Orig Life Evol Biosph
DOI 10.1007/s11084-014-9394-1

ORIGIN OF LIFE

Two Perspectives on the Origin of the Standard Genetic Code

Supratim Sengupta · Neha Aggarwal · Ashutosh Vishwa Bandhu

Received: 30 September 2014 / Accepted: 30 September 2014

Abstract The origin of a genetic code made it possible to create ordered sequences of amino acids. In this article we provide two perspectives on code origin by carrying out simulations of code-sequence coevolution in finite populations with the aim of examining how the standard genetic code may have evolved from more primitive code(s) encoding a small number of amino acids. We determine the efficacy of the physico-chemical hypothesis of code origin in the absence and presence of horizontal gene transfer (HGT) by allowing a diverse collection of code-sequence sets to compete with each other. We find that in the absence of horizontal gene transfer, natural selection between competing codes distinguished by differences in the degree of physico-chemical optimization is unable to explain the structure of the standard genetic code. However, for certain probabilities of the horizontal transfer events, a universal code emerges having a structure that is consistent with the standard genetic code.

Keywords Origin · Genetic code · Natural selection · Horizontal gene transfer · Finite population

Introduction

Prebiotic, non-enzymatic synthesis of amino acid oligomers starting from a pool of monomers is extremely difficult to achieve. So far, the attempts to synthesize such oligomers have been restricted to tetramers catalysed by di-peptides like Ser-His (Gorlero et al. 2009). Syntheses of longer polypeptides up to 44-mers have been shown to be possible (Chessari et al. 2006) only by starting from a library of 10-mers

Paper presented at the conference Open Questions on the Origin of Life, July 12–13, 2014, International Institute for Advanced Studies (IIAS), Kizugawa, Kyoto, Japan.

OQOL Q01: How can we make ordered sequences of amino acids and mononucleotides by prebiotic means?

S. Sengupta (✉)
Department of Physical Sciences, Indian Institute of Science Education and Research Kolkata,
Mohanpur 741246, India
e-mail: supratim.sen@iiserkol.ac.in

N. Aggarwal · A. V. Bandhu
School of Computational & Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, India

Published online: 15 January 2015

and using the method of stepwise fragment condensation of the randomly generated 10-mers available in the library. Even in such cases, yields can be low depending upon the conditions, with the synthesized oligomers showing no similarity with known proteins. These results highlight the difficulty of achieving prebiotic synthesis of functionally useful polypeptides. Nevertheless, despite the difficulties associated with prebiotic oligomerization reactions, the GADV hypothesis of the origin of life (Ikehara 2009) presented in this workshop also relies on the emergence of a GADV protein world through random polymerization and pseudo-replication of GADV amino acid chains.

The establishment of the standard genetic code (SGC) led to a remarkably efficient mechanism for producing ordered, functional sequences of amino acids. However, the origin of the SGC remains one of the most challenging problems in molecular evolution. The SGC provides a recipe for synthesizing proteins from DNA sequences. But the pattern of associations between codons and amino acids as specified in the SGC is a consequence of sophisticated molecular machinery that requires proteins as well as RNA. Hence the origin of the SGC poses a classic chicken and egg problem that makes it difficult to resolve. Some interesting hypotheses have been put forward (Wolf and Koonin 2007; Bernhardt and Tate 2010) suggesting a successively advantageous step-wise mechanism for the evolution of the complex molecular machinery responsible for translation starting from functionally simpler components that existed in the RNA world. These rely on the presence of tRNA-like molecules, amino-acylating ribozymes and proto-mRNA's that were the precursors of the modern day tRNA, amino-acyl synthetase and the mRNA molecules.

A primitive code (or codes) may have been a lot simpler than the SGC that organisms use to synthesize proteins. Such codes would have made it possible to produce an ordered sequence of amino acids, albeit with a far smaller amino acid vocabulary than is possible using the SGC. Even a primitive code that encodes a small number of early amino acids requires a mechanism of associating those amino acids with the set of 64 codons. Such an association might have been made possible either by direct stereo-chemical association between codons and amino acids (Yarus 2000) or by a primitive and possibly error prone translation machinery that did not have the complexity and fidelity of the current translation machinery. Nevertheless, even a primitive and error prone mechanism of association between codons and amino acids would facilitate the formation of ordered and possibly functional sequences of amino acids to a significantly greater extent than would have been possible by random ligation of amino acids selected from a diverse collection of monomers.

We expect that the establishment of primitive genetic code(s) marks the first in a series of steps that ends with the emergence of a universal and optimized SGC. The multiple genetic codes, encoding a small number of earliest amino acids (Trifonov 2000; Higgs 2009) would then compete with one another, co-evolve along with the sequences and gradually incorporate more amino acids as and when they became available. As the code expands via incorporation of new amino acids and the translated sequences become more robust to mutational and translational errors due to refinements in the translation machinery, there would be a manifold increase in the number of ordered and functionally diverse sequences. In this manner the primordial evolution of the genetic code from one which encodes a small number of amino acids to the SGC would have greatly facilitated the formation of ordered sequences of amino acids of increasing diversity in a stepwise manner that culminated with the appearance of sequences made up of the 20 biologically encoded amino acids. In this article we will present the results of a study that attempts to explain how a physico-chemically optimized (Woese 1965; Freeland and Hurst 1998) and universal code like the SGC may have emerged as a consequence of competition between finite populations of code-sequence sets.

Model and Results

In the finite population code-sequence co-evolution model (see (Bandhu et al. 2013) for details), competition between a set of primordial codes occurs through the sequences they translate. We do not discuss the nature of the translation machinery but assume that primitive translation machinery existed around the time the earliest primordial genetic codes appeared, perhaps along the lines described in (Bernhardt and Tate 2010). We study the effect of competition between codes encoding different numbers of amino acids ranging from 4 to 10. Following the 4-column theory of genetic code origin proposed by Higgs (Higgs 2009), we start from a scenario in which the earliest code which encoded 4 of the early amino acids (Val, Ala, Asp and Gly) possessed primitive translation machinery that could distinguish only between bases at the second codon position. The code gradually co-evolved in stages along with the translation machinery as newer amino acids were incorporated into the code by taking over codon blocks originally associated with the early amino acids.

Our aim is to address the following questions in the absence and presence of HGT. How does the emergent universal code compare with the SGC? Do sub-optimal codes also get fixed with significant probability? To determine the effect of the composition of codes present in the population on the structure and fixation probability of the emergent universal code, we considered two different sets of competing codes. In one set (the physico-chemically constrained set) the alternative codes had similar levels of physico-chemical optimization. The other (unconstrained) set consists of randomly generated codes along with at least one code that belongs to the physico-chemically constrained set.

Figure 1 shows the fixation probabilities of different codes obtained using a model where *all* the alternative codes in the constrained set are allowed to compete with each other only *after* the sequences associated with each code have achieved mutation-selection equilibrium. We find that several codes with similar levels of physico-chemical optimization (including many sub-optimal codes) have significant fixation probabilities. Figure 2 shows the structure of 2 of those codes that get fixed with the highest fixation probability. The third code is the one that is most consistent with the SGC (labelled CSGC) which got fixed only twice out of thousand trials. Similar results were obtained when we used a model where new codes from the pool were gradually introduced into the population with a fixed probability until all the codes in the pool are sampled and one of them gets fixed. However, the population dynamics in the two models are quite different. In the former, most of the codes present in the original pool are quickly eliminated by the selection process and only a few codes remain in the population after some generations. These codes compete with each other with one gradually increasing its frequency at the cost of others and eventually getting fixed in the population. In the latter model, usually one code is present in the population with a high frequency and many

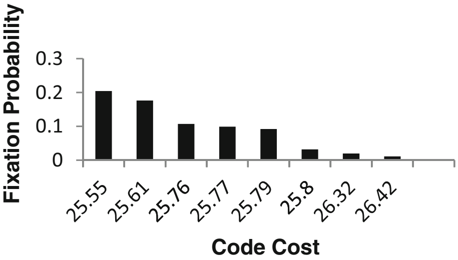
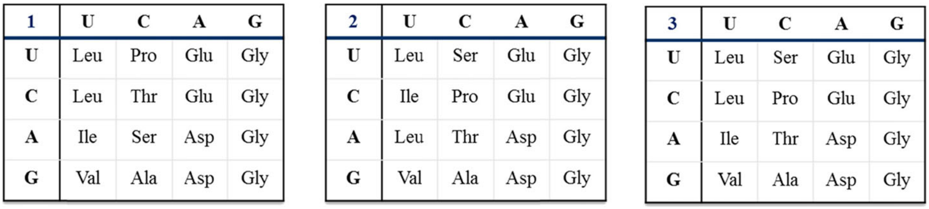


Fig. 1 The fixation probability of codes having the 8 highest fixation probabilities in the constrained set vs code cost. Higher cost implies a less physico-chemically optimized code



1	U	C	A	G
U	Leu	Pro	Glu	Gly
C	Leu	Thr	Glu	Gly
A	Ile	Ser	Asp	Gly
G	Val	Ala	Asp	Gly

2	U	C	A	G
U	Leu	Ser	Glu	Gly
C	Ile	Pro	Glu	Gly
A	Leu	Thr	Asp	Gly
G	Val	Ala	Asp	Gly

3	U	C	A	G
U	Leu	Ser	Glu	Gly
C	Leu	Pro	Glu	Gly
A	Ile	Thr	Asp	Gly
G	Val	Ala	Asp	Gly

Fig. 2 Structures of two of the codes with the highest fixation probabilities and the CSGC. Parameters used: Number of sequences per code (N)=1000, sequence length (L)=732, mutation rate per site (μ)=0.0001, selection coefficient (s)=0.05

other codes are present with low frequencies at any given time. Effects of stochasticity arising from finite population effects are more dominant in the latter model. In both cases, we find that the most optimized code does not get fixed with a substantially larger fixation probability than other codes. More significantly, the structure of the 10-amino acid codes that get fixed with significant probability can differ markedly from CSGC which has a lower level of optimization than many of the codes present in the population. However, if a physico-chemically optimized code (like CSGC) competes with codes from the unconstrained set, it gets fixed with a significantly higher probability than any of the randomly generated and therefore less optimized codes. We conclude that natural selection to increase the levels of physico-chemical optimization could not have been the sole factor in explaining the emergence of the SGC.

A primordial world that existed prior to the establishment of the SGC may have been characterized by leaky protocells allowing for rampant horizontal gene transfer (HGT) between them. Such exchanges may have significantly facilitated the emergence of innovations (Vetsigian et al. 2006) in biological information encoding which eventually culminated in the establishment of the SGC. By allowing HGT to occur between any two sequences translated by same or different primordial codes, we investigated whether HGT can facilitate the emergence of a single universal and physico-chemically optimized code and the effect it has on the structure of the emergent universal code. In addition to sequences undergoing mutations with a fixed rate, HGT between sequences can also occur due to transfer of a sequence segment from a randomly chosen donor sequence to the sequence under consideration (called the acceptor sequence). Figure 3 is a pictorial representation of the rule used for updating the code of the acceptor sequence after a HGT event. A change in code associated with a sequence is accepted only if the fitness of the sequence increases or remains unchanged when it is calculated using the new code. After evolving all the sequences in the population in this fashion, the population is updated by selecting sequences for the next generation from

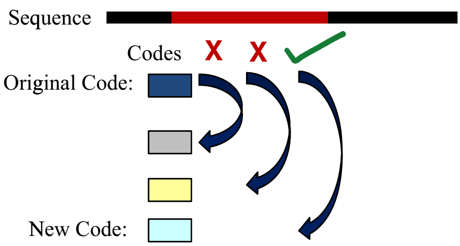


Fig. 3 Update rule used to update the code of a sequence in the model with HGT. The red segment represents a segment that has been horizontally transferred from another sequence. Codes in the population are sampled to update the code of the new sequence. The first two codes sampled are rejected but the third one is accepted and is subsequently used to translate the sequence

sequences in the current generation with a probability proportional to their fitness. Remarkably, we found that for probabilities of HGT above a certain threshold, the universal code that emerges from this evolutionary dynamics has a structure that is consistent with the SGC. A crucial role is played by HGT of the translational components which facilitates sampling of different codes and eventual selection of a code which is better adapted to the sequence. Hence, HGT provides a more efficient mechanism to search for a code that optimizes the fitness of the sequences they translate relative to the target protein. Such a code is characterized by amino acid assignments that are consistent with the SGC. The code-sequence coevolutionary dynamics eventually allows the population of sequences to converge on such a code leading to its fixation.

Conclusions

We have argued that the appearance of a genetic code, even a primitive one encoding a small number of amino acids, offers the best possible scenario for producing multiple copies of an ordered and functional sequence of amino acids. The functional diversity of proteins synthesized by any living cell is due to the 20 amino acid alphabet that is encoded in the SGC. Hence, it is important to understand to what extent the structure of the SGC arose as a consequence of coevolution of code-sequence sets encoding a smaller number of amino acids. Our results suggest that selection to minimize the effect of mutational and translational errors (espoused in the *physico-chemical hypothesis* of code origin) can explain the emergence of the SGC only if unrestricted HGT between leaky protocells is taken into account. In the absence of HGT, many codes with similar levels of physico-chemical optimization can get fixed in the population with significant probability. In such a situation, the emergence of one of those codes can only be attributed to stochastic fluctuations, a conclusion that is reminiscent of Crick's "frozen accident" hypothesis of code origin.

References

Bandhu AV, Aggarwal N, Sengupta S (2013) Revisiting the physico-chemical hypothesis of code origin: an analysis based on code-sequence coevolution in a finite population. Orig Life Evol Biosph 43:465–489. doi: 10.1007/s11084-014-9353-x

Bernhardt HS, Tate WP (2010) The transition from noncoded to coded protein synthesis : did coding mRNAs arise from stability-enhancing binding partners to tRNA ? Biol Direct 5:1–18. doi:10.1186/1745-6150-5-16

Chessari S, Thomas R, Polticelli F, Luisi PL (2006) The production of de novo folded proteins by a stepwise chain elongation: a model for prebiotic chemical evolution of macromolecular sequences. Chem Biodivers 3: 1202–1210. doi:10.1002/cbdv.200690121

Freeland SJ, Hurst LD (1998) The genetic code is one in a million. J Mol Evol 47:238–248. doi:10.1007/PL00006381

Gorlero M, Wieczorek R, Adamala K et al (2009) Ser-His catalyses the formation of peptides and PNAs. FEBS Lett 583:153–156. doi:10.1016/j.febslet.2008.11.052

Higgs PG (2009) A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biol Direct 4:16. doi:10.1186/1745-6150-4-16

Ikehara K (2009) Pseudo-Replication of [GADV] -Proteins and Origin of Life. 1525–1537. 10.3390/ijms10041525

Trifonov EN (2000) Consensus temporal order of amino acids and evolution of the triplet code. Gene 261:139–151. doi:10.1016/S0378-1119(00)00476-5

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

Woese CR (1965) Order in the genetic code. Proc Natl Acad Sci U S A 54:71–75

Wolf YI, Koonin EV (2007) On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. Biol Direct 2:14. doi:10.1186/1745-6150-2-14

Yarus M (2000) RNA-ligand chemistry: a testable source for the genetic code. RNA 6:475–484