Journal of Theoretical Biology 408 (2016) 237–242

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

The neutral emergence of error minimized genetic codes superior to the standard genetic code

Steven E. Massey

Department of Biology, University of Puerto Rico – Rio Piedras, San Juan, PR 00931, USA

HIGHLIGHTS

- Error minimization may arise from code expansion.
- Genetic codes better than the standard genetic code are easily produced.
- This is a form of self-organization at the coding level.

ARTICLE INFO

Article history:
Received 30 January 2016
Received in revised form
1 August 2016
Accepted 15 August 2016
Available online 17 August 2016

Keywords:
Error minimization
Genetic code
Neutral emergence
Codon reassignment
Gene duplication
Robustness

ABSTRACT

The standard genetic code (SGC) assigns amino acids to codons in such a way that the impact of point mutations is reduced, this is termed 'error minimization' (EM). The occurrence of EM has been attributed to the direct action of selection, however it is difficult to explain how the searching of alternative codes for an error minimized code can occur via codon reassignments, given that these are likely to be disruptive to the proteome. An alternative scenario is that EM has arisen via the process of genetic code expansion, facilitated by the duplication of genes encoding charging enzymes and adaptor molecules. This is likely to have led to similar amino acids being assigned to similar codons. Strikingly, we show that if during code expansion the most similar amino acid to the parent amino acid, out of the set of unassigned amino acids, is assigned to codons related to those of the parent amino acid, then genetic codes with EM superior to the SGC easily arise. This scheme mimics code expansion via the gene duplication of charging enzymes and adaptors. The result is obtained for a variety of different schemes of genetic code expansion and provides a mechanistically realistic manner in which EM has arisen in the SGC. These observations might be taken as evidence for self-organization in the earliest stages of life.

1. Introduction

While much is now known about protein translation, the forces that led to the assignment of the 20 canonical amino acids to the triplet codons of the SGC have proven elusive. One problem has been the lack of extant intermediate codes, and organisms with partially evolved translation systems, meaning that assertions regarding genetic code evolution have largely relied on indirect evidence, or deductive reasoning. A feature of the SGC that may provide the key to understanding its evolution is EM. EM means that physicochemically similar amino acids have a tendency to be assigned to codons that differ by only one nucleotide. Remarkably, the SGC is near optimal for the property compared to large numbers of randomly generated codes (Alff-Steinberger et al., 1969; Freeland et al., 1998, 2000; Gilis et al., 2001; Goodarzi et al., 2004), meaning that amino acids are assigned to codons in such a way that the impact of point mutations leading to amino acid substitutions is reduced. The 'physicochemical' theory (Sonneborn, 1965; Woese et al., 1965; Epstein et al., 1966; Goldberg and Wittes, 1966) proposes that this property was directly selected for, given its beneficial nature. A key problem however is that in order to select an error minimized genetic code, the space of alternative less optimal genetic codes needs to be searched. Given that the number of alternative codes is vast (5.908×10^{45}, calculated by Buhrman et al. (2011)), this implies a heuristic search was necessary in order to traverse code space to find a code with a high degree of EM. However, the physicochemical theory implies that codon reassignments were necessary in order to search for codes with progressively higher levels of EM, which is problematic given that codon reassignments are likely to be disruptive to the proteome. This latter consideration is a key tenet of the Frozen Accident theory of Crick (Crick, 1968), which proposes that the near

E-mail address: stevenemassey@gmail.com

http://dx.doi.org/10.1016/j.jtbi.2016.08.022
0022-5193/© 2016 Elsevier Ltd. All rights reserved.

universality of the SGC is due to the difficulty of changing amino acid – codon assignments due to this disruptive effect. Consistent with this, the occurrence of codon reassignments in some genomes has been linked to reduced proteome size, whereby this effect is reduced, allowing limited reassignments to occur (Massey and Garey, 2007; Massey, 2015) .

Alternatively, a number of workers have proposed that physicochemical similarities between neighboring amino acids in the genetic code could be due to the duplication of charging enzymes during code expansion (Crick, 1968; Calvalcanti et al., 2000; Stoltzfus and Yampolsky, 2007), and using simulation a degree of EM can be shown to neutrally arise via this process, when a similarity threshold is used to add amino acids similar to the parent amino acid, to codons related to those of the parent amino acid (Massey, 2008, 2015) . Here, we make a modification to this procedure by selecting the most similar daughter amino acid to the parent amino acid out of the unassigned amino acids. Remarkably, alternative genetic codes with EM superior to the SGC frequently arise. The result is obtained utilizing different pathways of code expansion and amino acid similarity matrices, and offers a mechanistically realistic explanation for how EM was produced.

2. Methods

The error minimization (*EM*) value is defined as (Massey, 2008):

$$EM = \left(\sum_{n=1}^{61} \sum_{i=1}^{9} \frac{Vc_{ni}}{9} \right) /61 \tag{1}$$

where c is a sense codon, n is the index for the 61 sense codons, i is the index for the 9 codons c_i that are separated from c_n by a single point mutation, Vc_{ni} is the similarity between the amino acids coded for by codon c_n and c_i, obtained from an amino acid similarity matrix. In order to calculate the *EM* value of a code, each sense codon of the code is systematically subjected to all nine potential point mutations, and the resulting amino acid distances of the mutated codon compared to the original codon averaged. Mutations to stop codons are not included in the calculation. In order to calculate the *EM* value with a transition/transversion (Ts/Tv) ratio of 2 (based on the value from the human nucleus (Bainbridge et al., 2011)), a two fold weighting for transitions was used, thus transitions were multiplied by 1.5, while transversions were multiplied by 0.75. While the choice of this value is somewhat arbitrary, due to biochemical principles it seems reasonable that $Ts/Tv > 0.5$ (random) in early life.

Use of amino acid similarity matrices derived from sequence alignments were avoided as they are expected to reflect the structure of the SGC (Di Giulio, 2001). The following amino acid similarity matrices were used: the polar requirement matrix (measures the affinity of amino acids for nucleic acid) (Woese et al., 1966), the Grantham matrix (uses a combination of three parameters, volume, polarity and composition) (Grantham, 1974), the Miyata matrix (uses a combination of three parameters, volume, polarity and relative rate of substitution) (Miyata et al., 1979), the EMPAR matrix (combines topological propensities and polarity) (Rao, 1987), the HDSM matrix (derived from protein structural superimpositions of distant homologs) (Prlic et al., 2000*)* and the EX matrix (measures fitness effects on proteins resulting from amino acid substitution) (Yampolsky and Stoltzfus, 2005). None of these matrices is perhaps ideal, for example the polarity requirement matrix reflects affinities of nucleic acids for amino acids, which may not have been important mechanistically during code evolution, while the EX matrix measures fitness changes resulting from amino acid substitutions in extant proteins which are comprised of the 20 canonical amino acids; this may not necessarily be directly applicable to fitness effects on primordial proteins that were comprised of a subset of the 20 canonical amino acids. Notwithstanding these considerations, a comparison of different matrices allows commonalities to be determined.

Three schemes of genetic code evolution were implemented; the '213' model (Massey, 2006) (similar schemes independently formulated by Higgs (2009) and Francis (2013); Fig. 1a), the ambiguity reduction model (Fitch and Upper, 1987) (Fig. 1b) and a scheme consistent with the coevolution theory (Wong, 1975). The 213 model proposes that the second codon position of the triplet codon became informational first, followed by the first codon position and lastly the third codon position. This reflects the relative levels of redundancy observed between the different codon positions ($2 < 1 < 3$), and the relative fidelity of translation of the anticodon triplet ($2 > 1 > 3$). The ambiguity reduction model involves a gradual acquisition of coding specificity resulting in more precise codon – amino acid mapping during the development of the code. The coevolution theory proposes that amino acids were added to the expanding code as metabolism developed, so that product amino acids were added to codons related to those of the respective precursor amino acids. Hence, the assignment of amino acids to codons is expected to reflect the pathways of amino acid metabolism.

The coevolution theory was modeled as follows. Amino acid precursor and product relationships are as originally described (Wong, 1975): E → Q, E → P, E → R, D → N, D → T, D → K, S → W, S → C, V → L, F → Y, Q → H, T → I, T → M. Notably, the precursors of H, I and M are themselves the products of precursor amino acids, thus there are five original precursors E, D, S, V, F. The 20 codon blocks used for the simulation were as in the SGC. Initially, the five original precursor amino acids (E, D, S, V, F) were randomly assigned to individual codon blocks. Then, the 10 product amino acids (Q, P, R, N, T, K, W, C, L, Y) were added to codon blocks related to the codon block of the respective precursor amino acid if they separated by a point mutation at codon position 1 or 2. This step mimics the addition of amino acids to the evolving genetic code, assuming that they were added to codons related to that of the precursor amino acid, which would result from a process of adaptor (tRNA) gene duplication. Then, the three additional product amino acids (H, I, M), which are products of products of the original five precursor amino acids, were added by the same process of duplication of codon blocks. Lastly, two remaining amino acids (A, G), which do not have precursor amino acids, were randomly added to the remaining two codon blocks.

3. Results and discussion

As expected, the SGC shows a high degree of optimization when compared to randomly generated codes (Table 1, *EM* values calculated using Ts/Tv =2). For the simulations of genetic code expansion, random initial starting amino acids were used for both the 213 and ambiguity reduction schemes (Table 2a and b). Hydrophobic, hydrophilic and the amino acids G,A,D,V (which were those originally described for the model) were also used for the 213 scheme, and a combination of hydrophobic and hydrophilic residues were also utilized as starting amino acids for the ambiguity reduction scheme (Supplementary Tables 1 and 2). Subsequently, amino acids were added to parent amino acids by choosing the most closely related amino acid according to the similarity matrix used, from the unassigned set, adhering to the schemes illustrated in Fig. 1. For each simulation, 10000 repetitions were conducted and an average *EM* value, $\bar{EM}$, calculated. The results show that genetic codes superior to the SGC for the property of EM can easily arise under reasonable biochemical