Translational Selection and Yeast Proteome Evolution

Hiroshi Akashi[1]

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802

Manuscript received October 11, 2002
Accepted for publication March 14, 2003

ABSTRACT

The primary structures of peptides may be adapted for efficient synthesis as well as proper function. Here, the *Saccharomyces cerevisiae* genome sequence, DNA microarray expression data, tRNA gene numbers, and functional categorizations of proteins are employed to determine whether the amino acid composition of peptides reflects natural selection to optimize the speed and accuracy of translation. Strong relationships between synonymous codon usage bias and estimates of transcript abundance suggest that DNA array data serve as adequate predictors of translation rates. Amino acid usage also shows striking relationships with expression levels. Stronger correlations between tRNA concentrations and amino acid abundances among highly expressed proteins than among less abundant proteins support adaptation of both tRNA abundances and amino acid usage to enhance the speed and accuracy of protein synthesis. Natural selection for efficient synthesis appears to also favor shorter proteins as a function of their expression levels. Comparisons restricted to proteins within functional classes are employed to control for differences in amino acid composition and protein size that reflect differences in the functional requirements of proteins expressed at different levels.

THE predominant view of protein evolution considers fitness effects of amino acid changes that arise from gene-specific relationships between the primary structures of encoded polypeptides and their particular function(s) (Nei 1975; Kimura 1983; Li 1997). Critical properties of proteins (*i.e.*, specificity, activity, or stability) depend on particular amino acids in specific regions of their structures. Mutation pressure and genetic drift determine encoded amino acids and their evolutionary divergence at sites where protein function is more tolerant to amino acid replacements.

Selection pressures related to efficient synthesis, rather than to proper function, of proteins are less firmly established. Amino-acid-altering mutations could affect fitness through physiological effects that are independent of their effects on protein function. Amino acids may vary in the energetic costs of their biosynthesis (Richmond 1970; Karlin and Bucher 1992; Lobry and Gautier 1994; Dufton 1997; Craig and Weber 1998; Garat and Musto 2000; Jansen and Gerstein 2000; Akashi and Gojobori 2002; Zavala *et al.* 2002), the complexity of their biosynthetic pathways (Karlin and Bucher 1992; Dufton 1997; Craig and Weber 1998), requirements for limiting resources (Mazel and Marlière 1989; Craig *et al.* 2000; Baudouin-Cornu *et al.* 2001), or the speed and accuracy with which their isoacceptors are translated (Eigen and Schuster 1979; Trifonov 1987; Shpaer 1989; Lobry and Gautier 1994; Akashi 1996; Gutiérrez *et al.* 1996; Percudani *et al.* 1997). Although individual amino acid mutations are likely to have small effects on overall cellular physiology, global evolutionary forces could underlie proteome-wide patterns of amino acid composition as well as variation in rates of protein evolution.

Among microbes, as well as multicellular eukaryotes, synonymous codon usage is coadapted with tRNA pools to enhance the efficiency of protein synthesis (reviewed in Andersson and Kurland 1990; Sharp *et al.* 1993; Akashi 2001). Among codons recognized by different aminoacyl tRNAs (aa-tRNAs), translationally preferred codons tend to be recognized by the more abundant isoacceptor. Among codons recognized by the same isoacceptor (through "wobble" pairing), preferred codons generally have intermediate codon-anticodon stability (Grosjean and Fiers 1982; Ikemura 1985; Yamao *et al.* 1991; Kanaya *et al.* 1999; Percudani and Ottonello 1999). In *Escherichia coli*, translation of major codons occurs 3- to 6-fold more quickly (Robinson *et al.* 1984; Varenne *et al.* 1984; Sorensen *et al.* 1989) and 10-fold more accurately (Precup and Parker 1987) than translation of minor codons. Thus, major codons allow efficient use of ribosomes and reduce the cost of GTP-dependent "proofreading" or rejection of noncognate isoacceptors. In addition, accurate translation reduces the costs of producing dysfunctional peptides resulting from misincorporations and processivity errors (frameshifting and premature termination). Stronger codon usage bias in highly expressed genes reflects increases in the fitness benefits of major codons with the number of translation events at a given codon. Major codon preference among synonymous codons was ap-

[1] *Address for correspondence:* Department of Biology, 208 Mueller Laboratory, Pennsylvania State University, University Park, PA 16802. E-mail: akashi@psu.edu

parent from early examinations of small numbers of yeast genes (Bennetzen and Hall 1982; Ikemura 1982; Sharp *et al.* 1986) and is consistent with large population sizes and a close relationship between growth rate and fitness in these microbes.

Among tRNAs carrying different amino acids, variation in either cellular concentrations or codon-anticodon stability could lead to translation selection both within and *among* synonymous families (Shpaer 1989; Lobry and Gautier 1994; Akashi 1996; Percudani *et al.* 1997; Morton and So 2000; Akashi 2001). Amino acid composition is related to gene expression in prokaryotes (Shpaer 1989; Yamao *et al.* 1991; Lobry and Gautier 1994; Gutiérrez *et al.* 1996; Akashi and Gojobori 2002; Zavala *et al.* 2002), yeast (Ikemura 1982; Percudani *et al.* 1997; Jansen and Gerstein 2000), *Giardia lamblia* (Garat and Musto 2000), *Caenorhabditis elegans* (Duret 2000), and plant chloroplasts (Morton and So 2000). Furthermore, amino acids represented by abundant tRNAs tend to be preferentially encoded in highly expressed genes (Shpaer 1989; Yamao *et al.* 1991; Lobry and Gautier 1994; Percudani *et al.* 1997; Duret 2000). However, highly expressed proteins fall into particular functional categories (*i.e.*, energy metabolism and protein synthesis) and tRNA pools could simply be adjusted to match the amino acid requirements for proper functioning of these proteins (Garel 1974; Ikemura 1982; Yamao *et al.* 1991; Xia 1998; Duret 2000).

In multicellular eukaryotes, tissue-specific tRNA abundances have been found in tissues committed to high expression of a small number of genes. tRNA concentrations match amino acid usage of fibroin in the posterior silk gland of the silkworm *Bombyx mori* L. and crystallines in the calf lens (Garel 1974) and hemoglobin in rabbit and human reticulocytes (Hatfield *et al.* 1982). Garel (1974) proposed "functional adaptation of tRNA"; selection for efficient translation regulates cellular tRNA isoacceptor concentrations to match the amino acid requirements of highly expressed proteins but does not affect their composition or evolutionary rates.

This study attempts to distinguish between unidirectional adjustments of tRNA pools to the amino acid requirements of highly expressed genes and coadaptation of both isoacceptor concentrations *and* amino acid usage in the budding yeast, *Saccharomyces cerevisiae.* Strong associations between synonymous codon usage and oligonucleotide DNA array estimates of mRNA levels suggest that estimates of transcript abundance provide informative predictors of the translation rates of genes. Usage of several amino acids shows associations with gene expression, and changes in amino acid composition result in stronger correlations between amino acid usage and tRNA abundances in highly expressed genes than in less expressed loci. Similar relationships within protein functional categories suggest that the primary structures of proteins reflect, at least in part, natural selection to enhance the rate and accuracy of their synthesis. Selection for efficient biosynthesis may also constrain protein size; among proteins in the same broad functional category, proteins encoded by highly expressed genes are consistently smaller than those encoded by less expressed loci.

MATERIALS AND METHODS

Yeast gene sequences: *S. cerevisiae* protein-coding sequences and descriptions (Goffeau *et al.* 1996) were obtained from ftp://genome-ftp.stanford.edu/pub/yeast/. Mitochondrial DNA-encoded genes, short coding regions (<100 codons), and genes identified as originating from phage or transposable elements were excluded from the analysis. In addition, genes with recent common ancestors (paralogs) were identified by performing unfiltered BLAST (Altschul *et al.* 1990) searches among all pairs of proteins encoded in the genome. Pairs of protein sequences showing alignments with ≥60% identity over ≥60 amino acids were formed into clusters and one gene from each cluster was included in the analysis. To maintain the sample size of highly expressed loci, the gene with the highest estimate of transcript abundance (see below) was chosen from each cluster.

Yeast expression data and functional categorization of proteins: Transcript abundance data from Holstege *et al.* (1998) were obtained from http://web.wi.mit.edu/young/expression/transcriptome.html and protein abundance measures (from 2D gel data) were taken from Futcher *et al.* (1999). Functional categorizations of gene products were obtained from the Yeast Protein Database (YPD; Costanzo *et al.* 2000; https://www.incyte.com/proteome/YPDcategories/Functional_Categories.html). Composite categories were constructed for 1571 genes listed in more than one category. Using this criterion, I found that yeast proteins fall into 259 different functional categories (including "unknown") and 128 of the categories contain a single gene. Only genes listed in both the YPD and the Holstege transcript abundance database were included in the analyses. Of the 6310 predicted yeast protein-coding genes, 5483 were included in the final data set (see supplemental material at http://www.genetics.org/supplemental/).

Sequence and expression data for *C. elegans*, *Drosophila melanogaster*, *Bacillus subtilis*, and *E. coli*: Coding sequences and estimates of transcript abundances [from matches to expressed sequences tag (EST) libraries] for *C. elegans* and *D. melanogaster* (Marais *et al.* 2001) were obtained from http://pbil.univ-lyon1.fr/datasets/Marais2001/data.html. Short coding regions (<100 codons) and genes identified as originating from phage or transposable elements were excluded from the analysis. Single members of each family of paralogs were included as described above. Genes that were not listed in the Marais *et al.* (2001) expression data files were not included in the analysis. A total of 11,546 and 11,864 predicted genes were analyzed from *D. melanogaster* and *C. elegans*, respectively. *E. coli* and *B. subtilis* data are described in Akashi and Gojobori (2002).

Identification of major codons: Major codon usage (MCU) was calculated as (number of major codons)/(number of major codons + number of minor codons). Identities of major codons for *S. cerevisiae* were taken from Kanaya *et al.* (1999) except for glutamic acid [the major codon was identified as GAA in Ikemura (1982)]. Major codons were taken from Akashi (1995) for *D. melanogaster*, Duret (2000) for *C. elegans*, and Kanaya *et al.* (1999) for *E. coli* and *B. subtilis* [with modifications described in Akashi and Gojobori (2002)]. Serine codons were divided into fourfold and twofold families so that each synonymous family is composed of codons that encode the same amino acid and that are connected by single mutational steps. Major codons for the two serine families were identified as those showing significant positive correlations with either gene expression [*S. cerevisiae* (Table 1) and *C. elegans*] or major codon usage for nonserine families (*E. coli* and *B. subtilis*).

Analyses of whole-genome data: Spearman rank correlations were employed in the whole-genome analyses. Because abundances for some codons and amino acids are quite low, analyses were conducted on binned data. Genes were ranked by the Holstege *et al.* (1998) estimates of transcript abundance and data were pooled for genes with low to high transcript abundance until 5000 codons were reached for each bin (all genes with identical expression estimates were included in the same bin). The numbers of genes in low expression bins were elevated by large numbers of identical estimates of transcript abundance. For statistical analyses, codon and amino acid usages were compared among 65 expression classes containing an average of 84 genes each. Bins of 50,000 codons were employed for visualization of trends (Figures 2, 3, 4, and 6).

Analyses within functional categories: Within functional categories, amino acid usage was compared between genes falling above and below a cutoff of one transcript per cell. Amino acid abundances were compared in 2 × 2 contingency tables; the columns of the tables were the high and low expression classes and the rows consisted of the counts of a particular amino acid and the pooled counts for all other amino acids. The Mantel-Haenszel procedure (Snedecor and Cochran 1989) was employed to calculate an overall probability for departures from equal amino acid usage among low and high expression genes across contingency tables from different functional categories. Thirty-one functional categories containing ≥10 genes in both expression classes were included in the analysis. "Unknown" was not included as a functional category.

The proportion of amino acids falling within "low complexity" regions was analyzed in a similar manner. The rows of 2 × 2 contingency tables consisted of the numbers of codons that fall within and outside of low complexity regions identified by the SEG and SEGN programs (Wootton and Federhen 1996) using default parameter settings. The columns were the low- and high-expression classes and Mantel-Haenszel tests were conducted as described above.

Mann-Whitney *U*-tests (Snedecor and Cochran 1989) were employed to test for differences in the mean gene lengths of highly and lowly expressed genes within each functional category. *Z* values of Mann-Whitney *U*-tests were assigned positive and negative signs for higher and lower mean ranks of size among highly expressed proteins. To test for a consistent trend among categories, a Wilcoxon ranked signs test statistic was calculated for the signed *Z* values. A null distribution of the test statistic was generated by 10^6 iterations of random assignment of sign (with $P = 0.5$) to each *Z* value and recalculation of the test statistic.

Yeast tRNA abundances: Ikemura (1982) quantified cellular abundances for 22 yeast tRNAs using 2D gel electrophoresis. The correlation between these measurements and the copy numbers of the corresponding tRNA genes in the *S. cerevisiae* genome (Percudani *et al.* 1997) is remarkably high ($r^2 = 0.803$). To include all 41 isoaccepting tRNAs in the analyses, the gene copy numbers of Percudani *et al.* (1997) were employed as estimates of tRNA abundances.

RESULTS

Transcript abundance and translation rates in yeast: Yeast cells growing at log phase under standard laboratory conditions contain ~15,000 poly(A)-mRNA molecules (Hereford and Rosbash 1977). Figure 1 plots the distribution of transcript abundances among yeast genes from the high-density oligonucleotide array data of Holstege *et al.* (1998) from yeast cells grown to mid-log phase in YPD media. The distribution of transcript abundance is strongly skewed toward low values; >80% of genes are represented by ≤2 mRNA molecules and only 3.5% of genes have transcript abundances of ≥10 mRNA molecules per cell. Futcher *et al.* (1999) found good correspondence between 2D gel quantifications of protein concentrations (ranging from 200 to 2 × 10^6 molecules/cell) and estimates of transcript abundances. They estimate a rate of protein synthesis of roughly 4000 proteins/transcript for genes represented by ≥1 mRNA molecule/cell. For proteins represented by <1 mRNA/cell, they suggest post-transcriptional regulation; *i.e.*, mRNA abundances are not informative predictors of translation rates.

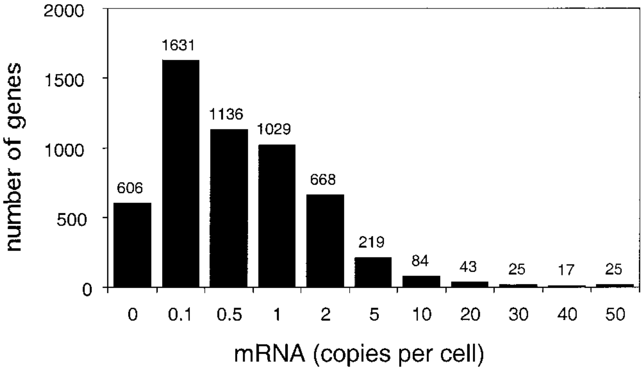


Figure 1.—Distribution of transcript abundances of yeast genes. Data are from Holstege *et al.* (1998). The first column plots the number of genes with no detectable transcripts. Other columns plot the numbers of genes with transcript abundances between a lower limit shown below the columns and an upper limit shown beneath the column to the right (the last column has no upper limit).

Synonymous codon usage was examined among expression classes to determine the strength of correspondence between GeneChip estimates of transcript abundances and the translation rates of genes. Under major codon preference, the fitness benefit of a major codon is strongly dependent on the number of translation events experienced at a given codon. Coghlan and Wolfe (2000) found that <40% of the variation in ranks of measures of codon bias among yeast genes was explained by transcript abundance (Spearman rank correlation, $r_s = 0.62$). However, correlations between major codon usage and transcript abundance are remarkably high in comparisons among bins of genes with similar expression estimates (Akashi 2001; Table 1; Figure 2). Similar patterns among synonymous families that differ in the favored nucleotide in the third codon position (G-favored: Lys, Leu; A-favored: Pro, Gln; T-favored: Gly; C-favored: Phe, Tyr, His, Asn, Asp) suggest both that estimates of transcript abundance are informative predictors of average translation rates

TABLE 1

Transcript abundance and synonymous codon usage in *S. cerevisiae*

aa	Cod	r_s	Pref	aa	Cod	r_s	Pref	aa	Cod	r_s	Pref	aa	Cod	r_s	Pref
F	TTT	−0.937	u	S_4	TCT	0.942	p	Y	TAT	−0.904	u	C	TGT	0.887	p
	TTC		p		TCC	0.872	p		TAC		p		TGC		u
L	TTA	−0.430	u		TCA	−0.944	u	*	TAA			*	TGA		
	TTG	0.958	p		TCG	−0.884	u	*	TAG			W	TGG		
	CTT	−0.882	u	P	CCT	−0.503	u	H	CAT	−0.754	u	R	CGT	0.580	p
	CTC	−0.913	u		CCC	−0.931	u		CAC		p		CGC	−0.817	u
	CTA	−0.635	u		CCA	0.966	p	Q	CAA	0.914	p		CGA	−0.910	u
	CTG	−0.856	u		CCG	−0.952	u		CAG		u		CGG	−0.854	u
I	ATT	0.536	p	T	ACT	0.940	p	N	AAT	−0.956	u	S_2	AGT	0.073	—
	ATC	0.898	p		ACC	0.868	p		AAC		p		AGC		
	ATA	−0.978	u		ACA	−0.944	u	K	AAA	−0.945	u	R	AGA	0.879	p
M	ATG				ACG	−0.951	u		AAG		p		AGG	−0.934	u
V	GTT	0.885	p	A	GCT	0.959	p	D	GAT	−0.748	u	G	GGT	0.975	p
	GTC	0.932	p		GCC	0.653	p		GAC		p		GGC	−0.886	u
	GTA	−0.965	u		GCA	−0.970	u	E	GAA	0.893	p		GGA	−0.969	u
	GTG	−0.853	u		GCG	−0.938	u		GAG		u		GGG	−0.924	u

Spearman rank correlation coefficients, r_s, are shown for usage of each codon within its synonymous family *vs.* average transcript abundance among genes grouped by expression estimates (bin size, 5000 codons). All correlations were statistically significant after Bonferroni sequential correction for multiple tests (Rice 1989) except for codons in the S_2, serine twofold, family. p, preferred codons, those that increase in frequency in highly transcribed genes; u, unpreferred codons, those that decrease in frequency.

(among binned genes) and that translation selection is sufficient to overcome mutational biases associated with transcription (Datta and Jinks-Robertson 1995; Morey *et al.* 2000) and substitutional biases associated with gene conversion (Gerton *et al.* 2000; Birdsell 2002). Codon preferences determined using microarray estimates of transcript abundance are consistent with those established through correspondence analysis of codon usage (Sharp and Cowe 1991; Kanaya *et al.* 1999).

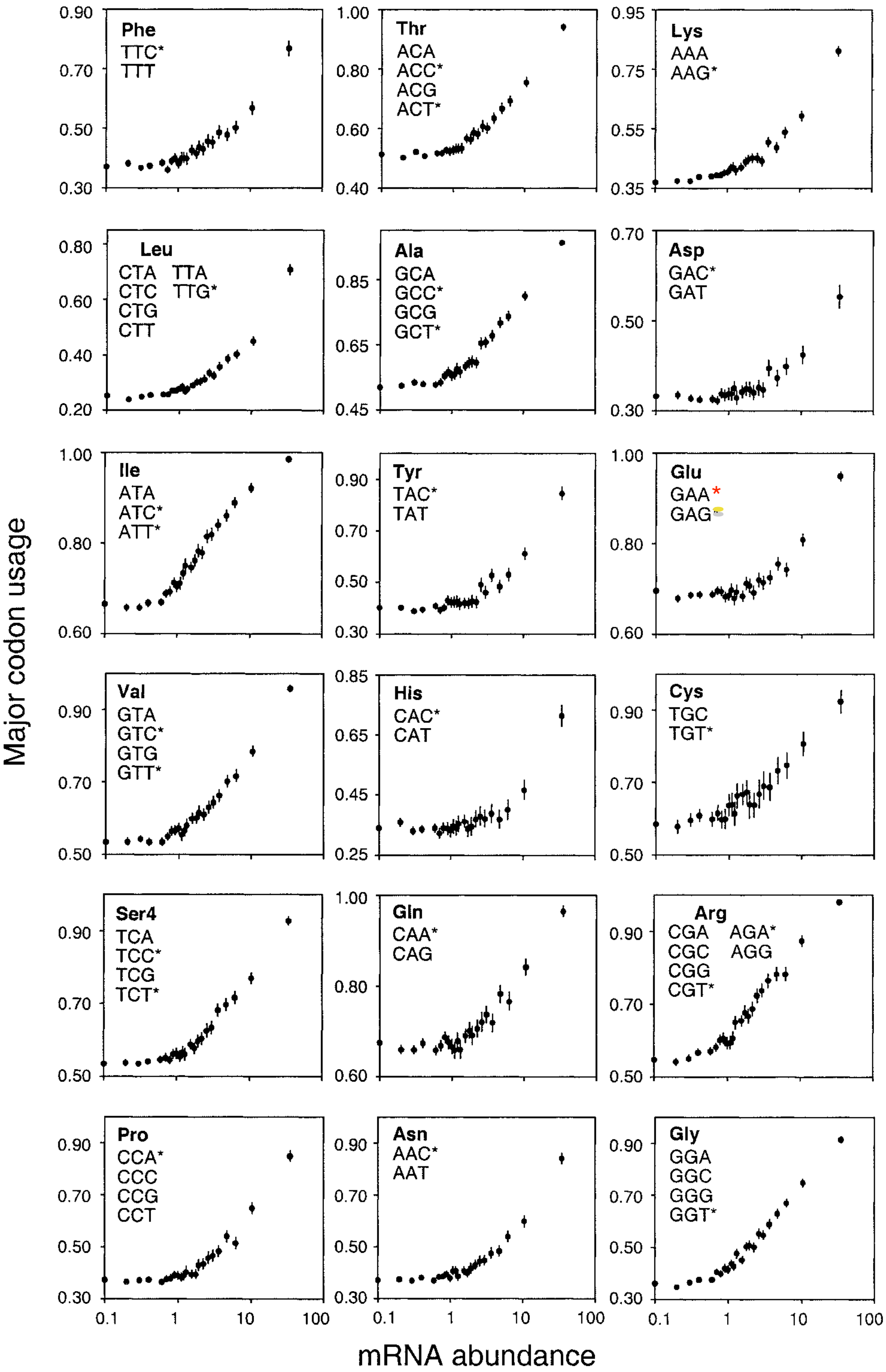
Seven of the twofold synonymous families (all NNY types) are recognized by a single isoacceptor through wobble pairing at the third codon position. Six of these families (Asn, Asp, Cys, His, Phe, and Tyr) show steady increases in usage of a single codon in highly transcribed genes (Table 1; Figure 2). Such patterns are consistent with translational selection for codon-anticodon stability (Grosjean *et al.* 1978; Percudani and Ottonello 1999).

Codon usage for amino acids encoded by sixfold redundant codons provide the clearest evidence for translational preferences related to tRNA abundances. Third codon position wobble rules for eukaryotes are ambiguous (Percudani 2001), but wobble pairing is not known to occur at the first codon position. Thus, fourfold and twofold redundant families for Leu, Arg, and Ser are recognized by nonoverlapping sets of tRNAs. Usage of twofold codons for Leu ($r_s = 0.933$, $P < 10^{-5}$) and Arg ($r_s = 0.558$, $P < 10^{-5}$) increases dramatically whereas twofold codons for Ser decrease ($r_s = -0.897$, $P < 10^{-5}$) within their synonymous families in highly expressed genes (Figure 3). These patterns correspond to differences in the numbers of tRNA genes (presumably resulting in higher tRNA abundances) that recognize the twofold and fourfold families for these amino acids. Translational selection appears to discriminate among synonymous codons recognized by nonoverlapping groups of tRNAs. Similar preferences may also bias the usage of codons that encode different amino acids.

Gene expression and amino acid composition in yeast: Abundances for a number of amino acids are strongly correlated with gene expression levels (Table 2). The magnitude of changes in abundance can be quite large; alanine usage increases by greater than twofold in highly expressed genes and serine twofold codons are only one-third as abundant in highly expressed genes (Figure 4).

In *S. cerevisiae*, estimates of transcript abundance show strong positive associations with the frequency of meiotic double-strand breaks, a measure of local recombination rate (Gerton *et al.* 2000; Birdsell 2002). The latter is also positively correlated with both intergenic and third position GC content (Birdsell 2002), suggesting that biased gene conversion elevates G + C content. However, associations between amino acid usage and estimates of recombination rate are considerably smaller than associations with transcript abundance (data available from H. Akashi).

The interpretation of relationships between gene expression and amino acid composition is less straightforward than that for similar associations among synonymous codons. Changes in amino acid usage could reflect



error in fig
GAA pref

FIGURE 2.—Transcript abundance and major codon usage. Data are graphed for expression categories (bin sizes ≥50,000 codons). Abundances are within synonymous families.

differences in the functional roles of proteins expressed at different levels (GAREL 1974; IKEMURA 1982; YAMAO *et al.* 1991; XIA 1998; DURET 2000). For example, JANSEN and GERSTEIN (2000) showed that transcript abundances are higher for cytosolic proteins than for membrane proteins in yeast. Greater usage of hydrophobic residues in less expressed proteins may reflect a greater abundance of transmembrane regions. To control for differences in the functional requirements of proteins, amino acid composition was compared among genes classified into common functional categories in the Yeast Proteome Database (COSTANZO *et al.* 2000). Table 3 shows the 31 different categories containing at least 10 genes in both the low (one or less transcript per cell) and the high (more than one transcript per cell) expression classes. There are clear differences in expression patterns among these categories; transcription factors and membrane proteins/transporters tend to be

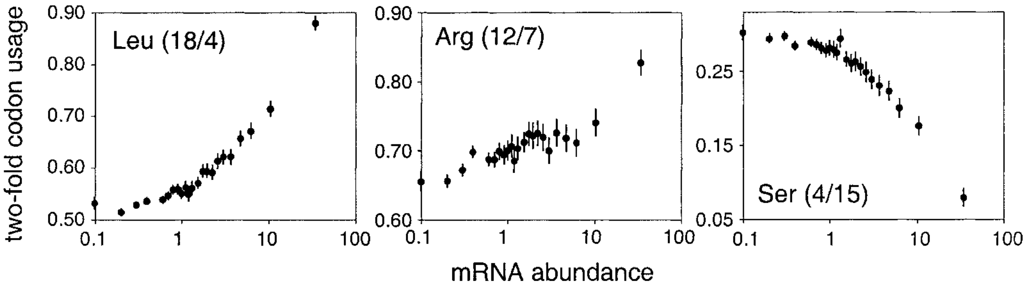


FIGURE 3.—Transcript abundance and codon usage in sixfold redundant families. The numbers of tRNAs that recognize twofold and fourfold degenerate codons are shown (twofold/fourfold). Data are graphed for expression categories (bin sizes ≥5 × 10^4 codons). Abundances are within synonymous families.

represented by few transcripts. Two classes of more abundant transcripts encode oxidoreductases, including enzymes of central metabolism and amino acid biosynthesis, and proteins involved in translation, such as ribosomal proteins and elongation factors. Transcript abundances for these classes are consistent with estimates of 200,000 ribosomes/cell in rapidly growing yeast (WARNER 1999) and 2,000,000 molecules/cell for some glycolytic enzymes (FUTCHER *et al.* 1999). Thus, amino acids that are employed more often in highly expressed genes may simply be those required for proper functioning of ribosomal proteins and cytosolic enzymes.

For analyses within the 31 functional categories, amino acid usage was compared between low- and high-expression classes in 2 × 2 contingency tables. Table 2 shows, for each amino acid, the number of individually significant 2 × 2 tests as well as the probability of the overall trend across tables. With cutoff values of two, three, and four transcripts per cell to divide high- and

TABLE 2

Gene expression and amino acid composition in yeast

Amino acid	tRNA gene no.	Codons	Usage	All genes (r_s)	Fun cat (31)	
					G test	*Z*
Ala	16	GCN	0.054	0.930*	22/0	24.99*
Gly	21	GGN	0.049	0.849*	20/0	21.74*
Val	18	GTN	0.056	0.847*	9/0	11.06*
Thr	16	ACN	0.058	0.140	1/1	0.81
Lys	21	AAR	0.074	0.133	1/3	−0.38
Glu	16	GAR	0.066	0.096	5/3	2.83*
Tyr	8	TAY	0.033	−0.055	1/6	−4.55*
Met	5	ATG	0.019	−0.106	3/0	1.21
Arg	19	CGN, AGR	0.045	−0.123	2/4	−2.94*
Pro	12	CCN	0.043	−0.159	2/4	−1.83
Trp	6	TGG	0.010	−0.179	2/0	0.41
Asp	15	GAY	0.059	−0.245	5/1	3.83*
Phe	10	TTY	0.045	−0.325	2/2	−2.93*
Ser_4	15	TCN	0.066	−0.359*	2/4	−2.57
Cys	4	TGY	0.013	−0.478*	2/5	−5.52*
His	10	CAY	0.022	−0.508*	1/4	−2.8*
Gln	7	CAR	0.040	−0.508*	3/3	1.3
Ile	15	ATH	0.066	−0.540*	0/8	−7.97*
Leu	21	TTR, CTN	0.097	−0.847*	0/8	−11.76*
Ser_2	4	AGY	0.025	−0.892*	0/18	−15.48*
Asn	10	AAY	0.062	−0.911*	0/17	−15.42*

tRNA gene copy numbers are from PERCUDANI *et al.* (1997). The numbers are pooled among genes encoding isoacceptors for each amino acid. Frequency of usage of amino acids in the genome is shown. Amino acids are listed in order of decreasing Spearman rank correlations, r_s, in all gene analyses among 65 expression classes (bin size ≥5000 codons). Functional category (fun cat) analyses were conducted among 31 categories with at least 10 genes in the high- and low-expression groups, using a cutoff of one transcript per cell. Amino acid composition was compared in 2 × 2 contingency tables for each amino acid within each functional category. *G* test shows the numbers of tables with higher/lower abundance in the high expression class (*G* tests with significance level of $P < 0.05$ uncorrected for multiple tests). The *Z* statistic of the Mantel-Haenszel procedure for the data pooled across functional categories is also shown for each amino acid. *$P < 0.05$ after Bonferroni sequential correction for multiple tests (RICE 1989).

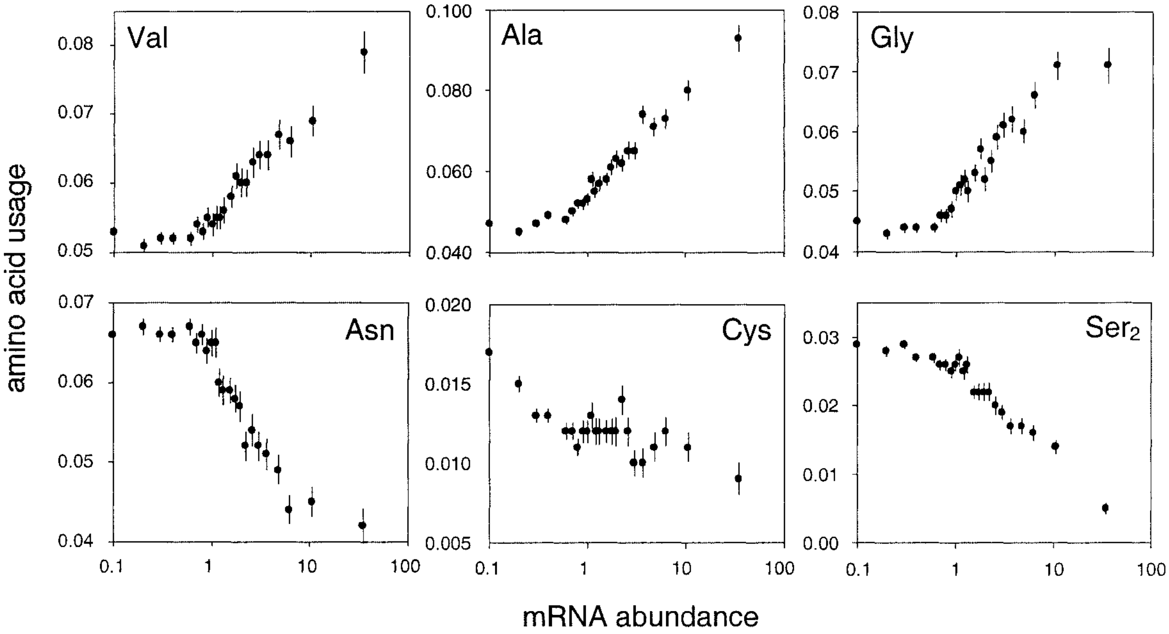


FIGURE 4.—Transcript abundance and amino acid usage. Data are graphed for expression categories (bin sizes ≥5 × 10^4 codons). Abundances are among all codons.

low-expression classes, the numbers of functional categories with at least 10 genes in each expression class reduce to 21, 15, and 11, respectively. However, the main trends of amino acid usage are robust to these cutoff values; Mantel-Haenszel test statistics remain significantly positive for Val, Ala, Gly, and Glu and negative for Phe, Leu, Ile, His, Asn, Cys, and Ser_2 for cutoff values between one and four transcripts per cell.

Several amino acids show strong statistical associations with expression levels in both whole-genome and within-category analyses. Ala, Val, and Gly show strong increases in abundance in highly expressed genes, whereas Leu, Ser_2, and Asn show strong declines (Figure 4). For such amino acids, changes in the relative abundances of different types of proteins in different expression classes are unlikely to explain relationships between amino acid usage and expression levels. These patterns are consistent with JANSEN and GERSTEIN's (2000) findings through comparison of amino acid composition of the yeast genome and transcriptome (amino acid usage for a given gene was weighted by estimates of its transcript abundance). However, results for some amino acids (Gln, Ser_4, Arg, Glu, and Tyr) differ between the all-gene and within-category analyses. Such patterns could reflect differences in the functional requirements of genes in different expression classes or differences in the statistical power of the two approaches. In either case, these amino acids show small differences in abundance between lowly and highly expressed genes.

Amino acid usage and tRNA abundances: Relationships between amino acid usage and tRNA gene numbers for yeast genes with low, intermediate, and high transcript abundance are shown in Figure 5. Codons that experience few translation events will be under little or no selection for translationally preferred codons (among either synonymous or nonsynonymous alternatives). Thus, relationships between tRNA abundance and amino acid usage should show substantial scatter. However, under translational selection, the magnitude of fitness differences among codons recognized by rare and common isoacceptors should increase as a function of gene expression levels. Such fitness differences may exist among nonsynonymous as well as synonymous alternatives. In highly expressed genes, translational selection at positions of peptides otherwise determined largely by mutation drift will result in greater correspondences between amino acid usage and tRNA abundances.

Figure 6 shows that the Pearson product-moment correlation coefficient between amino acid usage and tRNA gene numbers increases steadily as a function of gene expression levels (5000 codons/bin, $r_s = 0.68$, $Z = 6.39$, $P < 10^{-5}$). Plots are shown for bins of 50,000 codons. Stronger correlations between amino acid usage and tRNA gene copy numbers in highly expressed proteins within functional categories (Table 3) support the contribution of translational selection in determining the amino acid composition of proteins (Wilcoxon ranked signs test, $P < 10^{-5}$).

Gene expression and protein size: Given some tolerance of protein function to insertion/deletion variation, translational selection will favor reductions in protein size (AKASHI 1996). Eliminating codons from a given gene will reduce the amount of time that ribosomes spend translating its transcripts and enhance the overall rate of proteins synthesized per ribosome per time. En-

TABLE 3

Analyses within yeast protein functional categories: expression levels and tRNA vs. amino acid usage and protein sizes

	Gene no.		[tRNA] *vs.* [amino acid]		Protein size (avg.)		
Functional category	[mRNA] < 1	[mRNA] ≥ 1	Low *r*	High *r*	Low	High	MWU *Z*
Active transporter, secondary + major facilitator superfamily + transporter	34	17	0.711	0.718	559.2	587.4	−1.40
Active transporter, secondary + transporter	53	47	0.768	0.758	600.4	535.8	1.05
ATPase + helicase + hydrolase + RNA-binding protein	12	19	0.867	0.910	1100.4	624.1	2.84
Chaperones	15	21	0.759	0.878	476.1	366.8	1.35
Complex assembly protein	22	19	0.743	0.866	577.7	400.1	1.22
DNA-binding protein	45	17	0.789	0.875	651.0	267.0	4.96
DNA-binding protein + transcription factor	50	15	0.626	0.631	505.3	514.8	0.42
GTP-binding protein/GTPase + hydrolase	13	27	0.855	0.888	452.3	289.4	1.92
Hydrolase	61	40	0.862	0.863	509.9	399.9	1.33
Hydrolase + other phosphatase	13	13	0.814	0.878	501.5	408.0	0.08
Hydrolase + protease (other than proteasomal)	34	29	0.777	0.835	693.4	554.2	1.83
Hydrolase + protein phosphatase	15	17	0.703	0.804	551.3	434.1	1.51
Inhibitor or repressor	12	15	0.703	0.825	467.8	486.7	0.39
Ligase	16	26	0.820	0.903	712.6	768.5	0.60
Ligase + RNA-binding protein + tRNA synthetase	15	21	0.841	0.885	551.6	694.5	−2.05
Lyase	26	40	0.886	0.915	507.7	458.4	−0.12
Nuclear import/export protein	22	27	0.709	0.737	1023.9	764.0	1.75
Other kinase + transferase	25	25	0.814	0.897	677.2	521.0	2.22
Oxidoreductase	53	108	0.885	0.899	509.8	400.8	2.42
Protein conjugation factor	16	14	0.754	0.825	511.5	394.6	1.21
Protein kinase + transferase	87	20	0.780	0.827	754.4	532.6	2.89
Receptor (protein translocation)	14	18	0.779	0.850	610.9	336.9	2.36
Regulatory subunit	16	12	0.654	0.833	592.6	366.6	2.79
RNA-binding protein	37	42	0.697	0.865	579.0	471.9	1.66
RNA-binding protein + ribosomal subunit	17	110	0.876	0.940	301.3	194.4	3.92
RNA-binding protein + spliceosomal subunit	35	13	0.823	0.826	417.0	381.3	2.80
Structural protein	38	29	0.729	0.834	641.7	407.8	2.41
Transcription factor	137	51	0.711	0.784	670.3	563.2	2.18
Transferase	91	136	0.827	0.856	496.5	484.5	1.63
Translation factor	11	17	0.834	0.871	460.6	483.4	−0.05
Transporter	13	15	0.721	0.730	596.4	404.1	0.62

Functional categorizations are from the Yeast Protein Database (Costanzo *et al.* 2000). The numbers of genes in low (less than transcript per cell) and high (one or more transcripts per cell) expression groups are shown. Pearson product-moment correlation coefficients, *r*, for relationships between tRNA gene copy numbers (Percudani *et al.* 1997) and amino acid abundances are shown for the two expression classes within each functional category. Average sizes of proteins were compared among the high- and low-expression classes. *Z* statistics from the Mann-Whitney *U*-test (MWU *Z*) are shown (positive values indicate lower mean ranks of size among highly expressed proteins).

ergy expenditure for both amino acid and protein synthesis will also be reduced. The magnitude of time and energy savings, and, consequently, the fitness advantage of protein size reduction will be a function of the number of times a given gene is translated.

Negative correlations between gene length and both synonymous codon bias (Moriyama and Powell 1998) and transcript abundance (Coghlan and Wolfe 2000; Jansen and Gerstein 2000; Pal *et al.* 2001b) have been found among yeast genes. However, previous studies did not control for differences in the functional requirements of proteins among expression classes, such as the lack of highly expressed transmembrane proteins. Among the 31 functional categories of yeast proteins, 27 show a higher mean rank of protein size among the less highly expressed proteins and 4 deviate in the opposite direction (Table 3). Mann-Whitney *U*-tests show significant departures at the 5% level (prior to correction for multiple tests) for 12 of the 31 classes; highly expressed proteins are smaller in 11 of these 12 classes. Overall, highly expressed genes tend to encode smaller proteins than do less expressed proteins in the same functional category (Wilcoxon ranked signs test, $P < 10^{-5}$). Protein length is also negatively correlated with 2D gel quantifications of abundance for 64 proteins (Futcher *et al.* 1999) from cells grown on both glucose ($r_s = -0.47$, $P < 10^{-4}$) and ethanol ($r_s = -0.33$, $P < 0.005$).

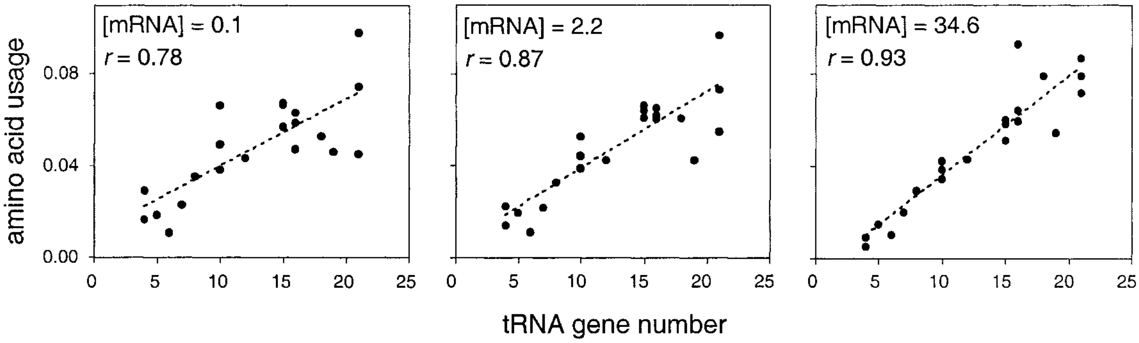


Figure 5.—Correlations between amino acid usage and tRNA gene copy numbers. tRNA gene copy numbers are from Percudani *et al.* (1997). The numbers are pooled among genes encoding isoacceptors for each amino acid.

To determine whether low complexity nucleotide sequences (including homonucleotide runs and short repeats) contribute to differences in protein sizes among expression classes, simple sequences were identified using the SEGN software (Wootton and Federhen 1996). Of the 31 functional categories, 21 show a higher percentage of simple sequences in less expressed proteins and 10 deviate in the opposite direction. A Mantel-Haenszel test shows significantly lower proportions of simple nucleotide sequences in highly expressed proteins ($Z = 12.29$, $P < 10^{-5}$). However, this reduction in simple sequences does not account entirely for the smaller sizes of highly expressed proteins; differences in the sizes remain highly significant after removal of low-complexity regions (Wilcoxon ranked signs test, $P < 10^{-4}$). Interestingly, the proportion of "low-complexity" amino acid sequences increases in highly expressed proteins within functional categories; 20 of 31 categories deviate in this direction and the overall pattern is highly statistically significant (Mantel-Haenszel test, $Z = 20.06$, $P < 10^{-5}$). This pattern may reflect a greater abundance of particular structural motifs (soluble folds with combinations of helices and sheets) represented among highly expressed proteins (Jansen and Gerstein 2000).

Gene expression and GNN usage: In yeast, GCN, GGN, and GTN codons for Ala, Gly, and Val, respectively, show the strongest increases in usage in highly expressed genes (Table 2; Figure 4). Figure 7 shows that such patterns are common to many prokaryotes and multicellular eukaryotes. GNN usage shows remarkably consistent increases in abundance with measures of translation rates (either estimates of transcript abundance or measures of synonymous codon usage bias) in *C. elegans*, *D. melanogaster*, *B. subtilis*, and *E. coli* (see also Gutiérrez *et al.* 1996), as well as yeast. Similar patterns have been noted in the genomes of plant chloroplasts (Morton and So 2000) and Buchnera (Palacios and Wernegreen 2002). GNN increases occur among cytosolic and membrane proteins encoded in plant chloroplast genomes (Morton and So 2000) and within functional categories of yeast (Table 2), *E. coli*, *B. subtilis* (Akashi and Gojobori 2002), and *C. elegans* (our unpublished data) proteins.

DISCUSSION

Establishing translational selection in protein evolution: Major codon preference posits adaptation of both tRNA concentrations and synonymous codon usage. Regulation of aa-tRNA abundances may result from relatively few, strongly selected mutations. However, codon usage bias results from weak selection at thousands of "silent" sites throughout the genome (reviewed in Andersson and Kurland 1990; Sharp *et al.* 1993; Akashi 2001). Although relationships between amino acid composition and gene expression have been established in yeast (Ikemura 1982; Percudani *et al.* 1997; Jansen and Gerstein 2000), translation selection in protein evolution has been difficult to substantiate because associations between amino acid composition and tRNA abundances can be explained by selection on tRNA concentrations to match the functional needs of

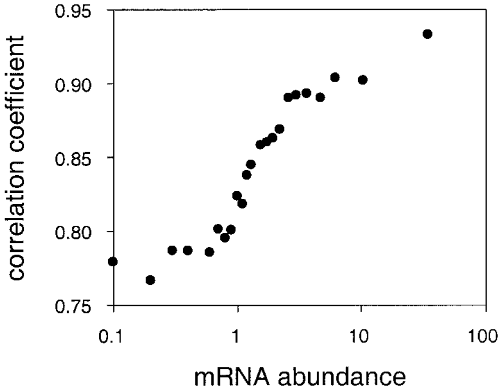


Figure 6.—Correlation between amino acid usage and tRNA gene copy numbers among expression classes. Data are graphed for expression categories (bin sizes ≥5×10^4 codons). Pearson product moment correlation coefficients between tRNA gene copy numbers and amino acid usage are plotted on the *y*-axis.

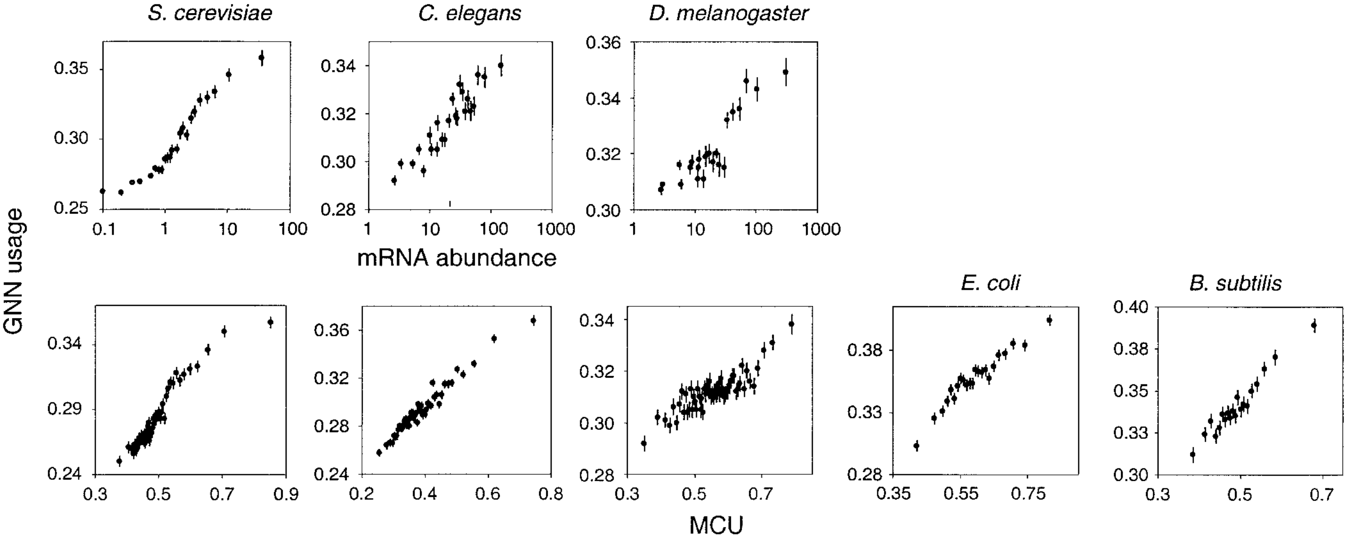


Figure 7.—Gene expression and GNN codon usage among highly diverged genomes. Data are described in the text. The pooled abundances of GNN codons (Val, Ala, Asp, Glu, and Gly) among all codons are plotted on the *y*-axis. mRNA abundances for *C. elegans* and *D. melanogaster* are counts of matches to EST libraries (Duret and Mouchiroud 1999) and are employed as estimators of translation rates. Bin sizes for ranking by mRNA abundances are 5×10^4 for each genome. MCU, major codon usage, is the number of major codons/(number of major + minor codons) and is also employed as an estimator of translation rates. Bins were constructed similarly to transcript abundance estimates (see text) for genes ranked by MCU. Bin sizes are 1×10^5 for *S. cerevisiae, C. elegans,* and *D. melanogaster* and 5×10^4 for *B. subtilis* and *E. coli.*

highly expressed proteins (Garel 1974; Ikemura 1982; Yamao *et al.* 1991; Xia 1998; Duret 2000).

Here, functional categorizations of proteins were employed to distinguish between associations between gene expression and amino acid composition that arise as a by-product of the functional requirements of proteins and those that reflect fitness benefits to translationally superior codons. Increases in the correlation between tRNA gene numbers and amino acid usage as a function of expression levels among all genes and similar patterns within broad functional categories support translational selection. Although the functional categorizations of proteins may be crude, it is unlikely that functional requirements explain consistent trends in amino acid usage in nonoverlapping groups of genes.

These findings do not exclude functional adaptation of isoacceptor abundances. tRNA pools may have been initially adjusted to match the functional requirements of highly expressed ribosomal proteins and enzymes of central metabolism. Translational selection would magnify amino acid usage biases beyond the initial functional needs of abundant proteins. The gradual increase in the correlation between amino acid usage and gene expression (Figure 6) supports a contribution of translational selection in the amino acid composition of even moderately expressed yeast genes. However, Table 2 shows relatively high overall usage of some amino acids that appear to be translationally less preferable (*i.e.*, Leu, Asn, and Ile), suggesting a balance among forces including translational selection, functional constraint, and mutation pressure.

Associations between mutational processes and transcription rates could contribute to relationships between gene expression and codon and amino acid usage, as well as protein length. In *E. coli*, transcription induces C → T transitions on the nontranscribed strand, presumably due to increased deamination of cytosine (Francino *et al.* 1996; Beletskii and Bhagwat 1998). Genetic experiments in yeast have shown >10-fold increases in rates of −1 frameshift reversion mutations with transcription rates (Datta and Jinks-Robertson 1995). However, such experiments screen for particular types of mutations and the dependence of the overall spectrum of mutations on transcription has not been determined. In *D. melanogaster* (Kliman and Hey 1994) and *C. elegans* (Duret and Mouchiroud 1999), intron base composition has been examined to determine whether mutational processes are transcription dependent. However, the small numbers of introns in the yeast genome (Davis *et al.* 2000) and their greater abundance (Ares *et al.* 1999; Lopez and Seraphin 1999) and lengths (Vinogradov 2001) in highly expressed transcripts suggest that intron base composition does not reflect mutational equilibrium.

Population genetic analyses of putative fitness classes of nonsynonymous mutations (Akashi 1995) may provide a means to distinguish between the contributions of translational selection and mutational biases in amino acid composition. In contrast to compositional studies that assume constant mutational processes among genes, such analyses assume constancy of mutational processes within genes over evolutionary time, and the statistical power to detect weak selection can be quite high (Akashi 1999). The compositional analysis undertaken here may provide putative fitness classes of amino acid changes for such studies. Rigorous support for translational selection in yeast protein evolution may require sequence data from within and between closely related species.

Properties of translationally preferred aa-tRNAs: The rate and accuracy of protein synthesis depend on both the abundances of aa-tRNAs and their intrinsic properties, such as the stability of codon-anticodon interactions. Recent studies have demonstrated conformational changes in aa-tRNAs, elongation factors, and ribosomes during protein synthesis (Ogle *et al.* 2001). Codon-anticodon interactions determine rate constants during the processes of tRNA selection, proofreading, and peptidyl transfer, which, in turn, determine both the speed and the accuracy of protein synthesis (reviewed in Rodnina and Wintermeyer 2001). Processing of aa-tRNAs also affects translation of neighboring codons; near-cognate isoacceptors bound at the ribosomal P site induce frameshift events at the A site (Farabaugh 2000). The analyses presented here have focused on aa-tRNA abundances, an extrinsic property of isoacceptors, because they can be estimated from gene numbers. However, codon preferences among nonsynonymous codons are likely to also reflect intrinsic properties of tRNAs or the amino acids that they carry. For example, the decline of Leu usage in highly expressed genes may reflect, in part, selection against usage of frameshift-prone codons (Farabaugh 2000).

Increasing GNN usage is the most prominent feature of associations between amino acid usage and gene expression in yeast as well as in a number of distantly related organisms. Three base nucleotide periodicities in protein-coding genes (Nassar and Cook 1976; Trifonov and Sussman 1980; Shepherd 1981) have been explained in light of theoretical studies of the early evolution of the genetic code (Crick *et al.* 1976; Eigen and Schuster 1979). Trifonov (1987, 1992) has argued that G:non-G:N codons may aid in maintenance of translational reading frame through interactions between mRNA and 16S rRNA during translation. Biochemical studies of protein synthesis will be required to determine whether GNN codons have special translational properties.

Intrinsic features of aa-tRNAs could also include properties of amino acids such as their requirements for limiting resources (Mazel and Marlière 1989; Craig *et al.* 2000; Baudouin-Cornu *et al.* 2001) or costs of biosynthesis or transport (Richmond 1970; Karlin and Bucher 1992; Lobry and Gautier 1994; Dufton 1997; Craig and Weber 1998; Garat and Musto 2000; Jansen and Gerstein 2000; Akashi and Gojobori 2002; Zavala *et al.* 2002). Natural selection may have elevated tRNA abundances for isoacceptors carrying metabolically favored amino acids so that translation selection acts in the same direction as such preferences (Akashi and Gojobori 2002).

Baudouin-Cornu *et al.* (2001) have shown that yeast proteins involved in sulfur and nitrogen transport and processing show reduced levels of amino acids requiring these atoms. Examination of the usage of S- and N-containing amino acids among proteins that are highly expressed during nitrogen or sulfur starvation would add support for nutrient limitation and protein evolution.

Calculations of the energetic costs of amino acid biosynthesis may differ between yeast and *E. coli* or *B. subtilis* due to both differences in amino acid biosynthetic pathways and alternative energy production pathways (alcohol fermentation and respiration). Such analyses may help to explain the identities of amino acids whose usage differs between highly and lowly expressed genes but are not undertaken here.

Translational selection and protein size: In the yeast genome, the smaller sizes of proteins encoded by highly expressed genes are consistent with selection favoring reductions in the metabolic costs of protein and/or amino acid biosynthesis. Relationships between gene length and expression levels in multicellular eukaryotes are less clear. Duret and Mouchiroud (1999) found no association between protein length and transcript abundance (measured by the numbers of matches of ESTs to predicted gene sequences) in *Arabidopsis thaliana* and positive correlations between protein size and expression in the *D. melanogaster* and *C. elegans.* In contrast, Castillo-Davis *et al.* (2002) employed DNA array estimates of mRNA abundance and found strong statistical support for reductions in length among highly expressed *C. elegans* genes. Biases in methods for estimating gene expression will need to be explored and patterns will need to be studied among proteins of related function to determine whether gene length and expression level are related in these organisms. A lack of negative relationships would be at odds with strong evidence of selection for metabolic efficiency at silent sites in these multicellular eukaryotes (reviewed in Sharp *et al.* 1993; Akashi 2001; Duret 2002).

Translational selection and protein divergence: Rates of protein divergence are negatively correlated with expression levels among yeast genes (Pal *et al.* 2001a). Translational selection in protein evolution could provide an explanation (Akashi 2001); in highly expressed genes, amino acid changes that may be neutral with respect to protein function will be selected against if they decrease the rate or accuracy of protein synthesis. Translationally unpreferred amino acids that are maintained in these genes may be restricted to those that serve critical roles in protein function and may also be evolutionarily conserved.

Rates of protein evolution are also negatively correlated with gene expression in plants (Wright *et al.* 2002), mammals (Duret and Mouchiroud 2000; Iida and Akashi 2000), and Drosophila (Betancourt and Presgraves 2002), suggesting that translation selection may be a factor in protein divergence among multicellular eukaryotes. However, Duret and Mouchiroud (2000) proposed that expression patterns are related to functional constraint. Proteins expressed in multiple tissues encounter a large number of chemical environments and their primary structures may be constrained to avoid physical interactions with other proteins (Hastings 1996). Alternatively, proteins expressed in a greater number of tissues or in a greater number of developmental stages may be more likely than tissue-specific proteins to affect fitness. Both ideas relate rates of protein evolution to constraints on function. If the interpretations proposed here are correct, then fitness effects of amino acid changes related to the overall physiology of cells, rather than the specific functions of proteins, should also contribute to patterns of protein divergence and amino acid compositional differences among taxa.

I am grateful to C. Castillo-Davis, F. Holstege, T. Ikemura, and W.-Y. Ko for valuable discussions and comments on the manuscript. This work was supported by a grant from the Alfred P. Sloan Foundation.

LITERATURE CITED

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

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution and larger proteins in *D. melanogaster.* Genetics **144:** 1297–1307.

Akashi, H., 1999 Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151:** 221–238.

Akashi, H., 2001 Gene expression and molecular evolution. Curr. Opin. Genet. Dev. **11:** 660–666.

Akashi, H., and T. Gojobori, 2002 Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis.* Proc. Natl. Acad. Sci. USA **99:** 3695–3700.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

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

Ares, M., Jr., L. Grate and M. H. Pauling, 1999 A handful of intron-containing genes produces the lion’s share of yeast mRNA. RNA **5:** 1138–1139.

Baudouin-Cornu, P., Y. Surdin-Kerjan, P. Marliere and D. Thomas, 2001 Molecular evolution of protein atomic composition. Science **293:** 297–300.

Beletskii, A., and A. S. Bhagwat, 1998 Correlation between transcription and C to T mutations in the non-transcribed DNA strand. Biol. Chem. **379:** 549–551.

Bennetzen, J. L., and B. D. Hall, 1982 Codon selection in yeast. J. Biol. Chem. **257:** 3026–3031.

Betancourt, A. J., and D. C. Presgraves, 2002 Linkage limits the power of natural selection in Drosophila. Proc. Natl. Acad. Sci. USA **99:** 13616–13620.

Birdsell, J. A., 2002 Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. **19:** 1181–1197.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin and F. A. Kondrashov, 2002 Selection for short introns in highly expressed genes. Nat. Genet. **31:** 415–418.

Coghlan, A., and K. H. Wolfe, 2000 Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae.* Yeast **16:** 1131–1145.

Costanzo, M. C., J. D. Hogan, M. E. Cusick, B. P. Davis, A. M. Fancher *et al.*, 2000 The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. Nucleic Acids Res. **28:** 73–76.

Craig, C. L., and R. S. Weber, 1998 Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by *Escherichia coli.* Mol. Biol. Evol. **15:** 774–776.

Craig, C. L., C. Riekel, M. E. Herberstein, R. S. Weber, D. Kaplan *et al.*, 2000 Evidence for diet effects on the composition of silk proteins produced by spiders. Mol. Biol. Evol. **17:** 1904–1913.

Crick, F. H., S. Brenner, A. Klug and G. Pieczenik, 1976 A speculation on the origin of protein synthesis. Origins Life **7:** 389–397.

Datta, A., and S. Jinks-Robertson, 1995 Association of increased spontaneous mutation rates with high levels of transcription in yeast. Science **268:** 1616–1619.

Davis, C. A., L. Grate, M. Spingola and M. Ares, Jr., 2000 Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. Nucleic Acids Res. **28:** 1700–1706.

Dufton, M. J., 1997 Genetic code synonym quotas and amino acid complexity: Cutting the cost of proteins? J. Theor. Biol. **187:** 165–173.

Duret, L., 2000 tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. **16:** 287–289.

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

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

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

Eigen, M., and P. Schuster, 1979 *The Hypercycle, a Principle of Natural Self-organization.* Springer-Verlag, Berlin/New York.

Farabaugh, P. J., 2000 Translational frameshifting: implications for the mechanism of translational frame maintenance. Prog. Nucleic Acid Res. Mol. Biol. **64:** 131–170.

Francino, M. P., L. Chao, M. A. Riley and H. Ochman, 1996 Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science **272:** 107–109.

Futcher, B., G. I. Latter, P. Monardo, C. S. Mclaughlin and J. I. Garrels, 1999 A sampling of the yeast proteome. Mol. Cell. Biol. **19:** 7357–7368.

Garat, B., and H. Musto, 2000 Trends of amino acid usage in the proteins from the unicellular parasite *Giardia lamblia.* Biochem. Biophys. Res. Commun. **279:** 996–1000.

Garel, J. P., 1974 Functional adaptation of tRNA population. J. Theor. Biol. **43:** 211–225.

Gerton, J. L., J. Derisi, R. Shroff, M. Lichten, P. O. Brown *et al.*, 2000 Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA **97:** 11383–11390.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.*, 1996 Life with 6000 genes. Science **274:** 546, 563–547.

Grosjean, H., and W. Fiers, 1982 Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18:** 199–209.

Grosjean, H., D. Sankoff, W. M. Jou, W. Fiers and R. J. Cedergren, 1978 Bacteriophage MS2 RNA: a correlation between the stability of the codon:anticodon interaction and the choice of code words. J. Mol. Evol. **12:** 113–119.

Gutiérrez, G., L. Marquez and A. Marin, 1996 Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. Nucleic Acids Res. **24:** 2525–2527.

Hastings, K. E., 1996 Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J. Mol. Evol. **42:** 631–640.

Hatfield, D., F. Varricchio, M. Rice and B. G. Forget, 1982 The aminoacyl-tRNA population of human reticulocytes. J. Biol. Chem. **257:** 3183–3188.

Hereford, L. M., and M. Rosbash, 1977 Number and distribution of polyadenylated RNA sequences in yeast. Cell **10:** 453–462.

Holstege, F. C., E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner *et al.*, 1998 Dissecting the regulatory circuitry of a eukaryotic genome. Cell **95:** 717–728.

Iida, K., and H. Akashi, 2000 A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene **261:** 93–105.

Ikemura, T., 1982 Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. **158:** 573–597.

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

Jansen, R., and M. Gerstein, 2000 Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. Nucleic Acids Res. **28:** 1481–1488.

Kanaya, S., Y. Yamada, Y. Kudo and T. Ikemura, 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238:** 143–155.

Karlin, S., and P. Bucher, 1992 Correlation analysis of amino acid usage in protein classes. Proc. Natl. Acad. Sci. USA **89:** 12165–12169.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK/New York.

Kliman, R. M., and J. Hey, 1994 The effects of mutation and natural selection on codon bias in the genes of Drosophila. Genetics **137:** 1049–1056.

Li, W.-H., 1997 *Molecular Evolution.* Sinauer Associates, Sunderland, MA.

Lobry, J. R., and C. Gautier, 1994 Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. **22:** 3174–3180.

Lopez, P. J., and B. Seraphin, 1999 Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition. RNA **5:** 1135–1137.

Marais, G., D. Mouchiroud and L. Duret, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98:** 5688–5692.

Mazel, D., and P. Marlière, 1989 Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. Nature **341:** 245–248.

Morey, N. J., C. N. Greene and S. Jinks-Robertson, 2000 Genetic analysis of transcription-associated mutation in *Saccharomyces cerevisiae.* Genetics **154:** 109–120.

Moriyama, E. N., and J. R. Powell, 1998 Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli.* Nucleic Acids Res. **26:** 3188–3193.

Morton, B. R., and B. G. So, 2000 Codon usage in plastid genes is correlated with context, position within the gene, and amino acid content. J. Mol. Evol. **50:** 184–193.

Nassar, R. F., and R. D. Cook, 1976 Non-randomness of nucleotide bases in mRNA codons. Genet. Res. **27:** 353–362.

Nei, M., 1975 *Molecular Population Genetics and Evolution.* North-Holland/American Elsevier, Amsterdam/New York.

Ogle, J. M., D. E. Brodersen, W. M. Clemons, Jr., M. J. Tarry, A. P. Carter *et al.*, 2001 Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science **292:** 897–902.

Pal, C., B. Papp and L. D. Hurst, 2001a Highly expressed genes in yeast evolve slowly. Genetics **158:** 927–931.

Pal, C., B. Papp and L. D. Hurst, 2001b Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. Mol. Biol. Evol. **18:** 2323–2326.

Palacios, C., and J. J. Wernegreen, 2002 A strong effect of AT mutational bias on amino acid usage in Buchnera is mitigated at high-expression genes. Mol. Biol. Evol. **19:** 1575–1584.

Percudani, R., 2001 Restricted wobble rules for eukaryotic genomes. Trends Genet. **17:** 133–135.

Percudani, R., and S. Ottonello, 1999 Selection at the wobble position of codons read by the same tRNA in *Saccharomyces cerevisiae.* Mol. Biol. Evol. **16:** 1752–1762.

Percudani, R., A. Pavesi and S. Ottonello, 1997 Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae.* J. Mol. Biol. **268:** 322–330.

Precup, J., and J. Parker, 1987 Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. **262:** 11351–11355.

Rice, W. R., 1989 Analyzing tables of statistical tests. Evolution **43:** 223–225.

Richmond, R. C., 1970 Non-Darwinian evolution: a critique. Nature **225:** 1025–1028.

Robinson, M., R. Lilley, S. Little, J. S. Emtage, G. Yarranton *et al.*, 1984 Codon usage can affect efficiency of translation of genes in *Escherichia coli.* Nucleic Acids Res. **12:** 6663–6671.

Rodnina, M. V., and W. Wintermeyer, 2001 Ribosome fidelity: tRNA discrimination, proofreading and induced fit. Trends Biochem. Sci. **26:** 124–130.

Sharp, P. M., and E. Cowe, 1991 Synonymous codon usage in *Saccharomyces cerevisiae.* Yeast **7:** 657–678.

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

Sharp, P. M., M. Stenico, J. F. Peden and A. T. Lloyd, 1993 Codon usage: Mutational bias, translational selection, or both? Biochem. Soc. Trans. **21:** 835–841.

Shepherd, J. C., 1981 Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. J. Mol. Evol. **17:** 94–102.

Shpaer, E. G., 1989 Amino acid composition is correlated with protein abundance in *Escherichia coli*: Can this be due to optimization of translational efficiency? Protein Seq. Data Anal. **2:** 107–110.

Snedecor, G. W., and W. G. Cochran, 1989 *Statistical Methods.* Iowa State University Press, Ames, IA.

Sorensen, M. A., C. G. Kurland and S. Pedersen, 1989 Codon usage determines translation rate in *Escherichia coli.* J. Mol. Biol. **207:** 365–377.

Trifonov, E. N., 1987 Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. J. Mol. Biol. **194:** 643–652.

Trifonov, E. N., 1992 Recognition of correct reading frame by the ribosome. Biochimie **74:** 357–362.

Trifonov, E. N., and J. L. Sussman, 1980 The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc. Natl. Acad. Sci. USA **77:** 3816–3820.

Varenne, S., J. Buc, R. Lloubes and C. Lazdunski, 1984 Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J. Mol. Biol. **180:** 549–576.

Vinogradov, A. E., 2001 Intron length and codon usage. J. Mol. Evol. **52:** 2–5.

Warner, J. R., 1999 The economics of ribosome biosynthesis in yeast. Trends Biochem. Sci. **24:** 437–440.

Wootton, J. C., and S. Federhen, 1996 Analysis of compositionally biased regions in sequence databases. Methods Enzymol. **266:** 554–571.

Wright, S. I., B. Lauga and D. Charlesworth, 2002 Rates and patterns of molecular evolution in inbred and outbred Arabidopsis. Mol. Biol. Evol. **19:** 1407–1420.

Xia, X., 1998 How optimized is the translational machinery in *Escherichia coli, Salmonella typhimurium* and *Saccharomyces cerevisiae*? Genetics **149:** 37–44.

Yamao, F., Y. Andachi, A. Muto, T. Ikemura and S. Osawa, 1991 Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. Nucleic Acids Res. **19:** 6119–6122.

Zavala, A., H. Naya, H. Romero and H. Musto, 2002 Trends in codon and amino acid usage in *Thermotoga maritima.* J. Mol. Evol. **54:** 563–568.

Communicating editor: W. Stephan