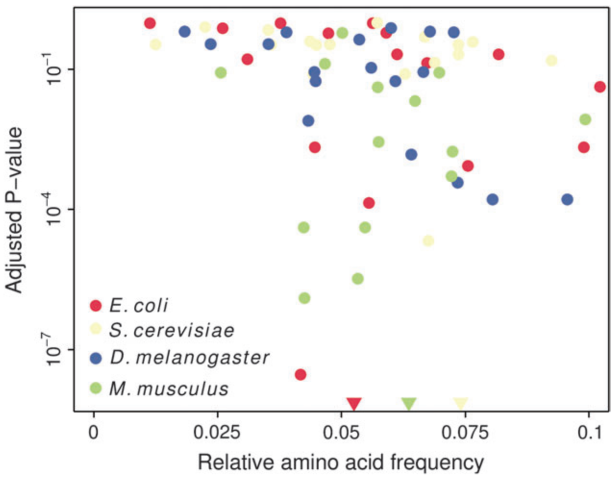

FIG. 2.—Mantel–Haenszel P value as a function of relative use frequency of each amino acid. The P value is adjusted for multiple testing (Benjamini and Hochberg 1995). Triangles indicate three data points that fall below the bottom of the graph.

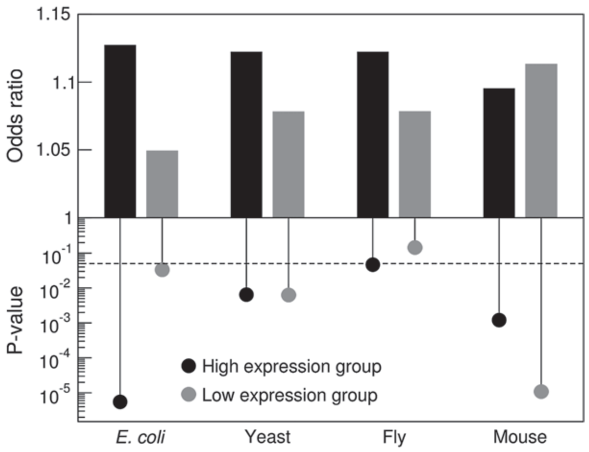

FIG. 3.—The overall joint odds ratio O_{joint} and the corresponding P value for the 25% highest and lowest expressed genes in all organisms. The dashed line denotes the significance level of $\alpha = 0.05$.

ratio for the highest expressed genes tended to be higher than the one for the lowest expressed genes. However, with the exception of fly, which had the smallest sample size, the overall odds ratio was significantly larger than 1.0 even for genes with low expression level. Thus, highly expressed genes seem to show a stronger association between optimal codons and buried sites than genes with low expression level, but even the latter genes do show a significant association. This result mirrors the general observation that evolutionary constraints appear to increase with gene expression level (Duret and Mouchiroud 2000; Pal et al. 2001; Lemos et al. 2005; Drummond et al. 2006; Wolf et al. 2006; Eames and Kortemme 2007; Drummond and Wilke 2008).

Codon Optimality Correlates Negatively with Solvent Accessibility

The Mantel–Haenszel procedure properly controls for potentially confounding effects of differing codon or amino acid usage frequencies among genes. Its main drawback is that it requires categorical data, such as a classification of all residues into buried or exposed, or of all codons into optimal or nonoptimal. Solvent accessibility and codon optimality are continuous quantities, and by forcing them into dichotomous categories, we may be losing statistical power.

We devised a new statistical test in the spirit of the Mantel–Haenszel procedure but that made use of the specific values of codon optimality and solvent accessibility for each residue. For each amino acid in each gene, we separately calculated the Pearson correlation coefficient between the codon optimality of all codons encoding this amino acid and the solvent accessibility of the amino acid in the 3D protein structure. As test statistic, we used the mean of all these correlation coefficients. We calculated the sampling distribution of this statistic by randomly permuting synonymous codons within each gene (see Materials and Methods). We then carried out one-tailed tests. Our alternative hypothesis was that the mean correlation coefficient should be more negative than expected by chance if optimal codons associate with low solvent accessibility.

When we combined the correlation coefficients for all amino acids, we found that, for all organisms, we could reject the null hypothesis of no significant association between codon optimality and solvent accessibility. We found $P < 0.001$ for *E. coli*, $P < 0.001$ for yeast, $P = 0.005$ for fly, and $P < 0.001$ for mouse (fig. 4). The analysis for individual amino acids largely mirrored our Mantel–Haenszel results (fig. S3). However, the statistical power of the association test on continuous variables seemed to be generally lower than that of the Mantel–Haenszel procedure. In general, if we found a significant result with the association test on continuous variables, we also found it in the Mantel–Haenszel procedure, but the reverse was not true in all cases.

The advantage of the association test on continuous variables is that we could employ it also for amino acids for which we could not clearly identify optimal codons. These amino acids include Lys for *E. coli* and Cys, Ala, and Tyr for mouse. Of these four, Cys in *E. coli* showed a marginally significant result and Ala in mouse showed a strongly significant result (fig. S3).

Codon Optimality Correlates with Structural Importance

So far, we tested for an association between optimal codons and buried sites. Our reasoning was that mutations at buried sites are more disruptive than mutations at exposed sites and that therefore buried sites should be more sensitive to translation errors. An alternative, and more direct, way of assessing whether a site is sensitive to mutation is to determine the distribution of free energy changes ($\Delta\Delta G$ values) that substitutions at this site effect. Mutations with negative $\Delta\Delta G$ stabilize the protein fold and should typically not have deleterious effects. Mutations with a small positive $\Delta\Delta G$ are mildly destabilizing. Whether such mutations will be disruptive or not is hard to predict. But once $\Delta\Delta G$ exceeds 2–3 kcal/mol, the mutation is almost certainly disruptive. Thus, as a measure of a site's

Table 3
Odds Ratio of Optimal Codon Usage between Buried and Exposed Sites

	Escherichia coli		*Saccharomyces cerevisiae*		*Drosophila melanogaster*		*Mus musculus*	
Amino Acid	Whole	Nonlinker	Whole	Nonlinker	Whole	Nonlinker	Whole	Nonlinker
Ala	0.92^{*}	0.93	1.06	1.06	1.13^{***}	$1.15^{**(*)}$	—	—
Arg	1.01	0.97	$1.19^{(*)}$	1.05	0.96	0.96	1.15^{***}	1.17^{***}
Asn	1.29^{***}	1.25^{***}	1.07	$1.16^{(*)}$	$1.10^{(*)}$	$1.12^{(*)}$	1.14^{***}	1.22^{***}
Asp	1.06	1.08	$0.89^{(*)}$	$0.86^{(*)}$	$0.93^{(*)}$	$0.89^{*(*)}$	1.09^{**}	1.14^{***}
Cys	0.99	0.99	1.19	1.30	0.96	0.91	—	—
Gln	$1.17^{**(*)}$	1.26^{***}	1.04	0.99	1.04	0.93	1.21^{***}	1.34^{***}
Glu	1.08	1.10	1.05	1.09	1.02	1.07	1.05	1.03
Gly	1.06	1.08	1.42^{***}	1.29^{***}	0.98	$0.92^{(*)}$	1.09^{***}	1.09^{**}
His	1.03	1.07	0.97	1.08	0.94	0.88	1.07	1.06
Ile	1.03	1.03	$1.17^{(*)}$	$1.27^{(*)}$	$1.14^{**(*)}$	$1.19^{**(*)}$	1.07^{*}	1.12^{*}
Leu	$1.11^{**(*)}$	1.08	1.10	$1.18^{(**)}$	1.15^{***}	$1.15^{**(*)}$	1.07^{**}	$1.11^{**(*)}$
Lys	—	—	1.07	1.07	$0.92^{(*)}$	0.91	1.08^{*}	$1.15^{**(*)}$
Phe	1.01	0.98	1.07	1.06	1.16^{**}	1.10	1.05	1.08
Pro	1.03	1.06	1.09	1.08	$1.08^{(*)}$	1.09	0.98	1.00
Ser	1.36^{***}	1.41^{***}	1.25^{***}	1.27^{***}	1.01	1.04	1.22^{***}	1.32^{***}
Thr	1.17^{***}	$1.15^{*(*)}$	1.00	1.02	1.06	1.03	1.13^{***}	1.12^{**}
Tyr	0.92	0.86	1.09	1.08	0.94	0.89	—	—
Val	0.88^{***}	0.89^{*}	1.10	$1.20^{(*)}$	1.16^{***}	$1.14^{*(*)}$	$1.09^{**(*)}$	1.14^{***}
Overall	1.06^{***}	1.07^{***}	1.10^{***}	1.11^{***}	1.04^{***}	1.03^{**}	1.10^{***}	1.13^{***}

NOTE.—Significance levels in parentheses disappear after correction for multiple testing.Whole, odds ratio for whole protein sequences; nonlinker, odds ratio for sequences without domain boundary region; —, no optimal codon.

$^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

sensitivity to substitutions, we calculated $\Delta\Delta G$ values for all 19 possible amino acid substitutions at that site and then determined the fraction of these substitutions with $\Delta\Delta G > 3.0$ kcal/mol. We refer to this fraction as the "structural importance" of the site because it assesses the likelihood that a mutation at this site is structurally disruptive.

Our hypothesis was that if selection for translational accuracy acts to minimize mistranslation-induced protein misfolding then sites with higher structural importance should associate with more optimal codons and vice versa. By classifying sites at which at least two mutations had $\Delta\Delta G > 3.0$ kcal/mol as important sites and all other sites as unimportant sites, we could employ the Mantel–Haenszel procedure to determine whether optimal codons associated with structurally important sites. Our results were similar to our previous results considering buried and exposed sites. In all organisms, the overall joint odds ratio was significantly larger than 1.0 (table S6). Our results for individual amino acids were also largely consistent with those obtained for buried/exposed sites, but there were some differences. For example, Glu in *E. coli* and Cys in yeast showed a significant association between optimal codons and structurally important sites but not between optimal codons and buried sites. By contrast, in fly and mouse, we mostly found the opposite result that several amino acids showed a significant association between optimal codons and buried sites but not between optimal codons and structurally important sites.

We also carried out the test of association between continuous variables. Because we expected codon optimality to increase with structural importance, we calculated one-tailed P values for the right tail of the sampling distribution of the mean correlation coefficient. We found a significant association between optimal codons and structurally important sites in yeast ($P < 0.001$) and mouse ($P < 0.001$) but not in *E. coli* ($P = 0.230$) or fly ($P = 0.230$) (fig. S4). When considering the results for individual amino acids (fig. S5), we found again that they largely agreed with those of the Mantel–Haenszel procedure on the same data (table S6) but that the association test on continuous variables seemed to be overall less powerful. However, there were a few notable exceptions. Gln in mouse showed a highly significant association when considering continuous variables but no association under the Mantel–Haenszel procedure. For Ser in fly, optimal codons seemed to associate with structurally unimportant sites when considering continuous variables (a right-tailed P value of 1 corresponds to a left-tailed P value of < 0.001), but we found no such signal under the Mantel–Haenszel procedure.

Nonrandom Pattern in Nonlinker Regions

Thanaraj and Argos (1996b) reported that nonoptimal codons are overrepresented in linker regions between domains. Because linker regions are generally solvent exposed and less sensitive to amino acid substitutions, an excess of nonoptimal codons in these regions could lead to an apparent excess of optimal codons at buried or structurally important sites. To exclude the possibility that our results are caused by abundant nonoptimal codons in interdomain linkers, we repeated all our analyses but excluded linker regions (see Materials and Methods). The new results were largely unchanged from the previous ones. Tables 3 and S6 and figures 4 and S4 show a side-by-side comparisons of the same results for entire proteins and proteins with interdomain linkers excluded. Figures S6 and S7 show the same analyses as figures S3 and S5, respectively, but with interdomain linkers excluded. There were some cases in which results disappeared, for example, in table 3 for Leu and Ala in *E. coli* and for Phe in fly. Yet by and

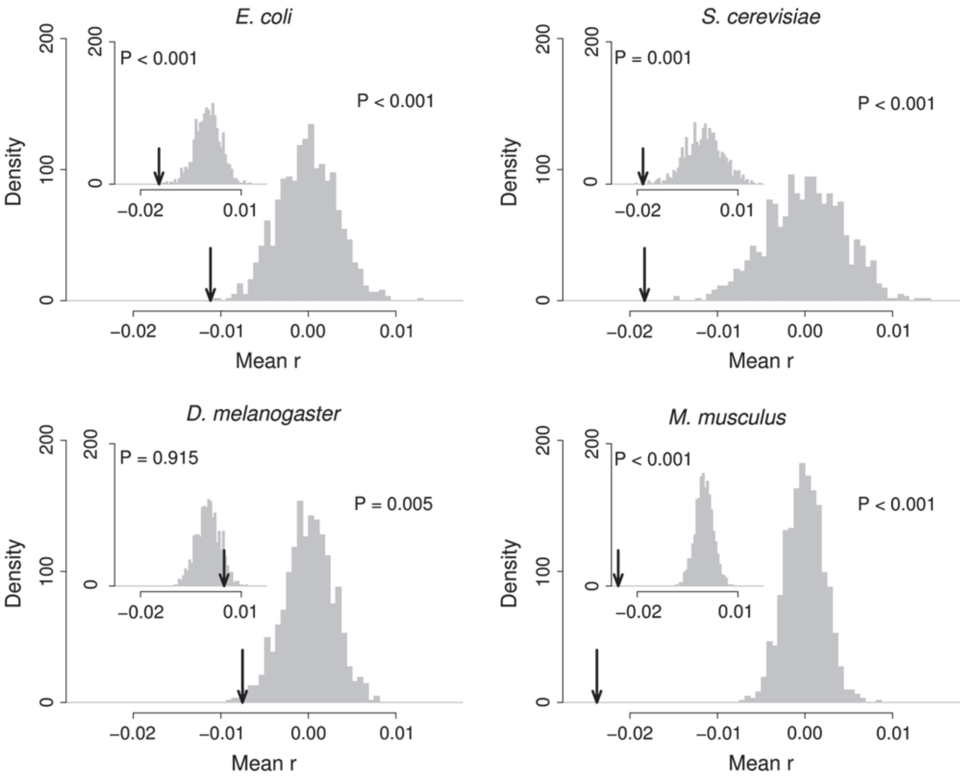

FIG. 4.—Test for association between codon optimality and solvent accessibility. The black arrows indicate the mean correlation coefficient between these two quantities over all amino acids and all genes. The gray histograms show the sampling distribution of the same quantity under the null hypothesis of no association. The main figure of each panel shows the results for complete genes, and the inset shows the results when interdomain linker regions are excluded.

large, we could confirm that even when we excluded interdomain linker regions, we found a significant association between optimal codons and either buried sites or structurally important sites.

Discussion

We examined the relationship between codon usage bias and protein structure in the genomes of four model organisms. We found that optimal codons tend to be associated with buried sites. In *E. coli*, yeast, and fly, the association was stronger in highly expressed genes than in genes with low expression level. The effect was present in most codon families in at least one organism. We found no clear relationship between the biophysical properties of the encoded amino acids and the presence or absence of an association between optimal codons and buried sites. Instead, the best predictor for a significant association was amino acid frequency. We also found that optimal codons tend to be associated with sites for which computational modeling predicts that substitutions are destabilizing.

This study is not the first to provide evidence that codon usage bias is affected by protein structure; previous studies include Thanaraj and Argos (1996a), Orešič and Shalloway (1998), Xie and Ding (1998), Orešič et al. (2003), and Gu et al. (2004). However, many of the previous studies suffer from statistical limitations such as not correcting for multiple testing or not controling for confounding effects of amino acid usage frequencies. Most previous studies suggest that the relationship between protein structure and codon usage bias is rather limited. For example, in a study of *E. coli* and human, Orešič and Shalloway (1998) found only a single codon each that associated with structural features of the encoded proteins. In human, this codon is GAU, and it is preferred at the N termini of α-helices. Most previous studies focused on protein secondary structure, whereas the structural features we considered here were solvent accessibility and structural importance (i.e., $\Delta\Delta G$ values). We believe that we found a comparatively strong and pervasive signal in part because of this choice. The extent to which mutations tend to disrupt protein folding and function depends only weakly on secondary structure and much more strongly on solvent accessibility (Goldman et al. 1998; Bustamante et al. 2000; Dean et al. 2002; Bloom, Labthavikul, et al. 2006).

What is the biophysical mechanism that links codon usage bias to protein structure? Experimental work has focused on the hypothesis that ribosome kinetics may affect protein folding (Komar et al. 1999; Cortazzo et al. 2002; Goymer 2007; Komar 2007; Kimchi-Sarfaty et al. 2007). A cluster of suboptimal codons might stall the ribosome and as a consequence either facilitate or disrupt cotranslational folding. Thanaraj and Argos (1996b) argued that rare codons associate with interdomain linkers because the slowdown of the ribosome when translating these linker regions would allow the individual domains to fold independently. Yet more recently, Widmann et al. (2008) showed that evolutionarily conserved (and thus presumably important) rare codons are not restricted to interdomain linker regions. To verify that our results were not caused by an excess of rare codons in solvent-exposed linker regions, we redid our analysis on a restricted data set that excluded all interdomain linker regions. We again found an association between optimal codons and structurally sensitive sites. Thus, although nonoptimal codons may or may not be preferred in interdomain linker regions, our results are not caused by the codon usage in these linkers.

Most previous work has focused on the interplay between translation and folding kinetics, but translation fidelity may be as important or more so in shaping synonymous codon usage. According to the mistranslation-induced misfolding hypothesis (Drummond and Wilke 2008), selection for translational accuracy should lead to an excess of high-fidelity codons at sites at which translation errors would be particularly destabilizing. Our results are broadly consistent with this hypothesis, even though our evidence is only indirect. Few studies have attempted the direct measurement of codon fidelity under translation or of the propensity of mutations at different sites to cause misfolding. Thus, we had no comprehensive data set for either of these quantities and had to use proxies for both.

We equated high-fidelity codons with optimal codons and identified as optimal those codons that were significantly more frequent in highly expressed genes than in genes with low expression level. Because the sets of optimal codons we determined in this way were largely consistent with sets of optimal codons determined by counts of tRNA genes (Ikemura 1985; Moriyama and Powell 1997; Man and Pilpel 2007; Drummond and Wilke 2008), we believe that our optimal codons are by and large the high-fidelity codons of the respective organism. However, we cannot be certain that we correctly identified high-fidelity codons in all cases. In fact, for the two amino acids for which optimal codons were associated with exposed sites in our analysis (Ala and Val in *E. coli*), the codons we determined to be optimal may be of low fidelity. One issue that may arise in this context is that of speed-accuracy tradeoffs. The most rapidly translated codon may not be the most accurately translated one or vice versa because speed should be determined primarily by the absolute number of tRNA copies in a cell, whereas accuracy should depend on the relative abundance of the cognate tRNA compared with competing tRNAs. If an organism experiences both selection for translation speed and translational accuracy then it is possible that the most rapidly translated codon is the most abundant one in highly expressed genes but that the most accurately translated codon is preferred at sites at which translation errors need to be avoided.

Our analysis also assumes that all genes in an organism have the same optimal codons. This assumption may be violated if tRNA pools vary substantially over time or among tissues in multicellular organisms. Such differences have indeed been reported (Dong et al. 1996; Dittmar et al. 2006), but they do not appear to be large enough to invalidate our approach. Nevertheless, a future study could try to obtain more accurate, gene-specific codon optimality values.

As a measure of the extent to which translation errors at a site may lead to misfolding, we used two quantities, the solvent-accessible surface area and the site's structural importance as measured by the computationally predicted stability effects ($\Delta\Delta G$ values) of mutations at that site. Both of these quantities can be calculated only if an accurate 3D crystal structure is available. Because the number of crystal structures for proteins of a specific organism remains limited, we augmented our data sets with crystal structures from homologous proteins. Consequently, for many genes in our data sets, solvent accessible surface areas and $\Delta\Delta G$ values are only estimates. To assess how reliable these estimates are, we carried out a controlled analysis of how these quantities change with decreasing sequence identity among homologs. We found that both remain highly predictive even if the homologs have diverged substantially (fig. S2). We also found that solvent-accessible surface area is generally more conserved among homologs than $\Delta\Delta G$ values are.

We predicted $\Delta\Delta G$ values using the default energy function of the Rosetta $\Delta\Delta G$ module. This energy function was not necessarily optimized for the wide range of different protein structures to which we applied it. However, because we were using the $\Delta\Delta G$ predictions only to find sites at which substitutions have large disruptive effects, such as would be caused by steric clashes of the side chains, the $\Delta\Delta G$ predictions should be reasonably accurate nevertheless.

Previous works (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008) studied selection for translational accuracy using the categorical Mantel–Haenszel procedure and we followed the same strategy here. But because the quantities we studied are not inherently categorical, we also devised an association test on continuous variables based on stratified correlation coefficients. The two statistical procedures yielded overall similar results, but the test on continuous variables seemed to be less powerful. Yet, this test has several advantages that make it worthy to pursue. First, the results of the categorical test depend on the arbitrary choice by which continuous variables are classified into categories. For example, we designated sites as buried if their relative solvent accessibility fell below a 25% cutoff. A different choice of cutoff led to somewhat (if only slightly) different results (table S3). The test on continuous variables does not suffer from this problem. Second, according to the criterion we chose to identify optimal codons, we could not select optimal codons for a few amino acids in some species. Thus, we had to exclude these amino acids from the categorical analysis but could include them in the continuous analysis.

The overall odds ratios we obtained from the Mantel–Haenszel procedure are of comparable magnitude, but generally somewhat smaller, than the odds ratios measuring the association between optimal codons and evolutionarily conserved sites (table 1 in Drummond and Wilke 2008). The biggest deviation arises in fly where we found an odds ratio of 1.04, whereas Drummond and Wilke (2008) found 1.36. Some of these differences are likely caused by the limitations on quality and quantity of structural data. Fly in particular had the lowest fraction of closely matching crystal structures (fig. S1). An alternative explanation for the smaller odds ratios in our study could be that a significant proportion of sites under translational accuracy selection are functionally important rather than structurally important and that the criterion of evolutionary conservation accurately identifies these sites. Future work could try to disentangle structural and functional constraints by assembling a data set of proteins for which structurally important sites are known and then testing whether optimal codons associate more strongly with structurally important sites, functionally important sites, or possibly the joint set of both structurally and functionally important sites.

Supplementary Material

Supplementary tables S1–S6 and figures S1–S7 are available online as supplementary material at Molecular Biology and Evolution (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank D. Allan Drummond for stimulating discussions. This work was supported in part by NIH grant R01 AI065960. The University of Texas provided computing resources.

Literature Cited

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

Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev. 8:688–693.

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.

Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol. 23:1751–1761.

Bloom JD, Labthavikul ST, Otey CR, Arnold FH. 2006. Protein stability promotes evolvability. Proc Natl Acad Sci USA. 103:5869–5874.

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005. Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci USA. 102:606–611.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. Mol Biol Evol. 17:301–308.

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.

Comeron JM. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics. 167:1293–1304.

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.

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. 2004. Integrating high-throughput and computational data elucidates bacterial networks. Nature. 429:92–96.

Creighton TE. 1992. Proteins: structures and molecular properties. 2 ed. New York: Freeman.

Dean AM, Neuhauser C, Grenier E, Golding GB. 2002. The pattern of amino acid replacements in α/β-barrels. Mol Biol Evol. 19:1846–1864.

Dittmar KA, Goodenbour JM, Pan T. 2006. Tissue-specific differences in human transfer RNA expression. PLoS Genet. 2:e221.

Dong HJ, 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.

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

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68–74.

Eames M, Kortemme T. 2007. Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. Structure. 15:1442–1451.

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

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 149:445–458.

Goymer P. 2007. Synonymous mutations break their silence. Nat Rev Genet. 8:92.

Greene LH, Lewis TE, Addou S, et al. (14 co-authors). 2007. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 35:D291–D297.

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.

Guo HH, Choe J, Loeb LA. 2004. Protein tolerance to random amino acid change. Proc Natl Acad Sci USA. 101: 9205–9210.

Hoede C, Denamur E, Tenaillon O. 2006. Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. PLoS Genetics. 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.

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.

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

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22:2577–2637.

Kawabata T, Fukuchi S, Homma K, Ota M, Araki J, Ito T, Ichiyoshi N, Nishikawa K. 2002. GTOP: a database of protein structures predicted from genome sequence. Nucleic Acids Res. 30:294–298.

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.

Komar AA. 2007. SNPs, silent but not invisible. Science. 315:466–467.

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.

Kortemme T, Baker D. 2002. A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci USA. 99:14116–14121.

Kortemme T, Kim DE, Baker D. 2004. Computational alanine scanning of protein-protein interfaces. Science STKE. 219:l2.

Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices. Protein Eng. 8:641–645.

Lavner Y, Kotlar D. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene. 345:127–138.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 22:1345–1354.

Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 257:342–358.

Man O, Pilpel Y. 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet. 39:415–421.

Mantel N. 1963. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. J Am Stat Assoc. 58:690–700.

Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 22:719–748.

Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH. 1994. Genetic studies of the *lac* repressor. xiv. analysis of 4000 altered *Escherichia coli lac* repressors reveals essential and non-essential residues, as well as “spacers” which do not require a specific sequence. J Mol Biol. 240:421–433.

Matthews BW. 1993. Structural and genetic analysis of protein stability. Annu Rev Biochem. 62:139–160.

Mirny LA, Shakhnovich EI. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol. 291: 177–196.

Mirny LA, Shakhnovich EI. 2001a. Evolutionary conservation of the folding nucleus. J Mol Biol. 308:123–129.

Mirny LA, Shakhnovich EI. 2001b. Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct. 30:361–396.

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

Noguchi T, Akiyama Y. 2003. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res. 31:492–493.

Noguchi T, Matsuda H, Akiyama Y. 2001. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). Nucleic Acids Res. 29:219–220.

Orengo CA, Michie AD, Jones DT, Swindells MB, Thornton JM. 1997. CATH: a hierarchic classification of protein domain structures. Structure. 5:1093–1108.

Orešič M, Dehn M, Korenblum D, Shalloway D. 2003. Tracing specific synonymous codon-secondary structure correlations through evolution. J Mol Evol. 56:473–484.

Orešič M, Shalloway D. 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol. 281:31–48.

Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics. 158:927–931.

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.

R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Sasidharan R, Chothia C. 2007. The selection of acceptable protein mutations. Proc Natl Acad Sci USA. 104:10080–10085.

Schueler-Furman O, Baker D. 2003. Conserved residue clustering and protein structure prediction. Proteins. 52:225–235.

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.

Stolc V, Gauhar Z, Mason C, et al. (12 co-authors). 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science. 306:655–660.

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.

Su AI, Wiltshire T, Batalov S, et al. (13 co-authors). 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA. 101:6062–6067.

Thanaraj TA, Argos P. 1996a. Protein secondary structural types are differentially coded on messenger RNA. Protein Sci. 5:1973–1983.

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

Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. 2007. The stability effects of protein mutations appear to be universally distributed. J Mol Biol. 369:1318–1332.

Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. Genome Res. 13:2260–2264.

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 sauge 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.

Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. Proc Biol Sci. 273:1507–1515.

Wright SI, Yau CB, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol Biol Evol. 21:1719–1726.

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.

Michele Vendruscolo, Associate Editor

Accepted March 27, 2009