J Mol Evol (2002) 55:52–64
DOI: 10.1007/s00239-001-0090-1

JOURNAL OF MOLECULAR EVOLUTION

Translation Initiation AUG Context Varies with Codon Usage Bias and Gene Length in *Drosophila melanogaster*

Hitoshi Miyasaka

Environmental Research Center, The Kansai Electric Power Company, 3-11-20 Nakoji, Amagasaki, Hyogo 661-0974, Japan

Received: 6 July 2001 / Accepted: 28 December 2001

Abstract. The relationship between the codon usage bias and the sequence context surrounding the AUG translation initiation codon was examined in 1100 *Drosophila melanogaster* mRNA sequences. The codon usage bias measured by the "codon adaptation index" (CAI), and the effectiveness of the AUG context for translation initiation assessed by the "AUG context adaptation index" (A_{UG}CAI), showed a significant positive relationship (correlation coefficient: $r = 0.34$, $p < 0.0001$), indicating that these two factors are evolutionally under a similar natural selection constraint at the translational level. The importance of each position of the AUG context in relation to codon usage bias was examined, and the preference for the nucleotide at the −13, −12, −11, −10, −7, −6, −5, −4, −3, −2, and −1 positions showed a significant positive correlation to the codon usage bias, suggesting the action of natural selection on these very specific positions of the *Drosophila* genome. The relationship between A_{UG}CAI value and gene length was also examined, and a significant negative relationship was found ($r = -0.15$, $p < 0.0001$), suggesting a general tendency of higher expressivity of shorter genes, and of lower expressivity of longer genes in *D. melanogaster*.

Key words: *Drosophila melanogaster* — Codon usage — AUG context — A_{UG}CAI — CAI — Selection-mutation-drift

Email: K448482@kepco.co.jp

Introduction

In unicellular organisms, such as *Saccharomyces cerevisiae* and *Escherichia coli*, efficiency of translation is highly affected by the choice of synonymous codons, causing a strong bias in codon usage (Sharp and Li 1986; Sharp and Cowe 1991). This form of codon usage bias is called "major codon preference"; codons recognized by the most abundant tRNA tend to be used preferentially, and among codons recognized by the same tRNA, those that optimize the codon-anticodon interaction energy are favored (Ikemura 1982, 1985; Grosjean and Fiers 1982; Percudani et al. 1997; Kanaya et al. 1999). Highly expressed genes exhibit a greater degree of codon usage bias than lowly expressed genes, reflecting the stronger selection constraint on highly expressed genes to optimize translation efficiency by the use of major codons (Bulmer 1988). In some multicellular organisms, such as *Drosophila melanogaster* (Shields et al. 1988; Moriyama and Powell 1997) and *Caenorhabditis elegans* (Duret 2000), the situation for codon usage bias is quite similar to that of unicellular organisms: relative tRNA abundance has a positive relationship with synonymous codon preference.

Synonymous codon usage could affect two different aspects of protein synthesis: accuracy, and amino acid incorporation rate (Kurland 1987; Sørensen et al. 1989). The use of preferred codon enhances the accuracy of translation by reducing the frequency of amino acid misincorporation. Akashi (1994) found that the frequency of preferred codons is significantly higher at the conserved amino acid positions than at

the non-conserved amino acid positions among different *Drosophila* species, indicating that natural selection favors synonymous codon usage to enhance the accuracy of protein synthesis. Major tRNA-encoding codons are also translated faster than their synonymous counterparts, because the waiting time before the arrival of the cognate tRNA is proportional to its abundance (Varenne et al. 1984). In highly expressed genes, natural selection both for speed and accuracy of translation acts more strongly than in lowly expressed genes, and causes more biased codon usage.

In eukaryotic organisms, beside the biased codon usage, there are many factors which affect the translation efficiency. Translation initiation is one of the most important regulating steps of translation, and a large number of proteins, eukaryotic translation initiation factors (eIF), are involved in this process (Kozak 1991; Pain 1996). After the binding of the 40S ribosomal subunit to the 5′ cap-site, the subunit scans the leader until the first AUG codon is encountered, at which point the 60S ribosomal subunit binds to the 40S ribosomal subunit, and the initiation of protein synthesis takes place (Kozak 1984). Translation initiates generally at the most 5′ proximal AUG codon in eukaryotic mRNA, but the context of the AUG is also recognized by the ribosomal subunit as an important signal to trigger a translation initiation event; the potential optimal sequences surrounding AUG initiation codons have been proposed in animals (Kozak 1987), in *Drosophila* (Cavener et al. 1991), in plants (Joshi et al. 1997; Ikeda and Miyasaka 1998), and in yeast (Hamilton et al. 1987), based on the results of compilation analysis of mRNA sequences.

Since both codon usage bias and AUG context affect the translation efficiency, although the action of these two factors on translation is completely different, one can postulate that natural selection might act on the AUG context to optimize translation initiation efficiency, the same way it acts on codon usage. A previous study (Miyasaka 1999) demonstrated that there is a significant positive relationship between these two factors (codon usage bias and AUG context) in *Saccharomyces cerevisiae*, indicating the action of natural selection at the translational level on the AUG context. A similar study showing the significant correlation between codon usage bias and Shine-Dalgarno (SD) sequence conservation in several prokaryotic organisms was also reported (Sakai et al. 2001), although the translation initiation mechanism is quite different from that of eukaryotic organisms.

In this study, 1100 mRNA sequences of *D. melanogaster* were compiled, and the relationship between codon usage bias and translation initiation AUG context was examined.

Materials and Methods

Indices for Codon Usage Bias and AUG Context

The degree of synonymous codon usage bias was measured by the "codon adaptation index" (CAI) (Sharp and Li 1987), and by the "effective number of codons" (ENC) (Wright 1990). The CAI estimates the extent of bias toward codons that are known to be preferred in highly expressed genes. A CAI value is between 0 and 1.0, and a higher value means a stronger codon usage bias. An ENC value ranges from 20, if only one codon is used for each amino acid, to 61, if all synonymous codons are used equally.

An AUG context adaptation index (A_{UG}CAI), devised in the previous study (Miyasaka 1999), was used for the assessment of effective AUG context for translation initiation. An A_{UG}CAI value is between 0 and 1.0, and a higher value means a more effective AUG context for translation initiation.

Data Sets

The data of the AUG context (−20 through −1, +4, +5, and +6 positions; +1 is the position of the translation starting nucleotide), gene length (coding-sequence length), CAI (codon adaptation index), ENC (effective number of codons), and GC3 (percent G+C in codon third position) of 1657 *D. melanogaster* sequences were downloaded through the World Wide Web from the TransTerm database (data source; GenBank version 106) (Dalphin et al. 1998, 1999, Jacobs et al. 2000).

Based on the simulation study of sampling errors reported by Moriyama and Powell (1998), only the genes of 300 bp or longer (n = 1578) were used for the analysis. All the entries of these sequences were retrieved from the GenBank database and checked manually. The genes with incomplete AUG context sequences were excluded from the data, and only the AUG context sequences with complete −20 through +6 nucleotides were included in the analysis. The transposon sequences, the identical genes with different accession numbers were also excluded from the data, and 917 mRNA sequences of *D. melanogaster* were finally selected. When a gene had alternatively spliced gene products, the data of the gene product with a higher CAI value was included in the analysis. In addition to the mRNA sequences, some genomic DNA sequences (n = 183) were also included in the analysis, when there was sufficient data to identify the mRNA sequence. In some genomic DNA sequences (n = 12), the exon/intron border is located within 20 bp upstream from the AUG codon, and the intron sequences are included in the AUG contexts of the TransTerm database; in this case the inadequate intron sequences were replaced manually with the proper exon sequences.

The data on gene expression level based on the ESTs (Expression Sequence Tags) relative abundance were downloaded through the World Wide Web from http://pbil.univ-lyon1.fr/datasets/Duret_Mouchiroud_PNAS_1999/data.html (Duret and Mouchiround 1999).

Statistical Methods

Least-squares linear regression analysis and Fisher's r to z transformation were used to examine the relationship between the factors of codon usage bias, that of AUG context, and gene length.

All compilation and calculation was performed using an Excel 98 program (Microsoft Corp., Redmond, WA, USA). Statistical analysis was done using a StatView v4.5 program (Abacus Concepts, Inc., Berkeley, CA, USA) on a Macintosh computer.

Results

Relationship Between Codon Usage Bias and AUG Context, and Between Gene Length and AUG Context in D. melanogaster

The relationship between codon usage bias, measured with CAI and ENC, and translation initiation AUG context, measured with $A_{UG}CAI$, was examined in 1100 mRNA sequences of *D. melanogaster*. First, for assessing the optimal AUG context in *D. melanogaster*, two different reference sequence sets of the potential highly expressed genes were selected. The sequences of the first reference set are the sequences with high CAI values (CAI > 0.4, $n = 39$), and those of the second set are the manually selected potentially highly expressed genes, such as ribosomal proteins, carbohydrate metabolism enzymes, amino acid metabolism enzymes, histons, and actins ($n = 43$, CAI > 0.4; 19 genes and CAI $\leqq$ 0.4; 24 genes). The AUG contexts (−20 through +6 positions), the CAI and ENC values, and the gene length of the reference genes are shown in Table 1. In the previous study with *Saccharomyces cerevisiae* (Miyasaka 1999), only the AUG contexts −9 through +6 positions were analyzed; the analysis was extended down to the −20 position of AUG context in this study, since the sequence data of AUG context down to the −20 position was provided from the TransTerm database.

The AUG contexts of the two reference gene sets were compiled, and the relative adaptiveness of the nucleotide (*wij*; $i = -20$ through −1, +4, +5, and +6, j = A, C, G, U) for each position was calculated (Table 2). The relative adaptiveness of a nucleotide (*wij*) is the frequency of the use of that nucleotide, at *i*th position, compared to the frequency of the optimal nucleotide at this position.

With these *wij* values, the $A_{UG}CAI$ (AUG context adaptation index) for the optimal AUG context was calculated. The concept of the $A_{UG}CAI$ is similar to that of CAI for codon usage bias, and is calculated as a geometric mean of the *wij* values as:

$$\begin{aligned} A_{UG}CAI = (&w-20j \times w-19j \times w-18j \times w-17j \\ &\times w-16j \times w-15j \times w-14j \times w-13j \\ &\times w-12j \times w-11j \times w-10j \times w-9j \\ &\times -8j \times w-7j \times w-6j \times w-5j \\ &\times w-4j \times w-3j \times w-2j \times w-1j \\ &\times w+4j \times w+5j \times w+6j)^{1/23} \end{aligned}$$

Note that the frequency of the nucleotide U at the −2 and −3 positions of the first reference set, and at the −4 and −3 positions of the second reference set, and C at the −3 position of the first reference set is 0, causing a problem for the calculation of $A_{UG}CAI$: a value of 0.5 is provisionally given to the frequencies of C and U at these positions to overcome this problem.

The $A_{UG}CAI$ values of 1100 mRNA sequences were calculated, and were plotted against the CAI values (Fig. 1). There is a significant ($p < 0.0001$) positive relationship between codon usage bias and AUG context (simple linear correlation coefficient: $r = 0.35$ [$r^2 = 0.12$] and $r = 0.30$ [$r^2 = 0.092$] between CAI and $A_{UG}CAI$ for the first and second reference sets, respectively), suggesting that these two factors are evolutionally under a similar natural selection constraint at the translational level. The $A_{UG}CAI$ and ENC values also showed a significant ($p < 0.0001$) negative relationship ($r = -0.31$ [$r^2 = 0.098$] and $r = -0.28$ [$r^2 = 0.080$] between ENC and $A_{UG}CAI$ for the first and second reference sets, respectively; plot data not shown). Since the results obtained from the two different reference sequence sets are quite similar, it seems likely that these sets of reference sequences can be used as a representative optimal AUG context for the present analysis, although there is a future possibility of minor change in the optimal context in accord with the increase of mRNA sequence data. The *wij* values of the first set of reference sequences ($n = 39$) was used in all subsequent works.

Since *Drosophila* is a multicellular organism, the expressivity of the genes differs from cell to cell. To see how the AUG context and codon usage bias are related to the expressivity of the genes, the $A_{UG}CAI$ value was plotted against the CAI value in the potential highly expressed genes (Fig. 2a) and in the potential lowly expressed genes (Fig. 2b, c). As highly expressed genes, non-tissue-specific highly expressed genes (the 43 sequences of the second reference set, closed circles in Fig. 2a) and the tissue specific highly expressed genes, such as yolk proteins, cuticle proteins, tubulins, muscle myosins, and larval serum proteins ($n = 9$, open circles in Fig. 2a) were selected (total 52 genes). As the lowly expressed genes, the regulatory proteins, such as protein kinases and phosphatase ($n = 67$, Fig. 2b), and transcription factors ($n = 78$, Fig. 2c), were selected: these regulatory genes were chosen as the potential lowly expressed genes because these genes are known to be lowly expressed even in unicellular organisms. As expected the potential highly expressed genes (both tissue-nonspecific and tissue specific genes) showed much higher $A_{UG}CAI$ and CAI values (Fig. 2a) than the lowly expressed genes (Fig. 2b, c). In the lowly expressed genes (protein kinase/phosphatase and transcription factor genes), the CAI values are mostly much lower than those of highly expressed genes, but the $A_{UG}CAI$ values were not always low: many lowly expressed genes have relatively high $A_{UG}CAI$ values (Fig. 2b, c), and as a result the relationship between $A_{UG}CAI$ and CAI values becomes very weak in these

Table 1. (A) AUG contexts of the reference sequences with high CAI values (CAI > 0.4, n = 39)

Accession No.	Description	−20	−19	−18	−17	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	AUG	4	5	6	CAI	ENC	Gene Length (bp)
Y07894	mitochondrial ATP synthase subunit alpha	A	C	A	G	U	U	A	A	A	A	A	G	C	A	U	U	A	A	A	A		U	C	G	0.605	27.5	1659
Y00795	mp20 gene for muscle-specific protein	C	C	A	A	A	A	A	C	C	A	A	G	A	A	U	C	A	A	A	C		U	C	U	0.588	27.8	555
M18280	vitelline membrane protein 26A-1	U	U	C	A	A	C	C	A	C	C	A	A	C	A	A	C	C	A	A	G		A	A	A	0.576	23.8	426
X13107	awd (abnormal wing disc)	U	U	U	C	U	G	U	A	A	U	C	U	C	G	G	C	G	A	C	A		G	C	G	0.571	26.7	462
AF006639	serine protease SER4	C	C	A	C	G	A	U	U	G	A	U	C	C	A	A	U	C	A	C	C		A	A	A	0.564	29.9	798
X06869	elongation factor 1-alpha	C	C	A	U	A	G	U	G	U	A	A	U	C	A	U	C	C	A	A	C		G	G	C	0.555	31.1	1392
X82782	ribosomal protein L7a	A	A	C	U	U	U	A	A	A	A	C	C	G	C	U	U	G	A	A	A		G	U	U	0.555	31.6	816
M77133	laminin receptor (k14)	A	G	G	A	G	G	A	C	G	A	C	A	U	C	A	C	C	A	A	G		U	U	G	0.547	29.1	762
AF016835	ribosomal protein L3	C	A	C	G	U	C	U	G	A	A	G	A	G	A	A	C	A	G	A	C		U	C	U	0.544	32.2	1251
U91524	ferritin subunit 1	A	C	U	A	C	G	U	U	C	G	A	C	G	A	U	C	A	A	A	G		G	U	G	0.534	30.8	618
Y00504	ribosomal protein 21C	A	U	U	A	G	C	C	A	C	A	G	C	A	C	U	U	C	G	A	C		U	C	C	0.533	21.9	339
M11254	glyceraldehyde-3-phosphate dehydrogenase-1	A	A	G	A	C	A	G	C	A	G	G	A	A	C	U	C	A	G	C	C		U	C	G	0.531	31.1	999
D10446	aldolase	A	C	U	A	C	A	C	U	C	G	A	A	U	C	U	C	A	A	A	A		A	C	C	0.523	31.3	1092
L02074	ribosomal protein S6	C	G	U	G	C	A	A	A	C	A	G	A	C	C	G	A	C	A	A	U		A	A	G	0.521	33.0	747
U96491	RACK1	G	C	A	A	A	U	A	A	U	A	U	A	A	A	C	U	C	A	A	G		U	C	C	0.517	29.0	957
X75339	ribosomal protein L18a	U	U	U	U	C	G	C	G	U	G	A	A	C	G	U	C	C	A	A	C		A	G	A	0.505	30.3	534
X57576	triosephosphate isomerase	C	C	G	A	C	A	G	C	A	A	C	A	G	C	A	A	C	A	A	C		A	G	C	0.493	28.4	744
X94613	ribosomal protein L9	C	U	A	A	C	G	G	A	G	C	U	G	U	C	A	U	C	A	A	G		C	G	U	0.479	35.2	573
U59146	vacuolar ATPase subunit A	A	A	A	A	A	A	C	A	G	A	A	U	A	A	A	G	C	A	A	A		U	C	C	0.478	33.1	1845
U01334	ribosomal protein S2	A	A	A	G	C	U	G	A	U	A	G	A	C	C	U	A	C	A	A	A		G	C	G	0.470	33.2	804
M11255	glyceraldehyde-3-phosphate dehydrogenase-2	A	G	A	U	A	A	A	C	A	A	A	A	A	G	U	U	A	A	C	C		U	C	G	0.469	35.2	999
AB003910	88F actin	A	A	A	G	A	U	A	A	A	C	A	A	C	U	G	C	C	A	A	G		U	G	U	0.462	31.7	1131
U25057	myosin light chain	A	U	A	G	C	U	A	A	A	A	A	C	A	A	A	U	C	G	A	U		G	C	A	0.462	28.7	444
U68038	lactate dehydrogenase	A	A	C	A	A	A	A	C	C	A	A	U	A	C	A	A	A	A	C	A		G	C	C	0.462	30.1	999
X12452	Act87E actin	G	A	A	A	A	C	A	G	C	C	A	G	U	A	G	C	C	A	A	G		U	G	U	0.460	32.0	1131
X14247	ribosomal protein S31	A	U	G	A	U	A	A	A	U	A	A	G	C	C	G	C	C	A	A	G		A	C	C	0.453	30.5	345
Y11314	glutamate dehydrogenase	G	A	A	U	C	C	A	A	G	C	G	A	U	U	A	G	C	A	G	C		U	A	U	0.449	30.8	1689
X91853	Minute(2)32A gene	U	C	A	A	A	A	A	C	U	A	G	A	C	U	A	U	C	A	A	G		G	G	U	0.441	31.3	456
X02497	chorion gene s18-1	C	U	A	A	C	C	A	A	C	U	C	A	G	C	C	U	C	A	G	A		A	U	G	0.431	30.9	519
L01498	heat shock protein (Hsc3)	U	U	U	C	C	U	G	A	G	A	G	A	U	C	U	U	C	A	A	U		A	A	G	0.429	34.6	1971
X67839	vacuolar ATPase B subunit	G	C	A	G	A	A	A	A	A	G	U	C	U	A	U	C	C	A	A	G		A	A	C	0.428	33.4	1473
U84756	cuticle protein LCP6	G	C	U	U	U	C	U	A	U	C	C	A	G	C	U	C	C	A	A	C		A	A	A	0.427	29.5	315
X76042	troponin-C	U	A	C	C	C	A	A	A	A	G	U	A	G	C	A	G	C	A	A	C		A	G	C	0.426	25.7	468
V00248	yolk protein I (vitellogenin)	C	C	A	A	A	U	C	C	C	A	A	A	U	C	C	G	A	A	C	C		A	A	C	0.424	31.3	1320
M14645	alpha-tubulin (alpha-3)	A	A	A	C	C	C	C	A	U	U	A	A	A	A	A	U	C	A	A	U		C	G	C	0.413	33.7	1353
M62398	CYP-1 protein	A	A	U	A	U	U	C	G	A	A	A	G	C	G	U	C	A	A	A	G		A	G	U	0.409	34.8	498
X13382	ribosomal protein L1	G	G	G	U	U	U	G	A	A	A	A	U	C	A	A	C	G	A	A	A		A	G	C	0.406	34.0	1224
X04754	yolk polypeptide (YP3)	A	A	U	U	C	C	G	G	A	U	U	U	G	C	A	C	C	A	A	A		A	U	G	0.401	33.2	1263
M25772	DNA repair protein (AP3)	G	U	C	C	C	U	A	A	U	A	C	A	C	A	A	U	U	A	A	A		G	U	U	0.401	31.8	954
	A=%	44	33	46	46	31	33	46	56	36	59	46	54	23	38	41	10	26	90	79	28		41	21	13			
	C=%	23	31	15	15	41	23	21	21	26	15	18	15	36	44	7.7	46	64	0	15	33		5.1	36	33			
	G=%	18	10	13	18	8	15	18	15	15	15	21	15	21	10	13	10	7.7	10	5.1	28		23	28	28			
	U=%	15	26	26	21	21	28	15	7.7	23	10	15	15	21	7.7	38	33	2.6	0	0	10		31	15	26			

Table 1. (B) AUG contexts of the manually selected reference genes (potentially highly expressed genes, $n = 43$)

Accession No.	Description	−20	−19	−18	−17	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	AUG	4	5	6	CAI	ENC	Gene Length (bp)
Y07894	mitochondrial ATP synthase alpha subunit	A	C	A	G	U	U	A	A	A	A	A	G	C	A	U	U	A	A	A	A		U	C	G	0.605	27.5	1659
X82782	ribosomal protein L7a	A	A	C	U	U	U	A	A	A	A	C	C	G	C	U	U	G	A	A	A		G	U	U	0.555	31.6	816
AF016835	ribosomal protein L3	C	A	C	G	U	C	U	G	A	A	G	A	G	A	A	C	A	G	A	C		U	C	U	0.544	32.2	1251
Y00504	ribosomal protein 21C	A	U	U	A	G	C	C	A	C	A	G	C	A	C	U	U	C	G	A	C		U	C	C	0.533	21.9	339
M11254	glyceraldehyde-3-phosphate dehydrogenase	A	A	G	A	C	A	G	C	A	G	G	A	A	C	U	C	A	G	C	C		U	C	G	0.531	31.1	999
D10446	aldolase	A	C	U	A	C	A	C	U	C	G	A	A	U	C	U	C	A	A	A	A		A	C	C	0.523	31.3	1092
L02074	ribosomal protein S6	C	G	U	G	C	A	A	A	C	A	G	A	C	C	G	A	C	A	A	U		A	A	G	0.521	33.0	747
X75339	ribosomal protein L18a	U	U	U	U	C	G	C	G	U	G	A	A	C	G	U	C	C	A	A	C		A	G	A	0.505	30.3	534
X94613	ribosomal protein L9	C	U	A	A	C	G	G	A	G	C	U	G	U	C	A	U	C	A	A	G		C	G	U	0.479	35.2	573
U01334	ribosomal protein S2	A	A	A	G	C	U	G	A	U	A	G	A	C	C	U	A	C	A	A	A		G	C	G	0.470	33.2	804
M11255	glyceraldehyde-3-phosphate dehydrogenase-2	A	G	A	U	A	A	A	C	A	A	A	A	A	A	G	U	U	A	A	C	C		U	C	G	0.469	35.2	999
AB003910	88F actin	A	A	A	G	A	U	A	A	A	C	A	A	C	U	G	C	C	A	A	G		U	G	U	0.462	31.7	1131
U25057	myosin light chain (nonmuscle cytoplasmic)	A	U	A	G	C	U	A	A	A	A	A	C	A	A	A	U	C	G	A	U		G	C	A	0.462	28.7	444
U68038	lactate dehydrogenase	A	A	C	A	A	A	A	C	C	A	A	U	A	C	A	A	A	A	C	A		G	C	C	0.462	30.1	999
X12452	87E actin	G	A	A	A	A	C	A	G	C	C	A	G	U	A	G	C	C	A	A	G		U	G	U	0.460	32.0	1131
X14247	ribosomal protein S31	A	U	G	A	U	A	A	A	U	A	A	G	C	C	G	C	C	A	A	G		A	C	C	0.453	30.5	345
Y11314	glutamate dehydrogenase	G	A	A	U	C	C	A	A	G	C	G	A	U	U	A	G	C	A	G	C		U	A	U	0.449	30.8	1689
X91853	ribosomal protein S13	U	C	A	A	A	A	A	C	U	A	G	A	C	U	A	U	C	A	A	G		G	G	U	0.441	31.3	456
X13382	ribosomal protein L1	G	G	G	U	U	U	G	A	A	A	A	U	C	A	A	C	G	A	A	A		A	G	C	0.406	34.0	1224
L27705	succinate dehydrogenase iron-subunit	C	A	G	C	A	C	A	A	A	C	C	G	C	A	A	C	G	A	A	A		U	U	G	0.392	29.8	894
U48394	ribosomal protein S5	C	G	A	U	U	U	C	U	U	C	U	G	U	G	A	C	A	A	A	C		G	C	C	0.385	37.0	687
X74776	ribosomal protein L19	G	C	C	A	C	G	A	C	G	A	G	G	U	C	G	A	C	A	G	C		A	G	U	0.374	40.1	612
X73153	ribosomal protein S19	A	U	U	C	G	G	C	U	U	U	A	A	C	G	A	G	A	A	A	G		C	C	A	0.368	34.8	471
U15643	ribosomal protein DL11	U	U	U	C	A	C	U	C	C	U	A	C	A	U	A	C	C	A	C	A		G	C	G	0.362	34.0	555
Y11119	ribosomal protein S20	A	C	C	A	G	G	G	A	A	A	U	U	G	C	U	A	A	A	U	A		G	C	U	0.325	32.3	363
U42587	ribosomal protein L22	C	G	U	G	G	U	U	U	U	C	G	A	C	U	G	A	C	A	A	G		G	C	U	0.321	31.3	900
M22428	ubiquitin	G	A	C	C	C	G	C	A	G	A	A	U	A	A	U	C	C	A	A	A		C	A	G	0.308	32.8	696
X53822	histone H3.3Q	A	A	A	A	A	A	A	A	A	G	C	U	A	A	G	A	A	A	A	C		G	C	U	0.280	31.5	411
Y00402	phosphoenolpyruvate carboxykinase	U	A	A	A	A	U	A	C	A	A	C	A	C	A	A	A	A	C	A	A		C	C	U	0.275	38.7	1944
Z19052	ribosomal protein S17	C	U	A	C	C	A	A	C	U	A	G	A	C	U	A	U	C	A	A	G		G	G	U	0.271	35.8	456
Z21673	ribosomal protein 15a	U	U	U	U	G	C	A	A	A	C	A	A	A	U	C	C	A	G	C	C		G	U	G	0.267	34.3	393
U44753	cytochrome P450	G	C	A	G	C	U	A	C	A	G	C	U	A	C	A	C	C	G	C	C		C	U	A	0.266	34.1	1617
X93090	NADPH-cytochrome P450 reductase	A	U	C	G	U	C	A	U	A	C	A	C	G	U	A	C	C	A	C	C		G	C	C	0.265	40.1	2040
X54848	42A actin	C	C	A	A	A	U	A	A	A	A	U	U	U	C	U	A	C	A	A	A		U	G	U	0.245	34.1	303
X52759	glutamine synthase (GS2)	G	U	G	C	A	U	U	G	A	A	A	A	C	A	A	U	C	A	G	U		U	C	C	0.243	38.4	1098
X82257	histone H3.3	A	C	U	U	A	A	U	U	A	C	C	A	G	G	U	A	A	A	C	G		G	C	A	0.235	31.8	411
X97437	histone H4r	C	U	U	U	A	C	G	A	A	G	C	A	A	C	U	G	A	G	A	A		A	C	U	0.234	28.5	312
X67650	sn-glycerol-3-phosphate dehydrogenase	A	U	C	G	A	C	A	A	C	A	A	U	A	A	C	A	A	A	A	U		G	C	G	0.197	41.3	1203
L27653	phosphofructokinase	A	U	A	G	U	G	C	A	A	A	A	G	A	U	U	G	A	A	A	A		A	A	U	0.160	44.3	2364
X07485	histone 2A	G	A	A	A	A	A	U	C	G	A	U	U	G	A	A	C	A	G	A	A		G	C	U	0.146	42.3	426
X52760	glutamine synthase (GS1)	A	C	G	A	G	C	A	G	A	G	C	A	C	G	A	A	C	C	A	G		G	C	A	0.098	49.0	1200
AF034971	ribosomal protein S3a	U	U	C	C	G	G	U	U	U	A	U	U	A	G	U	U	G	A	A	C		G	C	A	0.063	55.0	807
AF030251	ribosomal protein L15	C	U	G	U	C	U	A	U	C	A	A	A	U	U	G	C	A	G	A	G		G	G	G	0.023	46.9	615
	A=%	44	30	40	35	35	26	53	47	49	56	44	47	33	28	42	28	40	77	72	35		19	9.3	16			
	C=%	23	21	21	16	30	26	16	23	19	23	19	12	35	33	4.7	40	51	2.3	19	30		12	58	19			
	G=%	19	12	16	26	16	19	14	12	12	16	23	19	14	16	19	9.3	9.3	21	7.0	26		44	23	26			
	U=%	14	37	23	23	19	30	16	19	21	4.7	14	23	19	23	35	23	0	0	2.3	9.3		26	9.3	40			

Table 2. (A) Frequencies and relative adaptiveness (w) of the nucleotides in the AUG context of the reference genes (CAI > 0.4, n = 39)

	−20	−19	−18	−17	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3[a]	−2[a]	−1 AUG	+4	+5	+6
A	17	13	18	18	12	13	18	22	14	23	18	21	9	15	16	4	10	35	31	11	16	8	5
C	9	12	6	6	16	9	8	8	10	6	7	6	14	17	3	18	25	0	6	13	2	14	13
G	7	4	5	7	3	6	7	6	6	6	8	6	8	4	5	4	3	4	2	11	9	11	11
U	6	10	10	8	8	11	6	3	9	4	6	6	8	3	15	13	1	0	0	4	12	6	10
^{w}A	1.000	1.000	1.000	1.000	0.750	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.643	0.882	1.000	0.222	0.400	1.000	1.000	0.846	1.000	0.571	0.385
^{w}C	0.529	0.923	0.333	0.333	1.000	0.692	0.444	0.364	0.714	0.261	0.389	0.286	1.000	1.000	0.188	1.000	1.000	0.014	0.194	1.000	0.071	1.000	1.000
^{w}G	0.412	0.308	0.278	0.389	0.188	0.462	0.389	0.273	0.429	0.261	0.444	0.286	0.571	0.235	0.313	0.222	0.120	0.114	0.065	0.846	0.563	0.786	0.846
^{w}U	0.353	0.769	0.556	0.444	0.500	0.846	0.333	0.136	0.643	0.174	0.333	0.286	0.571	0.176	0.938	0.722	0.040	0.014	0.016	0.308	0.750	0.429	0.769

[a] For the calculation of w, a value 0.5 is given to the frequency C at the −3 position and U at the –3 and −2 positions to avoid the w value of 0.

(**B**) Frequencies and relative adaptiveness (w) of the nucleotides in the AUG context of the manually selected reference genes (potentially highly expressed genes, n = 43)

	−20	−19	−18	−17	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4[a]	−3[a]	−2	−1 AUG	+4	+5	+6
A	19	13	17	15	15	11	23	20	21	24	19	20	14	12	18	12	17	33	31	15	8	4	7
C	10	9	9	7	13	11	7	10	8	10	8	5	15	14	2	17	22	1	8	13	5	25	8
G	8	5	7	11	7	8	6	5	5	7	10	8	6	7	8	4	4	9	3	11	19	10	11
U	6	16	10	10	8	13	7	8	9	2	6	10	8	10	15	10	0	0	1	4	11	4	17
^{w}A	1.000	0.813	1.000	1.000	1.000	1.846	1.000	1.000	1.000	1.000	1.000	1.000	0.933	0.857	1.000	0.706	0.733	1.000	1.000	1.000	0.421	0.160	0.412
^{w}C	0.526	0.563	0.529	0.467	0.867	0.846	0.304	0.500	0.381	0.417	0.421	0.250	1.000	1.000	0.111	1.000	1.000	0.030	0.258	0.867	0.263	1.000	0.471
^{w}G	0.421	0.313	0.412	0.733	0.467	0.615	0.261	0.250	0.238	0.292	0.526	0.400	0.400	0.500	0.444	0.235	0.182	0.273	0.097	0.733	1.000	0.400	0.647
^{w}U	0.316	1.000	0.588	0.667	0.533	1.000	0.304	0.400	0.429	0.083	0.316	0.500	0.533	0.714	0.833	0.588	0.023	0.015	0.032	0.267	0.579	0.160	1.000

[a] For the calculation of w values, a value of 0.5 is given to the frequency of U at the −4 and −3 positions to avoid the w value of 0.

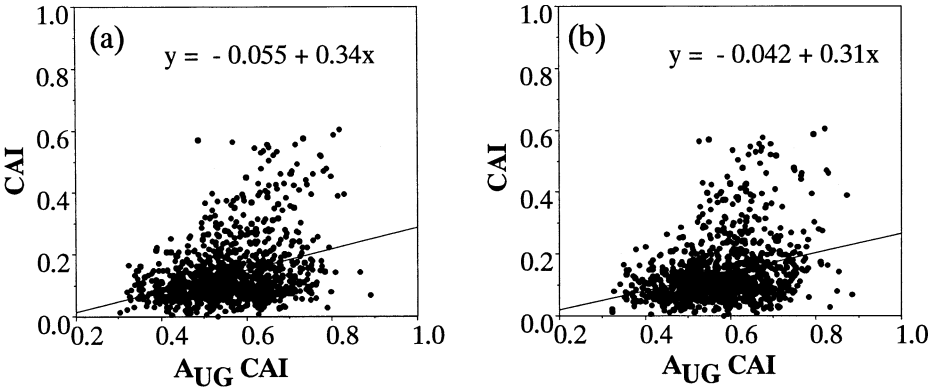

Fig. 1. The relationship between codon usage bias and translation initiation AUG context in 1100 mRNA sequences of *D. melanogaster*. The CAI values were plotted against the A_{UG}CAI values of the AUG context at the −20 through +6 positions. The relative adaptiveness of the nucleotide (*wij*; i = −20 through −1, +4, +5, and +6, j = A, C, G, U) calculated from the reference sequences with high CAI values (CAI ≧ 0.4, n = 39) was used for the calculation of A_{UG}CAI of **a**, and the *wij* values calculated from the manually selected reference sequences of the highly expressed genes (n = 43) were used for the calculation of A_{UG}CAI of **b**. The average and the variance of CAI and A_{UG}CAI values are 0.136 and 0.00981 for CAI, 0.575 and 0.0100 for A_{UG}CAI of **a**, and 0.556 and 0.00948 for A_{UG}CAI of **b**, respectively. Linear regression equations ($y = a + bx$) are also shown.

lowly expressed genes. The high A_{UG}CAI values of many lowly expressed genes can be explained by the fact that the mutation bias in *Drosophila* nuclear genes is toward A/T (Moriyama and Hartl 1993; Kliman and Hey 1994; Powell and Moriyama 1997); mutation-drift acts differently on AUG context from it acts on codon usage bias (for details see Discussion).

Since a significant negative relationship between codon usage bias and gene length has been reported in *D. melanogaster* (Powell and Moriyama 1997), the relationship between A_{UG}CAI value and gene length (bp) was also examined. There is a weak but significant negative relationship between these two factors (Fig. 3a; simple linear correlation coefficient: $r = -0.15$, $r^2 = 0.021$, $p < 0.0001$). The CAI values were also plotted against the gene length (Fig. 3b; simple linear correlation coefficient: $r = -0.26$, $r^2 = 0.066$, $p < 0.0001$). Thus in *D. melanogaster*, in general, there is a tendency that the shorter genes use a more effective AUG context, and show a more biased codon usage than the longer genes.

To observe the effect of the regional GC composition on AUG context, the GC content in 5′ upstream 20 nucleotides of AUG context at the −20 through −1 positions, GC(−20−1) was also calculated. The correlation matrix among A_{UG}CAI, CAI, ENC, gene length, GC3 (percent G+C in codon third position), and GC(−20−1) is shown in Table 3. Interestingly, there is a weak ($r = 0.12$, $r^2 = 0.014$) but a significant ($p < 0.0001$) positive relationship between gene length and GC(−20−1), while the relationship between gene length and GC3 is negative ($r = -0.23$, $r^2 = 0.055$, $p < 0.0001$). This contradictory relationship suggests that some factor other than the regional GC composition in the *Drosophila* genome might be involved in the relationship between gene length and AUG context, and that between gene length and codon usage bias (for details see Discussion).

Relationship Between AUG Context and Gene Expression at Transcriptional Level

Duret and Mouchiround (1999) demonstrated the significant positive relationship between codon usage bias and gene expression at the transcriptional level measured with the ESTs relative abundance. To see if there is also a positive relationship between AUG context and gene expression at the transcriptional level in *D. melanogaster*, the relationship between A_{UG}CAI value and ESTs relative abundance was examined. Of 1100 sample genes in this study, 870 genes found in the data set of the ESTs relative abundance were used for the analysis. There is a significant positive relationship between A_{UG}CAI values and ESTs relative abundance ($r = 0.20$, $r^2 = 0.041$, $p < 0.0001$; plot data not shown), indicating that the translation initiation efficiency and the amount of transcript are positively correlated in *D. melanogaster*. The correlation coefficient between CAI value and ESTs relative abundance was 0.34 ($r^2 = 0.12$, $p < 0.0001$) in the data set of the present study.

Importance of Each Nucleotide Position of AUG Context in Relation to Codon Usage Bias

To see the contribution of each position of the AUG context to the significant correlation between codon usage bias and AUG context, the relationship be-

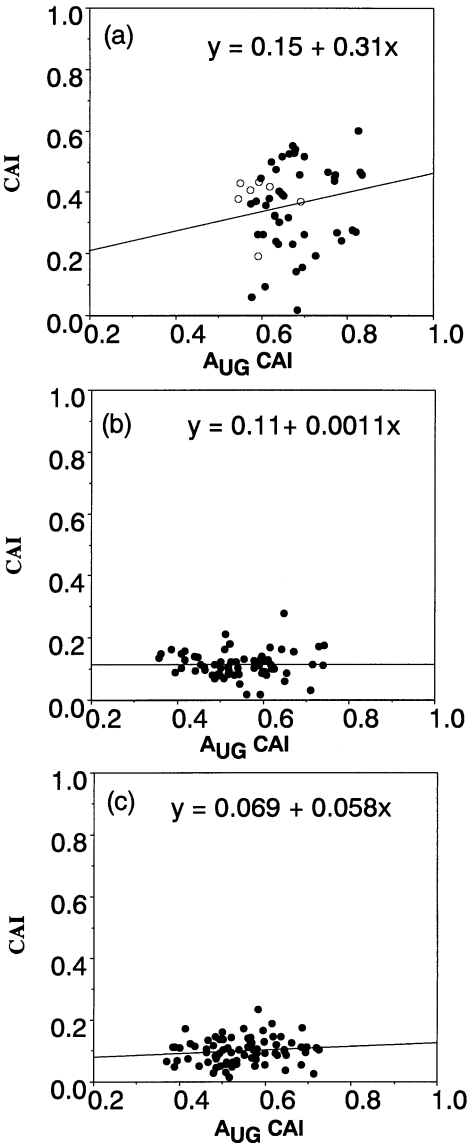

Fig. 2. The A_{UG}CAI and CAI values of the potential highly expressed genes and the potential lowly expressed genes. The CAI values were plotted against the A_{UG}CAI values in the (**a**) potential highly expressed genes ($n = 52$), and the potential lowly expressed genes: (**b**) protein kinase/phosphatase genes ($n = 67$) and (**c**) transcription factor genes ($n = 78$). Linear regression equations ($y = a + bx$) are also shown.

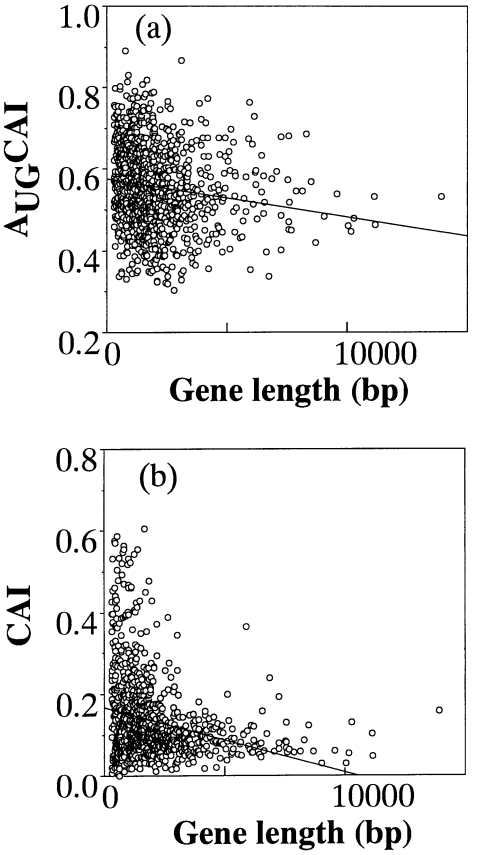

Fig. 3. The relationship between gene length (bp) and A_{UG}CAI and that between gene length and CAI in 1100 mRNA sequences of *D. melanogaster*. The gene length was plotted against the (**a**) A_{UG}CAI and (**b**) CAI values.

tween *wij* values of each position at −20 through +6, and the CAI and ENC values were examined by linear regression analysis and Fisher's *r* to *z* transformation. Figure 4 shows the correlation coefficient (*r*) between *wij* values of each position at −20 through +6, and CAI (Fig. 4a) and ENC (Fig. 4b).

For all the nucleotide positions −20 though +6, the positive correlations between *wij* and CAI values (Fig. 4a), and the negative correlations between *wij* and ENC values (Fig. 4b) were observed. Although the correlation was weak ($r < 0.2$), the *wij* values of the positions of −7, −4, −3, and −2 for CAI, and those of −7, −4, and −2 for ENC showed a significant correlation at the $p > 0.0001$ level, indicating the major contribution of these positions to the significant correlation between codon usage bias and AUG context. The positions of −6 for CAI and −5 for ENC showed a significant correlation at the $p < 0.001$ level, and the positions of −13, −11, and −10 for CAI and those of −12, −6, −3, and −1 for ENC showed a significant correlation at the $p < 0.01$ level. At the −20, −19, −18, −17, −16, −15, −14, −9, −8, +4, +5, and +6 positions, no significant (at the $p < 0.01$ level) correlation was observed between *wij* and codon usage bias indices.

Discussion

Since it is generally accepted that codon usage bias in *D. melanogaster* is under the natural selection constraint at the translational level (Moriyama and Powell 1997), the significant positive relationship between the AUG context and the codon usage bias indicates that the effectiveness of AUG context for translation initiation also reflects the natural selection constraint at the translational level, as well as codon

Table 3. Correlation matrix among CAI, ENC, $A_{UG}CAI$, gene length, GC3, and GC (−20−1)

	$A_{UG}CAI$	CAI	ENC	Gene length	GC3	GC(−20−1)
$A_{UG}CAI$	1	0.35	−0.31	−0.15	0.20	−0.39
CAI	[c]	1	−0.79	−0.26	0.66	−0.073
ENC	[c]	[c]	1	0.35	−0.73	0.049
Gene length	[c]	c	[c]	1	−0.24	0.12
GC3[a]	[c]	c	[c]	[c]	1	0.075
GC(−20−l)[b]	[c]	$p = 0.015$	$p = 0.10$	[c]	$p = 0.013$	1

[a] GC3; percent G + C in codon third position.
[b] GC(−20−l); percent G + C in 20 nucleotides of AUG context (−20 through −1, positions).
[c] $p < 0.001$.

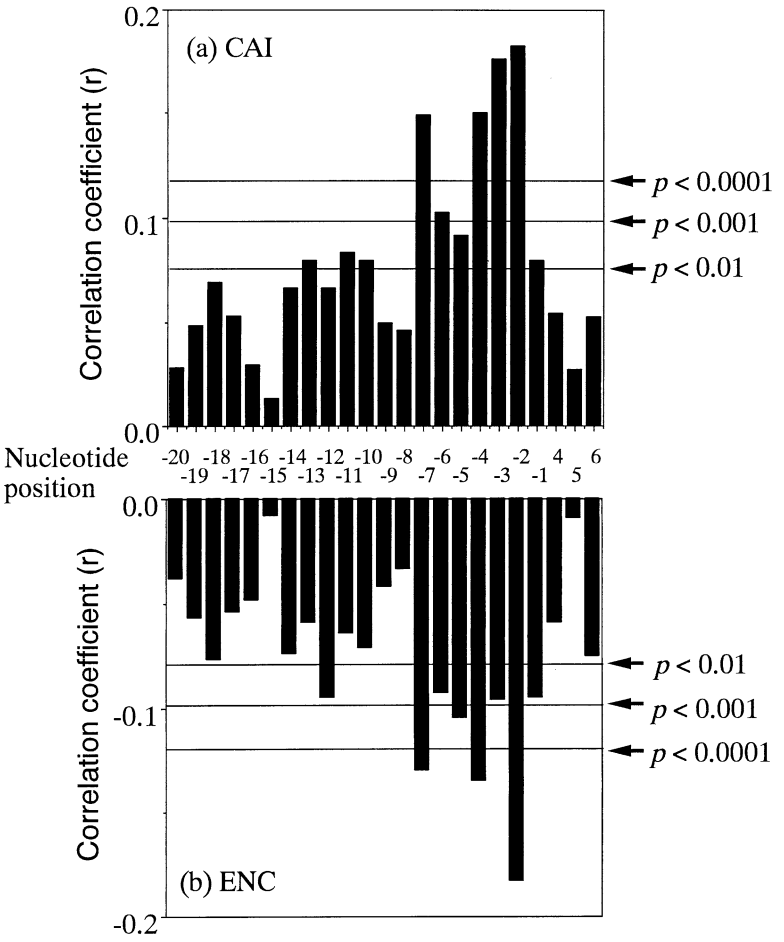

Fig. 4. Fisher's correlation coefficient (r) between wij values of the −20 through +6 positions and (**a**) CAI, (**b**) ENC. The correlation coefficient (r) values for $p < 0.0001$, $p < 0.001$, and $p < 0.01$ are indicated in the figures ($n = 1100$).

usage bias reflects the selection at the translational level. According to the "selection-mutation-drift" model for codon usage bias proposed by Sharp and Li (1986) and also by Bulmer (1988), synonymous codon usage bias simply reflects the balance between selection favoring optimal codons and mutation-drift allowing persistence of non-optimal codons: in highly expressed genes the selection dominates the codon usage bias, while in lowly expressed genes mutation-drift plays some role in determining codon usage (Kliman and Hey 1994). The significant positive relationship between codon usage bias and AUG context, therefore, suggests a possibility that the "selection-mutation-drift" model for codon usage bias can also be applied to the AUG context; namely, in the highly expressed genes the selection favoring the effective AUG context dominantly determines the AUG context, and in the lowly expressed genes there is room for mutation-drift to play some role allowing the persistence of non-effective AUG context.

When the CAI and $A_{UG}CAI$ values of the potential highly expressed genes and those of the potential lowly expressed genes (protein kinases/phosphatases and transcription factors) are compared (Fig. 2), the potential highly expressed genes generally show high CAI and $A_{UG}CAI$ values. On the other hand, in the lowly expressed genes, the codon usage bias is almost always low, but the $A_{UG}CAI$ value shows a greater

variation: many lowly expressed genes have relatively high $A_{UG}CAI$ values. This can be explained by the fact that the direction of mutation is biased toward A/T in *Drosophila* genome; the mutation-drift might act differently on AUG context from the way it acts on codon usage bias. Since the third position nucleotide of all the preferred codons, except for Asp, in *D. melanogaster* is C or G (Moriyama and Powell 1997), A/T-biased mutation-drift, which strongly acts on lowly expressed genes, inevitably lowers the codon usage bias by increasing the use of unfavored A/T-ending codons, resulting in a low codon usage bias in the lowly expressed genes. On the other hand, the same mutation-drift dose not necessarily lower the $A_{UG}CAI$ value, because the preferred nucleotides at the −20 through +6 positions is mainly A or C (see Table 1); this may result in many lowly expressed genes with relatively high $A_{UG}CAI$ values and a greater variation in $A_{UG}CAI$ values of the lowly expressed genes.

The positive relationship between $A_{UG}CAI$ values and relative abundance of ESTs demonstrated in this study is in a good degree of consistency with the previously found positive relationship between codon usage bias and ESTs abundance in *D. melanogaster* (Duret and Mouchiround 1999), suggesting that the transcriptional and translational activities are somehow correlated in *D. melanogaster*, although the correlation is not so strong. In yeast, a quantitative analysis of the proteome (measured with two-dimensional gel electrophoresis), and of the transcriptome (measured with the serial analysis of gene expression: SEGE), revealed a significant positive relationship between protein abundance and transcriptional activity of the corresponding genes (Futcher et al. 1999; Gygi et al. 1999); a similar analysis in the future in *D. melanogaster* might reveal the actual relationship between these two factors.

Generally the 5′ adjacent region of the AUG context is supposed to have an important role in translation initiation, and especially the occurrence of a purine nucleotide at the −3 position is the most important feature of the AUG context in all eukaryotic organisms (Kozak 1991). The correlation coefficient (r) between *wij* and codon usage bias indices (Fig. 4a and 4b) clearly indicates the major contribution of the nucleotides of the 5′ adjacent region (−7 through −1 positions) to the positive relationship between codon usage bias and AUG context, underlining the importance of this region for translation initiation. There are several mutagenesis experiments confirming the importance of this feature (Laz et al. 1987; Yun et al. 1996; Kozak 1997), but of *D. melanogaster* there is only one experimental study which reported on the effects of the AUG context on the translation initiation efficiency. Feng et al. (1991) examined the effects of mutations on the AUG context by introducing point mutations into the −9 to −1 region (GAAGUCACC*AUG*) of the alcohol dehydrogenase (*Adh*) gene of *D. melanogaster*. They constructed two mutants, one containing an A-to-T mutation at the critical −3 position, and the other containing five mutations (mutant AUG context: *C*AA*C*UC*UUUAUG*) designed to give the poorest predicted AUG context. The A-to-T mutation at the −3 position resulted in a 2.4-fold (58%) drop in translation of ADH protein at the adult stage, and the mutant AUG context with five mutations reduced the translation efficiency 12.5-fold (a 92% drop). The significant correlation between *wij* value, and CAI and ENC values at the −3 position, in the present study thus seems to show a good correlation to these experimental results.

In addition to the −7 through −1 region, a region with a relatively high correlation coefficient was observed between the −13 through −10 positions (Fig. 4a and 4b). Since this region is highly A-rich (see Table 2), it seems likely that the effect of this region on the translation initiation might be due to the avoidance of the secondary structure of mRNA, which reduces the translation initiation efficiency by inhibiting the ribosome scanning. The significant correlation between *wij* value and codon usage bias at the −13, −12, −11, −10, −7, −6, −5, −4, −3, −2, and −1 positions suggests the important roles of these positions in the translation initiation event and also the action of natural selection on these very specific positions of the *Drosophila* genome.

The significant negative relationship between $A_{UG}CAI$ value and gene length is in a good degree of consistency with the significant negative relationship between codon usage bias and gene length (higher codon usage bias in shorter genes) in *D. melanogaster* (Moriyama and Powell 1998), suggesting a general tendency of higher expressivity of shorter genes, and of lower expressivity of longer genes at least at the translational level. The significant negative relationship between codon usage bias and gene length was also reported in *C. elegans* (Marais and Duret 2001). In *E. coli* cells, unlike *Drosophila*, the codon usage bias is significantly positively correlated to the gene length (Eyre-Walker 1996), and this positive correlation can be quite reasonably explained with the selection constraint to avoid misincorporation errors during translation. Since the cost of producing a protein is proportional to its length, the selection in favor of codons which increase accuracy should be greater in longer genes than in shorter genes. On the other hand, in *D. melanogaster*, the negative relationship between these two factors is inexplicable with the model of selection on translation accuracy. This negative relationship is also inexplicable from the aspect of stop codon frequency. Since stop codons (TAA, TAG, and TGA) are AT-rich, one can expect

a general tendency of longer genes with a high GC content and shorter genes with a low GC content, if the stop codon frequency is the major determining factor for gene length. In *Drosophila*, however, the situation is quite opposite to this expectation: the relationship between gene length and GC3 (and codon usage bias) is significantly negative.

There are two possible explanations for this inconsistency in the relationship between codon usage bias and gene length in *D. melanogaster*. The first explanation is that the negative relationship between codon usage bias and gene length in *D. melanogaster* reflects a higher expressivity of shorter genes and a lower expressivity of longer genes, and the highly expressed shorter genes tend to use G/C-ending favored codons. The process which generated this tendency is not clear, but one possible explanation is the action of gene shortening selection constraint proposed by Moriyama and Powell (Moriyama and Powell 1997): if shorter proteins can perform similar functions to those of longer proteins, longer proteins become energy-expensive and disadvantageous, thus the selection constraint, which acts to reduce the size of highly expressed genes, dominantly determines the relationship between codon usage bias and gene length.

Another possible explanation proposed for this negative relationship between codon usage bias and gene length in *D. melanogaster* is that this relationship is not caused as a result of the natural selection but simply due to the regional GC composition of the *Drosophila* genome (Marin et al. 1998). The codon usage bias is strongly correlated to the GC content in third codon position (GC3) in *Drosophila* ($r = 0.66$, $p < 0.0001$ between CAI and GC3, and $r = -0.73$, $p < 0.0001$ between ENC and GC3, in the data set of present study), and the same negative relationship between gene length and GC3 has been observed even in mammals (Duret et al. 1995) in which the selection at the translational level is not supposed to work (in other words there is no relationship between codon usage bias and gene expressivity). Thus the negative correlation between GC3 (and also codon usage bias in *Drosophila*) and gene length might be a widespread property among eukaryote genome, no selection force is involved in this relationship, and there is no relationship between gene length and gene expressivity.

To answer this question from the aspect of AUG context, the relationship between gene length and GC3, and that between gene length and GC(−20−1), GC content in AUG context at −20 through −1 positions, were examined in this study. Suppose the negative relationship between codon usage bias (and GC3) and gene length is simply caused by the regional GC composition, and there is no relationship between gene length and gene expressivity, one can expect the same negative correlation between gene length and GC(−20−1). The relationship between gene length and GC(−20−1) shows, however, a weak ($r = 0.12$) but a significant ($p < 0.0001$) positive correlation (Table 3). This result suggests the involvement of some factor other than the regional GC composition in this relationship, and also indicates that there is a negative relationship between gene length and gene expressivity, at least, at the translation initiation level (AUG context level). In addition, since the relationship between codon usage bias and AUG context is reasonably explained with the natural selection at the translational level as shown this study, this result provides circumstantial evidence that the regional G/C content is not the sole determining factor for the negative relationship between gene length and codon usage bias in *D. melanogaster*.

The result of the present study suggests a general tendency of a higher expressivity of shorter genes and a lower expressivity of longer genes, at least at the translational level. Duret and Mouchiroud (1999) also found a negative relationship between codon usage bias and gene length in *D. melanogaster*. They also examined the relationship between gene expression level (estimated with the number of EST sequences) and gene length, but did not find any significant negative correlation between these two factors. There is also no significant correlation between gene length and ESTs relative abundance in the data set of present study (data not shown). These findings suggest at least that there is no evidence of higher transcriptional activity of shorter genes than of longer genes, and a further analysis might be required in the future to examine the relationship between gene expression level and gene length using a more precise expressivity datum estimated from both the transcriptional and the translational levels.

The use of AUG context as a measure, in a combination with codon usage bias, for the study on the natural selection at the translational level seems to have two advantageous features as follows: i) AUG context is much less influenced from the regional GC composition than codon usage bias, which is generally biased toward G/C-ending or A/T-ending codons in many organisms. For example, the nucleotide preference at the most important −3 position of AUG context is A/G (purine nucleotide) in all eukaryotic organisms, and ii) the AUG context is based on the ribosomal subunit scanning model which generally fits any of the eukaryotic organisms, while the "major codon preference" rule (the selection at the translational level on codon usage bias) is not the sole rule controlling the codon usage bias in all eukaryotic organisms; for example, regional GC content also greatly affect the codon usage bias in some organisms (Sharp et al., 1993). Thus the A_{UG}CAI might be applied to any of the eukaryotic organisms as a measure of selection at the translational level; for

example the relationship between gene length and $A_{UG}CAI$, as examined in this study, can be examined in any of the eukaryotic organisms.

In this study, the significant correlation between codon usage bias and the preference for the specific nucleotide at several positions (−13, −12, −11, −10, −7, −6, −5, −4, −3, −2, and −1 positions) of AUG context was shown for the first time in eukaryotic mRNA sequence. This result suggests that natural selection acts on these very specific positions of *Drosophila* genome, and the magnitude of natural selection differs from gene to gene depending on its expressivity. Since the most important feature of the AUG context, the occurrence of a purine nucleotide at the −3 position, is common in all eukaryotic organisms (Kozak 1991), if natural selection acts on this position as suggested in this study, the comparison of the nucleotide variation at the −3 position among the homologous genes from various eukaryotic organisms might possibly be used as a measure for the magnitude of natural selection in the future. For example, if the eukaryotic homologous genes of "gene A" show a more convergent pattern in the nucleotide variation at the −3 position than those of "gene B," it can be assumed that the action of natural selection at the −3 position is stronger in "gene A" than "gene B."

Acknowledgments. I thank Dr. G.W. Clendennen for his editorial revision of this manuscript.

References

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

Bulmer M (1988) Are codon usage patterns in unicellular organisms determined by selection-mutation balance? J Evol Biol 1:15–26

Cavener DR, Ray SC (1991) Eukaryotic start and stop translation sites. Nucleic Acids Res 19:3185–3192

Dalphin ME, Brown CM, Stockwell PA, Tate WP (1998) The translational signal database, TransTerm, is now a relational database. Nucleic Acids Res 26:335–337

Dalphin ME, Stockwell PA, Tate WP, Brown CM (1999) TransTerm, the translational signal database, extended to include lull coding sequences and untranslated regions. Nucleic Acids Res 27:293–294

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, Mouchiroud D (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, Mouchiroud D, Gautier C (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J Mol Evol 40:308–317

Eyre-Walker A (1996) Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol Biol Evol 13:867–872

Feng Y, Gunter LE, Organ ED, Cavener DR (1991) Translation initiation in *Drosophila melanogaster* is reduced by mutations upstream of the AUG initiator codon. Mol Cell Biol 11:2149–2153

Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI (1999) A sampling of the yeast proteome. Mol Cell Biol 19:7357–7368

Grosjean H, Fiers W (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

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19:1720–1730

Hamilton R, Watanabe CK, De Boer HA (1987) Compilation and comparison of the sequence context around the AUG start codons in *Saccharomyces cerevisiae* mRNAs. Nucleic Acids Res 15:3581–3593

Ikeda K, Miyasaka H (1998) Compilation of mRNA sequences surrounding the AUG translation initiation codon in the green alga *Chlamydomonas reinhardtii*. Biosci Biotechnol Biochem 62:2457–2459

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

Jacobs GH, Stockwell PA, Schrieber MJ, Tate WP, Brown CM (2000) Transterm: a database of messenger RNA components and signals. Nucleic Acids Res 28:293–295

Joshi CP, Zhou H, Huang X, Chiang VL (1997) Context sequences of translation initiation codon in plants. Plant Mol Biol 35:993–1001

Kanaya S, Yamada Y, Kudo Y, Ikemura T (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

Kliman RM, Hey J (1994) The effect of mutation and natural selection on codon bias in the genes of *Drosophila*. Genetics 137:1049–1056

Kozak M (1984) Selection on initiation sites by eukaryotic ribosome: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. Nucleic Acids Res 12:3873–3893

Kozak M (1987) An analysis of 5′-noncoding sequences from 699 vertebrates messenger RNAs. Nucleic Acids Res 15:8125–8148

Kozak M (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. J Biol Chem 266:19867–19870

Kozak M (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. EMBO J 16:2482–2492

Kurland CG (1987) Strategies for efficiency and accuracy in gene expression. Trend Biochem Sci 12:126–128

Laz T, Clements JM, Sherman F (1987) The role of messenger RNA sequences and structures in eukaryotic translation. In: Ilan J (ed) Translational regulation of gene expression. Plenum Press, New York, pp. 413–429

Marais G, Duret L (2001) Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J Mol Evol 52:275–280

Marin A, Gonzalez F, Gutierrez G, Oliver JL (1998) Scientific correspondence. Nucleic Acids Res 26:4540

Miyasaka H (1999) The positive relationship between codon usage bias and translation initiation AUG context in *Saccharomyces cerevisiae*. Yeast 15:633–637

Moriyama EN, Hartl DL (1993) Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics 134:847–858

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

Moriyama EN, Powell JR (1998) Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res 26:3188–3193

Pain VM (1996) Initiation of protein synthesis in eukaryotic cells. Eur J Biochem 236:747–771

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

Powell JR, Moriyama EN (1997) Evolution of codon usage bias in *Drosophila*. Proc Natl Acad Sci USA 94:7784–7790

Sakai H, Imamura C, Osada Y, Saito R, Washio T, Tomita M (2001) Correlation between Shine-Dalgarno sequence conservation and codon usage of bacterial genes. J Mol Evol 52:164–170

Sharp PM, Cowe E (1991) Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast 7:657–678

Sharp PM, Li W-H (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28–38

Sharp PM, Li W-H (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential application. Nucleic Acids Res 15:1281–1295

Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans 21:835–841

Shields DC, Sharp PM, Higgins DG, Wnght F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Sørensen MA, Kurland CG, Pedersen S (1989) Codon usage determines translation rate in *Escherichia coli*. J Mol Biol 207:365–377

Varenne S, Buc J, Lloubes R, Lazdunski C (1984) Translation is nonuniform process: effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol 180:549–576

Wright F (1990) The "effective number of codons" used in a gene. Gene 87:23–29

Yun DF, Laz TM, Clements JM, Sherman F (1996) mRNA sequence influencing translation and the selection of AUG initiator codons in the yeast *Saccharomyces cerevisiae*. Mol Microbiol 19:1225–1239