AVERAGE AND STANDARD DEVIATION OF THE ERROR FUNCTION FOR RANDOM GENETIC CODES WITH STANDARD STOP CODONS

Dino G. Salinas[1]

(1) Centro de Investigación Biomédica, Facultad de Medicina, Universidad Diego Portales, Santiago, Chile.

Email: dino.salinas@udp.cl

ORCID ID: 0000-0001-8152-1622

Keywords: origin; evolution; genetic code; random; error function

DECLARATIONS

Funding: Not applicable.

Conflicts of interest/Competing interests: The author declares that he has no conflict of interest.

Ethics approval: Not applicable.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and material: Not applicable.

Authors' contributions: The article has only one author.

ABSTRACT

The origin of the genetic code has been attributed in part to an accidental assignment of codons to amino acids. Although several lines of evidence indicate the subsequent expansion and improvement of the genetic code, the hypothesis of Francis Crick concerning a frozen accident occurring at the early stage of genetic code evolution is still widely accepted. Considering Crick's hypothesis, mathematical descriptions of hypothetical scenarios involving a huge number of possible coexisting random genetic codes could be very important to explain the origin and evolution of a selected genetic code. This work aims to contribute in this regard, that is, it provides a theoretical framework in which statistical parameters of error functions are calculated. Given a genetic code and an amino acid property, the functional code robustness is estimated by means of a known error function. In this work, using analytical calculations, general expressions for the average and standard deviation of the error function distributions of completely random codes with standard stop codons were obtained. As a possible biological application of these results, any set of amino acids and any pure or mixed amino acid properties can be used in the calculations, such that, in case of having to select a set of amino acids to create a genetic code, possible advantages of natural selection of the genetic codes could be discussed.

1. INTRODUCTION

All current natural genetic codes may have evolved from a single ancestral code. According to the Crick hypothesis in 1968, this ancestral code would have consisted of fixed random codon assignments for each encoded amino acid and the stop signal. This approach is known as "the frozen accident" (Crick 1968). An explanation for the original fixed codon assignments could be the deleterious effects of genetic changes. These effects would be increasingly catastrophic as the number of genes in organisms increased. However, in early evolution extensive horizontal gene transfer might have been useful because only one code survived, a requirement for the transition to the cellular level of complexity (Vetsigian et al. 2006). Given the above, the origin of LUCA (Last Universal Cellular Ancestor) (Weiss et al. 2018), the first common ancestor of all current organisms, but not the first cell, would have been a bottleneck resulting from this horizontal gene transfer, which would have resulted in the selection of a universal code (Koonin 2003; Koonin 2017; Vetsigian et al. 2006).

Thus, at an early stage a completely random universal code could be possible. However, this evolved in such a way that there were fewer reading and writing failures, diminishing the structural and functional consequences of the encoded proteins (i.e., the error function as a cost function) (Freeland and Hurst 1998). Moreover, considering that sometimes code errors could be important for developing new cellular adaptive properties, perhaps genetic code evolution, rather than a way of optimizing stability, tended to optimize the balance between stability and adaptability. According to an evolutionary increase in stability, it has been found that the errors associated with the standard genetic code are considerably smaller than most random codes, although its achieved stability is not the best possible (Błażej et al. 2018; Błażej et al. 2016; Buhrman et al. 2011; Freeland and Hurst 1998; Goldman 1993; Haig and Hurst 1991; Haig and Hurst 1999; Novozhilov et al. 2007; Salinas et al. 2016; Santos and Monteagudo 2010; Wnętrzak et al. 2018; Wnętrzak et al. 2019). The remarkable stability of the standard genetic code, besides being a driving force through the selection pressure, may have been a consequence of the expansion of the genetic code by mean of similar mechanism of codon assignments to physicochemically similar amino acids (Crick 1968; Koonin 2017). Thus, the hypothetical accidental nature of a selected ancestral genetic code is in agreement with subsequent genetic code extension and optimization mechanisms (Koonin and Novozhilov 2017). In this work, the average and standard deviation of error functions of random genetic codes with fixed standard stop codons were analytically obtained assuming that a primitive and completely random version of an ancestral genetic code may have been selected from a large set of random codes. The used error functions are different depending on a parameter indicating which codon bases (i.e., first, second or third) can be wrong. As a possible application of these results in future research, the deduced expressions of statistical parameters could be useful to select different sets of natural amino acids and many kinds of amino acid properties, either pure or mathematically combined. This approach, regarding the different statistical behaviors of the error function in a random Crick scenario, could allow a better understanding of the code stability as a selective pressure on the origin and evolution of the genetic code.

For calculations, the following mathematical formalism is introduced:

In the standard genetic code, from the 64 possible codons, there are 3 stop codons and 61 amino acid encoding codons, which encode the 20 standard amino acids. Hence, a genetic code can be described by a function, such that there are 61 different triplets ijk (with bases $i, j, k \in B = \{A, C, G, U\}$) termed codons, each one encoding one amino acid; E_p is the set of pairs of triplets indicated by $(ijk, i'j'k')$, that only differ in position p, with p = 1, 2 or 3, such that only codon pairs $(ijk, i'jk)$, $(ijk, ij'k)$ and (ijk, ijk') ($i \neq i', j \neq j'$ and $k \neq k'$ respectively) are considered. For p = 0, E_0 denotes the union $E_1 \cup E_2 \cup E_3$ (Buhrman et al. 2011).

Let r_{ijk} be the numeric value of property a_u as expressed by the standard amino acid u coded by triplet ijk codon (that is, in functional notation, $u = u(ijk)$)

$$r_{ijk} \equiv a_u = a_{u(ijk)} \qquad (1)$$

The r_{ijk} values of the six amino acid properties used in this work are shown in Table 1. Four of these properties are real properties taken from Haig and Hurst (Haig and Hurst 1991); however, two other properties are not real, but are only arbitrary values to increase the number of cases to test the theoretical results of this study.

To understand the robustness of the genetic code, the consequences of single-point changes in codons (either mutation or translation errors) have been studied. Hence, genetic code robustness can be inversely estimated by measuring a global error, basically a cost function associated with decoding mistakes. Such an error function (*MS*) is defined as follows:

$$MS_p \equiv \frac{1}{\left|E_p\right|} \sum_{(ijk,i'j'k') \in E_p} \left(r_{ijk} - r_{i'j'k'}\right)^2$$

(2)

where $\left|E_p\right|$ is the cardinality of set E_p.

We find that

$$\left|E_p\right| = \sum_{(ijk,i'j'k') \in E_p} 1$$

(3)

and verify that

$$|E_0| = |E_1| + |E_2| + |E_3| \qquad (4)$$

Considering 61 amino acid encoding codons and the 3 standard stop codons (UAG, UAA, and UGA), and since codon pairs with inner differences simultaneously in more than one position are not considered, we have $\left|E_0\right| = 263$, $\left|E_1\right| = 87$, $\left|E_2\right| = 88$, and $\left|E_3\right| = 88$ (Buhrman et al. 2011).

2. THEORETICAL FRAMEWORK AND RESULTS

Only *completely random models* of the genetic code with the fixed three standard stop codons (also named *the unrestricted structure model* (Wnętrzak et al. 2018)) were considered. Genetic codes were built fixing the three standard stop codons and using the other 61 codons to encode the 20 standard amino acids. Although the number of possible codes is finite (Novozhilov et al. 2007; Schönauer and Clote 1997), the number of randomly selected codes from the huge number of possible codes can be infinite. Hereinafter we will refer to an "infinite number of sampling cycles of random genetic codes" as "infinite random codes". Thus, let $\langle \quad \rangle_\infty$ be the average of infinite random codes. Then we denote the average of MS_p (Eq. 2) over infinite random genetic codes by $\langle MS_p \rangle_\infty$:

$$\langle MS_p \rangle_\infty = \langle \frac{1}{\left|E_p\right|} \sum_{(ijk,i'j'k') \in E_p} \left(r_{ijk} - r_{i'j'k'}\right)^2 \rangle_\infty$$

(5)

Let σ_p be the standard deviation of *MS*$_p$ over infinite random genetic codes. Then σ_p^2 is the variance given by

$$\sigma_p^2 = \langle \left(MS_p - \langle MS_p \rangle_\infty \right)^2 \rangle_\infty \qquad (6)$$

For the computer calculations, 100,000 randomly sampled genetic codes were obtained using randomly-generated non-overlapping block of codons with random assignment of the amino acids. In addition, six kinds of amino acid properties (Table 1) were used and their statistical properties were analyzed with respect to changes in a single position of the codon (*p* = 1, 2, or 3). The parameters $\langle MS_p \rangle_\infty$ and σ_p were numerically calculated using the Monte Carlo method. Subsequently, these parameters were analytically calculated using general expressions, which were obtained for any amino acid property and any set of encoded amino acids. Comparisons between the values obtained using numerical and analytical methods are shown in Tables 2 and 3. The analytical results are very similar to the numerical results of the Monte Carlo computational calculation. In fact, because the results of the analytical calculations are completely based on mathematical arguments, a numerical proof of these results is not necessary. However, Tables 2 and 3 are useful to show how derived statistical expressions can be applied and numerically contrasted.

2.1. Analytical calculation of $\langle MS_p \rangle_\infty$ (*p* = 1, 2, or 3) for infinite completely random genetic codes with the standard stop codons.

From Eq. 5, since summations of r_{ijk}^2 and $r_{i'j'k'}^2$ are equal, by exchanging summation and average operators ($\langle \sum \quad \rangle = \sum \langle \quad \rangle$), we obtain

$$\langle MS_p \rangle_\infty = \frac{2}{|E_p|} \left(\sum_{(ijk,i'j'k') \in E_p} \langle r_{ijk}^2 \rangle_\infty - \sum_{(ijk,i'j'k') \in E_p} \langle r_{ijk} r_{i'j'k'} \rangle_\infty \right) =$$

$$= \frac{2}{|E_p|} \left(\langle r_{ijk}^2 \rangle_\infty \sum_{(ijk,i'j'k') \in E_p} 1 - \langle r_{ijk} r_{i'j'k'} \rangle_\infty \sum_{(ijk,i'j'k') \in E_p} 1 \right)$$

(7)

where $\langle r_{ijk}^2 \rangle_\infty$ and $\langle r_{ijk} r_{i'j'k'} \rangle_\infty$ are constants because they are obtained from infinite selected random codes from all possible codes. That is, the averages $\langle \quad \rangle_\infty$ do not depend on the subscripts for r and therefore they can be written outside of the summations, as a factor.

Note that

$$\langle r_{ijk} r_{i'j'k'} \rangle_\infty = \langle r_{ijk} \rangle_\infty^2 \qquad (8)$$

Using Eqs. 1, 3 and 8 in Eq. 7 results in

$$\langle MS_p \rangle_\infty = 2(\langle a_u^2 \rangle_\infty - \langle a_u \rangle_\infty^2) \qquad (9)$$

Because each u-th amino acid has the same statistical weight in calculations of averages over infinite random genetic codes, in Eq. 9 we replace $\langle \quad \rangle_\infty$ with $\langle \quad \rangle_{Aa}$, that is the average over the 20 standard amino acids (i.e. $\langle \quad {}_u \rangle_{Aa} \equiv \sum_{u=1}^{20} \quad {}_u/20$)). Thus, we obtain

$$\langle MS_p \rangle_\infty = 2(\langle a_u^2 \rangle_{Aa} - \langle a_u \rangle_{Aa}^2) \qquad (10)$$

where p = 1, 2, or 3. The application results are shown in Table 2, and another demonstration of Eq. 10 is shown in Appendix A.

2.2. Analytical calculation of σ_p (p = 1, 2, or 3) for infinite completely random genetic codes with standard stop codons.

In Appendix B the following expression for the standard deviation is obtained

$$\sigma_p = 2\left[\frac{1}{|E_p|}\langle (a_u - \langle a_u \rangle_{Aa})^4 \rangle_{Aa}\right]^{\frac{1}{2}}$$

(11)

where p = 1, 2, or 3. The application results are shown in Table 3.

3. DISCUSSION

In the calculation of the averages over random genetic codes, each code has the same probability of being obtained by the Monte Carlo method. Therefore, the averages for infinite number of sampling cycles of random genetic codes are equal to the corresponding averages for the finite set of all possible genetic codes. Using standard stop codons in the completely random model of genetic code, we analytically obtained the average (Eq. 10) and standard deviation (Eq. 11) of error functions (Eq. 2) of infinite random codes selected. The formulae in Eq. 10 and 11 were exact and applicable for any kind of amino acid property, even for new properties resulting from combinations of some already known (e.g., a linear combination of several amino acid properties). Similarly, the set of encoded amino acids could also be redefined into the formulae. In computational experiments, using 100,000 random genetic codes, in addition to the 20 standard amino acids and 6 kinds of amino acid properties (4 real properties and the other 2 invented, for test purposes only) both statistical parameters (i.e., $\langle MS_p \rangle_\infty$ and σ_p (p = 1, 2, or 3) were obtained with values very similar to those predicted by the analytical calculations (See Tables 2 and 3).

It is interesting that the average of the error function of the code is proportional to the mean squared change (Eq. 10) of the encoded amino acid property, as long as the variance is proportional to the mean quartic change (Eq. B22). Such a simple result could avoid a large number of computational calculations and was capable of establishing a theoretical framework that could be applied to a random scenario prior to a universal code. For example, it seems plausible that genetic codes with small error function values are more competitive (i.e., genetic codes having a greater tolerance to errors of use) as σ_p decreases and $\langle MS_p \rangle_\infty$ increases. That could be achieved by a suitable selection of sets of amino acids and their properties (pure or mathematically combined) to give the appropriate parameters to the error function in primitive systems containing amino acids, such as in some meteorites or in primary organic soups (Burton et al. 2012; Cleaves 2010; Zaia et al. 2008). In this regard, the following question seems interesting: how optimal are the current standard amino acids and their selected properties in terms of the competitiveness of genetic codes within a system with more options of amino acids to be encoded? Therefore, the statistical parameters found here to describe the error in random genetic codes could be applied to the selection of sets of amino acids or to find more appropriated amino acid properties function, so that a few codes could be much more efficient (greater tolerance to error) than the rest, something very appropriate for a natural selection of a genetic code.

Despite the optimization patterns of the standard genetic code, Francis Crick's frozen accident theory still survives when combined with theories of genetic code expansion (Koonin 2017), although it has been said that the emphasis is on the frozen part (Kun and Radvanyi 2018). However, it seems important to consider random events in the earliest stages of the genetic code. Assuming a hypothetical early random scenario for the origin of the genetic code, in this approach the distribution of the error function for the completely random model was mathematically described under very general conditions, which may facilitate subsequent applications.

APPENDIX A

Alternative calculation of $\langle MS_p \rangle_\infty$ for infinite completely random genetic codes with standard stop codons.

From p = 3 into Eq. 5, we obtain

$$\langle MS_3 \rangle_\infty = \langle \frac{1}{|E_3|} \sum_{(ijk,ijk') \in E_3} \left(r_{ijk} - r_{ijk'} \right)^2 \rangle_\infty \tag{A1}$$

Using the Kronecker delta function δ_{xy} (i.e., given that x and y are positive integers, if $x = y$, then δ_{xy}= 1 and else δ_{xy}= 0), Eq. A1 becomes

$$\langle MS_3 \rangle_\infty = \langle \frac{1}{2|E_3|} \sum_{i,j,k,k' \in B} \left(r_{ijk} - r_{ijk'} \right)^2 \left(1 - \delta_{Ui}\delta_{Aj}\delta_{Gk} - \delta_{Ui}\delta_{Aj}\delta_{Ak} - \delta_{Ui}\delta_{Gj}\delta_{Ak} \right) \big(1 - \delta_{Ui}\delta_{Aj}\delta_{Gk'} - \delta_{Ui}\delta_{Aj}\delta_{Ak'} - \delta_{Ui}\delta_{Gj}\delta_{Ak'} \big) \rangle_\infty \tag{A2}$$

whereby

$$\begin{aligned}\langle MS_3 \rangle_\infty = \langle \frac{1}{2|E_3|} \sum_{i,j,k,k' \in B} \left(r_{ijk} - r_{ijk'} \right)^2 \big(1 &- \delta_{Ui}\delta_{Aj}\delta_{Gk'} - \delta_{Ui}\delta_{Aj}\delta_{Ak'} - \delta_{Ui}\delta_{Gj}\delta_{Ak'} \\ &- \delta_{Ui}\delta_{Aj}\delta_{Gk} + \delta_{Ui}\delta_{Aj}\delta_{Gk}\, \delta_{Ui}\delta_{Aj}\delta_{Gk'} + \delta_{Ui}\delta_{Aj}\delta_{Gk}\delta_{Ui}\delta_{Aj}\delta_{Ak'} \\ &+ \delta_{Ui}\delta_{Aj}\delta_{Gk}\delta_{Ui}\delta_{Gj}\delta_{Ak'} - \delta_{Ui}\delta_{Aj}\delta_{Ak} + \delta_{Ui}\delta_{Aj}\delta_{Ak}\, \delta_{Ui}\delta_{Aj}\delta_{Gk'} \\ &+ \delta_{Ui}\delta_{Aj}\delta_{Ak}\, \delta_{Ui}\delta_{Aj}\delta_{Ak'} + \delta_{Ui}\delta_{Aj}\delta_{Ak}\, \delta_{Ui}\delta_{Gj}\delta_{Ak'} - \delta_{Ui}\delta_{Gj}\delta_{Ak} \\ &+ \delta_{Ui}\delta_{Gj}\delta_{Ak}\, \delta_{Ui}\delta_{Aj}\delta_{Gk'} + \delta_{Ui}\delta_{Gj}\delta_{Ak}\, \delta_{Ui}\delta_{Aj}\delta_{Ak'} \\ &+ \delta_{Ui}\delta_{Gj}\delta_{Ak}\, \delta_{Ui}\delta_{Gj}\delta_{Ak'} \big) \rangle_\infty\end{aligned} \tag{A3}$$

Note that

$$\sum_{i,j,k,k'\in B} \left(r_{ijk} - r_{ijk'}\right)^2 \delta_{xi}\delta_{yj}\delta_{zk} = \sum_{i,j,k,k'\in B} \left(r_{ijk} - r_{ijk'}\right)^2 \delta_{xi}\delta_{yj}\delta_{zk'}$$

(A4)

$$\delta_{xy}\delta_{xy} = \delta_{xy} \qquad \text{(A5)}$$

and

$$\delta_{xy}\delta_{zy} = 0 \ \text{ given that } x \neq z \qquad \text{(A6)}$$

Moreover, since that Eq. A1 does not depend of any values of the r_{UAG}, r_{UAA} and r_{UGA}, we conveniently choose the values of those parameters as

$$r_{UAG} = r_{UAA} = r_{UGA} = 0 \qquad \text{(A7)}$$

Applying Eqs. A4-A7 in Eq. A3 results in

$$\langle MS_3\rangle_\infty = \frac{1}{2|E_3|}\langle \sum_{i,j,k,k'\in B} \left(r_{ijk} - r_{ijk'}\right)^2 \left(1 - 2\delta_{Ui}\delta_{Aj}\delta_{Gk} - 2\delta_{Ui}\delta_{Aj}\delta_{Ak} - 2\delta_{Ui}\delta_{Gj}\delta_{Ak}\right)\rangle_\infty$$

(A8)

which is

$$\langle MS_3\rangle_\infty = \frac{1}{2|E_3|}\langle \sum_{i,j,k,k'\in B} \left(r_{ijk} - r_{ijk'}\right)^2 - 4\sum_{k\in B} r_{UAk}^2 - 2\sum_{k\in B} r_{UGk}^2\rangle_\infty$$

(A9)

equivalently

$$\langle MS_3\rangle_\infty = \frac{1}{|E_3|}\langle \sum_{i,j,k,k'\in B} r_{ijk}^2 - \sum_{i,j,k\in B} r_{ijk}^2 - \sum_{\substack{i,j,k,k'\in B \\ (k\neq k')}} r_{ijk}r_{ijk'} - 2\sum_{k\in B} r_{UAk}^2 - \sum_{k\in B} r_{UGk}^2\rangle_\infty$$

(A10)

and this can be written as

$$\langle MS_3\rangle_\infty = \frac{1}{|E_3|}\langle 3\sum_{i,j,k\,\in\,B} r_{ijk}^2 - 2\sum_{k\,\in\,B} r_{UAk}^2 - \sum_{k\,\in\,B} r_{UGk}^2\rangle_\infty - \frac{1}{|E_3|}\langle \sum_{\substack{i,j,k,k'\,\in\,B\\(k\neq k')}} r_{ijk}r_{ijk'}\rangle_\infty \tag{A11}$$

Considering Eqs. A7 and A11, results from left to right in Eq. A11 summations having 61, 2, 3 and 176 amino acidic terms, respectively. Besides, the average of those terms are such that $\langle r_{ijk}^2\rangle_\infty = \langle r_{UAk}^2\rangle_\infty = \langle r_{UGk}^2\rangle_\infty$. Then, we obtain

$$\langle MS_3\rangle_\infty = \frac{(3\text{x}61 - 2\text{x}2 - 1\text{x}3)}{|E_3|}\langle r_{ijk}^2\rangle_\infty - \frac{176}{|E_3|}\langle r_{ijk}r_{ijk'}\rangle_\infty \tag{A12}$$

so that

$$\langle MS_3\rangle_\infty = \frac{176}{|E_3|}\langle r_{ijk}^2\rangle_\infty - \frac{176}{|E_3|}\langle r_{ijk}\rangle_\infty^2 \tag{A13}$$

From $|E_3| = 88$ and Eq. A13 we have

$$\langle MS_3\rangle_\infty = 2\big(\langle r_{ijk}^2\rangle_\infty - \langle r_{ijk}\rangle_\infty^2\big) \tag{A14}$$

Similar to that indicated to obtain Eq. 10 from Eq. 9, in Eq. A14 we replace r_{ijk} and $\langle\ \ \rangle_\infty$ by a_u and $\langle\ \ \rangle_{Aa}$, respectively. Thus, A14 becomes

$$\langle MS_3\rangle_\infty = 2(\langle a_u^2\rangle_{Aa} - \langle a_u\rangle_{Aa}^2) \tag{A15}$$

Similar demonstrations for *p* = 1 and 2 can be developed. Thus we obtain

$$\langle MS_p\rangle_\infty = 2(\langle a_u^2\rangle_{Aa} - \langle a_u\rangle_{Aa}^2) \tag{A16}$$

where *p* = 1, 2, 3, in agreement with Eq. 10.

APPENDIX B

Analytical calculation of σ_p (p = 1, 2, 3) for infinite completely random genetic codes with standard stop codons.

Alternatively, Eq. 6 can be written as

$$\sigma_p^2 = \langle MS_p^2 \rangle_\infty - \langle MS_p \rangle_\infty^2 \qquad \text{(B1)}$$

From p = 3 into Eq B1 we obtain

$$\sigma_3^2 = \langle MS_3^2 \rangle_\infty - \langle MS_3 \rangle_\infty^2 \qquad \text{(B2)}$$

Calculating the term $\langle MS_3^2 \rangle_\infty$, from p = 3 and Eq. 2, it becomes

$$\langle MS_3^2 \rangle_\infty = \frac{1}{|E_3|^2} \langle \left[\sum_{(ijk,ijk') \in E_3} \left(r_{ijk} - r_{ijk'} \right)^2 \right]^2 \rangle_\infty \qquad \text{(B3)}$$

equivalently

$$\langle MS_3^2 \rangle_\infty = \frac{4}{|E_3|^2} \langle \left[\sum_{(ijk,ijk') \in E_3} \left({r_{ijk}}^2 - r_{ijk} r_{ijk'} \right) \right]^2 \rangle_\infty \qquad \text{(B4)}$$

Calculating from Eq. B4

$$\langle MS_3^2 \rangle_\infty = \frac{4}{|E_3|^2} \langle \sum_{\substack{(ijk,ijk'), \\ (lmn,lmn') \\ \in E_3}} \left({r_{ijk}}^2 {r_{lmn}}^2 + r_{ijk} r_{ijk'} r_{lmn} r_{lmn'} - 2{r_{ijk}}^2 r_{lmn} r_{lmn'} \right) \rangle_\infty \qquad \text{(B5)}$$

which is

$$\langle MS_3^2 \rangle_\infty = \frac{4}{|E_3|^2} \langle \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\ \in E_3}} \left({r_{ijk}}^2 {r_{lmn}}^2 + r_{ijk} r_{ijk'} r_{lmn} r_{lmn'} \right) \rangle_\infty \\ - \frac{8}{|E_3|^2} \langle \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\ \in E_3}} \left({r_{ijk}}^2 r_{lmn} r_{lmn'} \right) \rangle_\infty$$

(B6)

On the other hand

$$\langle \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\ \in E_3}} \left({r_{ijk}}^2 {r_{lmn}}^2 + r_{ijk} r_{ijk'} r_{lmn} r_{lmn'} \right) \rangle_\infty \\ = \langle \sum_{\substack{(ijk,ijk')\\ \in E_3}} r_{ijk}^4 \quad + \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\ \in E_3\\ (ijk,ijk') \neq (lmn,lmn')}} {r_{ijk}}^2 {r_{lmn}}^2 + \sum_{\substack{(ijk,ijk')\\ \in E_3}} {r_{ijk}}^2 {r_{ijk'}}^2 \\ + \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\ \in E_3\\ (ijk,ijk') \neq (lmn,lmn')}} r_{ijk} r_{ijk'} r_{lmn} r_{lmn'} \rangle_\infty$$

(B7)

And then, reordering

$$\langle \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\\in E_3}} \left(r_{ijk}{}^2 r_{lmn}{}^2 + r_{ijk}r_{ijk'}r_{lmn}r_{lmn'}\right)\rangle_\infty$$

$$= \sum_{\substack{(ijk,ijk')\\\in E_3}} \langle r_{ijk}^4\rangle_\infty + \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\\in E_3\\(ijk,ijk')\neq(lmn,lmn')}} \langle r_{ijk}^2\rangle_\infty^2 + \sum_{\substack{(ijk,ijk')\\\in E_3}} \langle r_{ijk}^2\rangle_\infty^2$$

$$+ \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\\in E_3\\(ijk,ijk')\neq(lmn,lmn')}} \langle r_{ijk}r_{ijk'}r_{lmn}r_{lmn'}\rangle_\infty$$

(B8)

which is, considering Eq. 3 and Eq. B8,

$$\langle \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\\in E_3}} \left(r_{ijk}{}^2 r_{lmn}{}^2 + r_{ijk}r_{ijk'}r_{lmn}r_{lmn'}\right)\rangle_\infty$$

$$= |E_3|\langle r_{ijk}^4\rangle_\infty + |E_3|^2\langle r_{ijk}^2\rangle_\infty^2 + \sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\\in E_3\\(ijk,ijk')\neq(lmn,lmn')}} \langle r_{ijk}r_{ijk'}r_{lmn}r_{lmn'}\rangle_\infty$$

(B9)

Considering B9 in B6, we obtain

$$\langle MS_3^2 \rangle_\infty = \frac{4}{|E_3|}\langle r_{ijk}^4\rangle_\infty + 4\langle r_{ijk}^2\rangle_\infty^2 + \frac{4}{|E_3|^2}\sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\ \in E_3\\(ijk,ijk')\neq(lmn,lmn')}}\langle r_{ijk}r_{ijk'}r_{lmn}r_{lmn'}\rangle_\infty$$

$$-\frac{8}{|E_3|^2}\sum_{\substack{(ijk,ijk'),\\(lmn,lmn')\\ \in E_3}}\langle {r_{ijk}}^2 r_{lmn}r_{lmn'}\rangle_\infty$$

(B10)

and using constants $\alpha, \beta,$ and γ, Eq. B10 can be written as

$$\langle MS_3^2 \rangle_\infty = \frac{4}{|E_3|}\langle r_{ijk}^4\rangle_\infty + 4\langle r_{ijk}^2\rangle_\infty^2 + \alpha\langle r_{ijk}^2\rangle_\infty\langle r_{ijk}\rangle_\infty^2 + \beta\langle {r_{ijk}}^3\rangle_\infty\langle r_{ijk}\rangle_\infty + \gamma\langle r_{ijk}\rangle_\infty^4$$

(B11)

Similar to that indicated to obtain Eq. 10 from Eq. 9, in Eq. B11 we replace r_{ijk} and $\langle \quad \rangle_\infty$ by a_u and $\langle \quad \rangle_{Aa}$, respectively. Thus, B11 becomes

$$\langle MS_3^2 \rangle_\infty = \frac{4}{|E_3|}\langle a_u^4\rangle_{Aa} + 4\langle a_u^2\rangle_{Aa}^2 + \alpha\langle a_u^2\rangle_{Aa}\langle a_u\rangle_{Aa}^2 + \beta\langle {a_u}^3\rangle_{Aa}\langle a_u\rangle_{Aa} + \gamma\langle a_u\rangle_{Aa}^4$$

(B12)

Replacing Eqs. 10 and B12 into Eq. B2, we obtain

$$\sigma_3^2 = \frac{4}{|E_3|}\langle a_u^4\rangle_{Aa} + 4\langle a_u^2\rangle_{Aa}^2 + \alpha\langle a_u^2\rangle_{Aa}\langle a_u\rangle_{Aa}^2 + \beta\langle {a_u}^3\rangle_{Aa}\langle a_u\rangle_{Aa} + \gamma\langle a_u\rangle_{Aa}^4$$

$$-4(\langle a_u^2\rangle_{Aa} - \langle a_u\rangle_{Aa}^2)^2$$

(B13)

Let a change of variable given by

$$r'_{ijk} = r_{ijk} - \langle a_u \rangle_{Aa} \qquad \text{(B14)}$$

and

$$b_u = a_u - \langle a_u \rangle_{Aa} \qquad \text{(B15)}$$

(defining a new amino acid property)

Eqs. 1, B14 and B15 give

$$r'_{ijk} = b_u \qquad \text{(B16)}$$

Let MS'_3 be an error function calculated for the encoded b_u values of the new amino acid property

$$MS'_3 = \frac{1}{|E_3|} \sum_{(ijk,ijk') \in E_3} \left(r'_{ijk} - r'_{ijk'} \right)^2$$

(B17)

Let σ'_3 be the standard deviation of MS'_3 over infinite random genetic codes

$$\sigma_3'^2 = \langle (MS'_3 - \langle MS'_3 \rangle_\infty)^2 \rangle_\infty \qquad \text{(B18)}$$

(Similar to Eq. 6 with $p = 3$)

Then, considering the b_u values of the new amino acid property results

$$\sigma_3'^2 = \frac{4}{|E_3|} \langle b_u^4 \rangle_{Aa} + 4 \langle b_u^2 \rangle_{Aa}^2 + \alpha \langle b_u^2 \rangle_{Aa} \langle b_u \rangle_{Aa}^2 + \beta \langle {b_u}^3 \rangle_{Aa} \langle b_u \rangle_{Aa} + \gamma \langle b_u \rangle_{Aa}^4$$

$$-4(\langle b_u^2 \rangle_{Aa} - \langle b_u \rangle_{Aa}^2)^2$$

(B19)

(Similar to Eq. B13, but using $\sigma_3'^2$ and b_u instead of σ_3^2 and a_u, respectively)

Replacing $\langle b_u \rangle_{Aa} = 0$ (from B15) in Eq. B19, we obtain

$$\sigma_3'^2 = \frac{4}{|E_3|}\langle b_u^4\rangle_{Aa}$$

(B20)

Finally, from $\sigma_3 = \sigma_3'$ (because $r_{ijk} - r_{ijk'} = r'_{ijk} - r'_{ijk'}$), Eqs. B15 and B20, we have

$$\sigma_3^2 = \frac{4}{|E_3|}\langle(a_u - \langle a_u\rangle_{Aa})^4\rangle_{Aa}$$

(B21)

Similar demonstrations for *p* = 1 and 2 can be developed. Thus, we obtain

$$\sigma_p^2 = \frac{4}{|E_p|}\langle(a_u - \langle a_u\rangle_{Aa})^4\rangle_{Aa}$$

(B22)

where *p* = 1, 2, 3. Whereby the standard deviation is

$$\sigma_p = 2\left[\frac{1}{|E_p|}\langle(a_u - \langle a_u\rangle_{Aa})^4\rangle_{Aa}\right]^{\frac{1}{2}}$$

(B23)

Amino acid	Value of the amino acid property (a_u)	Amino acid property					
		Polar requirement	Hydropathy	Molecular volume .	Isoelectric point	Random (∈ [0, 10])	Binary (∈ {1, -1})
Ala	$a_1 =$	7.0	1.8	31	6.00	7.8	1
Arg	$a_2 =$	9.1	-4.5	124	10.76	5.1	1
Asn	$a_3 =$	10.0	-3.5	56	5.41	4.3	1
Asp	$a_4 =$	13.0	-3.5	54	2.77	5.3	1
Cys	$a_5 =$	4.8	2.5	55	5.07	1.9	1
Gln	$a_6 =$	8.6	-3.5	85	5.65	4.2	1
Glu	$a_7 =$	12.5	-3.5	83	3.22	6.3	1
Gly	$a_8 =$	7.9	-0.4	3	5.97	7.2	1
His	$a_9 =$	8.4	-3.2	96	7.59	2.3	1
Ile	$a_{10} =$	4.9	4.5	111	6.02	5.3	1
Leu	$a_{11} =$	4.9	3.8	111	5.98	9.1	-1
Lys	$a_{12} =$	10.1	-3.9	119	9.74	5.9	-1
Met	$a_{13} =$	5.3	1.9	105	5.74	2.3	-1
Phe	$a_{14} =$	5.0	2.8	132	5.48	3.9	-1
Pro	$a_{15} =$	6.6	-1.6	32.5	6.30	5.2	-1
Ser	$a_{16} =$	7.5	-0.8	32	5.68	6.8	-1
Thr	$a_{17} =$	6.6	-0.7	61	6.16	5.9	-1
Trp	$a_{18} =$	5.2	-0.9	170	5.89	1.4	-1
Tyr	$a_{19} =$	5.4	-1.3	136	5.66	3.2	-1
Val	$a_{20} =$	5.6	4.2	84	5.96	6.6	-1

Table 1. Values of amino acid properties used in this study. Four properties are taken from Haig and Hurst (Haig and Hurst 1991): polar requirement, hydropathy, molecular volume and isoelectric point. The other two amino acid properties are not real, but are arbitrary designations to increase the number of cases to apply the analytical and numerical calculation modes used in this study.

Amino acid property	P	Average of MS_p ($\langle MS_p\rangle_\infty$)		$Error$
		Analytical	*Numerical*	
Polar requirement	1	11.995	12.066	-0.6 %
	2	11.995	12.059	-0.5 %
	3	11.995	12.060	-0.5 %
Hydropathy	1	16.952	17.047	-0.6 %
	2	16.952	17.036	-0,5 %
	3	16.952	17.039	-0.5 %
Molecular Volume	1	3505.4	3522.2	-0.5 %
	2	3505.4	3526.2	-0.6 %
	3	3505.4	3522.4	-0.5 %
Isoelectric Point	1	5.9302	5.9626	-0.5 %
	2	5.9302	5.9674	-0.6 %
	3	5.9302	5.9651	-0.6 %
Random (∈ [0, 10])	1	8.156	8.2001	-0.5 %
	2	8.156	8.2010	-0.5 %
	3	8.156	8.1986	-0.5 %
Binary (∈ {1, -1})	1	2	2.0109	-0.5 %
	2	2	2.0115	-0.6 %
	3	2	2.0096	-0.5 %

Table 2. Analytical and numerical calculations of the average of *MS_p* over infinite completely random genetic codes with standard stop codons ($\langle MS_p\rangle_\infty$). MS_p (Eq. 2) is the error function of the genetic code. $\langle MS_p\rangle_\infty$ is calculated analytically using Eq. 10 and numerically by computational statistics over 100,000 completely random genetic codes with standard stop codons. The values of the amino acid properties are shown in Table 1. The $\langle MS_p\rangle_\infty$ error is defined as $Error = 100\,\%\big(Analytical\ \langle MS_p\rangle_\infty - Numerical\ \langle MS_p\rangle_\infty\big)/Analytical\ \ \langle MS_p\rangle_\infty$. Moreover, in addition to the two artificial properties of the table, more artificial random properties were applied (using 100 set of random properties and the same 1000 random codes for each one). As a result, the averages of $Error$ were equal to -0.5 % (0.6), -0.5 % (0.5) and -0.6 % (0.5), for non-binary random properties between 0 and 10 (∈ [0, 10]), and -0.4 % (0.5), -0.5 % (0.4) and -0.5 % (0.4), for binary random properties (∈ {1, -1}). In both cases, with p = 1, 2 and 3, respectively, and with the standard deviations of the $Error$ in parentheses.

Amino acid property	P	Standard deviation of MS_p (σ_p)		$Error$
		Analytical	*Numerical*	
Polar requirement	1	2.1320	2.1095	1.1 %
	2	2.1199	2.1066	0.6 %
	3	2.1199	2.1099	0.5 %
Hydropathy	1	2.3867	2.3666	0.8 %
	2	2.3731	2.3617	0.5 %
	3	2.3731	2.3636	0.4 %
Molecular Volume	1	568.03	560.83	1.3 %
	2	564.79	563.65	0.2 %
	3	564.79	559.59	0.9 %
Isoelectric Point	1	1.4088	1.3946	1.0 %
	2	1.4008	1.3989	0.1 %
	3	1.4008	1.3985	0.2 %
Random (∈ [0, 10])	1	1.3287	1.3101	1.4 %
	2	1.3211	1.3133	0.6 %
	3	1.3211	1.3169	0.3 %
Binary (∈ {1, -1})	1	0.2144	0.2128	0.8 %
	2	0.2132	0.2108	1.1 %
	3	0.2132	0.2112	0.9 %

Table 3. Analytical and numerical calculations of the standard deviation of MS_p over infinite completely random genetic codes with standard stop codons (σ_p). σ_p is calculated analytically using Eq. 11 and calculated numerically by computational statistics using the same genetic codes for Table 2. The σ_p error is defined as $Error = 100\,\%\left(Analytical\ \sigma_p\ - Numerical\ \sigma_p\ \right)/Analytical\ \ \sigma_p$. Moreover, for the same simulation of codes with additional artificial random properties described in the legend of Table 2, the calculated averages of $Error$ were equal to 0.9 % (2.1), 0.2 % (2.2) and 0.3 % (1.9), for non-binary random properties, and 0.6 % (2.3), 0.7 % (2.4) and 1.0 % (2.3), for binary random properties. In both cases, with p = 1, 2 and 3, respectively, and with the standard deviations of the $Error$ in parentheses.

REFERENCES

Błażej P, Wnętrzak M, Mackiewicz D, Mackiewicz P (2018) Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm PLOS ONE 13:e0201715 doi:10.1371/journal.pone.0201715

Błażej P, Wnętrzak M, Mackiewicz P (2016) The role of crossover operator in evolutionary-based approach to the problem of genetic code optimization Biosystems 150:61-72 doi:10.1016/j.biosystems.2016.08.008

Buhrman H, van der Gulik PT, Kelk SM, Koolen WM, Stougie L (2011) Some mathematical refinements concerning error minimization in the genetic code IEEE/ACM Trans Comput Biol Bioinform 8:1358-1372 doi:10.1109/tcbb.2011.40

Burton AS, Stern JC, Elsila JE, Glavin DP, Dworkin JP (2012) Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites Chem Soc Rev 41:5459-5472 doi:10.1039/c2cs35109a

Cleaves HJ, 2nd (2010) The origin of the biologically coded amino acids J Theor Biol 263:490-498 doi:10.1016/j.jtbi.2009.12.014

Crick FH (1968) The origin of the genetic code J Mol Biol 38:367-379 doi:10.1016/0022-2836(68)90392-6

Freeland SJ, Hurst LD (1998) The genetic code is one in a million J Mol Evol 47:238-248 doi:10.1007/pl00006381

Goldman N (1993) Further results on error minimization in the genetic code J Mol Evol 37:662-664 doi:10.1007/bf00182752

Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code J Mol Evol 33:412-417 doi:10.1007/bf02103132

Haig D, Hurst LD (1999) A quantitative measure of error minimization in the genetic code J Mol Evol 49:708 doi:10.1007/pl00006591

Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor Nat Rev Microbiol 1:127-136 doi:10.1038/nrmicro751

Koonin EV (2017) Frozen Accident Pushing 50: Stereochemistry, Expansion, and Chance in the Evolution of the Genetic Code Life (Basel) 7 doi:10.3390/life7020022

Koonin EV, Novozhilov AS (2017) Origin and Evolution of the Universal Genetic Code Annu Rev Genet 51:45-62 doi:10.1146/annurev-genet-120116-024713

Kun A, Radvanyi A (2018) The evolution of the genetic code: Impasses and challenges Biosystems 164:217-225 doi:10.1016/j.biosystems.2017.10.006

Novozhilov AS, Wolf YI, Koonin EV (2007) Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape Biol Direct 2:24-24 doi:10.1186/1745-6150-2-24

Salinas DG, Gallardo MO, Osorio MI (2016) Local conditions for global stability in the space of codons of the genetic code Bio Systems 150:73-77 doi:10.1016/j.biosystems.2016.08.007

Santos J, Monteagudo Á (2010) Study of the genetic code adaptability by means of a genetic algorithm Journal of Theoretical Biology 264:854-865 doi:https://doi.org/10.1016/j.jtbi.2010.02.041

Schönauer S, Clote P (1997) How optimal is the genetic code? In: Frishman D, Mewes HW, editors. Computer Science and Biology Proceedings of the German Conference on Bioinformatics (GCB'97). Sep 21–24, 1997. p. 65-67. http://clavius.bc.edu/~clote/pub/geneticCode.pdf.

Vetsigian K, Woese C, Goldenfeld N (2006) Collective evolution and the genetic code Proc Natl Acad Sci U S A 103:10696-10701 doi:10.1073/pnas.0603780103

Weiss MC, Preiner M, Xavier JC, Zimorski V, Martin WF (2018) The last universal common ancestor between ancient Earth chemistry and the onset of genetics PLoS Genet 14:e1007518 doi:10.1371/journal.pgen.1007518

Wnętrzak M, Błażej P, Mackiewicz D, Mackiewicz P (2018) The optimality of the standard genetic code assessed by an eight-objective evolutionary algorithm BMC Evol Biol 18:192 doi:10.1186/s12862-018-1304-0

Wnętrzak M, Błażej P, Mackiewicz P (2019) Optimization of the standard genetic code in terms of two mutation types: Point mutations and frameshifts Biosystems 181:44-50 doi:10.1016/j.biosystems.2019.04.012

Zaia DA, Zaia CT, De Santana H (2008) Which amino acids should be used in prebiotic chemistry studies? Orig Life Evol Biosph 38:469-488 doi:10.1007/s11084-008-9150-5