Gene 424 (2008) 102–107

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

# Determining the relationship of gene expression and global mRNA stability in *Drosophila melanogaster* and *Escherichia coli* using linear models

Sebastian Eck [a,*], Wolfgang Stephan [b]

[a] Helmholtz Center Munich, German Research Center for Environmental Health, Institute of Human Genetics, D-85764, Munich-Neuherberg, Germany
[b] Section of Evolutionary Biology, Biocenter, University of Munich, D-82152, Planegg-Martinsried, Germany

ARTICLE INFO

*Article history:*
Received 1 April 2008
Received in revised form 1 July 2008
Accepted 28 July 2008
Available online 5 August 2008

Received by I. King Jordan

*Keywords:*
Gene expression
mRNA stability
GC content
Codon usage bias
*D. melanogaster*
*E. coli*

ABSTRACT

There are several sequence-dependent factors regulating gene expression. Some of them have been extensively studied, among the most prominent are GC content and codon usage bias. Other factors hypothesized to have an impact on gene expression are gene length and the thermodynamic stability of mRNA secondary structure. In this work, we analyzed two different microarray datasets of *Drosophila melanogaster* gene expression and one dataset of *Escherichia coli*. To investigate the relationship between gene expression, codon usage bias and GC content of first, second and third codon position, gene length and mRNA stability we employed a multiple regression analysis using a comprehensive linear model. It is shown that codon usage bias and GC content of the first, second and third codon position show a significant influence on gene expression, whereas no significant effect of mRNA secondary structure stability is observed.

## 1. Introduction

Gene expression, the conversion of genetic information stored in DNA, through the intermediate messenger RNA, to functional proteins is one of the central processes of all organisms. It is highly regulated at several different stages. For example regulation of gene expression can take place at the DNA sequence level with activators and repressors enabling or repressing expression. Sequence properties like GC content and codon composition also have a major influence on protein abundance as highly expressed genes use a set of preferred optimal codons and thus display a high codon usage bias. It is thought that this effect is a consequence of the optimal codons having high relative proportions of isoaccepting tRNA's. The general belief is that codon usage bias originates from a balance of mutation and weak selection on synonymous codons and that optimal codons help to achieve faster translation rates and higher accuracy. As a result, translational selection is expected to be stronger in highly expressed genes (Akashi, 1994).

When a gene is transcribed and processed into mature, single stranded mRNA it has the ability of adopting a unique secondary structure through forming Watson–Crick base pairs (Watson and Crick, 1953). This leads to several recognizable secondary structure elements like hairpin loops, bulges and internal loops, with even more complex arrangements like pseudoknots possible. The secondary structure of RNA molecules can be predicted computationally by calculating the minimum free energy structure for all different combinations of hydrogen bondings and domains using the well known algorithms of the Vienna RNA package (Hofacker, 2003), RNAfold (Hofacker et al., 1994) and RNAalifold (Hofacker et al., 2002), and it can also be validated experimentally (Parsch et al., 1997; Chen et al., 2003). The mRNA secondary structure predicted by these programs can be described by its thermodynamic stability, and the fact that compact structures take more energy to unfold may play a role in regulating the expression of genes. However, the effect of RNA secondary structure on gene expression is generally not very well understood.

With the growing amount of expression data that are available from microarray experiments, genome-wide studies of gene expression are now possible. Previous approaches showed a strong correlation between codon usage bias and gene expression (Akashi, 1994; Moriyama et al., 1997; Akashi et al., 1998; Duret et al., 1999; Duret, 2000; Kanaya et al., 2001; Stenøien et al., 2005). Yet the relationship between the thermodynamic stability of the mRNA secondary structure of a gene and its expression remains controversial. Carlini et al. (2001) hypothesized that the stability of secondary structural elements (hairpins) has a negative influence on gene expression by analyzing two related *drosophilid genes*. Jia and Li

*Abbreviations:* $F_{OP}$, frequency of optimal codons; EST, expressed sequence tag.
[*] Corresponding author.
*E-mail address:* Sebastian.Eck@helmholtz-muenchen.de (S. Eck).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.07.033

(2005) reached a similar conclusion based on a study of microarray data from *Escherichia coli*. They estimated the folding free energies by applying RNAfold to short sequences (50 nucleotides) in a sliding window fashion. In contrast, Stenøien and Stephan (2005) found no association between global mRNA stability and gene expression in a *Drosophila melanogaster* dataset. In the latter study, gene expression was measured as transcript abundance in EST databases, and global mRNA stability was estimated by applying RNAfold to complete-length mRNAs. To resolve some of these controversies, we follow the general approach of Stenøien and Stephan (2005) and investigate here the possible effects of global mRNA secondary structure on *D. melanogaster* and *E. coli* gene expression. However, we employ a different statistical method (multiple regression analysis using linear models) that we apply to both *D. melanogaster* and *E. coli* data, and we use microarray data (instead of EST hits) as well as an improved prediction algorithm of RNA secondary structures based on multiple sequence alignments. At the same time, we re-visit other sequence-dependent influences on gene expression, such as sequence length, codon usage bias, and GC content.

## 2. Materials and methods

### 2.1. Datasets

*D. melanogaster* sequences were downloaded from Michael Eisen's lab at the Lawrence Berkeley National Lab (LBNL) and the University of California at Berkeley (UCB) (http://rana.lbl.gov/drosophila/wiki/index.php/Datasets). The coding gene alignments were produced using T-COFFEE (Notredame et al., 2000).

For this analysis the alignments of *D. melanogaster* genes with its five relatives *D. simulans, D. sechelia, D. yakuba, D. erecta* and *D. ananassae* were downloaded. This yielded 12,300 multiple sequence alignments ranging in length from 96 to 15,966 nucleotides, which were organized in a SQL database. The complete protocol for building the alignments can by found at (http://rana.lbl.gov/drosophila/wiki/index.php/Datasets).

To remove potential bias of gene length inherently present in EST libraries, gene expression data is based on microarray experiments and is measured as normalized transcript abundance. The first gene expression dataset used was generated by Gibson et al. (2004). It consisted of expression values for 11,604 *D. melanogaster* genes measured relatively to the genome-wide average gene expression. The expression values for male, female and both sexes were obtained using Aligent microarrays (Gibson et al., 2004). These data were also organized in a SQL database and 6137 genes were identified having both a multiple sequence alignment as well as expression data available by matching their flybase identifiers (www.flybase.org).

To study the influence of thermodynamic stability we decided to remove the probably strong influence of sex bias from the dataset (Hambuch et al., 2005). Therefore only genes with a ratio of male to female expression ranging from 0.5 to 2.0 were selected. Furthermore, all sequences having a poor alignment, i.e. low quality alignment over at least 10% of the sequence length or gaps over at least 30% of the sequence length were removed. This produced the final dataset of 3389 unbiased genes with reasonable multiple alignment quality. In the following, this dataset will be referred to as dataset 1.

The second dataset for *D. melanogaster* gene expression was generated by Hutter et al. (2008). The platform used was a genome-wide *D. melanogaster* microarray obtained from the Drosophila Genomic Research Center (DGRC, Bloomington, Indiana, USA). The microarray chip is known as DGRC-1 and consists of 11,895 unique genes, the equivalent of 88% of the genome (based on genome annotation 4.1). The experiment provided 9131 expression values relative to reference gene *Actin5*. From these, 5660 were identified as having a reliable multiple sequence alignment. Note that, in contrast to the expression data of Gibson et al. (2004), only male flies were used, thus no further streamlining of the dataset to reduce possible sex bias was necessary. This dataset will be referred to as dataset 2.

*E. coli* coding sequences and corresponding mRNA expression levels were obtained from the ASAP database (Glasner et al., 2003). *E. coli* K-12 MG1655 (GenBank Accession No. U00096 (Blattner et al., 1997)) was chosen for the analysis and the 4212 mRNA sequences were downloaded. For this wild-type *E. coli* strain five separate microarray experiments under standard growth conditions using an Affymetrix microarray chip were available. Each mRNA with non-zero expression values for all five experiments was selected for further analysis, and the expression values were averaged over the five experiments yielding the final *E. coli* dataset consisting of 4058 mRNA sequences and their corresponding expression values. This dataset will be referred to as dataset 3 in the following analysis. We have also performed studies using the five datasets separately, but this did not produce different results (data not shown).

### 2.2. RNAfold and RNAalifold

This work focuses on the influence of RNA stability on gene expression. The stability of a RNA sequence is defined as the folding free energy of the predicted RNA secondary structure (Zucker et al., 1999; Mathews et al., 1999). Several methods are available to predict RNA secondary structures (Zucker et al., 1999; Parsch et al., 2000; Hofacker et al., 1994; 2002). Here, these predictions were performed using the RNAfold and RNAalifold programs of the Vienna RNA package (Hofacker, 2003) (http://www.tbi.univie.ac.at/ivo/RNA/).

These methods minimize the folding free energy of the molecule. The RNAalifold program extends the basic RNAfold algorithm by expanding the prediction with information from the multiple sequence alignment by introducing a distance measure $\partial$ for the columns of the alignment. Assuming a correct alignment, columns with different nucleotides, indicating compensatory substitutions ("covariations"), are rewarded while sequence inconsistencies (gaps) are penalized.

### 2.3. Regression analysis

To assess the influence of different factors on gene expression we employed a multiple regression analysis to construct a comprehensive linear model (Fahrmeir et al., 2004). The standard model of linear regression with multiple influence variables was used.

$$Y_i = \alpha_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, i = 1, \ldots, n. \tag{1}$$

with

$Y_1,\ldots,Y_n$ expression data

$X_{1j},\ldots,X_{nj}$ deterministic values, e.g. global mRNA stability, sequence length and GC content

$\varepsilon_{1,\ldots,}\varepsilon_n$ unobserved random variable, error or random component, identically distributed with $E(\varepsilon_i)=0$ and $Var(\varepsilon_i)=\sigma^2$.

The regression coefficients $\beta_0,\ldots, \beta_p$ and the error variance $\sigma^2$ are unknown parameters, which are estimated from the data $(Y_i,x_{i1},\ldots,x_{ip})$, $i=1,\ldots,n$.

After computing the model we incorporated a step of model diagnosis. This includes statistical means to determine whether the assumptions of the standard model are– at least approximately– met or if significant discrepancies occur. In addition to formal tests (e.g. the Shapiro–Wilk test and Kolmogorov–Smirnov test for normality), mainly graphic model diagnosis based on residual analysis was used. On the residual plot the residuals $\varepsilon_i$, i.e. the difference of the real datapoints from the model estimations, are plotted. They should ideally show no systematic pattern (homoscedasticity). However,