J Mol Evol (2005) 61:306–314
DOI: 10.1007/s00239-004-0271-9

JOURNAL OF MOLECULAR EVOLUTION

Global mRNA Stability Is Not Associated with Levels of Gene Expression in *Drosophila melanogaster* But Shows a Negative Correlation with Codon Bias

Hans K. Stenøien,[1]* Wolfgang Stephan[2]

[1] Plant Ecology/Department of Ecology and Evolution, Evolutionary Biology Centre, Uppsala University, SE-752 36 Uppsala, Sweden
[2] Section of Evolutionary Biology, Biocenter, University of Munich, D-82152 Planegg-Martinsried, Germany

Received: 2 September 2004 / Accepted: 16 March 2005 [*Reviewing Editor:* Dr. Martin Kreitman]

Abstract. A multitude of factors contribute to the regulation of gene expression in living cells. The relationship between codon usage bias and gene expression has been extensively studied, and it has been shown that codon bias may have adaptive significance in many unicellular and multicellular organisms. Given the central role of mRNA in post-transcriptional regulation, we hypothesize that mRNA stability is another important factor associated either with positive or negative regulation of gene expression. We have conducted genome-wide studies of the association between gene expression (measured as transcript abundance in public EST databases), mRNA stability, codon bias, GC content, and gene length in *Drosophila melanogaster*. To remove potential bias of gene length inherently present in EST libraries, gene expression is measured as normalized transcript abundance. It is demonstrated that codon bias and GC content in second codon position are positively associated with transcript abundance. Gene length is negatively associated with transcript abundance. The stability of thermodynamically predicted mRNA secondary structures is not associated with transcript abundance, but there is a negative correlation between mRNA stability and codon bias. This finding does not support the hypothesis that codon bias has evolved as an indirect consequence of selection favoring thermodynamically stable mRNA molecules.

**Present address:* Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.
Correspondence to: Hans K. Stenøien; *email:* stenoien@bio.ntnu.no

Key words: mRNA secondary structure — Mutation — Regulation — Selection — Translation

Introduction

Gene expression in protein-coding genes is the conversion of genetic information encoded in a gene into mRNA and protein, by transcription of a gene into mRNA and the subsequent translation of mRNA to produce a protein. Gene expression must be regulated in order to control what genes will be producing proteins in various cell types and at various developmental stages. There is a stepwise conversion of DNA information into proteins, and regulation may take place in any of these steps. For instance, genes can be regulated by differential transport of spliced mRNAs into cytoplasm or by differential rates of mRNA translation.

There have been considerable efforts to understand the relationship between gene expression and non-random usage of synonymous codons, i.e., codon usage bias. Codon bias is generally thought to be the result of the balance between mutation and weak selection on synonymous codons (Li 1987; Bulmer 1991; Akashi and Eyre-Walker 1998; Duret and Mouchiroud 1999). Differences in the codon bias between genes in the same organism are attributed to mutational bias and the variability in guanine + cytosine (GC) content throughout the genome (Ikemura 1985), partly due to the dispersion of large isochores

homogeneous for GC content (Bernardi et al. 1985). Moreover, in many organisms codon usage in highly expressed genes is shown to be dependent on the relative proportions of isoaccepting tRNAs. Thus, the degree of codon bias for individual genes is associated with the level of translation in a wide range of unicellular (Ikemura 198la, 1985; Gouy and Gautier 1982; Andersson and Kurland 1990; Sharp and Matassi 1994; Kanaya et al. 1999) and multicellular organisms (Fennoy and Bailey-Serres 1993; Akashi 1994; Moriyama and Powell 1997; Akashi and Eyre-Walker 1998; Duret and Mouchiroud 1999; Duret 2000; Kanaya et al. 2001; Musto et al. 2001; Stenøien 2005). In contrast, mutation pressure has been shown to be the dominant factor shaping both codon usage bias and base composition in, e.g., several mammalian genomes (Wolfe et al. 1989; Sharp et al. 1993; Francino and Ochman 1999). The lack of translational selection in mammals and some *Drosophila* species has been explained by their relatively small effective population sizes, meaning that genetic drift will dominate the evolutionary dynamics of mutations that only differ marginally in fitness (Shields et al. 1988; Sharp et al. 1995; Akashi 1997; Jenkins and Holmes 2003). Also other factors are associated with codon usage, including gene length (Comeron et al. 1999; Duret and Mouchiroud 1999; Coghlan and Wolfe 2000), recombination levels across genomes (Kliman and Hey 1993; Hey and Kliman 2002), gene density (Hey and Kliman 2002), and gene structure (Comeron and Kreitman 2002). A functional relationship between mRNA structure stability and codon usage has also been proposed (Antezana and Kreitman 1999; Duan and Antezana 2003).

The degradation of mRNAs is equally important for the regulation of gene expression as the synthesis of mRNAs (Grunberg-Manago 1999; Tinoco and Bustamante 1999; Guhaniyogi and Brewer 2001; Katz and Burge 2003). Thus, the turnover rate of mRNAs may be an important means for regulating gene expression. Experiments have shown that there is considerable variation between and within organisms in mRNA turnover (Brown 1999). For instance, bacterial mRNAs are generally turned over very rapidly, their half-lives being rarely longer than a few minutes. This reflects the rapid changes in protein synthesis patterns that can occur in an actively growing bacterium with a generation time of 20 min or so. Eukaryotic mRNAs, on the other hand, are longer lived, with average half-lives of 10–20 min for yeast and several hours for mammals. Within individual cells, the variation is almost equally striking: some yeast mRNAs have half-lives of only 1 min whereas for others it may be up to 35 min (Tuite 1996; Brown 1999).

There are several factors regulating the rate of mRNA turnover (Grunberg-Manago 1999; Tinoco and Bustamante 1999; Guhaniyogi and Brewer 2001). One possible factor contributing to the rate of mRNA decay is the intrinsic stability of the mRNA molecule, as expressed in the free-folding energies of its secondary and tertiary structures. This might regulate gene expression in one of two ways; either as a positive regulator, in which mRNA stability is associated with longer turnover rates of mRNA molecules within a cell and therefore increased translational rate, or, on the other hand, as a negative regulator, in which high mRNA stability is associated with high translational costs, leading to natural selection for less energetically stable mRNA transcripts of highly expressed genes. Some studies have been conducted on how stability of local mRNA hairpins is associated with levels of gene expression (Carlini et al. 2001; Katz and Burge 2003), but no studies have been conducted aimed at understanding the relationship between global mRNA stability and gene expression.

The main aim of this project is to study the joint effects of putative pre- and post-transcriptional factors, which may be associated with levels of gene expression in the eukaryotic model organism *Drosophila melanogaster*. Of particular interest are the relationships between global mRNA stability, codon bias, and GC content, and their effects on estimated levels of gene expression, measured by transcript abundance. We also study the associations between gene length and the other variables. Gene length is included because a negative correlation between codon bias and gene length has been reported in several multicellular organisms (Moriyama and Powell 1997; Duret and Mouchiroud 1999), and it has proven difficult to find satisfactory explanations for this observation (Akashi 2001). Our results show that codon bias, GC content, and possibly gene length are associated with transcript abundance in this species. Global stability of mRNA molecules is not associated with estimated levels of expression but is negatively associated with codon bias.

Materials and Methods

D. melanogaster sequences were downloaded from GenBank release 130.0 (http://www.ncbi.nlm.nih.gov/) and EMBL release 72 (http://www.ebi.ac.uk/) by using the NCBI Entrez (http://www.ncbi.nlm.nih.gov/Entrez/index.html) and ACNUC (http://pbil.univ-lyon1.fr/search/query_fam.php; Gouy et al. 1985) retrieval systems. Only complete and fully annotated protein-coding sequences (CDS) from nuclear genes, being less than 800 nucleotides long, were used (for reasons discussed below).

Identical or nearly identical sequences were removed from the data set in order to promote independence among data points. A local database was created for the data set with the software package DNATools version 6.0.122 (http://www.seqtools.dk/, S. W. Rasmussen, unpublished data). A Blastn comparison (Altschul et al. 1990) was performed for each sequence on itself and the self-score value was extracted. Then, all sequences were compared to each other with Blastn, and the resulting score values were divided

by the respective self-score values. If these normalized score values were larger than 0.90, then two sequences were classified as similar, and one of them was removed from the data set. Both strands were compared in all of the Blastn runs. The final data sets consisted of 1438 *D. melanogaster* sequences.

Gene expression was estimated as the number of hits when comparing a sequence with publicly available expressed sequence tags (ESTs) for this species (see Duret and Mouchiroud 1999; Marais and Duret 2001; Hey and Kliman 2002; Miyasaka 2002). All available cDNA libraries in release 101102 of the NCBI dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/; Boguski et al. 1993) were used, comprising a total of 256,583 *D. melanogaster* ESTs. Each sequence was compared with the available cDNA libraries using MegaBlast (http://www.ncbi.nlm.nih.gov/blast/). Both strands were compared for each sequence, and segments within query sequences showing low compositional complexity were masked. Blastn alignments showing at least 95% identities were counted as sequence match when the significance level was set to $E = 0.001$. It has been shown that expression levels as estimated by EST databases may not be constant compared to expression levels more reliably estimated by microarrays, but rather systematically biased with gene length (Munoz et al. 2004). To remove this potential bias, we developed a normalized measure of gene expression by dividing the number of EST hits by gene length for each gene tested.

The term mRNA stability has been used to describe the capacity of mRNA molecules to resist degradation. We define mRNA stability more specifically as thermodynamic stability, estimated as the minimum free-folding energy (ΔG) for a predicted RNA secondary structure (Zuker et al. 1999; Mathews et al. 1999). Thermodynamic stability is not necessarily positively correlated with overall degradation resistance (Meister and Tuschl 2004). The Zipfold program (being part of the Mfold server) was used to predict free-folding energies for each sequence, using default parameters (http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form4.cgi). In this approach, every base is first compared to every other base by a type of analysis similar to the dot matrix analysis. Just as a diagonal in a two-sequence comparison, a row of matches in the RNA matrix indicates a succession of complementary nucleotides that can potentially form a double-stranded region. The energy of each predicted structure is estimated by the nearest-neighbor rule by summing the negative base-stacking energies for each pair of bases in double-stranded regions, and by adding the estimated positive energies of destabilizing regions such as loops at the end of hairpins, bulges within hairpins, internal bulges, and other unpaired regions. To evaluate all the different possible configurations and to find the most energetically favorable, several types of scoring matrices are used. The complementary regions are evaluated by a dynamic programming algorithm to predict the most energetically stable molecule, i.e., that with the lowest ΔG (for an overview, see Mount 2001). ΔG is used as a measure of global stability, with decreasing negative values representing increasing mRNA stability. For comparison, we also estimated minimum free-folding energies for the various sequences by using the Vienna algorithm (Hofacker et al. 1994).

Thermodynamically predicted stability measures increase with gene length due to increased number of base pairs and thus increased opportunities for stable structure formation (Pervouchine et al. 2003). This dependency between ΔG and gene length prohibits the joint statistical analysis of these variables because of the underlying assumptions in a multiple regression analysis. Thus, one needs to remove this dependency, and this we have done in two ways. First, we analyzed sequences of approximately the same length separately, i.e., sequences of length 100–199 nt were grouped together, sequences of 200–299 were grouped together etc., comprising a total of eight categories analyzed separately. Second, we calculated a normalized ΔG (termed $N\Delta G$) removing confounding effects of length of studied sequences, and enabling us to analyze the whole data set together. In this second approach, we generated 100 randomizations of the nucleotide sequence order using the algorithm Shuffleseq implemented in EMBOSS (http://ngfn-blast.gbf.de/emboss.html; Rice et al. 2000), yielding a total of 143,800 individual randomized sequences. ΔG was estimated for each randomized sequence using Zipfold, and the average ΔG for all 100 randomized sequences ($A\Delta G$) was calculated for each gene. Normalized ΔG was calculated as $N\Delta G = \Delta G/A\Delta G$ for each gene, implying that $N\Delta G$ is a positive number that increases with increasing stability.

Algorithms based on free-energy minimization perform well for short sequences (Walter et al. 1994), but become unreliable as sequence length increases because of the large amount of possible structures for longer sequences (Konings and Gutell 1995). Therefore, only sequences shorter than 800 nucleotides were used in this study. This could possibly yield low statistical power in correlation analyses of, e.g., codon usage and gene length, since it is known that significant associations are mostly caused by genes with more than 500–600 codons (Comeron et al. 1999; Duret and Mouchiroud 1999).

Codon usage bias was estimated by the frequency of optimal codons (Fop; Ikemura 1981b), i.e., the ratio of preferred to synonymous codons, when "preferred" codons are defined as those found to occur significantly more often in highly than in lowly expressed genes. Favored codons have been identified by multivariate analysis in *D. melanogaster* (Sharp and Lloyd 1993). Fop values were estimated with the CodonW software (http://www.molbiol.ox.ac.uk/cu/, J. Peden, unpublished data). Various measures of GC content, i.e., GC content overall and GC content in first (N1), second (N2), and third (N3) codon positions, were measured with DnaSP version 3.53 (Rozas and Rozas 1999).

Multiple regressions were performed on the data set using either transcript abundance or normalized transcript abundance as the dependent variable, and codon bias, normalized mRNA stability, gene length, and GC content in N1, N2, and N3 positions as independent variables. Both transcript abundance and normalized transcript abundance are not necessarily statistically independent of gene length, the first due to a possible systematic bias in EST databases with gene length (Munoz et al. 2004), and the latter because gene length is used to define this variable (see above). Strictly speaking, then, it is not appropriate to do any statistical association tests between these two measures of gene expression and gene length since all available tests assume independency among studied variables. We have, therefore, also performed regression analyses of the same dependent and independent variables but excluding gene length. Eight multiple regression analyses were performed on the subset of sequences with approximately the same lengths, using transcript abundance as dependent variable, and codon bias, mRNA stability, and GC content in N1, N2, and N3 positions as independent variables. The significance of the regression models were tested with analysis of variance (ANOVA). Kolmogorov-Smirnov and Shapiro-Wilk tests on studentized residuals revealed departures from normality; dependent variables were therefore log-transformed. Multicollinearity was investigated with the tolerance statistic, i.e., the proportion of variability of a given independent variable not explained by its linear relationships with other independent variables in the model. Pearson correlation coefficients and partial correlation coefficients were obtained to further investigate the association among individual variables. All statistical analyses were performed using the SPSS package (Windows version 11.5.1, SPSS Inc., Chicago, IL).

Results

We get essentially the same results when employing mRNA stability measures based on the Mfold and

Table 1. Results of multiple regression analyses

Variables included in model	Standardized Beta coefficients	*T*	*P*
Codon bias (Fop)	0.833	11.757	***
GC content in N3	−0.516	−7.495	***
GC content in N2	0.144	5.641	***
Gene length	−0.108	−4.377	***

Note: Result of the multiple regression analysis using normalized transcript abundance as dependent variable, and normalized mRNA stability, codon bias, GC contents in codon positions 2 and 3, and gene length as independent variables. Only variables included in the regression models are presented. Codon bias and GC content in second codon position are positively associated with estimated transcript abundance. GC content in third codon position and gene length is negatively associated with high transcript abundance.
***$p < 0.001$.

the Vienna algorithms and for simplicity only mRNA stability estimates obtained from the Mfold program are reported below. The degree of linear combination among variables is investigated by the tolerance statistics. Tolerance levels are small for mRNA stability and gene length (<0.06), indicating that multicollinearity exists between mRNA stability and gene length (data not shown). We also observe a strong negative correlation between ΔG and gene length (Pearson correlation coefficient −0.930, $p < 0.001$), We, therefore, did not use mRNA stability as an independent variable but used normalized mRNA stability instead, i.e., a variable not showing multicollinearity to gene length (no significant correlation, $p = 0.541$). The tolerance levels are also low for codon bias and GC content in N3 (0.12), as shown by a strong positive correlation between these variables (see below).

Regression analysis was performed on the whole data set using normalized transcript abundance as a dependent variable (Table 1). The model is significant according to test of analysis of variance (ANOVA $p < 0.001$), and the proportion of variation in transcript abundance explained by the model is $R^2 = 0.166$. Gene length is negatively associated with normalized transcript abundance, while codon bias is positively associated with normalized transcript abundance. Thus, it seems that genes with high Fop values (i.e., high codon bias) have a high level of transcript abundance, as measured by the number of EST hits. Also, GC content in N3 position significantly explains variability in estimated transcript abundance, being negatively associated with transcript abundance levels, while GC content in the N2 position is positively associated with transcript level variability. We performed regression analyses using the same dependent and independent variables but excluding gene length (data not shown). The results were essentially the same as those reported above for codon bias and GC contents in N2 and N3. We also performed regression analysis on the whole data set using non-normalized transcript abundance as a dependent variable (data not shown). The results were essentially the same as for the analysis using normalized transcript abundance as dependent variable, except that while gene length is negatively associated with normalized transcript abundance, it is positively associated with non-normalized transcript abundance.

Table 2 shows correlation coefficients between pairs of variables, as well as partial correlation coefficients after removing variance explained by the other variables (i.e., variables defined as independent variables in the regression analyses above). There is a significant negative correlation between gene length and codon usage bias when controlling for mRNA stability and GC contents in the various codon positions (−0.100, $p < 0.001$). This negative correlation between gene length and codon bias remains significant if only either mRNA stability or GC contents are controlled for in a partial correlation test, and is not found in a simple correlation analysis between the two variables. There is a strong positive correlation between codon bias and GC content in N3 (0.932, $p < 0.001$). Normalized measures of mRNA stability show significant negative correlation with both codon bias and GC contents in N1 and N3. After controlling for the effects of the other variables, normalized mRNA stability is negatively correlated only with codon bias (partial correlation coefficient −0.070, $p = 0.004$), i.e., biased genes tend to encode unstable mRNAs.

We also performed eight multiple regression analyses on subsets of sequences of approximately the same lengths. In this way, we could directly estimate the putative effects of mRNA stability and the other variables on gene expression without any confounding effects of gene lengths. The results were basically the same as those reported above for the whole data set (data not shown). No independent variable could explain variability in sequences in size class 200–299 nt. However, codon bias (Fop) was significantly positively associated with transcript abundance in all other size classes. GC content in N3 was negatively associated with transcript abundance in four of eight analyses, and GC content in N2 was positively associated with transcript abundance in two of the analyses. Normalized mRNA stability and GC content in N1 were negatively associated with transcript abundance in one analysis.

Discussion

Methodology

There are several methodological pitfalls in studies using transcript abundance as a measure of protein

Table 2. Results from Pearson correlation and partial correlation analyses

Correlations/partial correlations	Codon bias	GC content in N1	GC content in N2	GC content in N3	Normalized ΔG	Gene length
Codon bias	—	0.226 (***)	−0.194 (***)	0.932 (***)	−0.147 (***)	NS
GC content in N1	0.278 (***)	—	0.130 (***)	0.148 (***)	−0.084 (***)	0.062 (*)
GC content in N2	−0.214 (***)	0.211 (***)	—	−0.150 (***)	NS	−0.130 (***)
GC content in N3	0.932 (***)	−0.203 (***)	0.140 (***)	—	−0.127 (t*f)	NS
Normalized ΔG	−0.070 (**)	NS	NS	NS	—	NS
Gene length	−0.100 (***)	0.099 (***)	−0.153 (***)	0.100 (***)	NS	—

Note: Pearson correlation coefficients (above diagonal) and partial correlation coefficients (below diagonal) between codon bias (Fop), GC content in the first (N1), second (N2), and third (N3) codon positions, normalized mRNA stability (NΔG), and gene length. Partial correlation coefficients quantify degree of association between pairs of variables after the effects of the other variables are removed. NS: not significant.
*$0.01 < p < 0.05$.
**$0.001 < p < 0.01$.
***$p < 0.001$.

expression levels. First, cell protein concentrations reflect both the rate of synthesis and protein turnover rates. Since assessments of mRNA concentrations do not take into account possible translational regulation, this could potentially cause bias in studies merely based on transcriptional abundances. Duret and Mouchiroud (1999) were the first to use sequence matches to EST libraries in their study of relationships between codon usage and mRNA abundance in *Arabidopsis thaliana*, *D. melanogaster*, and *Caenorhabditis elegans*. As these authors point out, mRNA abundance estimates may in themselves be error-prone because of biases in the tissues sampled, biases in cloning of mRNAs, and the normalization of cDNA libraries prior to sequencing, i.e., adjustment toward uniform concentrations of cDNAs from different genes, causing an underestimation of highly expressed genes.

Nevertheless, broad-scale positive associations between mRNA abundance and codon bias are in concordance with results from experimental studies (Shields et al. 1988; Stenico et al. 1994; Chiapello et al. 1998; Akashi 2001). Moreover, the imprecise estimates of gene expression and underestimation of highly expressed genes yield conservative statistical association tests. Associations between, e.g., codon bias and gene expression may therefore be even stronger than what is estimated in the present study, because at least some *D. melanogaster* cDNA libraries contained in the NCBI dbEST database are normalized. Another source of bias is that expression measures that require mRNA to be maintained for some time (e.g., EST and SAGE) are likely prone to a GC bias, since GC-rich sequences tend to decay slower than AT-rich ones (Margulies et al. 2001). Finally, it has been shown that transcript abundances within EST databases can be systematically positively biased with gene length, at least in *C. elegans* EST databases (Munoz et al. 2004). In the presence of different types of biases, it is appropriate to do multiple regression analyses in studies based on EST matching in order to separate the effects of the various variables and exclude spurious associations. Furthermore, by normalizing the estimates of transcript abundance with gene length one removes possible bias with increasing gene length. Other biases that may influence association tests include genome-wide over- and underrepresentation of short oligonucleotides, as well as possible variance of mutational patterns along the sequence (e.g., Karlin et al. 1998; Gentles and Karlin 2001).

In silico estimation of mRNA secondary structures by Mfold is likely inaccurate, causing uncertainties in the prediction of individual helices (e.g., Doshi et al. 2004; but see Mathews et al. 1999). However, the patterns revealed by statistical analyses of a large data set may nonetheless be correct. Low statistical power due to inaccurate mRNA secondary structure prediction, normalization within EST databases, GC bias, or other biases in the cDNA libraries may explain the relatively low R^2 values in the present regression analyses.

Codon Bias, GC Content, and Transcript Abundance

We find a positive association between codon bias and transcript abundance, similar to what is reported in other studies (Akashi 1994; Moriyama and Powell 1997; Akashi and Eyre-Walker 1998; Duret and Mouchiroud 1999; Duret 2000; Kanaya et al. 2001). This might reflect a mutational bias or coadaptation between codon usage and tRNAs abundance optimizing the efficiency of protein synthesis. Associations between mutation processes and transcription do not seem to explain correlations between transcript abundance and codon usage bias in this species. For instance, the GC content is found to be uniformly higher at silent sites in coding regions than in putatively neutrally evolving introns for *D. melanogaster* (Kliman and Hey 1994; Duret and Mouchiroud 1999). In addition, within alternatively spliced genes, constitutively translated exons show

higher major codon usage than alternatively spliced exons that are transcribed at the same rate but translated at lower levels (Iida and Akashi 2000).

Duret and Mouchiroud (1999) define optimal codons as codons found to occur more often in highly than in lowly expressed genes. In their study of *A. thaliana*, *D. melanogaster*, *C. elegans*, they find that GC content in N3 increases with expression level in genes containing optimal codons. We also find that genes with optimal codons tend to increase GC content in N3 in *D. melanogaster*. More surprisingly, we find that GC content in N3 is negatively associated with gene expression in this latter species. Thus, when controlling for the effect of codon bias, GC content in N3 is negatively associated with transcript abundance variability. This means that when the variability in transcript abundance caused by codon bias is removed from the analyses, then a significant amount of the mRNA transcript variability left is negatively associated with GC content in N3. The most likely explanation for this pattern is the strong multicollinearity between codon bias and GC content in N3 position, causing difficulties in both disentangling the relative effects of the correlated variables and determining the direction of association with the independent variable (Norusis 2000). Codon bias and GC content in second codon position are positively associated with transcript abundance in our study. The energetic cost of amino acids encoded by GC at second codon position has been shown to be much lower than the amino acids encoded by AT at that position in bacteria (Akashi and Gojobori 2002). It seems plausible that highly transcribed genes are encoded by "cheaper" amino acids also in *Drosophila*. Alternatively, proteins of different functional classifications may have different amino acid properties. Thus, selection on amino acid usage for speed and accuracy of protein synthesis (e.g., Akashi 2003) or other protein features could possibly yield specific GC patterns at the second position.

We find a negative correlation between codon bias and gene length in a partial correlation analysis, as previously reported in Moriyama and Powell (1997), Comeron et al. (1999), and Duret and Mouchiroud (1999).

Gene Length and Transcript Abundance

Munoz et al. (2004) found a systematic overrepresentation of mRNA transcripts from long genes in EST databases of *C. elegans*. The same problem could well apply also in *D. melanogaster* EST databases, and we have therefore normalized our transcript abundance measures on gene length to control for this. However, it may still not be appropriate to use gene length as an independent variable in a regression analysis using normalized transcript abundance as a dependent variable, since gene length and normalized transcript abundance are statistically dependent. The results of our regression analyses on normalized and non-normalized transcript abundance, including gene length as an independent variable, may therefore not be statistically valid. It is presently not possible to determine whether the association between gene length and measures of transcript abundance has biological relevance or if it is due to confounding factors. The same problem arises in any association test using nonindependent variables, including correlation and partial correlation tests, or nonparametric tests. Nevertheless, studies of *C. elegans* have found associations between gene length and gene expression more reliably estimated from microarray data (Munoz et al. 2004). We find it likely that gene length is in some way associated with gene expression in *D. melanogaster*, possibly by a negative association, but the precise relationship is difficult to assess in studies based on EST databases.

Duret and Mouchiroud (1999) found that genes with no EST hits tend to encode shorter proteins than genes encountering one or several EST hits in *D. melanogaster*. This is in contrast to our results that long *Drosophila* genes are seemingly less expressed than short genes, an observation also made in *C. elegans* when employing normalized transcription data (Munoz et al. 2004), as well as yeast (Akashi 2003) and humans (Comeron 2004). Nevertheless, the discrepancy between our and Duret and Mouchiroud's (1999) results could be due to the small gene lengths in our compared to their data set.

mRNA Stability and Transcript Abundance

This is the first genome-wide study of the association between the stability of *global* mRNA secondary structures and levels of gene expression. We find that in *D. melanogaster* genes, the normalized mRNA stability, as obtained from measures of free-folding energy of mRNA secondary structure, is not associated with levels of gene expression, i.e., transcript abundance. Thus, it does not seem to be translational selection for global mRNA stability in highly expressed genes, promoting either high or low translation rates. Our results do not rule out the possibility that stability of local mRNA hairpins is associated with gene expression. Even though Katz and Burge (2003) did not find such association in their studies of bacterial genes, Carlini et al. (2001) report a negative association between stability of local mRNA hairpins and overall gene expression for two drosophilid genes.

mRNA Stability and Codon Bias

The stability of mRNA molecules can be increased by means of the energetically more stable GC bindings.

If this is the case, then one should expect a positive correlation between mRNA stability, GC levels in various codon positions, and codon bias, at least in organisms where preferred codons encode G or C in the third codon position. In the present data sets, codons frequently employed mostly encode G or C in N3. However, we find a negative correlation between normalized mRNA stability and both GC content in N3 and codon usage bias. We do not find any correlation between GC content in N3 and mRNA stability in a partial correlation analysis controlling for codon bias and gene length. We do, however, find a significant negative partial correlation between normalized mRNA stability and codon bias. These results are essentially the same as reported by Carlini et al. (2001) for two drosophilid genes, since they also found a negative association between codon bias and local mRNA stability in *Adh* and *Adhr*. Carlini et al. (2001) suggested that this negative association may be explained by considering that not only the total amount of G + C influences mRNA stability, but also the distribution of Gs and Cs, meaning that stability increases if G and C are about equally abundant. Differences in G and C frequencies can be measured as the absolute value of G minus C frequencies. In our data, there is no significant correlation between mRNA stability and G and C frequency differences, or partial correlation between these variables when controlling for GC content and other independent variables (results not shown). Thus, G and C distribution does not per se explain the negative association between mRNA stability and codon bias. Another possible explanation is that Gs and Cs are not frequently occurring in mRNA helixes in codon-biased genes but are rather confined to the unpaired regions of secondary structural elements (e.g., internal loops) and are, therefore, not contributing to stabilize the mRNA molecule. Alternatively, mRNA stability may not be truly associated with GC content in N3, since the negative correlation observed may be an artifact of the strong multicollinearity between codon bias and GC content in N3. This is supported by the lack of association between mRNA stability and GC content in N3 in the partial correlation analysis controlling for codon bias and gene length (Table 2).

Several authors have pointed out that the translation selection hypothesis for explaining codon bias is not completely satisfactory (Antezana and Kreitman 1999; Duan and Antezana 2003). For instance, major codons within degenerate families are almost identical in widely diverged species, an observation difficult to explain by a simple tRNA pool mirroring initially random patterns of codon usage within a codon pool. These authors suggest that mRNA stability and codon usage are equilibrated, so that gains in mRNA stability due to higher codon bias might be selected because secondary structure can affect translatability. Thus, major codons might have functional advantages or disadvantages relative to minor codons related to mRNA stability but unrelated to tRNA abundance. We find a negative association between our measure of mRNA stability and codon bias, and no association between mRNA stability and transcript abundance. This does not support the mRNA stability hypothesis sensu Antezana and Kreitman (1999), which is based on the assumption that codon bias and mRNA stability are positively associated.

mRNA stability could be associated with gene expression even though we do not find any association with estimated transcript abundance. For instance, selection for temporal regulation of gene expression may favor mRNAs that are rapidly turned over. Genes being rapidly induced and shortly afterwards repressed, e.g., stress-induced heat shock proteins (Craig 1986; Morita et al. 2000; Bose et al. 2005), may be selected for internal factors promoting rapid translation (e.g., biased codon usage) and rapid degradation (e.g., low mRNA stability). Alternatively, even though mRNA stability may be unrelated to transcription rate, it may be negatively associated with translational speed and/or accuracy, i.e., being nonadaptive due to high translational costs. Thus, in both cases mRNA stability and codon bias might be relatively independent means to fine-regulate various aspects of gene expression (Carlini et al. 2001).

Acknowledgments. We thank John Parsch, Martin Kreitman, and anonymous reviewers for helpful comments. This work has been partly supported by the Norwegian Research Council grant no. 134800/410, and partly by the Swedish Research Council grant no. 621-2002-5896.

References

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

Akashi H (1997) Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. Gene 205:269–278

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

Akashi H (2003) Translational selection and yeast proteome evolution. Genetics 164:1291–1303

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

Akashi H, Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci USA 99:3696–3700

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Andersson SGE, Kurland CG (1990) Codon preferences in free-living microorganisms. Microbiol Rev 54:198–210

Antezana MA, Kreitman M (1999) The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J Mol Evol 49:36–43

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST: database for "expressed sequence tags." Nat Genet 4:332–333

Bose S, Dutko JA, Zitomer RS (2005) Genetic factors that regulate the attenuation of the general stress response of yeast. Genetics 169:1215–1226

Brown TA (1999) Genomes. John Wiley & Sons, New York

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907

Carlini DB, Chen Y, Stephan W (2001) The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. Genetics 159:623–633

Chiapello H, Fisacek F, Caboche M, Henaut A (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene 209:GC1–GC38

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

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

Comeron JM, Kreitman M (2002) Population, evolutionary and genomic consequences of interference selection. Genetics 161:389–410

Comeron JM, Kreitman M, Aguadé M (1999) Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics 151:239–249

Craig EA (1986) The heat shock response. CRC Crit Rev Biochem 18:239–280

Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics 5:105

Duan J, Antezana MA (2003) Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J Mol Evol 57:694–701

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

Fennoy SL, Bailey-Serres J (1993) Synonymous codon usage in *Zea mays* L. nuclear genes is varied by levels of C and G-ending codons. Nucleic Acids Res 21:5294–5300

Francino HP, Ochman H (1999) Isochores result from mutation not selection. Nature 400:30–31

Gentles AJ, Karlin S (2001) Genome-scale compositional comparisons in eukaryotes. Genome Res 11:540–546

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074

Gouy M, Gautier C, Attimonelli M, Lanave C, di Paola G (1985) ACNUC: portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput Appl Biosci 1:167–172

Grunberg-Manago M (1999) Messenger RNA stability and its role in control of gene expression in bacteria and phages. Annu Rev Genet 33:193–227

Guhaniyogi J, Brewer G (2001) Regulation of mRNA stability in mammalian cells. Gene 265:11–23

Hey J, Kliman RM (2002) Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. Genetics 160:595–608

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. Monatsh Chem 125:167–188

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

Ikemura T (1981a) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146:1–21

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

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

Jenkins GM, Holmes EC (2003) The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 92:1–7

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

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T (2001) Codon usage and tRNA genes in eukaryotes; correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol 53:290–298

Karlin S, Campbell AM, Mrazek J (1998) Comparative DNA analysis across diverse genomes. Annu Rev Genet 32:185–225

Katz L, Burge CB (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res 13:2042–2051

Kliman RM, Hey J (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol Biol Evol 10:1239–1258

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

Konings DA, Gutell RR (1995) A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. RNA 1:559–574

Li WH (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol 24:337–345

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

Margulies EH, Kardia SL, Innis JW (2001) Identification and prevention of a GC content bias in SAGE libraries. Nucleic Acids Res 29:E60

Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911–940

Meister G, Tuschl T (2004) Mechanisms of gene silencing by double-stranded RNA. Nature 431:343–349

Miyasaka H (2002) Translation initiation AUG context varies with codon usage bias and gene length in *Drosophila melanogaster*. J Mol Evol 55:52–64

Morita MT, Kanemori M, Yanagi H, Yura T (2000) Dynamic interplay between antagonistic pathways controlling for the σ32 level in *Escherichia coli*. Proc Natl Acad Sci USA 97:5860–5865

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

Mount DW (2001) Bioinformatics: sequence and genome analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Munoz ET, Bogarad LD, Deem MW (2004) Microarray and EST database estimates of mRNA expression levels differ: the protein length versus expression curve for *C. elegans*. BMC Genomics 5:30

Musto H, Cruveiller S, D'Onofrio G, Romero H, Bernardi G (2001) Translational selection on codon usage in *Xenopus laevis*. Mol Biol Evol 18:1703–1707

Norusis MJ (2000) SPSS 10.0 guide to data analysis. Prentice Hall, Upper Saddle River, NJ

Pervouchine DD, Graber JH, Kasif S (2003) On the normalization of RNA equilibrium free energy to the length of the sequence. Nucleic Acids Res 31:e49

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Sharp PM, Lloyd AT (1993) Codon usage. In: Maroni GP (ed) Atlas of *Drosophila* genes. Oxford University Press, New York, pp 378–397

Sharp PM, Matassi G (1994) Codon usage and genome evolution. Curr Opin Genet Dev 4:851–860

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

Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF (1995) DNA sequence evolution: the sounds of silence. Philos Trans R Soc Lond B 349:241–247

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

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

Stenøien HK (2005) Adaptive basis of codon usage in the haploid moss *Physcomitrella patens*. Heredity 94:87–93

Tinoco I, Bustamante C (1999) How RNA folds. J Mol Biol 293:271–281

Tuite MF (1996) RNA processing: death by decapitation for mRNA. Nature 382:577–579

Walter AE, Turner DH, Kim J, Lyttle MH, Muller P, Mathews DH, Zuker M (1994) Coaxial stacking of helixes enhances binding of oligonucleotides and improves predictions of RNA folding. Proc Natl Acad Sci USA 91:9218–9222

Wolfe K, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Zuker M, Mathews DH, Turner DH (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: Barciszewski J, Clark BFC (eds) RNA biochemistry and biotechnology. NATO ASI Series, Kluwer Academic Publishers, pp 11–43