Available online at www.sciencedirect.com

ScienceDirect

Gene 407 (2008) 30–41

GENE

www.elsevier.com/locate/gene

Amino acid and codon usage profiles: Adaptive changes in the frequency of amino acids and codons

Hani Goodarzi [a,*,1], Noorossadat Torabi [b,1], Hamed Shateri Najafabadi [b], Marco Archetti [c]

[a] Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA
[b] Department of Biotechnology, Faculty of Sciences, University of Tehran, Tehran, Iran
[c] Department of Zoology, Oxford University, South Parks Road, OX1 3PS Oxford, UK

Received 4 March 2007; received in revised form 29 May 2007; accepted 24 September 2007
Available online 5 October 2007

Abstract

In the work presented, the changes in codon and amino acid contents have been studied as a function of environmental conditions by comparing pairs of homologs in a group of extremophilic/non-extremophilic genomes. Our results obtained based on such analysis highlights a number of notable observations: (i) the overall preference of amino acid usages in the proteins of a given organism is significantly affected by major environmental factors. The changes in amino acid preferences (amino acid usage profiles) in an extremophile compared to its non-extremophile relative recurs in the organisms of similar extreme habitats. (ii) On the other hand, changes in codon usage preferences in these extremophilic/non-extremophilic pairs, lack such persistency not only in different genome-pairs but also in the individual genes of a specific pair. (iii) We have noted a correlation between cellular function and codon usage profiles of the genes in the studied pairs. (iv) Based on this correlation, we could obtain a decent prediction of cellular functions solely based on codon usage profile data. (v) Comparisons made between two sets of randomly generated genomes suggest that different patterns of codon usage changes in genes of different functional categories result in a partial resistance towards the changes in the concentration of a given amino acid. This buffering capacity might explain the observed differences in codon usage trends in genes of different functions. In the end, we suggest codon usage and amino acid profiles as powerful tools that can be utilized to improve function predictions and genome-environment mappings.

Keywords: Codon usage profiles; Genetic robustness; Extreme environments; Molecular evolution

1. Introduction

Many reported attempts for finding regulatory and functional regions in genomic DNA of the fully sequenced organisms conveys the message that "genomes are not simply a collection of coding sequences" (Tavazoie et al., 1999; Segal et al., 2003; Cliften et al., 2003; Beer and Tavazoie, 2004). In order to fully comprehend a genome, all the aspects upon which the natural selection might be acting should be identified and studied: gene locations, gene expressions, base compositions, repeated sequences and other yet unknown characteristics (Duret, 2002). Different organisms show different preferences for synonymous codons. The frequency of a codon is not simply the frequency of its corresponding amino acid divided by the number of its codons (Grantham et al., 1980, 1981). The cause of such variations in codon usages is largely debated among the researchers of this field; yet, the major hypotheses can be categorized into three points of view:

i. Mutation bias: a mutation bias towards GC or AT might drastically change the frequency of the commonly used synonymous codons in a genome (Sueoka, 1988; Knight et al., 2001).
ii. Translation efficiency: codon usage biases might match the tRNA abundances to maximize speed and efficiency at the level of translation (Ikemura, 1981, 1985; Sorensen et al., 1989; Bulmer, 1991; Akashi, 1994).
iii. Load minimization: codon usage preferences might be based on an error-minimizing selection at the protein level as similar amino acids result in a similar conformation and

Abbreviation: MD, Mean Dissimilarity.

* Corresponding author. Tel.: +1 609 258 3658; fax: +1 609 258 3980.
E-mail address: goodarzi@princeton.edu (H. Goodarzi).

[1] These authors have contributed equally to this study.

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.09.020

mutations in some codons are relatively less deleterious compared to others (Modiano et al., 1981; Ofria et al., 2003; Archetti, 2004; Najafabadi et al., 2005).

Principally, all the three hypotheses may be important in shaping the codon usage preferences that we observe in different organisms. In this work, we are chiefly interested in studying the evolutionary changes in the amino acid and codon usage preferences, mainly as a function of environmental factors (e.g. temperature and pH). To this end, phylogenetically related extremophilic/non-extremophilic pairs of organisms with fully sequenced genomes were chosen for characterizing differences in the codon and amino acid usages in both the individual genes and the genomes as a whole. Our goal was to find common trends towards a certain pattern of amino acid and codon usage changes that could be associated with a particular environment. In this study, we have defined profiles as the ratio of the frequency of a given amino acid or codon in an extremophile compared to its non-extremophile counterpart. Amino acid profiles (i.e. changes in amino acid usages as we go from a non-extremophilic genome to its extremophilic relative) show similar patterns for the comparisons of species in the same environments; whereas, codon usage profiles differ for each genome-pair. Besides, codon usage profiles show drastic variations among the genes of a single study as well; however, a significant overlap between the genes' cellular functions and their codon usage profiles has been noted in our results. Subsequently, we have shown that these patterns in codon usage changes are quite informative and might even be used to crudely predict functions. This functional enrichment (i.e. high frequency of a given function in the collection of genes with similar codon usage profiles) could not be addressed through prior models. Consequently, we have proposed a model at the level of translation efficiency which is based on a buffering capacity towards sudden changes in amino acid availabilities. This model can be speculated as a possible evolutionary drive for the appearance of such a correlation between codon usage and function in the genes of an organism.

In general, a daily increase in the number of available sequenced genomes calls for a need to find simple and informative characteristics based on which these genomes can be analyzed and compared. In this work, we have noted the potential usage of amino acid and codon usage profiles as tools for genome analyses. Defined as changes in the amino acid content of homologous genes in different genomes, amino acid usage profiles can aid us to identify the late-stage environments in which different genes have evolved. On the other hand, codon usage profiles seem promising as informative data for a crude prediction (although not too impressive on its own) of functions in the sequenced genes.

2. Methods

2.1. Building amino acid usage and codon usage profiles

For each of the extremophilic/non-extremophilic organism-pairs, the following steps were performed in order to build profiles:

1. Each gene in the extremophilic organism was compared to all the genes in the non-extremophilic one in search of a "best hit" and vice versa. The alignments were done using the bl2seq 2.2.14 which comes with the BLAST package (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/blast-2.2.15-ia32-linux.tar.gz). The reciprocal "best hits" with E-values greater than 10^{-4} were omitted from this study.
2. The frequency of each amino acid or codon was calculated for each of the genes and their homologs.
3. The following equations were used for building the profiles:

   $$M(g,g',a) = \log\frac{f(g,a)}{f(g',a)} \tag{1}$$

   where M is the propensity score of amino acid a in gene g (i.e. the extremophile gene) compared to its homolog g' (in the non-extremophile). f returns the frequency of amino acid a in the corresponding gene.

   $$M(g,g',c) = \log\frac{f(g,c)}{f(g',c)} \tag{2}$$

   where f returns the frequency of codon c for each gene and its homolog.
   The result would be a 1×20 matrix of amino acid propensity scores and a 1×64 matrix of codon usage propensity scores for every gene that we call "profiles".
4. The same equations were used to build total amino acid and codon usage profiles for the whole genomes (Fig. 1; Table S1 and S2 in supplementary materials).
5. Regression analysis was performed on the amino acid usage profiles of comparable genomes to study the consistency of amino acid usages in different environmental conditions (Table 2). The corresponding p-values are also reported in Table 2. The same analysis performed on codon usage profiles showed insignificant correlations (data not shows).

2.2. Measuring the genetic robustness in the genomes

Codon robustness is calculated by the mean dissimilarity (MD) between the amino acid coded by each codon and its possible mutants as in Archetti (2004) using a matrix based on chemical similarity (McLachlan, 1971). We assume no transition/transversion bias and either no CG/AT mutation bias ($c=1$) or a CG/AT mutation bias calculated according to the percent content (PCG) of C and G in the genome: ($c=100$/PCG-1; this is, therefore, the value expected if the observed codon usage bias was due entirely to mutation bias, not a true, measured mutation bias). This allows to measure genetic robustness under the two extreme assumptions that mutation bias does not affect at all, or is completely responsible for, codon usage bias. The similarity score of each amino acid with the termination signal is set to -10 (while the other similarity scores in McLachlan's matrix vary between 0 and 9) and multiple generations (10) are allowed as in Archetti (2004).

The level of genetic robustness (wRN — Archetti, 2004, 2006) of a coding sequence is measured by the mean, weighted (proportional to the frequency of the corresponding amino acid) value of the correlations between the MD values and the corresponding codon frequencies for each of the N synonymous