FEBS Letters 581 (2007) 5825–5830

On the origin of synonymous codon usage divergence between thermophilic and mesophilic prokaryotes

Surajit Basak, Sujata Roy, Tapash Chandra Ghosh*

Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, India

Received 19 September 2007; revised 14 November 2007; accepted 16 November 2007

Available online 29 November 2007

Edited by Takashi Gojobori

Abstract **Synonymous codon usage analysis between thermophilic and mesophilic prokaryotes has gained wide attention in recent years. Although it is known that thermophilic and mesophilic prokaryotes use different subset of synonymous codons, no reason for this difference is known so far. In the present communication, by analyzing a large number of thermophilic and mesophilic prokaryotes, we provide evidence that bias in the selection of synonymous codons between thermophilic and mesophilic prokaryotes is related to differential folding pattern of mRNA secondary structures. Moreover, we observe that error-minimizing property has significant influence in differentiating the synonymous codon usage between thermophilic and mesophilic prokaryotes. Biological implications of these results are discussed.**

Keywords: Optimal growth temperature; Synonymous codon usage; Free folding energy; Error minimization; *Z*-score

1. Introduction

Non-random usages of synonymous codons both within and between organisms are well documented in the literature [1–3]. Difference in synonymous codon usage may arise from various factors. It has been reported that mutational bias and/or selective forces are the main driving force for the variation of synonymous codon usage among genes in different organisms [4–9]. Variation in synonymous codon usage among genes from the same organism has been shown to depend on many parameters, including expression level [1,3,10], amino acid composition [11–14], gene length [15,16], mRNA structure [17–19], and protein level noise [20]. Global forces can also differentiate the synonymous codon usage between different organisms, e.g. an organism's optimal growth temperature influences the codon usage of its genes [21]. Most of these global forces are thought to be mutational, acting on all DNA sequences, although it has also been argued that growth temperature exerts a selective force on mRNA structure [22] and on codon bias [21].

It has been suggested that biased codon usage due to natural selection could enhance the translational efficiency of protein synthesis [15]. Moreover, biased codon usage is indicative of the differential speed of translation of mRNA [23] and topological features of the encoded proteins [24]. Translational efficiency has two interrelated factors: translational speed and accuracy. Both these factors are influenced by codon usage, and it is difficult to separate the effects of codon usage on each [5,25]. In bacteria and yeast, the correspondence of tRNA abundance with the genome codon usage indicates that high-level expression results in the depletion of internal tRNA pools. Consequently, the translation of an unbiased mRNA is delayed. Most abnormal translation occurs during the waiting time for the "search" for the ternary complex (aminoacyl-tRNA-elongation factor Tu-GTP in bacteria) that matches the codon being translated; the longer the waiting time, the higher the probability of abnormality [26,27]. Hence genes translated rapidly are also translated more accurately.

Recently it has been reported that thermophilic organisms have a different pattern of synonymous codon usage compared to mesophilic organisms [21]. However, no obvious explanation has ever been proposed for the selective advantage of certain codons among other synonymous alternatives under high temperature conditions. It was speculated that synonymous codon usage difference between thermophilic and mesophilic prokaryotes might be related to the mRNA stability [21], i.e., thermodynamically more stable mRNA secondary structure having minimum free energy. The expression level of genes has also been shown to be dependent on RNA secondary structure [28]. However, it was argued that, within the cell, co-transcriptional folding is important in controlling the speed of transcription and thereby influencing both the folding pathway and the functional secondary structure of the mRNA molecule [29]. It has also been demonstrated that RNA sequences can simultaneously encode functional RNA structures as well as proteins and can be analysed through RNA-Decoder [30]. Apart from this, studies on noncoding RNA (ncRNA) genes producing functional RNAs instead of encoding proteins has become more common than previously thought [31].

In the present work, we investigated the variation of free folding energy of original and randomized transcripts of a large number of genomic sequences to assess the influence of mRNA stability on the synonymous codon usage difference between thermophilic and mesophilic prokaryotes. Since the genetic code is degenerate, most amino acids are encoded by several synonymous codons. The theory of error minimization for the evolution of the genetic codes postulates that the codons are arranged in the code in a way that reduces errors. We analyzed the influence of error-minimizing property of

*Corresponding author. Fax: +91 33 2355 3886.
E-mail address: tapash@bic.boseinst.ernet.in (T.C. Ghosh).

0014-5793/$32.00
doi:10.1016/j.febslet.2007.11.054

the coding sequences in differentiating the synonymous codon usage of thermophilic and mesophilic prokaryotes.

2. Materials and methods

The complete genome sequences of all the 37 microorganisms have been downloaded from ftp.ncbi.nlm.nih.gov/genbank/genomes. These genomes have been chosen in such a way as to include a wide variation in genomic G + C content and optimal growth temperature. The same criteria were previously used for synonymous codon usage analysis between thermophilic and mesophilic prokaryotes [21]. Correspondence analysis [32] available in CodonW 1.4.2 (J. Peden, 2000; http://www.molbiol.ox.ac.uk/cu/) was used to investigate the major trend in relative synonymous codon usage variation among the genes. For each native mRNA sequence, 60 random sequences were generated using the randomization protocols, CodonShuffle and DicodonShuffle [33]. The CodonShuffle protocol randomly permutes synonymous codons in codon degenerate family, preserving the exact count of each codon and order of encoded amino acids as in the original transcript. In this protocol, the dinucleotide composition at the (1,2) and (2,3) positions of codons (first/second bases and second/third codon bases, respectively) of the native sequence is preserved, because it preserves codon usage. However, it does not preserve the dinucleotide composition at (3,1) positions; that is, dinucleotides formed by the last base of one codon and the first base of the next. The DicodonShuffle algorithm preserves the dinucleotide composition at (3,1), (1,2), and (2,3) positions, as well as the same encoded amino acid sequence and codon usage of the native mRNA. The important idea of this algorithm is to make only those synonymous codon swaps which either preserve (3,1) dinucleotide composition by themselves, or which can be paired with another reciprocal synonymous codon swap, such that simultaneous swapping of both codon pairs results in no net change in the (3,1) dinucleotide composition. The difference between the two shuffling procedures is that while CodonShuffle protocol preserves dinucleotide composition at the (1,2) and (2,3) positions, the DicodonShuffle algorithm preserves the dinucleotide composition at (3,1), (1,2), and (2,3).

The mfold program was used to predict free folding energies for each native mRNA sequence and the corresponding shuffled sequence available at http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form4.cgi. The difference in the free energy of folding between the native sequence and the corresponding random sequences was measured by the Z-score, given by $Z\text{-score} = \{E_{\text{native}} - \langle E_{\text{random}}\rangle\}/\text{STD}$, where E_{native} denotes the folding free energy of native mRNA sequence, $\langle E_{\text{random}}\rangle$ denotes the average folding free energy over a large number of randomized sequences generated from the native sequence and STD denotes its standard deviation. A positive Z-score indicates that the native sequence has a higher folding free energy than the average of the randomized sequences and therefore is thought to have a less stable secondary structure compared to that for the random sequence.

The degree of error minimization for each genome has been calculated using the method suggested by Archetti [34]. For each pair of amino acids, we measured $D_{\text{AA/AA}^*} = \omega_{\text{AA/AA}} - \omega_{\text{AA/AA}^*}$ from McLachlan's matrix of chemical similarity [35], where $\omega_{\text{AA/AA}}$ is the similarity of amino acid AA with itself and $\omega_{\text{AA/AA}^*}$ is the similarity of AA to the mutant amino acid AA*, produced after an error is introduced at one of the three positions of the original codon. Thus $D_{\text{AA/AA}^*}$ is the distance (dissimilarity) between the original (AA) and the mutant (AA*) amino acids. There are three possible mutants for each codon position and hence there are nine measures of $D_{\text{AA/AA}^*}$ for each codon. Their mean value is taken as a measure of distance (dissimilarity) between the original codon and its possible mutants. This measure is called the mean distance (MD). According to this method proposed by Archetti [34], we have calculated the mean distance (MD) for each synonymous codon based on the McLachlan's [35] matrix of chemical similarity. To calculate the degree of error minimization of a coding sequence, the correlation between the MD values and the corresponding codon frequencies (RSCU) is calculated for each synonymous family. If N is the number of degenerate synonymous codon families on which the correlation is calculated and R is the sum of the correlations, the degree of error minimization is measured by $R_N = R/N$ (R_N ranging between -1 and 1). Since MD is a measure of dissimilarity, the lower the value of R_N, the higher the degree of error minimization.

3. Results and discussion

3.1. Synonymous codon usage variation between thermophilic and mesophilic prokaryotes

Correspondence analysis on relative synonymous codon usage (RSCU) was performed by combining all the genes of an individual genome taken in this study. Similar to the observation made by Lynn et al. [21], we also found that thermophilic and mesophilic genes are completely separated along the second major axes on the basis of RSCU (Fig. 1a). The analysis of relative synonymous codon usage may not detect any constraint imposed by amino acid composition [36]. To examine if amino acid compositions exert any constraint on synonymous codon usage we also performed correspondence analysis on codon usage. The positions of genes along the first and the second major axes produced by correspondence analysis on codon usage (Fig. 1b) are very similar to the figure produced by correspondence analysis on RSCU values (Fig. 1a). Thus it is evident that amino acid composition does not exert any constraint in separating genes according to their synonymous codon usage.

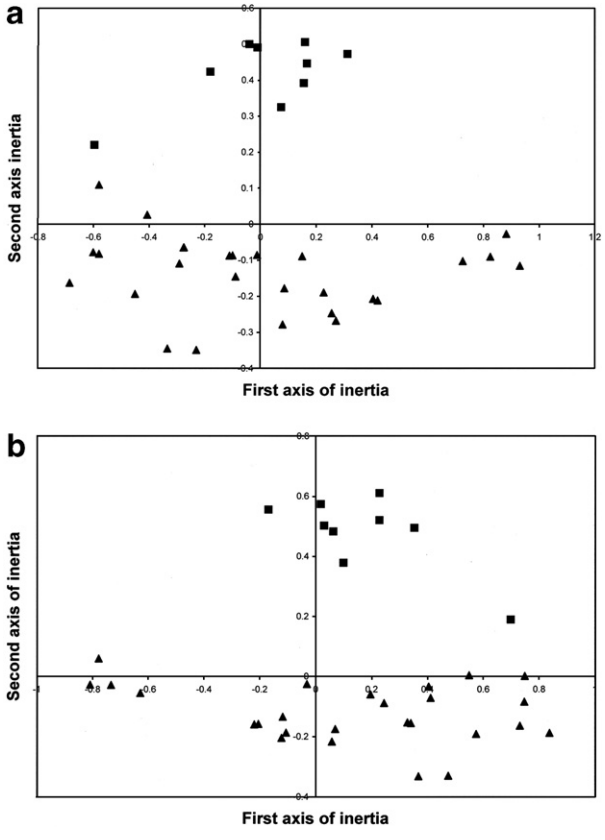

Fig. 1. (a) Positions of the genomes along the first two major axes in the correspondence analysis based on relative synonymous codon usage (RSCU) of all the thermophilic and mesophilic genomes taken in this study. Each square corresponds to one thermophilic genome and each triangle corresponds to one mesophilic genome. (b) Positions of the genomes along the first two major axes in the correspondence analysis based on codon usage of all the thermophilic and mesophilic genomes taken in this study. Each square represents one thermophilic genome and each triangle represents one mesophilic genome.