BioSystems 150 (2016) 73–77

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Local conditions for global stability in the space of codons of the genetic code

Dino G. Salinas*, Mauricio O. Gallardo, Manuel I. Osorio

Centro de Investigación Biomédica, Facultad de Medicina, Universidad Diego Portales, Avda. Ejército 141, Santiago, Chile

ARTICLE INFO

Article history:
Received 16 April 2015
Received in revised form 1 June 2016
Accepted 11 August 2016
Available online 12 August 2016

ABSTRACT

The polar requirement is an attribute of amino acids that is a major determinant of the structure and function of the proteins, and it plays a role in the flexibility and robustness of the genetic code. The viability of an organism depends on flexibility, which allows the exploration of new functions. However, robustness is necessary to protect the organism from deleterious changes derived from misreading errors and single-point mutations. Compared with random codes, the standard genetic code is one of the most robust against such errors. Here, using analytical and numerical calculations and the set of amino acid-encoding codons, we have proposed some local conditions that are necessary for the optimal robustness of the genetic code, and we explored the association between the local conditions and the robustness. The localness of the proposed conditions and the underlying evolutionary mechanism, which begins with a random code and progresses toward more efficient codes (e.g., the standard code), might be biologically plausible.

1. Introduction

A genetic code is a relationship between a set of codons and a set of amino acids and one stop signal; each codon corresponds to either one amino acid or the stop signal. Through this code, the sequences of amino acids that form proteins can be encoded as sequences of codons. Errors in this process can modify the functions of proteins, producing organisms with either increased or decreased viability. Therefore, it is interesting to study two necessary attributes, robustness and flexibility, of the genetic code (Maeshiro and Kimura, 1998). Specifically, with respect to the robustness, here we study local conditions in the space of the codons of the genetic code; these conditions are derived from a global stability condition.

Given a known molecular environment, the function of proteins results from the structural and dynamic consequences of several physical and chemical attributes of amino acids. In this regard, the polar requirement of amino acids, which increases with hydrophobicity (Woese et al., 1966), is one of the most important factors that is evaluated in studies of evolution of the genetic code. To investigate the robustness of the genetic code, the codon-associated changes in amino acid polar requirement are usually considered; these changes can result from a misreading error or a single-point mutation, such that a codon is replaced by any of its neighboring codons. Then, the stability of any genetic code can be estimated by a measurement error, specifically the average of the squared differences of polar requirements (Haig and Hurst, 1991; Freeland and Hurst, 1998). As we will see below, based on the formalism showed by Buhrman et al. (2011), a graph can be associated with the genetic code to provide a theoretical framework for these calculations.

Let $G = (V, E_p)$ be an undirected graph with a set V of 61 codons as vertices and a set E_p composed of the all edges between any two codons having differences in just a single position (i.e., adjacent codons), such as a single-point mutation. By convention, in the case $p = 0$, the difference between neighboring codons is at any position of the codon. Instead, if the single-point mutation is restricted to the first, second or third position of the codon, the cases are termed $p = 1$, $p = 2$ and $p = 3$, respectively. Thus, we define E_0 as the set of edges between any two adjacent codons. Similarly, E_1, E_2 and E_3 are the sets of edges between two adjacent codons but only for changes in the first, second and third position of the codon, respectively. These sets are disjoint, and $E_0 = E_1 \cup E_2 \cup E_3$ (Buhrman et al., 2011).

Codons are denoted by the triplet ijk, with $i, j, k \in B = \{A, C, G, U\}$, that is, the set of nucleotides represented by their component bases. Let F be a function mapping each amino acid-encoding codon ijk to one amino acid $F(ijk)$, and let r be a

* Corresponding author.
E-mail addresses: dinosalinas@gmail.com, dino.salinas@udp.cl (D.G. Salinas), mauricio.gallardor@mail.udp.cl (M.O. Gallardo), manuel.osorio@mail.udp.cl (M.I. Osorio).

http://dx.doi.org/10.1016/j.biosystems.2016.08.007
0303-2647/

Table 1
Comparison between the LE(p) values of the standard genetic code and the codes for either the completely random model or the fixed standard block model. Mean values are shown with their variance. Calculations are from the same data as used in Fig. 1.

	LE(0)	LE(1)	LE(2)	LE(3)
Completely random model	7.04 ± 1.17	8.85 ± 1.42	9.07 ± 1.46	9.07 ± 1.46
Fixed standard block model	4.67 ± 0.85	8.81 ± 1.78	9.43 ± 1.91	3.57 ± 1.08
Standard genetic code	2.0400	4.8093	7.5956	0.6482

function mapping each amino acid $F(ijk)$ to a polar requirement $r_{ijk} \equiv r(F(ijk))$. In other words, for 20 standard amino acids and 61 non-stop codons, each ijk codon is associated with an r_{ijk} value, which is equal to the polar requirement of the amino acid $F(ijk)$. In this article, for a genetic code (i.e., F), the r_{ijk}values are taken from the values of polar requirement in Table 1 of the work of Haig and Hurst (1991).

Then, the error function MS(p) can be defined as

$$\mathrm{MS}(p) = \frac{1}{|E_p|} \sum_{\{ijk,i'j'k'\} \in E_p} \left(r_{ijk} - r_{i'j'k'}\right)^2 \tag{1}$$

with $|E_p|$ equal to the cardinality of the set E_p. Then, for only the 61 codons encoding amino acids in the standard genetic code, $|E_0| = 263$, $|E_1| = 87$, $|E_2| = 88$, and $|E_3| = 88$ (Buhrman et al. 2011).

The function MS defined above is a measure that can be applied to random or natural genetic codes, such as the standard genetic code. For a code, the codons encoding the same amino acid or stop signal form a set of synonymous codons or *block* (Here we use the word "block" as in Goldman (1993), Freeland and Hurst (1998), Freeland et al. (2000) and Buhrman et al. (2011), and not as in other works, such as in Novozhilov et al. (2007)). A *completely random model* corresponds to genetic codes formed by randomly generated blocks with random assignment of the amino acids and stop signal. Instead, in a more restricted case, blocks can be fixed (for any degeneration and composition of codons), although amino acids and the stop signal are still randomly assigned; this model is known as the *fixed block model* (Goldman, 1993). Using MS(p) values, comparisons have been performed elsewhere between the standard code and codes that follow a particular fixed block model; this type of model employs the same standard block structure as that of the standard code. We named the latter model of genetic code a *fixed standard block model*. All genetic codes used in this article consider 61 and 3 codons that encode 20 amino acids and one stop signal, respectively, regardless of the selected model of code. However, the 3 stop codons are not included in the calculations.

The MS measure has been previously studied by Haig and Hurst (1991, 1999) using the standard code and 10,000 randomly generated codes following the fixed standard block model. The researchers studied all cases of single-nucleotide substitutions corresponding to any of the three positions of codons ($p = 0$). Moreover, they studied changes restricted to a fixed single position of codons (depending on whether $p = 1$, 2 or 3), and they considered other attributes of amino acids besides the polar requirement, such as the molecular volume, hydropathy and isoelectric point. For the polar requirement withMS, the standard code had one of the higher values for the robustness of error (i.e., one of the lower MS values) compared to those of randomly generated codes, and this robustness was greater than that for the other amino acid attributes. Freeland and Hurst (1998) corroborated these results using one million randomly generated codes. They found that for the measure of error MS(0), 114 genetic codes had lower values than the standard genetic code. That is, these codes were more robust than the standard genetic code. They also evaluated changing a codon by a single nucleotide at a fixed position of the codon, i.e., they also applied the MS(p,) measure on the graph $G = (V, E_p)$ for $p = 1$, 2 or 3. For the MS(p) measure, their results showed that the robustness of the standard genetic code is better in the third position of the codon and worse in the second position of the codon. Through a heuristic method, Goldman (1993) found a code with a lower value of MS(p) than the values for any of the constructed codes. Buhrman et al. (2011) demonstrated that this heuristic code corresponds to the global optimum, that is, the most robust code.

Knowledge about the conditions required for a genetic code to have maximal stability could be useful for theories about the evolution of the genetic code. Moreover, many artificial amino acids can currently be used to synthesize proteins; thus, knowing the conditions of maximum stability would be useful for generating new and more robust genetic codes.

In this article, we deduce some local conditions for the space of codons, and we determine whether these conditions are sufficient and necessary for the minimum of the MS(p) function, which is associated with a stable genetic code.

2. Theoretical framework and results

For a minimum of MS(p), we have

$$\frac{\partial \mathrm{MS}(p)}{\partial r_{abc}}\Big|_{r_{abc} = r^*_{abc}} = 0 \quad \forall\ a,\ b,\ c \in B \tag{2}$$

From Eq. (2), it is possible to deduce a general local stability condition for $r_{abc} = r^*_{abc}$.

Case p = 0

From Eq. (1) and $p = 0$,

$$\mathrm{MS}(0) = \frac{1}{2|E_0|}\left[\sum_{i,j,k,i' \in B}\left(r_{ijk} - r_{i'jk}\right)^2 + \sum_{i,j,k,j' \in B}\left(r_{ijk} - r_{ij'k}\right)^2 + \sum_{i,j,k,k' \in B}\left(r_{ijk} - r_{ijk'}\right)^2\right] \tag{3}$$

The derivative of MS(0) is taken:

$$\frac{\partial \mathrm{MS}(0)}{\partial r_{abc}} = \frac{1}{|E_0|}\sum_{i,j,k,i' \in B}\left(r_{ijk} - r_{i'jk}\right)\frac{\partial}{\partial r_{abc}}\left(r_{ijk} - r_{i'jk}\right) + \frac{1}{|E_0|}\sum_{i,j,k,j' \in B}\left(r_{ijk} - r_{ij'k}\right)\frac{\partial}{\partial r_{abc}}\left(r_{ijk} - r_{ij'k}\right) + \frac{1}{|E_0|}\sum_{i,j,k,k' \in B}\left(r_{ijk} - r_{ijk'}\right)\frac{\partial}{\partial r_{abc}}\left(r_{ijk} - r_{ijk'}\right) \tag{4}$$

The first summation is calculated as

$$\begin{aligned}\sum_{i,j,k,i' \in B}\left(r_{ijk} - r_{i'jk}\right)\frac{\partial}{\partial r_{abc}}\left(r_{ijk} - r_{i'jk}\right) &= \\ = 2\sum_{i,j,k,i' \in B} r_{ijk}\frac{\partial}{\partial r_{abc}}\left(r_{ijk}\right) - 2\sum_{i,j,k,i' \in B} r_{ijk}\frac{\partial}{\partial r_{abc}}\left(r_{i'jk}\right) &= 2\sum_{i' \in B} r_{abc} - 2\sum_{i \in B} r_{ibc} = \\ = 8r_{abc} - 2r_{abc} - 2\sum_{i \neq a,\ i \in B} r_{ibc} &= 6r_{abc} - 2\sum_{i \neq a,\ i \in B} r_{ibc}\end{aligned} \tag{5}$$