Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 327 (2005) 4–7
BBRC
www.elsevier.com/locate/ybbrc

Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*

Regina M. Goetz[a], Anders Fuglsang[b,c,*]

[a] *Structural Biology Program, Skirball Institute of Biomolecular Medicine, New York University School of Medicine, USA*
[b] *Danish University of Pharmaceutical Sciences, 2 Universitetsparken, DK-2100 Copenhagen Ø, Denmark*
[c] *TPR-Group ApS, 3 Puggaardsgade 1th, DK-1573 Copenhagen V, Denmark*

Received 7 November 2004
Available online 8 December 2004

Abstract

Although codon usage is often represented by a 61-dimensional vector, the ability of determining the codon bias in a gene relies on a uni-dimensional vector which measures the total bias in usage of synonymous codons. Codon usage is receiving more and more focus because codon biases might be valuable tools to predict and optimize gene/protein expression. How good any of these measures is for correlating codon usage with gene and protein expression has yet to be investigated. In this study, we correlated gene transcript levels in *Escherichia coli* with codon usage, using a number of different codon bias measures. We found that there is a significant correlation between transcript levels and codon bias measures, suggesting that these measures can be used to assess or predict gene expression. The codon bias measure performing best in this context was the codon adaptation index.

Keywords: Codon usage; Bias calculation; Transcriptome; *E. coli*

In the early 1980s, Grantham et al. [1] developed the 'genome hypothesis,' stating that the usage of codons is dependent on a phenomenon that varies from organism to organism. Shortly afterwards, Gouy and Gautier [2] demonstrated in a study using *Escherichia coli* that this phenomenon is, at least in part, related to gene expressivity. Ikemura's studies then significantly advanced the understanding of the phenomenon of biased codon usage. These studies established that codon usage is correlated with tRNA levels in *E. coli*. Ikemura's [3,4] achievements contributed greatly to the understanding of the phenomenon by reporting that codon usage in highly expressed genes implies preferential usage of the codons for which the tRNA levels are high. Soon after, the practical consequences of these findings became obvious as it was reported that heterologous expression could be severely impaired when the introduced gene contained low-usage codons such as AGA or AGG (arginine), while the problem could be compensated by exchanging rare codons for more frequently used synonymous codons or by providing additional copies of the corresponding rare tRNAs (good examples are given in [5,6]). Thus, a very important question, and one with enormous practical potential, is if we can describe the relationship between protein levels and codon usage in quantitative terms. Ultimately, because the process from gene to protein involves an intermediary RNA step, the protein yield of a gene is dependent on many additional factors, including nutritional status of the producing organism, promotor strength and activity, interaction between the 16s rRNA and the region upstream of start codons (Shine–Dalgarno sequence), etc. On this basis, we realize that a deeper understanding of the quantitative relationship

* Corresponding author. Fax: +45 35306020.
E-mail address: anfu@dfuni.dk (A. Fuglsang).

0006-291X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2004.11.134

between codon usage and protein levels must include studies on the relationship between codon usage and mRNA levels, and subsequently also the relationship between mRNA levels and protein levels. In this brief work, we have focused on the first aspect.

Materials and methods

Choice of dataset. The *E. coli* expression data of Bernstein et al. [7] were chosen in particular for three different reasons:

1. *E. coli* is a highly relevant organism, being one of the most widely used hosts for homologous and heterologous expression of peptides and proteins.
2. We need expression data (mRNA levels) for as many members of the transcriptome as possible. Bernstein et al. have essentially quantified the mRNA abundance for several thousands of *E. coli* genes.
3. We need the data to be available in a format that allows easy extraction to custom-designed bioinformatic computer programs for use in conjunction with codon bias measurements.

Furthermore, the data of Bernstein et al. contain expression data for both a rich medium (LB) as well as a defined minimal medium (M9). It is well known that the expression of many genes is heavily influenced by the chemical environment; the *lac* operon is a classical example, where the lactose utilization genes are activated only when lactose is present and glucose is absent. Thus, any difference in correlation observed between the rich medium and the minimal medium gives potentially important information regarding the general applicability of the estimators.

Classical codon bias measures. The most widely used estimators of codon bias are Wright's 'effective number of codons' ($\hat{N}c$, [8]), the Sharp and Li's codon adaptation index (CAI, [9]), and Ikemura's 'frequency of optimal codons' (F_{op}, [3]). These three as well as two recent variations of $\hat{N}c$ are all included in this study and are briefly introduced in the following.

Wright's method is a species-independent method, which assigns a number between 20 and 61 to a gene, estimating the degree by which the entire genetic code is used in a gene. If all codons of the genetic code are used equally, $\hat{N}c$ will assume a value close to 61. If there is an extreme restriction on the usage of codons in a particular gene, i.e., when just one codon is used for each of the 20 amino acids, $\hat{N}c$ will have a value close to 20. That way, $\hat{N}c$ has no intrinsic relation to expression levels (but may in practice be well correlated with expressivity, see later). Recently, an improved estimator based on the concepts of Wright was proposed in this journal ($\hat{N}c^{*}$ [10]). This led Marashi and Najafabadi [11] to question certain aspects of this method, leading to another estimator ($\hat{N}c^{**}$, see elsewhere in this issue of Biochemical and Biophysical Research Communications).

The CAI is, in contrast to Wright's method, a species-dependent codon bias measure, and has been used as an empirical measure for gene expressivity in studies investigating mutational and selectional components of codon usage. With this methodology, a set of genes known to be highly expressed are characterized with respect to their codon usage. To each codon a 'weight' is calculated as each codon's abundance (raw count) divided by the abundance of the most common synonym, and the codon adaptation index for a given gene is then given by

$$\text{CAI} = \left(\prod_{i=1}^{L} W_i \right)^{1/L},$$

where L is the number of codons from synonymous families in the gene.

Ikemura's F_{op} is somewhat related to the CAI in that it evaluates the fraction of optimal codons, i.e., the fraction of codons that presumably have a good Watson–Crick basepairing between mRNA and tRNA.

Novel codon bias measures. Knowledge about the translational selection on codon usage in a species such as *E. coli* can be applied to heterologous gene expression; that is, codon usage can be optimized in foreign genes in order to improve expression in the host cell, particularly when those genes contain rare codons in amounts causing ribosomal stalling. Codon optimization is typically accomplished by generating artificial genes and exchanging rare codons for frequently used ones. Alternatively, rare tRNAs are co-expressed at increased levels. Rare codons that can become rate limiting due to very low expression levels of corresponding tRNAs are AGA/AGG (coding for arginine), ATA (coding for isoleucine), CTA (coding for leucine), CCC (coding for proline), and GGA (coding for glycine). AGA, ATA, and CTA are particularly found in genes from AT-rich genomes. *E. coli* strains bearing additional copies for some or all of those tRNAs are commercially available and allow codon optimization without the need for generating artificial genes.

Because of the apparent impact of these few codon families on gene expressivity, we thought that codon usage measures emphasizing these families might represent better estimators of gene expressivity. We modified existing codon bias measures to that extent and introduced two new codon adaptation indices termed CAI_{RIL} and $\text{CAI}_{\text{RILPG}}$, which are calculated the same way as mentioned before but with the modification that only codons for arginine + isoleucine + leucine are included in the case of CAI_{RIL}, and for the case of $\text{CAI}_{\text{RILPG}}$ proline + glycine codons are also included. $\overline{W}_{\text{RILPG}}$ is defined as the average weight of arginine + isoleucine + leucine + proline + glycine codons in a gene. f_{nop} is the frequency of non-optimal codons, i.e., the number of AGG/AGA/ATA/CTA/CCC/GGA codons divided by the total number of codons. f_{nop} is of course inspired by Ikemura's f_{op} however, because of the way these two bias parameters are calculated it should be emphasized that $f_{\text{nop}} \neq 1 - f_{\text{op}}$. In addition, several studies have demonstrated that the negative effect of rare codons on gene expression is more pronounced when those codons occur in clusters, that is, right next to each other in a given gene [12–14]. Therefore, we introduced another codon bias measure, fcl, which is the frequency of non-optimal codon clusters.

Correlation analysis. The different codon bias measures were correlated with mRNA levels using Spearman's rank correlation analysis. This non-parametric correlation analysis was used, since there is no reason to assume linearity between the bias measures and the mRNA levels. P values corresponding to less than 5% chance were considered statistically significant.

Results and discussion

The results are listed in Table 1. First and foremost, it is clear that CAI gives the best correlation with mRNA levels for both the LB medium and the M9 medium. The principal conclusion of this work is therefore that among the different codon bias measures CAI is the one that has the best correlation with mRNA levels in *E. coli*. Second, in the M9 medium the correlation is in all cases worse than in the LB medium. CAI (and the other codon bias measures) are thus better correlated with mRNA levels under conditions of an optimized nutritional environment, at least in *E. coli*. It would be very relevant to repeat this kind of study with other organisms in order to examine if this is a general princi-

Table 1
Correlation of mRNA levels with codon bias measures[a]

Bias measure	Correlation with mRNA levels in LB medium	Correlation with mRNA levels in M9 medium
$\hat{N}c$	$r_s = -0.2198$	$r_s = -0.0786$[b]
$\hat{N}c^*$	$r_s = -0.2440$	$r_s = -0.0981$[c]
$\hat{N}c^{**}$	$r_s = -0.3900$	$r_s = -0.2721$
CAI	$r_s = 0.4303$	$r_s = 0.2800$
f_{op}	$r_s = 0.3631$	$r_s = 0.2192$
f_{cl}	$r_s = -0.1702$	$r_s = -0.1218$
CAI_{RIL}	$r_s = 0.3796$	$r_s = 0.2087$
CAI_{RILPG}	$r_s = 0.3866$	$r_s = 0.2277$
f_{nop}	$r_s = -0.3300$	$r_s = -0.2315$
$\overline{W}_{RILPG}$	$r_s = 0.3677$	$r_s = 0.2133$

[a] $P < 0.0001$ unless otherwise stated.
[b] $P < 0.01$.
[c] $P < 0.001$.

ple. We know that the transcription of many genes is influenced by the chemical composition of the medium. There is no general rule about the change in mRNA levels (up or down) for an *E. coli* on restricted diet, as mRNA levels are regulated primarily through promotor activities, and these activities in turn are regulated in a complex way. Classical examples that emphasize the complexity of the subject are the nutritionally regulated operons like the his operon or lac operon. Basically, a variation in the chemical environment may change the transcription of one or more genes, while it does not affect the codon usage of any gene. The effects of codon usage therefore must be anticipated to be stronger when a gene is experiencing full promotor activity. To put it differently: the limiting potential of rare codons on translation becomes negligible when the mRNA level approaches zero, therefore the codon effects are thought to be most clearly visible when the promoter activity is maximal.

Although the effect of rare codon clusters has been documented as having an exceptionally strong influence on translation, the codon bias parameter f_{cl} has the lowest correlation coefficient in LB. Likewise, weak correlations were obtained for Wright's codon bias measure, $\hat{N}c$, and a modification of it, $\hat{N}c^*$. These therefore are probably not the ones that pave the way forward in an attempt to improve the predictiveness value of codon biases. A surprising aspect of the data in Table 1 is that $\hat{N}c^{**}$ correlates almost as well as CAI with transcript levels, both in LB and in M9. After all, CAI-like Ikemura's F_{op} is a species-dependent bias measure, i.e., its calculation depends on knowledge of codons in putatively highly expressed genes, while the three variants of the effective number of codons are species-independent. A priori we would anticipate some correlation of CAI and $\hat{N}c$, $\hat{N}c^*$, and $\hat{N}c^{**}$, since a gene having a CAI-value of 1.0 uses only one codon for each of the 18 amino acids having synonymous codons, so there will be 20 codons effectively used. We can see here that the correlation coefficients obtained with $\hat{N}c^*$ and $\hat{N}c$ are not nearly as good as that obtained with $\hat{N}c^{**}$. The reason for this observation is at present unclear, but it might be worth the effort to look more into it. None but a few studies have investigated the relationship between codon usage and gene expressivity in *E. coli*. Using multivariate statistics, dos Reis et al. [15] found a positive correlation between mRNA levels and the codon bias measure CAI. Although this is in agreement with the results of our study, differences in the study design preclude further comparisons. Our study indicates that codon bias measures are fair to good predictors of gene expressivity in *E. coli*. The CAI appears to be the most predictive of all codon bias measures in this context, both in Lb and M9.

References

[1] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pave, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1980) 49–62.
[2] M. Gouy, C. Gautier, Codon usage in bacteria: correlation with gene expressivity, Nucleic Acids Res. 10 (1982) 7055–7074.
[3] T. Ikemura, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes, J. Mol. Biol. 146 (1981) 1–21.
[4] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 13–34.
[5] U. Brinkmann, R.E. Mattes, P. Buckel, High-level expression of recombinant genes in *Escherichia coli* is dependent on the availability of the dnaY gene product, Gene 85 (1989) 109–114.
[6] R. Seetharam, R.A. Heeren, E.Y. Wong, S.R. Braford, B.K. Klein, S. Aykent, C.E. Kotts, K.J. Mathis, B.F. Bishop, M.J. Jennings, Mistranslation in IGF-1 during over-expression of the protein in *Escherichia coli* using a synthetic gene containing low frequency codons, Biochem. Biophys. Res. Commun. 155 (1988) 518–523.
[7] J.A. Bernstein, A.B. Khodursky, P.H. Lin, S. Lin-Chao, S.N. Cohen, Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays, Proc. Natl. Acad. Sci. USA 99 (2002) 9697–9702.
[8] F. Wright, The 'effective number of codons' used in a gene, Gene 87 (1990) 23–29.
[9] P.M. Sharp, W.-H. Li, The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (1987) 1281–1295.
[10] A. Fuglsang, The 'effective number of codons' revisited, Biochem. Biophys. Res. Commun. 317 (2004) 957–964.
[11] S.A. Marashi, H.S. Najafabadi, How reliable re-adjustment is: correspondence regarding A. Fuglsang, "The 'effective number of codons' revisited", Biochem. Biophys. Res. Commun. 324 (2004) 1–2.
[12] I. Ivanov, R. Alexandrova, B. Dragulev, A. Saraffova, M.G. AbouHaidar, Effect of tandemly repeated AGG triplets on the translation of CAT-mRNA in *E. coli*, FEBS Lett. 307 (1992) 173–176.
[13] G.T. Chen, M. Inouye, Role of the AGA/AGG codons, the rarest codons in global gene expression in *Escherichia coli*, Genes Dev. 8 (1994) 2641–2652.

[14] A.H. Rosenberg, E. Goldman, J.J. Dunn, F.W. Studier, G. Zubay, Effects of consecutive AGG codons on translation in *Escherichia coli*, demonstrated with a versatile codon test system, J. Bacteriol. 175 (1993) 716–722.

[15] M. dos Reis, L. Wernisch, R. Savva, Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome, Nucleic Acids Res. 31 (2003) 6976–6985.