Synonymous Codon Usage in *Drosophila melanogaster*: Natural Selection and Translational Accuracy

Hiroshi Akashi

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received June 30, 1993
Accepted for publication November 3, 1993

ABSTRACT

I present evidence that natural selection biases synonymous codon usage to enhance the accuracy of protein synthesis in *Drosophila melanogaster*. Since the fitness cost of a translational misincorporation will depend on how the amino acid substitution affects protein function, selection for translational accuracy predicts an association between codon usage in DNA and functional constraint at the protein level. The frequency of preferred codons is significantly higher at codons conserved for amino acids than at nonconserved codons in 38 genes compared between *D. melanogaster* and *Drosophila virilis* or *Drosophila pseudoobscura* ($Z = 5.93, P < 10^{-6}$). Preferred codon usage is also significantly higher in putative zinc-finger and homeodomain regions than in the rest of 28 *D. melanogaster* transcription factor encoding genes ($Z = 8.38$, $P < 10^{-6}$). Mutational alternatives (within-gene differences in mutation rates, amino acid changes altering codon preference states, and doublet mutations at adjacent bases) do not appear to explain this association between synonymous codon usage and amino acid constraint.

SYNONYMOUS mutations are good candidates for selectively neutral evolution because they do not affect the primary structure of proteins (KIMURA 1968a; KING and JUKES 1969). Nonrandom patterns of codon usage, however, strongly suggest the action of natural selection on "silent" changes in DNA (GRANTHAM *et al.*, 1980, 1981; KIMURA 1983; IKEMURA 1985). Mutational biases or selection for genomic base composition have been suggested as important factors determining codon usage in some species (reviewed in SHARP 1989). In *Saccharomyces cerevisiae* and *Escherichia coli*, natural selection appears to discriminate among synonymous codons at the level of translation. This form of codon usage bias, called "major codon preference," has three characteristics. First, codons recognized by the most abundant tRNA for a given amino acid tend to be used preferentially. Among codons recognized by the most abundant tRNA, the codon forming the natural Watson-Crick pairing with the tRNA anticodon is generally favored (IKEMURA 1982; BENNETZEN and HALL 1982). Second, the degree of codon usage bias varies considerably between genes and correlates positively with gene expression levels (GRANTHAM *et al.* 1981; GOUY and GAUTIER 1982; GROSJEAN and FRIERS 1982; BENNETZEN and HALL 1982). Finally, silent divergence between species is inversely related to codon usage bias implying greater purifying selection on synonymous changes in highly biased genes (SHARP and LI 1987).

Codon usage in *Drosophila melanogaster* appears similar, in pattern, to that found in yeast and *E. coli* (SHIELDS *et al.* 1988). Codon usage is biased toward mostly C-ending codons and the scant data on tRNA levels show a positive relationship between favored codon usage and tRNA abundances. Although relative expression levels are difficult to quantify in multicellular organisms, highly expressed genes such as ribosomal proteins and *Adh* are highly biased while genes with limited or low expression such as *Adhr* show less biased codon usage. Mutational biases in the genome do not explain these patterns; silent position base composition is not correlated with presumably selectively neutral intron base composition (SHIELDS *et al.* 1988; MORIYAMA and HARTL 1993). Silent divergence between Drosophila species is inversely related to codon usage bias (SHARP and LI 1989; CARULLI *et al.* 1993).Although the action of natural selection on silent sites in *E. coli*, yeast and Drosophila is well established, the underlying cause of this selection remains ambiguous. Synonymous codon usage could affect at least two different aspects of protein synthesis. In *E. coli*, major tRNA-encoding codons are translated 3–6-fold faster than their synonymous counterparts (ROBINSON *et al.* 1984; VARENNE *et al.* 1984; SORENSEN *et al.* 1989). Enhancing elongation rates will increase the overall rate of protein synthesis by elevating cellular concentrations of free ribosomes. The fitness effect of changing elongation rates, through its impact on the free ribosome pool, will be a positive function of the number of times a gene is translated. Selection for faster growth rates will favor codons which increase elongation rates in highly expressed genes (KURLAND 1987a).

Preferred codons also enhance the accuracy of translation. In *E. coli*, favored codons can reduce the frequency of amino acid misincorporations found in protein products by approximately 10-fold over nonpreferred codons for the same amino acid (PRECUP and PARKER 1987). Enhancing the fidelity of protein synthesis

TABLE 1

Preferred codons in *D. Melanogaster*

Amino		Frequencies		Amino		Frequencies		Amino		Frequencies		Amino		Frequencies	
acid	Codon	Low	High	acid	Codon	Low	High	acid	Codon	Low	High	acid	Codon	Low	High
Phe	UUU			Ser	UCU			Tyr	UAU			Cys	UGU		
	UUC[a]	42.9	91.9		UCC[a]	29.2	61.1		UAC[a]	48.6	87.7		UGC[a]	59.7	90.4
Leu	UUA				UCA			ter	UAA			ter	UGA		
	UUG				UCG[a]	19.8	26.1	ter	UAG			Trp	UGG		
Leu	CUU			Pro	CCU			His	CAU			Arg	CGU[a]	16.6	36.7
	CUC[a]	8.8	14.7		CCC[a]	23.4	65.4		CAC[a]	46.8	82.4		CGC[a]	18.3	52.8
	CUA				CCA			Gln	CAA				CGA		
	CUG[a]	22.1	69.3		CCG				CAG[a]	51.5	91.9		CGG		
Ile	AUU			Thr	ACU			Asn	AAU			Ser	AGU		
	AUC[a]	25.8	80.9		ACC[a]	26.2	80.3		AAC[a]	44.9	87.5		AGC		
	AUA				ACA			Lys	AAA			Arg	AGA		
Met	AUG				ACG				AAG[a]	50.7	95.5		AGG		
Val	GUU			Ala	GCU			Asp	GAU			Gly	GGU		
	GUC[a]	19.8	36.5		GCC[a]	28.9	69.8		GAC[a]	35.3	61.7		GGC[a]	27.0	47.7
	GUA				GCA			Glu	GAA				GGA		
	GUG[a]	32.5	47.8		GCG				GAG[a]	45.1	91.5		GGG		

[a] "Preferred" codons within each synonymous family (SHARP and LLOYD 1993, Table 37.2). Frequencies of preferred codons within synonymous families in 44 genes showing the lowest and highest codon usage bias among 438 *D. melanogaster* genes are shown. SHARP and LLOYD Define preferred codons as those showing a statistically significant increase in frequency between the low and high bias groups in heterogeneity χ^2 tests using $P = 0.01$ as the critical level for significance.

at preferred codons could involve more accurate amino acid acylation of preferred anticodon-encoding tRNAs, greater fidelity in the initial discrimination step of protein synthesis at preferred codons, or more effective proofreading in the subsequent step at these codons. All three processes will decrease the frequency of misincorporated amino acids in the products of protein synthesis. Since the metabolic cost of misincorporations will be a function of the number of nonfunctional proteins synthesized, selection for translational accuracy should act more strongly in highly expressed genes. More accurate tRNA acylation or greater fidelity in the initial discrimination step will also increase the synthesis rate of functional proteins resulting in fitness benefits similar to increasing elongation rates (see above). More efficient proofreading will enhance translational accuracy at a cost of both decreasing elongation rates (KURLAND 1987b) and an added energetic cost to proofreading (GAVRILOVA *et al.* 1984). For the purpose of this analysis, the term "translational accuracy" will encompass all aspects of translation which lower the frequency of errors in protein products.

Experimental observations establish that major codon bias could result from either (or both) selection for translational elongation rates or translational accuracy. The relationship between favored codons and tRNA abundances and the between-gene correlations of codon usage bias and expression levels are consistent with both explanations. Here, I present evidence identifying natural selection for translational accuracy as a determinant of synonymous codon usage in *D. melanogaster*. Selection to enhance the accuracy of protein synthesis predicts a relationship between the strength of selection at silent sites in DNA and functional constraint at the protein level. A comparative statistical approach reveals a highly statistically significant association between synonymous codon usage and protein functional constraint in *D. melanogaster* DNA sequences.

MATERIALS AND METHODS

Codon families: I define synonymous codon families as two or more codons encoding the same amino acid. Codons with no synonyms, AUG, UGG, and termination codons are excluded from the analysis.

Preferred codons: Preferred codons in *D. melanogaster* increase in relative frequency as selection acts to bias codon usage. SHARP and LLOYD (1993) compare the frequencies of synonymous codons in *D. melanogaster* genes showing the lowest and highest codon usage bias among 438 genes. Codons showing a statistically significant increase in frequency between these classes are defined as "preferred" (Table 1). In the following analysis, interspecific data are used only to classify codons as conserved or diverged for amino acids; I make no assumptions about codon preference in *Drosophila virilis* or *Drosophila pseudoobscura*.

Interspecific comparisons: All available genes sequenced in *D. melanogaster* and either *D. virilis* or *D. pseudoobscura* were drawn from GenBank or directly from the literature (Table 2). A total of 39 genes were available for analysis. To ensure evolutionary independence of DNA sequence evolution between all genes, *Rh2* was eliminated from the data set because of extensive homology (>70% DNA sequence identity) with *Rh4*. Amino acid sequences of homologous genes were aligned using the "gap" program in the Genetics Computer Group Sequence Analysis Software Package Version 7.2. Only homologous amino acids, those which align with other amino acids in the between-species comparison, were used in

TABLE 2

Genes used in between species comparisons

					Codon		Accession numbers	
Gene	Full name/product	Species	K_a	K_s	no.	Z	Mel	Other
ade3	Adenosine-3	*p*	0.10	1.27	1265	1.53	J02527	X06285
Adh	Alcohol dehydrogenase	*p*	0.06	0.56	192	1.55	M36580	X62181
Adhr	Alcohol dehydrogenase related protein	*p*	0.05	1.49	179	1.34		X62181
Amy-d	α-Amylase	*p*	0.07	0.37	385	-0.02	X04569	
Antp	Antennapedia	*v*	0.04	1.01	252	-1.47	M20705	M95826
Aprt	Adenine-phosphoribosyltransferase	*p*	0.09	1.27	157	0.38	M18432	L06281
bcd	Bicoid	*p*	0.12	1.04	421	0.39	X07870	X55735
boss	Bride of sevenless		0.13	1.74	804	1.11	L08133	L08132
Cp15	Chorion protein S15	*v*	0.55	0.92	84	0.93	X02497	X53421
Cp16	Chorion protein S16	*v*	0.24	0.78	119	1.08	X16715	X53421
Cp18	Chorion protein S18	*v*	0.27	0.83	140	-0.56	X02497	X53421
Cp19	Chorion protein S19	*v*	0.34	0.75	145	-0.01	X02497	X53421
Cp36	Chorion protein S36	*v*	0.16	0.58	269	1.43	X05245	X51343
elav	Embryonic lethal, abnormal vision	*v*	0.03	1.38	213	0.57	M21152	M61748
en	Engrailed	*v*	0.14	1.01	426	1.65*	M10017	X04727
Est6/5a	Esterase 6/esterase 5a	*p*	0.26	2.40	502	1.01	M15961	M55908
Fmrf	FMRFamide-related	*v*	0.28	1.13	281	-0.39	J03232	M32643
Gld	Glucose dehydrogenase	*p*	0.06	1.12	534	-0.13	M29298	M29299
Gpdh	Glycerol-3-phosphate dehydrogenase	*v*	0.03	1.35	132	0.97	J04567	X59076
h	Hairy	*v*	0.08	0.88	254	0.08	X15905	S63793
hb	Hunchback	*v*	0.14	1.11	658	2.88**	Y00274	X15359
Hsp83	Heat shock protein 83	*v*	0.00	0.71	57	0.66	X03810	X02813
mam	Mastermind	*v*	0.15	0.97	1364	-0.03	X54251	M92914
Pcp	Pupal cuticle protein	*p*	0.16	0.97	170	0.16	J02527	X06285
per	Period	*v*	0.25	1.05	927	0.37	M30114	X13877
Rh4	Rhodopsin-4	*v*	0.05	1.61	208	0.56	M17730	M77281
ro	Rough	*v*	0.29	1.53	294	-0.54	M23629	M35372
ry	Rosy: xanthine dehydrogenase	*p*	0.07	1.25	1261	0.52	Y00308	M33977
sev	Sevenless	*v*	0.31	1.29	2293	3.40***	J03158	M34545
sina	Seven in absentia	*v*	0.01	0.88	83	0.83	M38384	M77282
slbo	Slow border cells	*v*	0.17	1.09	363	0.29	L00632	L00725
su(Hw)	Suppressor of Hairy wing		0.23	1.45	842	1.91*	Y00228	Z25520
tll	Tailless	*v*	0.04	1.17	269	-0.09	L04954	L04955
tra	Transformer	*v*	0.53	1.52	128	1.58	M17478	X66528
tub	Tube	*v*	0.36	1.60	413	1.35	M59501	L20449
Ubx	Ultrabithorax	*p*	0.06	0.70	156	0.60	X05724	X05179
Uro	Urate oxidase	*v*	0.17	1.38	328	2.18*	X51940	X57114
z	Zeste	*v*	0.27	1.28	528	1.24	Y00049	M76700

All genes were drawn from the GenBank/EMBL DNA sequence library (GenBank release 78.0) except *D. pseudoobscura amy-d* (Brown *et al.* 1990), and *D. melanogaster Adhr* (Kreitman and Hudson 1991). Gene names follow FlyBase (1993). "Species" refers to the species used in the interspecific comparison to *D. melanogaster*. "*p*" denotes *D. pseudoobscura*, and "*v*" refers to *D. virilis*. "Codon no." denotes the total number of codons from each gene used in the analysis. *Z* values are calculated as described in the text. K_s and K_a were calculated according to Nei and Gojobori (1986). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

the analysis. Amino acids in *D. melanogaster* proteins aligning with gaps in the between-species comparisons were excluded from the analysis.

Statistical test: The frequency of preferred codons at conserved and non-conserved amino acid positions in the *D. melanogaster* genes were compared in 2 × 2 contingency tables. Independent tables were constructed for each synonymous codon family within each gene (see Table 3 for example). A joint probability for all tables (for each gene and in the whole data set) was calculated according to the Mantel-Haenszel procedure (Mantel and Haenszel 1959; Mantel 1963). This statistical test takes into account both the magnitude and direction of deviations within independent contingency tables to test for an overall association in the data. Because the direction of deviations was predicted prior to the analysis, all probabilities were calculated as one-tailed tests.

The same procedure was applied to the 38 genes in Table 2 using two restricted subsets of diverged amino acid sites. In the first subset, only non-conserved codons for which all intermediate states in all reconstructed mutational pathways between the extant codons have the same possible silent position bases and the same favored silent base(s) were used. In the second restricted subset, only diverged amino acid sites at which both bases immediately flanking the silent position(s) are conserved between species were used. For 2-, 3- and 4-fold redundant codons, those at which both the second position within the codon and the first position in the 3′ codon are conserved between species are included in the analysis. For 6-fold redundant sites, only those with conserved bases at the third position of the 5′ codon as well as the second base within the codon and the first base 3′ to the codon are included in the analysis. Non-conserved amino acid codons satisfying these criteria were compared to conserved amino acid positions as described above.

Comparisons within transcription-factor encoding genes: The relationship between codon usage bias and functional constraint was also examined in 28 *D. melanogaster* homeodomain and zinc-finger genes drawn from GenBank (Table 4). Rather than using amino acid conservation to infer functional constraint, however, I compare the frequency of favored codons between the coding sequence for the putative DNA-binding domains and the rest of each gene. Homeodomain

TABLE 3

Preferred codon frequencies at glutamic acid codons in the *D. melanogaster sev* gene

Codon	Conserved sites	Non-conserved sites
GAG	79	57
GAA	23	25
% preferred codon	77.5	69.5

The frequency of the preferred codon, GAG, is higher at conserved amino acid positions than at non-conserved amino acid positions between *D. melanogaster* and *D. virilis*.

regions were delimited as defined by SCOTT *et al.* (1989) and zinc-finger motifs were defined as by the authors.

RESULTS

Rationale of the statistical test: Amino acid substitutions at different positions within a protein vary in their effect on protein function. Changes at some peptide positions will have little or no effect on function whereas substitutions at other sites will disrupt protein structure or activity. The fitness cost of translational errors, if natural selection acts to increase the rate of synthesis of functional peptides or to lower the metabolic cost of wasted synthesis, will be highest at codons where misincorporations disrupt protein function. Translational errors resulting in fully functional proteins will entail little if any fitness cost to the organism. If natural selection biases codon usage to enhance the accuracy of protein synthesis, then preferred codon usage will be stronger at functionally constrained amino acid positions than at less constrained sites. Other forms of translational selection do not predict such an association. Selection for increased translational elongation rates predicts stronger codon bias at rate-limiting regions within genes. There is no reason to expect within-gene variation in translational elongation rates to correlate with functional constraint at the protein level.

Testing the predicted association between codon usage and protein functional constraint requires some knowledge of the functional consequences of amino acid substitutions. Two different kinds of information can be used to determine relative tolerance to amino acid changes at different peptide positions–evolutionary conservation and protein functional analysis. Under the neutral theory of molecular evolution, purifying selection will prevent divergence of functionally important amino acids over evolutionary time while mutation and drift will lead to divergence of less constrained sites (KIMURA 1983). Codon selection to enhance the fidelity of translation predicts greater preferred codon usage at evolutionarily conserved amino acid positions than at diverged sites.

Alternatively, regional constraint in proteins can be determined from functional studies. Homeodomains and zinc-finger motifs are well characterized, functionally important DNA-binding regions found in numerous transcription factors (reviewed in SCHLIEF 1988). Since constraint is known to be high in these regions, selection for translational accuracy predicts stronger selection at silent sites in DNA-binding domains than elsewhere in transcription factor-coding genes. Since functional rather than interspecific data are used to classify codons, this prediction does not require the assumption that most amino acid changes between species are selectively neutral or nearly neutral.

To test for an association between functional constraint and codon bias, I attempt to control for other causes of variability in selection intensity at silent sites. The fitness cost of translational errors at a given amino acid position will be a function of the error frequency at the site, the effect of particular misincorporations on protein function, and the expression level of the gene. Error rates depend on both the sequence of a codon and its neighboring bases. Separate comparisons of codon usage for each synonymous family control for variation in codon-sequence dependent misincorporation rates. This also controls for the chemical similarity of misincorporated amino acids, since the same amino acids will be misincorporated at a given codon. The nature of codon context effects on translational error rates is unknown, so neighboring base effects could not be taken into account. PRECUP and PARKER's (1987) study showed an approximately 10-fold effect on misincorporation frequency of silent sites within a codon and only about a 2-fold effect of neighboring bases. Context effects should not be strong relative to codon identity and (unless context dependent error rates are associated with functional constraint at the protein level) will not bias the test. Finally, to eliminate variation in fitness effects due to expression level differences, I compare only codons within genes. A given triplet codon will err to the same amino acids at similar rates and will be expressed at the same level as the same codon elsewhere in a gene. To test whether selection to enhance the accuracy of translation has resulted in a higher frequency of favored codons at constrained amino acid positions, I compare the frequency of preferred codons (a measure of the strength of selection) between constrained and less constrained amino acid sites for each synonymous codon family within each gene.

Application of the test using interspecific comparisons: Comparison of favored codon frequencies in 38 *D. melanogaster* genes reveals a highly statistically significant association between codon usage and amino acid conservation as determined by comparisons with *D. virilis* or *D. pseudoobscura* sequences ($Z = 5.93$, $P < 10^{-6}$). Preferred codons occur more frequently at conserved (presumably constrained) residues than at non-conserved positions for synonymous families within the same gene. A total of 525 contingency tables (13,179 conserved and 3,917 non-conserved codons) were combined in the analysis. Analyzed separately, 29 of the 38

TABLE 4

Genes used in comparisons within homeodomain and Zn-finger encoding genes

Gene	Full name/product	Functional domain(s)	Codon no.	Z	Accession no.
abd-A	Abdominal A	Hom	262	1.53	X54453
Antp	Antennapedia	Hom	264	-0.09	M20705
ap	Apterous	Hom	389	-1.84	M92841
B	Bar	Hom	778	0.24	M73079
bcd	Bicoid	Hom	457	0.51	X07870
cad	Caudal	Hom	334	4.47***	M21070
Cf2	Chorion factor 2	Zn-f	214	0.87	X53380
ci	Cubitus interruptus	Zn-f	1285	-0.14	X54360
ct	Cut	Hom	1976	0.09	X07985
Dfd	Deformed	Hom	473	2.24*	X05136
en	Engrailed	Hom	464	1.24	M10017
eve	Even skipped	Hom	331	1.19	M14767
ftz	Fushi tarazu	Hom	310	1.37	X00854
gl	Glass	Zn-f	546	3.12***	X15400
H2.0	Homeodomain protein 2.0	Hom	374	2.30*	Y00843
hb	Hunchback	Zn-f	683	3.05**	Y00274
inv	Invected	Hom	501	1.51	X05273
Kr	Kruppel	Zn-f	427	0.42	X03414
lab	Labial	Hom	386	-0.05	X13103
prd	Paired	Hom	517	1.79*	M14548
ro	Rough	Hom	279	1.96*	M35372
Scr	Sex combs reduced	Hom	274	2.86**	X14475
sna	Snail	Zn-f	361	3.21***	Y00288
sry-b	Serendipity b	Zn-f	324	0.60	X03121
su(HW)	Suppressor of Hairy wing	Hom	612	1.19	Y00228
ttk	Tramtrak	Zn-f	518	1.07	X17121
zfh1	Zn-finger homeodomain protein 1	Hom, Zn-f	968	3.90***	M63449
zfh2	Zn-finger homeodomain protein 2	Hom, Zn-f	2813	2.56**	M63450

All genes were drawn from the GenBank/EMBL DNA sequence library (GenBank release 78.0). Gene names follow FlyBase (1993). "Hom" denotes homeodomain and "Zn-f" denotes zinc-finger protein. Zinc-finger coding regions were defined as in: Cf2 (Shea *et al.* 1990), *ci* (Orenic *et al.* 1990), *gl* (Moses *et al.* 1989), *hb* (Tautz *et al.* 1987), *Kr* (Rosenberg *et al.* 1986), *sna* (Boulay *et al.* 1987), *sry-b* (Vincent *et al.* 1985), *ttk* (Harrison and Travers 1990), *zfh1*, *zfh2* (Fortini *et al.* 1991). "Codon no." refers to the number of codons from each gene used in the analysis. Z values were calculated as described in the text. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

genes deviate in the predicted direction and 5 are significant at the 5% level (Table 2). Longer genes, with larger sample sizes, tend to show stronger associations. The association between codon usage and amino acid conservation remains highly significant when diverged sites are limited to those which have retained the same silent bases as well as the same favored third position base through all intermediates in all mutational pathways between extant codon sequences ($Z = 3.49$, $P = 2.4 \times 10^{-4}$). A total 228 contingency tables (6,847 conserved and 717 non-conserved codons) were combined in the analysis. The association between codon usage and amino acid conservation also remains significant when diverged codon positions are limited to those for which synonymous sites are flanked on both sides by conserved bases ($Z = 2.48$, $P = 6.6 \times 10^{-3}$). A total of 355 contingency tables (10,350 conserved and 1,076 non-conserved codons) were combined in this analysis.

Application of the test using comparisons within transcription factors: Comparison of favored codon frequencies in 28 *D. melanogaster* homeodomain and zinc-finger motif genes reveals a highly statistically significant association between codon usage and putative DNA-binding domains ($Z = 8.38$, $P < 10^{-6}$). Preferred codon usage bias is significantly greater within the more constrained homeodomain and zinc-finger coding areas than elsewhere in these genes. A total of 457 contingency tables (3,098 motif region and 14,022 non-motif region codons) were combined in the analysis. Analyzed separately, 24 of the 28 genes deviate in the predicted direction and 11 are significant at the 5% level. Again, genes with more codons tend to show the strongest associations.

DISCUSSION

Silent divergence, K_s, is greater than one fixation/silent site between *D. melanogaster* and both *D. virilis* and *D. pseudoobscura* in genes with low codon bias (Table 2). If these rates of substitution reflect values close to the neutral rate, then the majority of unconstrained amino acid positions will also have changed between these species. The strong association between preferred codon usage and amino acid divergence in the 38 *D. melanogaster* genes examined here supports the translational accuracy hypothesis–natural selection biases codon usage to enhance the fidelity of protein synthesis. Selection to increase elongation rates does not predict such a relationship.

Three different mutational alternatives, however, could explain this observation. If mutation rates vary between codons within a gene, then, in the absence of selective differences, increased amino acid divergence and weaker codon usage bias will occur at codons with higher mutation rates. Since the neutral rate of substitution is proportional to the mutation rate (KIMURA 1968b), higher mutation rates will elevate the rate of amino acid divergence. If codon usage is in mutation-selection balance (SHARP and LI 1986; BULMER 1988, 1991), a higher mutation rate, given the same selection intensity, will cause a shift away from preferred codon usage at silent sites. The observed pattern could reflect an association between mutation rates at different codons with both replacement site divergence and silent site mutation-selection balance. Examination of codon usage in genes with experimental evidence for functional constraint tests this possibility. Homeodomains and zinc-finger motifs are well characterized, functionally important DNA-binding regions found in numerous transcription factors in all eukaryotes examined. Mutagenesis experiments show that the DNA-binding function is highly sensitive to amino acid changes in these regions (reviewed in SCHLEIF 1988). Although these regions may not be the only functionally important or constrained regions of the proteins, overall, the deleterious effects of amino acid substitutions are likely to be higher in these areas. As predicted by codon selection for translational accuracy, preferred codon usage is significantly higher within putative homeodomains and zinc-finger coding regions than elsewhere in 28 *D. melanogaster* genes ($Z = 8.38, P < 10^{-6}$). Mutational differences do not explain this result unless mutation rates tend to be lower in DNA-binding motif encoding regions.

LIPMAN and WILBUR (1985) point out a second mutational explanation for an association between codon usage and conservation of amino acids. When preferred silent bases differ between synonymous families, replacement mutations can shift peptide positions from a favored to unfavored codon state. For example, a replacement mutation from GUG to GCG in *D. melanogaster* will change the preferred codon for valine to a nonfavored codon for alanine. This essentially adds to mutation pressure away from preferred codons at less constrained peptide positions. In *D. melanogaster*, only 36 of the possible 196 (18.4%) single base amino acid replacement mutations change codon preference states. Given that the majority of replacement changes retain the favored silent base and the preponderance of silent substitutions over amino acid replacements in most genes (Table 2), the possibility that such mutations will significantly lower preferred codon usage at diverged amino acid sites seems unlikely.

To test whether such an effect could explain the observed associations, I examined a subset of the data to eliminate replacement substitutions that alter codon preference state in the non-conserved class of codons. I limited diverged sites to those that have retained the same silent bases and the same favored silent base(s) in all reconstructed pathways between the extant codon sequences. For such codons, regardless of replacement substitutions at first and second positions, synonymous evolution at the third position will always involve mutations between the same silent bases and selection for the same preferred base(s). Although the sample size reduces from 3,917 to 717 non-conserved codons, the association between codon usage bias and amino acid constraint remains highly significant ($Z = 3.49$, $P = 2.4 \times 10^{-4}$).

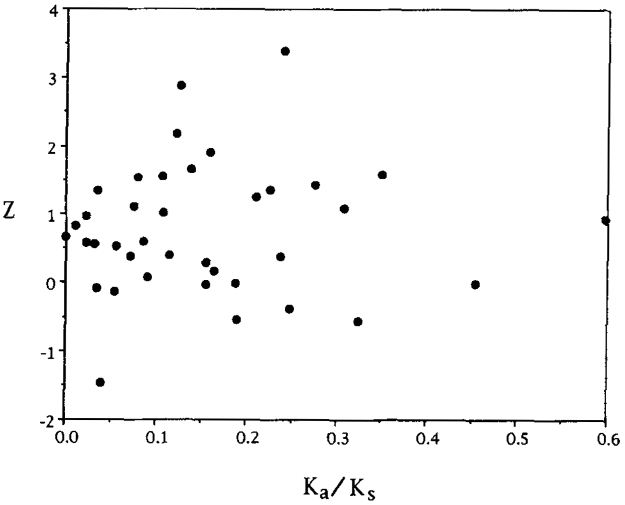

FIGURE 1.—Lack of relationship between the ratio of replacement, K_a, to synonymous, K_s, divergence and the association between codon usage and amino acid conversion. Data from Table 2.

LIPMAN and WILBUR's suggestion also predicts a relationship between the strength of association between amino acid conservation and codon usage bias and the relative contributions of replacement and synonymous changes in a given gene. The effect of replacement changes altering codon preference states will be greater in genes showing a higher ratio of replacement to silent mutations between species. Natural selection for translational accuracy does not predict such a relationship. Figure 1 illustrates the lack of a correlation between the ratio of K_a to K_s and the strength of the association between preferred codon usage and amino acid conservation ($r = 0.05$, $P = 0.73$). Replacement mutations changing codon preference states do not appear to be the sole cause of lower codon usage bias at non-conserved amino acid codons.

WOLFE and SHARP (1993) propose a third mutational alternative to explain the observed association between codon usage and amino acid conservation. Comparisons of both coding and non-coding DNA sequences between rat and mouse show an excess of tandem substitutions between adjacent bases. To explain this pattern and to account for the positive correlation between silent and replacement divergence between genes in mammals, WOLFE and SHARP suggest that a proportion of mutations arise and go to fixation as "doublets" between adjacent bases. Such mutations, if they occur with appreciable frequency relative to single base changes, could elevate both silent and replacement mutation pressure at weakly constrained peptide positions. The contribution of doublet mutations in Drosophila molecular evolution has not been established.

To eliminate the possible contribution of doublet mutations in the between species test of codon usage and amino acid conservation, I limit non-conserved amino acid codons to those at which synonymous sites are flanked on both sides by conserved bases. At such sites, silent base composition cannot be attributed to the fixation of doublet mutations. Although the predicted association between codon usage and amino acid constraint is weakened by eliminating many of the least constrained peptide positions where the amino acids are most divergent chemically (those which differ at the second codon position) and by reducing the sample size of non-conserved amino acid codons from 3,917 to 1,076, the higher frequency of preferred codons at conserved amino acids sites remains significant ($Z = 2.48$, $P = 6.6 \times 10^{-3}$). The fixation of doublet mutations, although a possible contributor, does not explain the observed excess of preferred codons at conserved amino acid positions.

Although mutational alternatives provide formally valid explanations of greater preferred codon usage at conserved amino acid positions, differences in mutation rates, replacement mutations altering codon preference states, and doublet mutations do not account for the observed association between codon usage and functional constraint in 38 *D. melanogaster* genes. Natural selection at silent sites to enhance the accuracy of protein synthesis appears to be necessary to explain the patterns found in this data. The underlying cause of this selection remains to be established. *In vivo* measurements of translational misincorporations in *E. coli* estimate the frequency of errors at 10^{-4} to 10^{-3} per site (reviewed in PARKER 1989). Preferred codons can reduce translational misincorporation rates by about 10-fold over non-preferred synonymous codons (PRECUP and PARKER 1987). Reducing error rates can increase the rate of production of functional peptides and lower the metabolic cost of synthesizing dysfunctional products. Selection may also operate against functional interference caused by proteins with misincorporated amino acids. The existence of dominant negative mutations suggest that such changes can occur. Positive correlations between codon usage bias and gene expression levels supports selection for synthesis rate or metabolic efficiency but it is unclear how these alternative mechanisms can be distinguished.

Some evidence relating codon bias and amino acid conservation has also been found in mammals (KAFATOS *et al.* 1977; LIPMAN and WILBUR 1985; TICHER and GRAUR 1989). However, the study of codon usage in mammals is complicated by nonuniform patterns of base composition and these studies do not exclude mutational alternatives to selection at silent sites. Mammalian genomes are mosaics of regions of varying G/C richness (BERNARDI *et al.* 1985), and codon usage within genes correlates with the base content of introns and adjacent flanking regions (BERNARDI and BERNARDI 1985; AOTA and IKEMURA 1986). Mammalian silent evolution may reflect mutational differences or selection on regional base composition rather than translational selection among synonymous codons.

An association between codon usage and protein functional constraint could explain the unexpected relationship between silent and replacement substitution rates in several lineages (SHARP 1991; TICHER and GRAUR 1989). Both prokaryotes (SHARP and LI 1987) and mammals (GRAUR 1985; LI *et al.* 1985; WOLFE and SHARP 1993) show statistically significant positive correlations between K_s and K_a. Given the same expression levels, selection for translational accuracy will act more strongly to bias codon usage toward preferred codons and conserve silent sites in highly constrained proteins where purifying selection will also conserve replacement sites. Since codon bias and silent divergence depend on both the expression level and overall constraint in a given protein, expression level differences between genes will weaken the association between K_s and K_a. Selection will not act very strongly to bias codon usage in highly constrained genes with low expression levels or in unconstrained genes that are highly expressed. The correlations between synonymous and non-synonymous rates of evolution in prokaryotes and mammals raise the possibility that selection may bias codon usage to enhance translational accuracy in groups other than Drosophila. Alternatively, the same correlation in different taxonomic groups could have different causes such as doublet mutations in mammals (WOLFE and SHARP 1993).

Surprisingly, in the 38 Drosophila genes used in this study, the correlation between K_s and K_a is not statistically significant ($r = 0.19$, $P = 0.24$). The five chorion genes appear to violate the general trend; they show a relatively high amino acid divergence and low synonymous divergence among these genes (Table 2 and Figure 2). When the five chorion genes, *Cp15*, *Cp16*, *Cp18*, *Cp19* and *Cp36*, are eliminated from the analysis, the correlation between K_s and K_a becomes statistically significant ($r = 0.43$, $P = 0.013$). These proteins are structural components of the extracellular assembly of the Drosophila eggshell. The chorion gene clusters are amplified 20–80 fold and are very highly expressed during oogenesis in female Drosophila (KAFATOS *et al.* 1987). If the high rate of amino acid evolution in these genes

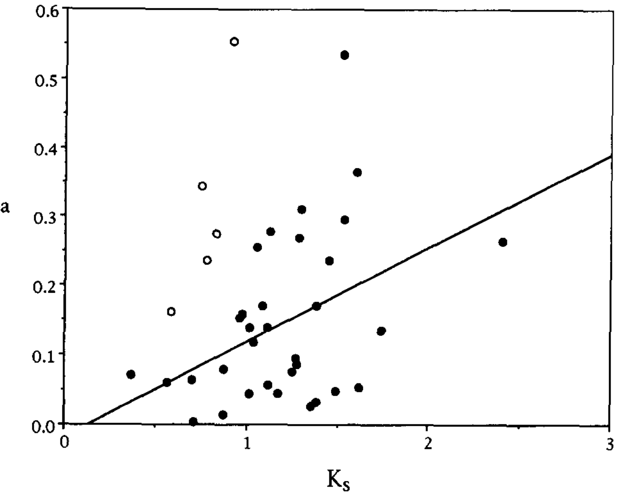

FIGURE 2.—Relationship between replacement, K_a, and synonymous, K_s, divergence in 38 genes sequenced in *D. melanogaster* and either *D. virilis* or *D. pseudoobscura*. Data from Table 2. Open circles represent chorion genes *Cp15*, *Cp16*, *Cp18*, *Cp19* and *Cp36*.

reflects low functional constraint, then selection for translational accuracy should be less effective. Strong codon bias in highly expressed but weakly constrained genes could reflect selection at silent sites to enhance translational elongation rates. The results presented here have no bearing on whether, or how strongly, factors other than translational accuracy contribute to fitness differences between synonymous codons. However, in the absence of independent criteria for identifying translation-rate limiting regions within protein-coding DNA sequences, statistical tests to establish codon selection for elongation rates will be difficult to develop.

I am grateful to M. KREITMAN for many valuable discussions. I also thank S. ORZACK, J. H. MCDONALD, J. BERGELSON, M. BULMER, P. SHARP and an anonymous reviewer for comments and criticism. H.A. is a Howard Hughes Medical Institute Predoctoral Fellow.

LITERATURE CITED

AOTA, S., and T. IKEMURA, 1986 Diversity in G+C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. **14:** 6345–6355.

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

BERNARDI, G., and G. BERNARDI, 1985 Codon usage and genome composition. J. Mol. Evol. **22:** 363–365.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL and F. RODIER, 1985 The mosaic genome of warm-blooded vertebrates. Science **228:** 953–958.

BOULAY, J. L., C. DENNEFELD and A. ALBERGA, 1987 The *Drosophila* developmental gene *Snail* encodes a protein with nucleic acid binding fingers. Nature **330:** 395–398.

BROWN, C. J., C. F. AQUADRO and W. ANDERSON, 1990 DNA sequence evolution of the amylase multigene family in *Drosophila pseudoobscura*. Genetics **126:** 131–138.

BULMER, M., 1988 Codon usage and intergenic position. J. Theor. Biol. **133:** 67–71.

BULMER, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

CARULLI J. P., D. E. KRANE, D. L. HARTL and H. OCHMAN, 1993 Compositional heterogenaity and patterns of molecular evolution in the Drosophila genome. Genetics **134:** 837–845.

FLYBASE, 1993 THE DROSOPHILA GENETIC DATABASE. AVAILABLE FROM THE FTP.BIO.INDIANA.EDU NETWORK SERVER AND GOPHER SITE.

FORTINI, M. E., Z. LAI and G. M. RUBIN, 1991 The *Drosophila* Zfh-1 and Zfh-2 genes encode novel proteins containing both zinc-finger and homeodomain motifs. Mech. Dev. **34:** 113–122.

GAVRILOVA, L. P., D. G. KAKHNIASHVILI and S. K. SMAILOV, 1984 Stoichiometry of GTP hydrolysis in a poly(U)-dependent cell-free translation system. FEBS Lett. **178:** 283–287.

GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7064.

GRANTHAM, R., C. GAUTIER, M. GOUY, R. MERCIER and A. PAVE, 1980 Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8:** r49–79.

GRANTHAM, R., C. GAUTIER, M. GOUY, M. JACOBZONE and R. MERCIER, 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9:** r43–79.

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

GROSJEAN, H., and W. FREIRS, 1982 Preferential codon usage in procaryotic genes: the optimal codon-anti-codon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18:** 199–209.

HARRISON, S. D., and A. A. TRAVERS, 1990 The *Tramtrack* gene encodes a *Drosophila* finger protein that interacts with the *Ftz* transcriptional regulatory region and shows a novel embryonic expression pattern. EMBO **9:** 207–216.

IKEMURA, T., 1982 Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. 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.

KAFATOS, F. C., A. EFSTRATIADIS, B. G. FORGET and S. M. WEISSMAN, 1977 Molecular evolution of human and rabbit β-globin mRNAs. Proc. Natl. Acad. Sci. **74:** 5618–5622.

KAFATOS, F. C., N. SPOEREL, S. A. MITSIALIS, H. T. NGUYEN, C. RAMANO, J. R. LINGAPPA, B. D. MARIANI, G. C. RODAKIS, R. LECANIDOU and S. G. TSITILOU, 1987 Developmental control and evolution in the chorion gene families of insects. Adv. Genetics **24:** 223–242.

KIMURA, M., 1968a Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. **11:** 247–269.

KIMURA, M., 1968b Evolutionary rate at the molecular level. Nature **217:** 624–626.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KING, J. L., and T. H. JUKES, 1969 Non-Darwinian evolution. Science **164:** 788–798.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

KURLAND, C. G. 1987a Strategies for efficiency and accuracy in gene expression. 1. The major codon preference: a growth optimization strategy. Trends Biochem. Sci. **12:** 126–128.

KURLAND, C. G. 1987b Strategies for efficiency and accuracy in gene expression. 2. Growth optimized ribosomes. Trends Biochem. Sci. **12:** 169–171.

LI, W.-H., C.-I. WU and C.-C. LUO, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

LIPMAN, D. J., and W. J. WILBUR, 1985 Interaction of silent and replacement changes in eucaryotic coding sequences. J. Mol. Evol. **21:** 161–167.

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

MANTEL, N., and W. HAENSZEL, 1959 Statistical aspects of the analysis of data from the retrospective analysis of disease. JNCI **22:**719.

MORIYAMA, E. N., and D. L. HARTL, 1993 Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics **134:** 847–858.

MOSES, K., M. C. ELLIS and G. M. RUBIN, 1989 The *Glass* gene encodes a zinc-finger protein required by *Drosophila* photoreceptor cells. Nature **340:** 531–536.

NEI, M., and T. GOJOBORI, 1986 Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

ORENIC, T. V., D. C. SLUSARSKI, K. L. KROLL and R. A. HOLMGREN, 1990 Cloning and characterization of the segment polarity gene cubitus interruptus dominant of *Drosophila*. Genes Dev. **4:** 1053–1067.

PARKER, J., 1989 Errors and alternatives in reading the universal genetic code. Microbiol. Rev. **53:** 273–298.

PRECUP, J., and J. PARKER, 1987 Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. **262:** 11351–11356.

ROBINSON, M., R. LILLEY, S. LITTLE, J. S. EMTAGE, G. YAMAMOTO, P. STEPHENS, A. MILLICAN, M. EATON and G. HUMPHREYS, 1984 Codon usage can effect efficiency of translation of genes in *Escherichia coli*. Nucleic Acids Res. **12:** 6663–6671.

ROSENBERG, U. B., C. SCHRODER, A. PREISS, A. KIENLIN, S. COTE, I. RIEDE and H. JACKLE, 1986 Structural homology of the product of the *Drosophila Kruppel* gene with *Xenopus* transcription factor IIIA. Nature **319:** 336–339.

SCHLEIF, R., 1988 DNA binding by proteins. Science **241:** 1182–1187.

SCOTT, M. P., J. W. TAMKUN and G. W. HARTZELL, III, 1989 The structure and function of the homeodomain. Biochim. Biophys. Acta **989:** 25–48.

SHARP, P. M., 1989 Evolution at 'silent' sites in DNA, pp. 23–32 in *Evolution and Animal Breeding: Reviews in Molecular and Quentitative Approaches in Honour of Alan Robertson*, edited by W. G. HILL and T. F. C. MACKAY. CAB International, Wallingford, U.K.

SHARP, P. M., 1991 Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. **33:** 23–33.

SHARP, P. M., and W.-H. LI, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

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

SHARP, P. M., and W.-H. LI, 1989 On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Biol. **28:** 398–402.

SHARP, P. M., and A. T. LLOYD, 1993 Codon Usage, pp. 378–397 in *An Atlas of Drosophila Genes: Sequences and Molecular Features*, edited by G. MARONI. Oxford University Press, New York.

SHEA, M. J., D. L. KING, M. J. CONBOY, B. D. MARIANI and F. C. KAFATOS, 1990 Molecular cloning of the cDNAs for two *Drosophila* zinc finger proteins that bind to chorion cis-regulatory elements: a new C2H2 protein and a C2C2 steroid receptor-like component. Genes Dev. **4:** 1128–1140.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS and F. WRIGHT, 1988 "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

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

TAUTZ, D., R. LEHMANN, H. SCHNUERCH, R. SCHUH, E. SEIFERT, A. KIENLIN, K. JONES, and H. JAECKLE, 1987 Finger protein of novel structure encoded by *hunchback*, a second member of the gap class of *Drosophila* segmentation genes. Nature **327:** 383–389.

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

TOMLINSON, A., B. E. KIMMEL and G. M. RUBIN, 1988 Rough, a *Drosophila* homeobox gene required in photoreceptors R2 and R5 for inductive interactions in the developing eye. Cell **55:** 771–784.

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. Biol. Chem. **180:** 549–576.

VINCENT, A., H. V. COLOT and M. ROSBASH, 1985 Sequence and structure of the serendipity locus of *Drosophila melanogaster*, a densely transcribed region including a blastoderm-specific gene. J. Mol. Biol. **186:** 149–166.

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

Communicating editor: A. G. CLARK