Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 345 (2005) 127–138

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Codon bias as a factor in regulating expression via translation rate in the human genome

Yizhar Lavner, Daniel Kotlar*

Department of Computer Science, Tel Hai Academic College, Upper Galilee 12210, Israel

Received 13 September 2004; received in revised form 10 November 2004; accepted 11 November 2004
Available online 24 December 2004
Received by H.E. Roman

Abstract

We study the interrelations between tRNA gene copy numbers, gene expression levels and measures of codon bias in the human genome. First, we show that isoaccepting tRNA gene copy numbers correlate positively with expression-weighted frequencies of amino acids and codons. Using expression data of more than 14,000 human genes, we show a weak positive correlation between gene expression level and frequency of optimal codons (codons with highest tRNA gene copy number). Interestingly, contrary to non-mammalian eukaryotes, codon bias tends to be high in both highly expressed genes and lowly expressed genes. We suggest that selection may act on codon bias, not only to increase elongation rate by favoring optimal codons in highly expressed genes, but also to reduce elongation rate by favoring non-optimal codons in lowly expressed genes. We also show that the frequency of optimal codons is in positive correlation with estimates of protein biosynthetic cost, and suggest another possible action of selection on codon bias: preference of optimal codons as production cost rises, to reduce the rate of amino acid misincorporation. In the analyses of this work, we introduce a new measure of frequency of optimal codons (FOP′), which is unaffected by amino acid composition and is corrected for background nucleotide content; we also introduce a new method for computing expected codon frequencies, based on the dinucleotide composition of the introns and the non-coding regions surrounding a gene.

Keywords: *Homo sapiens*; Codon bias; Gene expression; Translation efficiency; Optimal codon; Biosynthetic cost

1. Introduction

Codon bias, the unequal use of synonymous codons for encoding amino acids (Grantham et al., 1980; Moriyama, 2003), has been found in many organisms, both prokaryotes and eukaryotes. This bias varies considerably among organisms and even within the genes of the same organism. The bias was found to be in relation with many genomic factors, such as gene length, GC-content, recombination rate, gene expression level, and density of genes (Duret and Mouchiroud, 1999; Kreitman and Comeron, 1999; Duret, 2000; Marais et al., 2001; Urrutia and Hurst, 2001, 2003; Hey and Kliman, 2002; Versteeg et al., 2003), or with other regularities in the genetic code (Karlin and Mrazek, 1996). In different species, codon bias was found to be in weak correlation with gene expression level (Ikemura, 1981; Sharp et al., 1986; Duret and Mouchiroud, 1999; Urrutia and Hurst, 2003). Two main processes were proposed to explain codon bias: natural selection acting on silent changes in DNA, mutational bias, or both. In unicellular organisms, such as *E. coli* and *S. cerevisiae*, it was found that the codons translated by the most abundant tRNA are the most frequently used (Ikemura, 1981, 1982). In multicellular organisms, such as *C. elegans* (Duret, 2000) and *Drosophila* (Akashi, 1995; Moriyama and Powel, 1997), similar findings were found, namely, that codon bias favoring codons with high tRNA gene copy number rises with expression level, thus supporting the action of selection on codon bias to improve translation efficiency. This idea has not been confirmed in

Abbreviations: CB, codon bias; ENC, effective number of codons; FOP, frequency of optimal codons; MCB, maximum likelihood codon bias.

* Corresponding author. Tel.: +972 4 6952965; fax: +972 4 6952899.
E-mail address: danny-k@actcom.co.il (D. Kotlar).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.11.035

mammals (Kanaya et al., 2001). Although a weak correlation between gene expression level and codon bias has been observed in the human genome (Urrutia and Hurst, 2003), this relation has not been linked to tRNA abundance. Recently, Comeron (2004) showed that in the human genome, in the majority of amino acids with degeneracy greater than one, the codons with the most abundant tRNA gene copy numbers, also exhibit an increase in frequency in highly expressed genes compared to lowly expressed genes.

In this study, we introduce new methods for computing the frequency of optimal codons (FOP) and for correcting codon bias for background nucleotide content. Using these methods, we show evidence indicating that the human genome translation efficiency, as estimated using tRNA gene copy numbers, is in weak positive correlation with expression level, and that codon bias has a role in this relation, although not the simple role it has in the model described above: on the one hand, we found that codon bias favors codons with high tRNA gene copy number in highly expressed genes, and on the other hand, based on the evidence presented here, we suggest that codon bias may act as a gene expression regulator by favoring codons with low tRNA gene copy numbers in lowly expressed genes. This supports a mechanism proposed by Fiers and Grosjean (1979) and supported by Konigsberg and Godson (1983) for rare codons in regulatory genes of *E. coli*. Zhang et al. (1991) also proposed this regulatory mechanism for several organisms, including primates. In addition, we present evidence that selection might act on codon bias to prefer optimal codons, possibly to reduce the rate of amino acid misincorporation as protein production cost rises.

2. Materials and methods

2.1. Frequency weighted by expression

The count c_a of each amino acid a is calculated as follows:

$$c_a = \sum_g c_a(g)E(g)$$

where $c_a(g)$ is the count of a in the gene g, $E(g)$ is the expression level of g (average of expression; see below), and the sum is taken over all the relevant genes (either the highly expressed genes or all expressed genes). The expression-weighted frequency f_a^{ex} of the amino acid a is given by

$$f_a^{\mathrm{ex}} = \frac{c_a}{\sum_a c_a} \tag{1}$$

where the sum in the denominator is over all the amino acids. This calculation is similar to the one performed by Duret (2000) for *C. elegans*. In a similar manner, we compute the expression-weighted frequency of a codon.

2.2. Estimating translation efficiency

2.2.1. Gene copy numbers data

Gene copy number data was taken from Lander et al. (2001) and from the tRNA-scan site (http://www.rna.wustl.edu/GtRDB/Hs/Hs-summary.html). In these data, pseudogenes have already been removed. We use tRNA gene copy numbers as an assumed estimate of cellular tRNA abundance (see explanation for this at the beginning of the Results section).

2.2.2. Frequency of optimal codons (FOP)

The optimal codon of an amino acid is defined here as the codon with the highest number of tRNA genes for its anticodon, among its synonymous codons. The simplest way to compute the frequency of optimal codons (FOP) of a gene is to count the number of appearances of optimal codons in the gene, and divide it by the total number of codons in the gene (excluding the stop codons):

$$\mathrm{FOP_s}(g) = \frac{1}{N}\sum_i n_i(g) \tag{2}$$

where $n_i(g)$ is the count of the codon i in the gene g, N is the total number of codons in g, and the sum is taken over all the optimal codons. The subscript s stands for "simple". This FOP measure is affected by amino acid usage. If synonymous codon usage is random, a gene composed only of amino acids of degeneracy two would have FOP of 0.5, whereas a gene composed of amino acids of degeneracy four would have FOP of 0.25. In order to obtain a measure which is independent of amino acid composition, we multiply each codon count in Eq. (2) by the corresponding amino acid degeneracy:

$$\mathrm{FOP}(g) = \frac{1}{N}\sum_i \mathrm{syn}(i)n_i(g). \tag{3}$$

Here, syn(i) is the degeneracy of the amino acid coded by i. This way a gene with close to random synonymous codon usage will have FOP value close to 1, regardless of its amino acid composition. To see that this is a sensible measure, we write Eq. (3) in a slightly different way:

$$\mathrm{FOP}(g) = \sum_i \frac{n_{aa(i)}(g)}{N}\,\frac{n_i(g)/n_{aa(i)}(g)}{1/\mathrm{syn}(i)} \tag{4}$$

where $n_{aa(i)}(g)$ is the count of the amino acid coded by i in g. Assigning $f_i(g)=n_i(g)/n_{aa(i)}(g)$ and $f_{aa(i)}(g)=n_{aa(i)}(g)/N$, we have:

$$\mathrm{FOP}(g) = \sum_i f_{aa(i)}(g)\,\frac{f_i(g)}{1/\mathrm{syn}(i)} \tag{5}$$

Now, the second multiplier is just the *relative synonymous codon usage*, or RSCU, of the codon i in the gene g (Sharp et al., 1986). Hence, the FOP measure is a weighted