BioSystems 117 (2014) 77–81

Contents lists available at ScienceDirect

BioSystems

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

Probable relationship between partitions of the set of codons and the origin 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 17 March 2013
Received in revised form
26 December 2013
Accepted 24 January 2014

Keywords:
Genetic code
Origin
Evolution
Entropy
Integer partition
Stirling numbers of the second kind

ABSTRACT

Here we study the distribution of randomly generated partitions of the set of amino acid-coding codons. Some results are an application from a previous work, about the Stirling numbers of the second kind and triplet codes, both to the cases of triplet codes having four stop codons, as in mammalian mitochondrial genetic code, and hypothetical doublet codes.

Extending previous results, in this work it is found that the most probable number of blocks of synonymous codons, in a genetic code, is similar to the number of amino acids when there are four stop codons, as well as it could be for a primigenious doublet code. Also it is studied the integer partitions associated to patterns of synonymous codons and it is shown, for the canonical code, that the standard deviation inside an integer partition is one of the most probable.

We think that, in some early epoch, the genetic code might have had a maximum of the disorder or entropy, independent of the assignment between codons and amino acids, reaching a state similar to "code freeze" proposed by Francis Crick. In later stages, maybe deterministic rules have reassigned codons to amino acids, forming the natural codes, such as the canonical code, but keeping the numerical features describing the set partitions and the integer partitions, like a "fossil numbers"; both kinds of partitions about the set of amino acid-coding codons.

1. Introduction

The genetic code might have originated in a prebiotic environment, with very simple structures capable of division and autocatalytic growth, perhaps as encapsulated RNA into lipid vesicles (Chen et al., 2005; Deamer and Weber, 2010). A complete theory of the genetic code evolution, besides to be applied since its origin, should explain the canonical genetic code and its natural variations (Koonin and Novozhilov, 2009) and why the standard amino acids have been selected to be encoded (Weber and Miller, 1981).

There are many theoretically possible kinds of amino acids (Grützmann et al., 2011) and there are some arguments to explain why the number of encoded amino acids must be limited, such as assumptions about the topology of codon space determining the maximal numbers of amino acids: 11 for 4-base doublets and 25 for 4-base triplets (Tlusty, 2010). However, these arguments do not explain the exact number of encoded amino acids for the canonical genetic code (i.e., the standard genetic code) and for other natural genetic codes (Fernandez, 2004; Wu et al., 2005). In our previous work, we found a possible explanation for the number of amino acids (Salinas et al., 2012). There, we calculated the number of possible set partitions of a set of n elements (codons encoding to k amino acids) into k subsets (or blocks, each one consisting of the synonymous codons for a particular amino acid). Such number of partitions is known as the *Stirling number of the second kind* ($S(n, k)$). We have demonstrated that the maximum value of $S(n, k)$ is for $k = 19$ or 20, depending on the number of stop codons previously fixed. We think that these values of k could determine the number of currently encoded amino acids in both canonical genetic code and other natural genetic codes (Sammet et al., 2010). Such analysis is not contrary to theoretical models determining the selection and adaptation of the genetic code (Freeland et al., 2000; Novozhilov et al., 2007; Vetsigian et al., 2006).

In this paper, we extend our previous study to two cases: the triplet code with four stop codons, as in mammalian mitochondria, and, in the second case, the hypothetical doublet code at the more early origin of the genetic code. Moreover, here we study integer partitions formed by the cardinalities of the blocks formed by synonymous amino acid-coding codons. For example, in the canonical genetic code, amino acid-coding codons are partitioned into 20 blocks of synonymous codons, whose cardinalities form the integer partition {1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 4, 4, 4, 4, 4, 6, 6, 6}. Interestingly, in entropically governed systems, the probability to get

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

0303-2647/$ – see front matter © 2014 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.biosystems.2014.01.007

any state increases with the disorder. For this reason, here, we also study whether disorder or entropy could determine, in the genetic code, both to the number of encoded amino acids and the kind of integer partition derived from the pattern of synonymous codon.

2. Theoretical framework and results

Let $C \equiv \{c_1, \ldots, c_n\}$, a set of codons to encode k amino acids, according to a genetic code (Nieselt-Struwe and Wills, 1997). To formal purposes, the stop signal is considered like an amino acid, but here n corresponds only to the total of the amino acid encoding codons. That is, total codons minus the stop codons. We define a set partition $\mathcal{P}$ of C as $\mathcal{P} \equiv \{C_1, \ldots, C_k\}$, with $C = \bigcup_{i=1}^{k} C_i$, $C_i \neq \phi$ and, given $i \neq j$, $C_i \bigcap C_j = \phi$. Since $\mathcal{P}$ is a set partition into k blocks, we say that $\mathcal{P}$ is a k-partition. In this work, C_i is named *block of synonymous codons* and $\mathcal{P}$ is named *pattern of synonymous codons* (*pattern of degeneration*, in our previous work (Salinas et al., 2012)). The number of different kinds of $\mathcal{P}$ is given by $S(n, k)$, the *Stirling number of the second kind* (Comtet, 1974).

2.1. Determining the probability to get a pattern of k indistinguishable blocks of synonymous distinguishable codons

Given a set of n codons encoding k amino acids, $\mathcal{P}$ represents a pattern of k indistinguishable blocks of synonymous distinguishable codons. If k is an integer variable, such that $1 \leq k \leq n$, then, the total number of possible $\mathcal{P}$ is called the *Bell number*, which is represented by B_n (Comtet, 1974). Defining

$$P(n, k) \equiv \frac{S(n, k)}{B_n} \tag{3}$$

and if each one of the B_n different set partitions have equal probability to be selected, then, like in a random model, $P(n, k)$ represents the probability to get a k-partition.

The maximal of $S(n, k)$ is obtained at the $k = k_n$ values; having only one or two k_n values (Harper, 1967) into a determined range (Yu, 2009). In our previous work, we have found the values of k_n, corresponding to the maxima value of $P(n, k)$ (Eq. (3)) for the triplet code mode, with $n = 61$–64, depending of the number of stop codons (Salinas et al., 2012). Here, similarly, the calculus is extended to two new cases:

- In the first case, for the triplet code, with $n = 60$, it is obtained $k_n = 19$, with $P(60, 19) = 0.199$.
- In the second case, for the hypothetical doublet code mode, employing four different nucleotide bases, for $n = 15$ (i.e., one predetermined stop codon) it is obtained $k_n = 6$ and $P(15, 6) = 0.304$ (very near to $P(15, 7) = 0.296$); and for $n = 16$ (without stop codon), it is obtained $k_n = 7$ and $P(16, 7) = 0.313$.

Results from these two cases, and some ones from our previous work, are shown in Fig. 1.

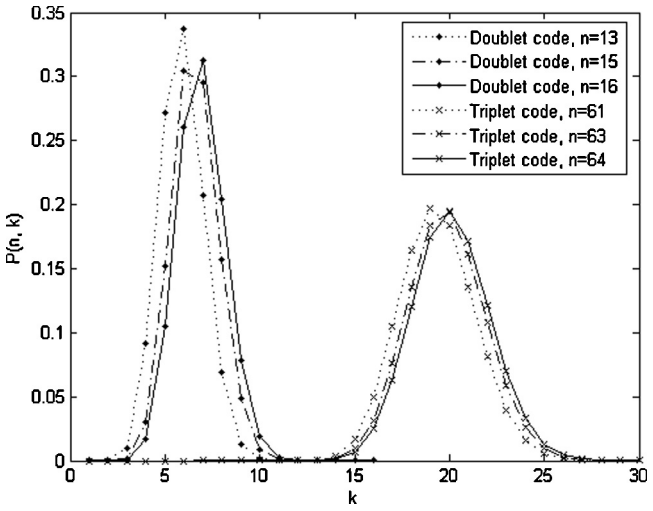

Fig. 1. Probability $P(n, k)$ to get a k set partition (partition into k blocks) from the all possible set partitions, considering the kind of encoding system (doublet code or triplet code) and different numbers of stop codons.

2.2. Integer partitions of patterns of synonymous codons

2.2.1. Generating an integer partition from a set partition

From each $\mathcal{P}$ on a set of n codons, using the cardinalities of the k blocks (i.e., the numbers of degeneration inside the blocks of synonymous codons), we obtain an *integer partition of n*. That is, a set of k positive numbers, such that their sum is n. The allocation of distinguishable entities in indistinguishable states has been studied elsewhere (Niven, 2007) and here, in this section, we apply a similar analysis to a problem of a set of n codons partitioned into k nonlabeled blocks.

Let's suppose a realization in which n distinguishable codons can be allocated into k indistinguishable blocks (indistinguishable amino acids), such allocation being described by a specific set partition $\mathcal{P}$ that is named $\mathcal{P}ij$, such that $\mathcal{P}ij$ determines an *integer partition* p_i, given by $p_i = \{n_1, n_2, \ldots, n_k\}$, with n_μ $(1 \leq \mu \leq k)$ the cardinality in the μth block of synonymous codons and, then, $\sum_{\mu=1}^{k} n_\mu = n$ (the order inside p_i is not matter, and we annotate n_μ in ascending order).

Fig. 2 shows how different k-partitions (e.g., $\mathcal{P}ij$ and $\mathcal{P}i'j'$, whose associated integer partitions are p_i and $p_{i'}$, respectively) could generate either similar $(i = i')$ or different $(i \neq i')$ integer partitions. There, a set partition corresponds to the canonical code and the other ones are random set partitions.

Let W_i denotes the number of possible realizations (the statistical weight) of $\mathcal{P}ij$, $1 \leq j \leq W$, to obtain p_i. Similarly, W_i can be represented by $\left\{\overset{n}{\left\{n_1, n_2, \ldots, n_k\right\}}\right\}$. Then (Niven, 2007)

$$S(n, k) = \sum_{\text{all } p_i} W_i = \sum_{\text{all } p_i} \left\{\overset{n}{\left\{n_1, n_2, \ldots, n_k\right\}}\right\} \tag{6}$$

In the example of Niven (2007), in which 5 distinguishable elements can be allocated into 3 indistinguishable blocks:

$$W_1 = \left\{\overset{5}{\left\{1, 2, 2\right\}}\right\} = 15 \quad \text{and} \quad W_2 = \left\{\overset{5}{\left\{1, 1, 3\right\}}\right\} = 10 \tag{7}$$

$$\text{Then} \quad S(5, 3) = \left\{\overset{5}{\left\{1, 2, 2\right\}}\right\} + \left\{\overset{5}{\left\{1, 1, 3\right\}}\right\} = 25 \tag{8}$$

$$\text{Note that} \quad \left\{\overset{5}{\left\{3, 1, 1\right\}}\right\} = \left\{\overset{5}{\left\{1, 3, 1\right\}}\right\} = \left\{\overset{5}{\left\{1, 1, 3\right\}}\right\} \quad \text{(the order does not matter)}$$

W_i is given by (Niven, 2007)

$$W_i = \left\{\overset{n}{\left\{n_1, n_2, \ldots, n_k\right\}}\right\} = \left(n_1, n_2, \overset{n}{\ldots}, n_k\right)\left(\prod_{j=1}^{n} \frac{1}{r_j!}\right)$$

$$= \frac{n!}{\left(\prod_{i=1}^{k} n_i!\right)\left(\prod_{j=1}^{n} r_j!\right)} \tag{9}$$

with $r_j \geq 0$, r_j is the number of occurrences of an integer j in the set p_i.