Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 362 (2005) 133 – 140
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

On the coevolution of genes and genetic code

Hani Goodarzi *, Hamed Shateri Najafabadi, Noorossadat Torabi

Department of Biotechnology, Faculty of Science, University of Tehran, Tehran, Iran

Received 29 March 2005; received in revised form 17 July 2005; accepted 3 August 2005
Available online 5 October 2005
Received by T. Ikemura

Abstract

The canonical genetic code acts efficiently in minimizing the effects of mistranslations and point mutations. In the work presented we have also considered the effects of single nucleotide insertions and deletions on the optimality of the genetic code. Our results suggest that the canonical genetic code compensates for the ins/del mutations as well as mistranslations and point mutations.

On the other hand, we highlighted the point that ins/del mutations have a lesser impact on the selected genes of *Saccharomyces cerevisiae* compared to randomly generated ones. We hypothesized that the codon usage preferences in *S. cerevisiae* genes are responsible for the higher efficiency of translation machinery in this organism. Our results support the conjecture that codon usage preferences render the genetic code more effective in minimizing the effects of ins/del mutations.

Keywords: Genetic code; Gene evolution; Mutational stability; Fitness function; Ins/Del mutations

1. Introduction

Once the canonical genetic code was thought to be a "frozen accident" (Crick, 1968). Yet, the discovery of nonstandard genetic codes led to the conclusion that the genetic code can be modified (Osawa, 1995). Thus, the contemporary structure of the genetic code requires explanation regarding many unique symmetries and characteristics. Two sets of hypotheses were introduced; (i) the codon assignments result from natural selection favoring those codes that minimize the effects of mutations and mistranslations, that is minimizing the chemical distances between amino acids (Sonneborn, 1965; Woese, 1965; Knight et al., 1999), and (ii) the structure of the genetic code reflects the biosynthetic pathways of amino acids through time and the error minimization at protein level is just a consequence of this process (Wong, 1975; Szathmary, 1993; Di Giulio, 1997, 1999, 2000; Di Giulio and Medugno, 2000). The debate seems to be unresolved (Freeland et al., 2000; Di Giulio, 2000, 2001).

In any case, the canonical genetic code is known to be highly efficient in minimizing the effects of mutations (Epstein, 1966; Alff-Steinberger, 1969; Ardell, 1998; Freeland, 2002) and mistranslations (Woese, 1965; Goldberg and Wittes, 1966; Haig and Hurst, 1991; Freeland and Hurst, 1998; Gilis et al., 2001; Goodarzi et al., 2004, in press). Herein, we have essayed to measure the efficiency of the genetic code in compensating for the single nucleotide insertions or deletions. While our results suggest that the canonical genetic code is among the very best ones with the capability of minimizing the effects of ins/del mutations, the importance of codon frequencies in the corresponding genes seems to be additionally important. Although our results are mainly derived from *Saccharomyces cerevisiae*, the same approach might be used in order to study other organisms as well.

2. Materials and methods

2.1. Defining a novel fitness function

Fitness functions are mathematical functions devised in order to quantitatively measure the robustness of a given genetic code (i.e. not necessarily the canonical one). These functions assign a score to any given code based on the types of errors and

Abbreviations: Ins/Del mutations, single nucleotide insertion and deletion mutations.

* Corresponding author. Tel.: +98 21 8058210; fax: +98 21 8040284.
E-mail address: hani.goodarzi@gmail.com (H. Goodarzi).

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.08.005

their corresponding costs (Haig and Hurst, 1991; Freeland and Hurst, 1998; Gilis et al., 2001; Goodarzi et al., 2004; Goodarzi et al., in press). For example Gilis et al. (2001) introduced the following function:

$$\varphi^{\mathrm{faa}} = \sum_{c=1}^{64} \frac{p(a(c))}{n(a(c))} \sum_{c'=1}^{64} p(c'|c) \cdot g(a(c), a(c')) \quad (1)$$

where $p(a(c))$ returns the relative frequency of the amino acid a (c) coded by codon c (Gilis et al., 2001), $n(a(c))$ is an integer representing the number of synonymous codons that amino acid $a(c)$ possesses, and $g(a(c), a(c'))$ is a cost measure function which illustrates the deleterious effect of the amino acid substitutions. $p(c'|c)$ stands for the probability of codon c being misinterpreted as codon c' obtained using the transition/transversion weightings and codon position biases chosen by Freeland and Hurst (1998) and tabulated in Table 1.

In the work presented, similar functions were devised to measure the efficiency of a code in minimizing the effects of single nucleotide insertions or deletions:

$$\varphi^{\mathrm{InsDel}} = \sum_{c=1}^{L} g(a(c_3), a(c_4)) + g(a(c_3), a(c_2)). \quad (2)$$

This function scans the input genes from the first codon (c) up to the last (L) one. Each codon (c_3) is compared with two codons obtained from moving the frame one nucleotide forward (c_4) or one nucleotide backward (c_2) which represent a single nucleotide insertion or deletion, respectively. Then the costs of these mutations are summed up for every codon and the resulting value is returned as the fitness of the applied genetic code. Low values of $\varphi^{\mathrm{InsDel}}$ indicate the robustness of the corresponding code regarding the insertion and deletion mutations in the selected gene(s).

Another fitness function was devised to score a given code without providing a gene as an input:

$$\varphi^{\mathrm{InsDel}}_{\mathrm{Total}} = \sum_{c=1}^{64} \mathrm{CU}(c) \sum_{c'=1}^{64} p(c'|c) \cdot f(c'_N) \cdot g(a(c), a(c')) \quad (3)$$

where CU(c) returns the codon usage of codon c in the organism of interest (in the case of *S. cerevisiae* the codon usages were downloaded from http://www.kazusa.or.jp/codon). $p(c'|c)$ is 1 if the two last codon positions or the two first codon positions are similar, otherwise the returned value is 0 (e.g. when c is AUG and c' is UGC or c is AUG and c' is CAU). $f(c'_N)$ is the frequency of the nucleotide N at the first or the third codon position of c' depending on the nature of the corresponding mutation (i.e. insertion or deletion).

Table 1
Mistranslational probabilities ($p(c'|c)$) as previously chosen by Freeland and Hurst (1998)

Condition	Corresponding probability
c and c' differ only in the third base	$1/N$
c and c' differ only in the first base and cause a transition	$1/N$
c and c' differ only in the first base and cause a transversion	$0.5/N$
c and c' differ only in the second base and cause a transition	$0.5/N$
c and c' differ only in the second base and cause a transversion	$0.1/N$
Otherwise	0

N is the normalization factor so that $\sum p(c'|c) = 1$.

Table 2
Different measures of cost of amino acid substitutions used in this work, the references in which they are introduced, and some other works which used them to compute the efficiency of the genetic code in reducing the effects of translational errors

Cost measure	Introduced in	Used in	Corresponding g function (in this work)
PAM_{74-100}	Benner et al. (1994)	Freeland et al. (2000), Gilis et al. (2001), Goodarzi et al. (2004)	$g(a_1, a_2) = -h(a_1, a_2)$
Mutation	Gilis et al. (2001)	Gilis et al. (2001), Goodarzi et al. (2004)	$g(a_1, a_2) = -h(a_1, a_2)$
Polar requirement	Woese et al. (1966)	Haig and Hurst (1991), Freeland and Hurst (1998), Gilis et al. (2001)	$g(a_1, a_2) = [h^{a}(a_1) - h(a_2)]^2$
Hydrophobicity scale #1	Engelman et al. (1986)	Zhu et al. (2003)	$g(a_1, a_2) = [h(a_1) - h(a_2)]^2$
Hydrophobicity scale #2	Nozaki and Tanford (1971)	Zhu et al. (2003)	$g(a_1, a_2) = [h(a_1) - h(a_2)]^2$
Hydropathic character	Kyte and Doolittle (1982)	Zhu et al. (2003)	$g(a_1, a_2) = [h(a_1) - h(a_2)]^2$

The corresponding g functions are also represented.
[a] $h(a)$ returns the value in the corresponding index assigned to amino acid a.

2.2. The cost measure matrices

The abovementioned cost measure function (g in Section 2.1) represents the deleterious effect of each amino acid substitution. In this work we have used two amino acid substitution matrices and four amino acid indices that are tabulated in Table 2 along with their corresponding functions and references.

2.3. Comparing the canonical genetic code with randomly generated codes

As a tangible measure for the optimality of the genetic code, the canonical genetic code is compared to a number of randomly generated codes (the exact numbers are defined for each of the experiments, respectively). In this measure two methods have been used for generating random codes.

2.3.1. The standard method

Freeland and Hurst (1998) used the following rules to generate random codes as a measure of optimality by comparing the canonical genetic code with the randomly generated ones:

1. The "codon space" is divided into 21 non-overlapping sets of codons observed in the canonical code, each set specifying an amino acid in the natural genetic code (one set consists of stop codons).