IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Finite population analysis of the effect of horizontal gene transfer on the origin of an universal and optimal genetic code

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2016 Phys. Biol. 13 036007

(http://iopscience.iop.org/1478-3975/13/3/036007)

View the table of contents for this issue, or go to the journal homepage for more

Download details:

IP Address: 132.72.237.4
This content was downloaded on 27/05/2016 at 18:49

Please note that terms and conditions apply.

Physical Biology

RECEIVED
29 April 2016

ACCEPTED FOR PUBLICATION
9 May 2016

PUBLISHED
27 May 2016

PAPER

Finite population analysis of the effect of horizontal gene transfer on the origin of an universal and optimal genetic code

Neha Aggarwal[1], **Ashutosh Vishwa Bandhu**[1,2] **and Supratim Sengupta**[3]

[1] School of Computational & Integrative Sciences, Jawaharlal Nehru University, New Delhi—110067, India
[2] Department of Physics, ARSD College, University of Delhi (South Campus), New Delhi—110021, India
[3] Department of Physical Sciences, Indian Institute of Science Education and Research, Kolkata, Mohanpur—741246, India

E-mail: supratim.sen@iiserkol.ac.in

Keywords: genetic code, origin, evolutionary dynamics, horizontal gene transfer, finite population

Supplementary material for this article is available online

Abstract

The origin of a universal and optimal genetic code remains a compelling mystery in molecular biology and marks an essential step in the origin of DNA and protein based life. We examine a collective evolution model of genetic code origin that allows for unconstrained horizontal transfer of genetic elements within a *finite* population of sequences each of which is associated with a genetic code selected from a pool of primordial codes. We find that when horizontal transfer of genetic elements is incorporated in this more realistic model of code-sequence coevolution in a *finite* population, it can increase the likelihood of emergence of a more optimal code eventually leading to its universality through fixation in the population. The establishment of such an optimal code depends on the probability of HGT events. Only when the probability of HGT events is above a critical threshold, we find that the ten amino acid code having a structure that is most consistent with the standard genetic code (SGC) often gets fixed in the population with the highest probability. We examine how the threshold is determined by factors like the population size, length of the sequences and selection coefficient. Our simulation results reveal the conditions under which sharing of coding innovations through horizontal transfer of genetic elements may have facilitated the emergence of a universal code having a structure similar to that of the SGC.

1. Introduction

The standard genetic code (SGC) established prior to the appearance of the last universal common ancestor (LUCA) is nearly universal and the pattern of associations between the 61 sense codons and the 20 amino acids is non-random [1, 2]. Attempts to explain the origin of the genetic code started nearly 40 years ago with the *physico-chemical or adaptive theory* of code origin. The *adaptive* theory is to be distinguished from the *stereochemical theory* [3–7] of code origin that posits that the mapping between codons and amino acids arose originally from the physicochemical affinity between amino acids and nucleotide triplets. According to the former hypothesis, the genetic code evolved to minimize the effects of mutational [8, 9] and translational errors [10–14]. Woese [15] was the first to identify the existence of a pattern in the distribution of amino acids across codons. He pointed out the similarity in polarity values encoded by amino acids in the first and second column of the SGC naturally leads to reduction in translational errors arising from the replacement of an amino acid by another belonging to the same column. This qualitative feature was first quantified [1, 2] by defining a cost function associated with different code structures characterized by distinct patterns of associations between codons and amino acids. The code cost function is defined as $\Phi = \sum_i \sum_j F_i p_{ij} g(a_i, a_j)$ where F_i denotes the frequency of the i'th codon, p_{ij} is the probability that the codon i is mistranslated as the codon j; a_i is the amino acid associated with codon i; a_j is the amino acid associated with codon j and $g(a_i, a_j)$ is the cost of replacing amino acid a_i by amino acid a_j. F_i is given in terms of the frequency $P(a_i)$ of the amino acid a_i and the number of codons $n(a_i)$ associated with that amino acid in the code through the relation $F_i = P(a_i)/n(a_i)$ where $P(a_i)$ is determined from the

© 2016 IOP Publishing Ltd

mean fraction of each amino acid in coding sequences of modern organisms [16].

The physico-chemical hypothesis then posited that selection acts to minimize the cost function since codes having lower costs are better optimized to withstand the effect of translational and mutational errors. Following the work of Hurst and collaborators [1, 2, 17], several studies that were robust to refinements in cost function [18–20], the amino-acid substitution matrices [16, 21] and incorporation of nonsense mutations [22] provided strong support for the physico-chemical hypothesis by finding that the SGC is highly optimized relative to many other non-canonical codes. A common theme in all these studies was to advocate the primacy of code cost in the fixation of a code with a more physico-chemically optimized (less costly code) always getting preferentially fixed. However, all such studies were based on infinite population models and did not take into account the effect of fluctuations arising from finite population size on code fixation probability.

Several studies have also questioned the extent to which selection acting to minimize code cost can explain the structure of the genetic code. An early study [23] used genetic algorithms to find codes that were optimized according to a variety of physico-chemical as well as structural criteria (such as code redundancy, codon mutability etc) and compared them to real codes. They found that codes that were similar to real codes were obtained by optimizing structural and not physico-chemical properties. Analysis of robustness [24] of non-canonical codes as well as those codes that differ from the SGC by only one codon assignment have found many codes that are better adapted to mutational errors than the SGC. In a recent work [25], we have shown that in finite populations, the physico-chemical hypothesis of code evolution is insufficient to explain the emergence of the SGC since codes which are slightly sub-optimal can also get fixed in the population with significant probability. Our results indicate that natural selection acting on a collection of codes with similar levels of optimality is incapable of discriminating between such codes. Hence alternatives to the physico-chemical hypothesis need to be explored in order to understand emergence and universality of the SGC.

The *co-evolution theory* [26–29] which provides one such alternative has been extensively studied and refined by Di Giulio and collaborators [30–34]. According to this theory, the pattern of codon amino acid associations observed in the SGC was determined by the metabolic pathways [35] of amino acid biosynthesis (see also [36] for an alternative theory of genetic code origin). Initially, the code encoded only a few early amino acids called the precursor amino acids that could be synthesized abiotically. Subsequently, as the late (product) amino acids were synthesized from the precursor amino acids, the latter ceded some of its codon blocks to the former, thereby increasing the encoding capability of the code. Di Giulio [37] has also argued that the predictions of the co-evolution theory is consistent with the optimal character of the SGC.

Codon reassignments leading to alternative genetic codes (AGCs) in the post-LUCA phase have also been observed in many mitochondrial genomes and in a few nuclear and bacterial genomes [38, 39]. The factors affecting codon reassignments in the pre-LUCA and post-LUCA phase [40] are quite different and have been discussed extensively in [41].

Most of the literature on genetic code origin is based on models which assume vertical descent even though the importance of HGT on the evolution of the earliest life forms has been pointed out by several researchers [42–44]. An alternative model of code evolution was proposed in an insightful paper by Vetsigian *et al* [45] where they argued that a primordial world was most likely dominated by rampant transfer of genetic elements across leaky protocells. The high degree of horizontal transfer renders a communal character to the evolutionary process and makes it impossible to trace back a common ancestry. It therefore becomes imperative to understand the evolution of a community of sequences and codes.

Vetsigian *et al* [45] described a general scenario in which different code communities (termed as 'innovation pools') coexist in a population. HGT occurs between sequences within each code community (strong HGT) as well as between sequences belonging to different code communities albeit with a lesser probability (weak HGT). Such transfer of genetic elements is instrumental in giving rise to new innovations in coding protocols. It was argued that such an evolutionary dynamics would eventually lead to the dominance of the code community that had access to the largest number of coding innovations at the cost of other less innovative code communities. Further competition within this code community expedited by strong HGT would lead to the emergence of a single and perhaps ambiguous code because of ambiguity in codon amino acid associations. Such ambiguous codes result in the synthesis of statistical proteins since the same coding sequence can give rise to multiple amino acid sequences because codons are ambiguously associated with two or more amino acids. As primordial organisms evolved greater levels of complexity, ambiguity in the meaning of codons would be disadvantageous and ambiguous codes would be gradually replaced by an unambiguous, optimal and universal genetic code. This type of communal evolutionary dynamics in which competition between different code communities distinguished by different degrees of coding innovations brought about largely via HGT is characteristic of Lamarckian evolution. The transition from Lamarckian to Darwinian evolution is characterized by increasing levels of organismal complexity which necessitates significant reductions in HGT, the elimination of coding

ambiguity and culminates in the appearance of the LUCA thereby marking the onset of vertical descent.

Vetsigian *et al* [45] used an *infinite* population model of code-sequence coevolution that was originally developed by Ardell and Sella [46, 47] to explore a simplified version of the above scenario where HGT is allowed between any two codes with equal probability regardless of their degree of similarity. They were able to show that optimality and universality of the genetic code emerges as a consequence of the communal evolution characterized by HGT. However, they did not explore the role of various parameters in shaping the co-evolutionary dynamics and the structure of the universal code that emerges in the presence of HGT.

We envisage a similar scenario where genetic elements such as genes as well as components of the translation machinery can be freely exchanged among a community of sequences and code. Such HGT events can induce a sequence to change its code in order to better adapt to the changed genome. Change in code is facilitated through HGT of translational components which can also be freely exchanged between members of the community. We have developed a more realistic *finite* population model of code-sequence co-evolution that enables us to explore in greater detail the consequences of communal evolution of code-sequence sets in the presence of unconstrained HGT and allows us to explicitly ascertain the structure of the code that emerges from the competition between several code-sequence sets.

In our simulations, we have considered population sizes that differ by an order of magnitude with the smallest population having approximately a thousand sequences. In finite populations, fluctuations in quantities like the mean fitness and optimum code fitness (see 'Results' section for definition) would vary as $O(1/\sqrt{N})$ where N is the population size. Moreover, the stochastic nature of code-sequence co-evolution results in more than one code getting fixed with a finite probability. As, N increases such fluctuations would decrease and the evolutionary dynamics would eventually become deterministic with only one code getting fixed with a probability close to unity. It is difficult to predict what a realistic population size would be for protocells that were present during a primordial epoch prior to the existence of LUCA. The population sizes we have chosen clearly highlight finite population effects on the code-sequence co-evolutionary dynamics.

Even though there are estimates of HGT rates in prokaryotes [48], such values are not relevant in the primordial epoch that we focus on. This is because the HGT rates in the pre-LUCA and post-LUCA epoch is likely to be quite different. Vogan and Higgs [44] have argued that modern (post-LUCA regime) organisms should evolve to reduce HGT rates that depend on rates of gene loss during genome replication. As replication fidelity increases, HGT becomes less favourable and HGT rates gradually decrease in contrast to the pre-LUCA phase where gene-loss rates are expected to be high and therefore higher HGT rates would aid in the rapid spread of useful genomic innovations. This suggests that HGT is likely to play a significant role only during the pre-LUCA epoch of genetic code evolution.

In this paper we investigate how our conclusions regarding the emergence of an optimal and universal code resulting from competition between a set of primordial code-sequence sets, are affected when the probability of horizontal gene transfer events, selection coefficient, the initial number of sequences associated with each code and the length of the sequence are varied.

We find that for a range of values of the probability of horizontal gene transfer ($P_{\rm hgt}$), the ten amino acid code that gets fixed in the population with the highest probability is the one that is most consistent with the SGC. However, that is no longer the case when $P_{\rm hgt}$ falls below a threshold value that depends on other parameters in the model. In the latter case, several codes have similar fixation probabilities. By modifying the code-update rule after an HGT event, we further examined how a lower frequency of transfer of translational components affect the threshold value of $P_{\rm hgt}$ and the structure of the code that gets fixed in the population. We also study the effect of changing the benchmark code (that determines the fitness of a sequence) on the structure of the emergent code.

2. Model

We first specify a set ($N_{\rm C}$) of primordial codes that compete with one another through the sequences they translate. A plausible pool of codes was most likely constrained by the nature and diversity of the translational machinery. It therefore seems reasonable to believe that a primordial pool of codes would differ in few codon assignments due to small changes in the nature of the translation machinery. Moreover, the set of primordial codes is selected based on the assumption that only a few amino acids were encoded in a very primitive code which gradually evolved from being able to distinguish a single base at the second codon position (4-column code) to being able to distinguish between bases at the second as well as at the first codon position. The pool is primarily constrained by the physico-chemical similarity observed between amino acids in the first and second columns. New codes are generated from the 4-column code via reassignment of one or more synonymous codon blocks to a previously unassigned amino acid. The rationale for choosing this set is discussed in detail in [25]. Here we summarize the constraints that were imposed to obtain these of alternative codes. Reassignments do not occur in codon blocks whose amino acid assignments are consistent with the SGC. A 4-codon block can be reassigned to another amino acid only if its physico-

chemical properties are similar to the original amino acid. Reassignments can occur only within the set of ten earliest amino acids [49, 50] since we consider the early phase of code evolution where the code encoded a limited number of amino acids not larger than ten. In generating alternative codes, we do not change the assignment of the third or fourth columns since we assume the amino acid composition of these columns primarily involve the ten late stage amino acids which were incorporated after the translation apparatus acquired the capability of distinguishing between purines and pyrimidines at the third position. The third column is therefore assigned to Glu (first eight codons) and Asp (last eight codons) both of which belong to the set of ten earliest amino acids. A 4-codon block that is reassigned once cannot be reassigned later back to its original meaning and can only be reassigned to a new amino-acid. These constraints lead to a set of 216 codes characterized by similar levels of physicochemical optimization. While this set may not be exhaustive, it seems plausible to assume that the set is representative of a pool of codes that may have existed in certain ecological niches in a primordial world.

A finite set of sequences ($N_{\mathrm{seq/C}}$) is associated with each code with the number being same for each code initially. The total number of sequences in the population is $N = N_{\mathrm{C}} \times N_{\mathrm{seq/C}}$. The initial population is made up of identical DNA sequences of length L containing a variety of sense codons. Each sequence undergoes mutation with a fixed rate μ per site per generation. The Jukes–Cantor model is used as a model for base substitutions. The number of mutations is determined from a Poisson distribution having a mean of μL. The fitness of a sequence (and hence indirectly, the fitness of a code) is estimated by comparing the amino acid sequence obtained by translating the RNA sequence according to the associated code to the target protein [51] obtained by translating the original RNA sequence according to the SGC.

As mentioned earlier, a cost can be associated with each code independent of sequences it translates and has been previously used to determine the optimization level of a code in infinite population models. Since, we are using codes with less than 20 amino acids, such a code cost was calculated using the function $\Phi = \sum_{\alpha}\sum_{i}\sum_{j} P(\alpha)\phi_i(\alpha)p_{ij}g(\alpha, a_j)$ proposed by Higgs [19], where α labels the site type (α runs over all the 20 biological encoded amino acids) and $\phi_i(\alpha)$ gives the frequency of the i'th codon at sites of type α. If we suppose that in the absence of the amino acid associated with a site-type α, the genome uses the best possible alternative amino acid available $B(\alpha)$ at that site, (i.e. $B(\alpha)$ is the amino acid with minimum $g(\alpha, B(\alpha))$ value), and synonymous codons occur with equal frequency; $\phi_i(\alpha) = \delta(a_i, B(\alpha))/n(a_i)$; where the δ-function is 1 if the two arguments are equal and is 0 otherwise. $g(a.b)$ values are calculated using the relation $g_{ab} = \sqrt{\sum_i w_i (z_{ia} - z_{ib})^2}$ where the sum is over amino acid properties, w_i is the weight associated with the i'th amino acid property (taken from table 1 of [19]) and z_{ia} and z_{ib} are the normalized values of the i'th property of amino acids a and b.

In code-sequence coevolution models, the code cost is not an appropriate measure of code fitness which depends on the fitness of the sequences a code translates. In our simulations, we calculate the fitness of a sequence by comparing it with a reference protein. We assume that the fitness of a sequence is given by the product of the fitness of individual codons making up the sequence. The fitness of the ith sequence is given by $w(i) = \prod_{k=1}^{L/3}(1 - sg(a_k, b_k))$ where s is the selection coefficient, a_k and b_k are amino acids associated with the k'th codon in the reference protein and the evolving sequence respectively. The values of $g(a.b)$ were taken from Higgs [19].

To set up the initial conditions for subsequent evolution in the presence of HGT, the sequences associated with each code are first allowed to attain mutation–selection equilibrium without having to compete with other code-sequence sets. After equilibration, the equilibrated sets of sequences associated with each code are allowed to compete with each other while evolving by undergoing mutations as well as HGT with the rate P_{hgt} per sequence per generation. HGT was implemented by randomly selecting a donor sequence from the population, taking a randomly selected segment from the donor sequence and pasting it at a randomly selected position in the acceptor sequence after removing a segment of equal length from the acceptor sequence. This manner of implementing HGT was adopted to ensure that the total length of the sequence remain unchanged after HGT. During the collective evolution epoch in a primordial world, sequence segments as well as translational components could be freely transferred between leaky protocells allowing them to explore diverse coding strategies. Hence, following the procedure adopted in [45], the genetic code of the acceptor sequence was then changed by picking a code at random from the code population. The change was accepted if the fitness of the acceptor sequence calculated using the new code either matches or is greater than its original fitness calculated using the old code associated with it. If the fitness of the former is less than the latter, the new code is rejected; another code is randomly selected from the population with the process being repeated until all the codes in the pool have been sampled without successful acceptance of any new code. In each generation, this cycle is repeated for every sequence in the population that undergoes HGT. Each code-sequence pair in the population is then made to compete with each other, with the entire population being updated every generation by selecting sequences with a probability proportional to their fitness. The process continues until a code gets fixed in the population. Fixation of a code implies that all sequences in the population are translated using that code. The fixation probability of a

particular code is then calculated by running the simulation for N_T trials. In order to highlight the fact that code fixation probability is not inversely correlated with code cost, we have provided the code cost (independent of sequences) along with the fixation probability in various cases. The algorithm for the model is given in appendix 1 of the 'supplementary material' file. The effect of changing the code update rule on the code fixation probability was also investigated. To do so, we used a more constrained update rule after an HGT event. According to this update rule, a code change is attempted once per HGT event failing which the original code associated with the sequence is retained.

3. Results

We first discuss the case where the SGC serves as the benchmark code since it is used to translate the set of equilibrated sequences that serve as the benchmark RNA sequences to give the benchmark protein sequences. By varying $P_{\rm hgt}$ and selection coefficient (s) for sequences of different lengths (L), we study the effect of these factors on the emergent code structure. We have carried out simulations for several different values of the above parameters. The figures and tables provided are a representative sample. Table 1 below gives the cost of the codes that get fixed in the population with highest fixation probabilities for three different probabilities of HGT when sequences of length $L = 183$ nucleotides are used. The cost of a code is calculated according to the prescription provided by Higgs [19] for codes encoding less than 20 amino acids. Figure 1(a) shows the variation in the mean fitness of the population for a particular trial; figures 1(b)–(d) gives the fluctuations in the frequency of different codes present in the population (indicated by different colours) for three different values of $P_{\rm hgt}$ and figures 1(e) and (f) shows the change in the optimum code fitness over time for two different values of $P_{\rm hgt}$. We define the optimum code as the one whose associated sequences have a mean fitness that is larger than the mean fitness of sequences associated with any other code in the population in that generation. The structure of the codes that get fixed in the population with non-zero fixation probabilities for the parameter set of table 1(a) is shown in figure 2. We define the critical threshold of $P_{\rm hgt}$ as one for which the probability of fixation of CSGC is at least twice as large as the next highest fixation probability. A caveat worth noting is that this definition is somewhat ad hoc and the critical threshold will change with the definition for finite population simulations. Nevertheless, it is necessary to specify a criterion for identifying the critical threshold in finite population simulations. For the parameter set specified in table 1, the critical threshold of $P_{\rm hgt}$ is $P_{\rm hgt}^{\rm C} = 0.005$. At values of $P_{\rm hgt}$ equal to or larger than this threshold, we find that the code that gets fixed with a significantly higher probability than all other codes in the population is the one that is most consistent with the SGC (see figure 2 for the structure of the 10 amino-acid code that is most consistent with the SGC that is henceforth referred to as CSGC). Incidentally, the code with second highest fixation probability is also quite similar to CSGC and differs from CSGC only in mapping of codons AUN to amino acid leucine instead of isoleucine. As $P_{\rm hgt}$ decreases below the threshold, the probability of fixation of CSGC decreases and eventually CSGC no longer gets fixed with the highest probability.

As indicated by figure 1(a), the code-sequence population coevolves to gradually increase the mean fitness of the sequences in the population. Although the mean fitness sharply increases in the first few generations as the majority of codes with low fitness values are out-competed by the ones with greater fitness's, the rate of increase in the average fitness later becomes

Table 1. The code cost versus fixation probabilities for the codes with highest fixation probabilities for the parameter values: $s = 0.2$, $\mu = 0.001$, $L = 183$, $N_{\rm seq/C} = 5$, $N_T = 1000$, $N_{\rm C} = 216$. (a) $P_{\rm hgt} = 0.02$ (b) $P_{\rm hgt} = 0.005$ (c) $P_{\rm hgt} = 0.003$. For the cases (a) the non-zero fixation probabilities are given while for case (b) and (c) six highest fixation probabilities are shown. CSGC corresponds to the code in row 1 of 1(a), 1(b) and 1(c). Code labels are used only for those codes that correspond to one of the code structures shown in figures 2 and S3 (see supplementary material file).

Code label	Code cost	Fixation probability
CSGC	25.76	0.785
A	26.42	0.196
B	25.77	0.011
C	25.79	0.004
D	25.89	0.001
E	26.93	0.001
F	26.44	0.001
G	27.63	0.001

(a)

Code label	Code cost	Fixation probability
CSGC	25.76	0.550
A	26.42	0.248
J	25.80	0.091
B	25.77	0.036
	27.74	0.033
D	25.89	0.018

(b)

Code label	Code cost	Fixation probability
CSGC	25.76	0.305
J	25.80	0.255
A	26.42	0.157
	27.74	0.135
D	25.89	0.049
B	25.77	0.047

(c)

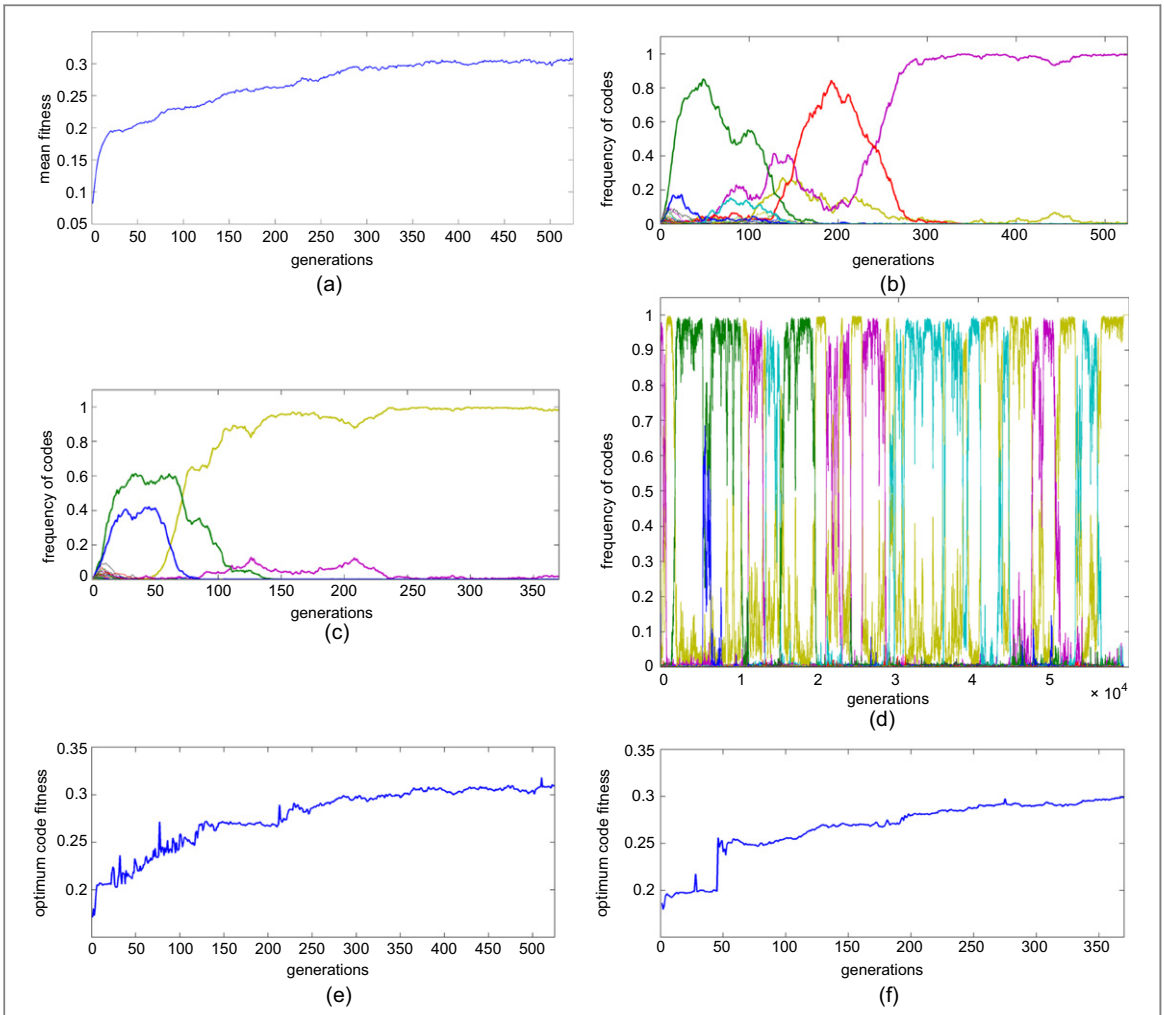

Figure 1. (a) Variation in the mean fitness of the population with time. (b) Frequency of the different codes (indicated by different coloured lines) in the population; for a single trial. The code indicated by the purple line eventually gets fixed in the population. (c) Same as in (b) but with $P_{hgt} = 0.003$. (d) Same as in (b) but with $P_{hgt} = 0.1$. (e) Variation of the optimum code fitness with time for $P_{hgt} = 0.02$ and (f) $P_{hgt} = 0.003$. Other parameters used: $s = 0.2$, $\mu = 0.001$, $L = 183$, $N_{seq/C} = 5$, $N_C = 216$, $P_{hgt} = 0.02$.

more gradual as the few codes remaining in the population (see figure 1(b)) compete and try to optimize their sequence fitness. The mean fitness eventually saturates when a code gets fixed in the population and subsequent fluctuations in the mean fitness can be attributed to sequences better adapting to the universal code that emerges from the competition. This trend is also roughly followed by the optimum code fitness (see figure 1(e)) albeit with a lot of fluctuations. These fluctuations occur when the optimum code changes over generations or when new sequences are added to be translated by the optimum code. We find that the optimum code as defined earlier is a more appropriate metric for understanding primordial code evolution than the code having the least cost which has been typically used in all previous infinite population analysis of standard code origin.

When P_{hgt} is increased while keeping the number of sequences per code fixed, the rate at which new codes appear in the population increases. Many of these codes are quickly eliminated in the competition with other codes leaving a few to compete with each other without any getting fixed in the population. This is manifest through large fluctuations in the frequencies of these codes (compare figure 1(d) with (b) and (c)) exhibiting shifting dominance of these few codes in the population over time. Eventually, one of those codes (represented by the olive green colour) gets fixed in the population. However, when HGT rates are high, the fixation of a code can be short-lived since another new code can appear in the population due to HGT subsequent events.

For probabilities of HGT that fall below the critical threshold, CSGC is no longer the code with the highest fixation probability. For such low values of P_{hgt}, HGT does not have a significant impact on shaping the outcome of code-sequence coevolution (see figure S1 of 'supplementary material' for the structures of the codes that gets fixed in this case). The result is consistent with our analysis of code origin in the absence of HGT [25] where we found that it becomes difficult to distinguish between codes with similar levels of physico-chemical optimization.

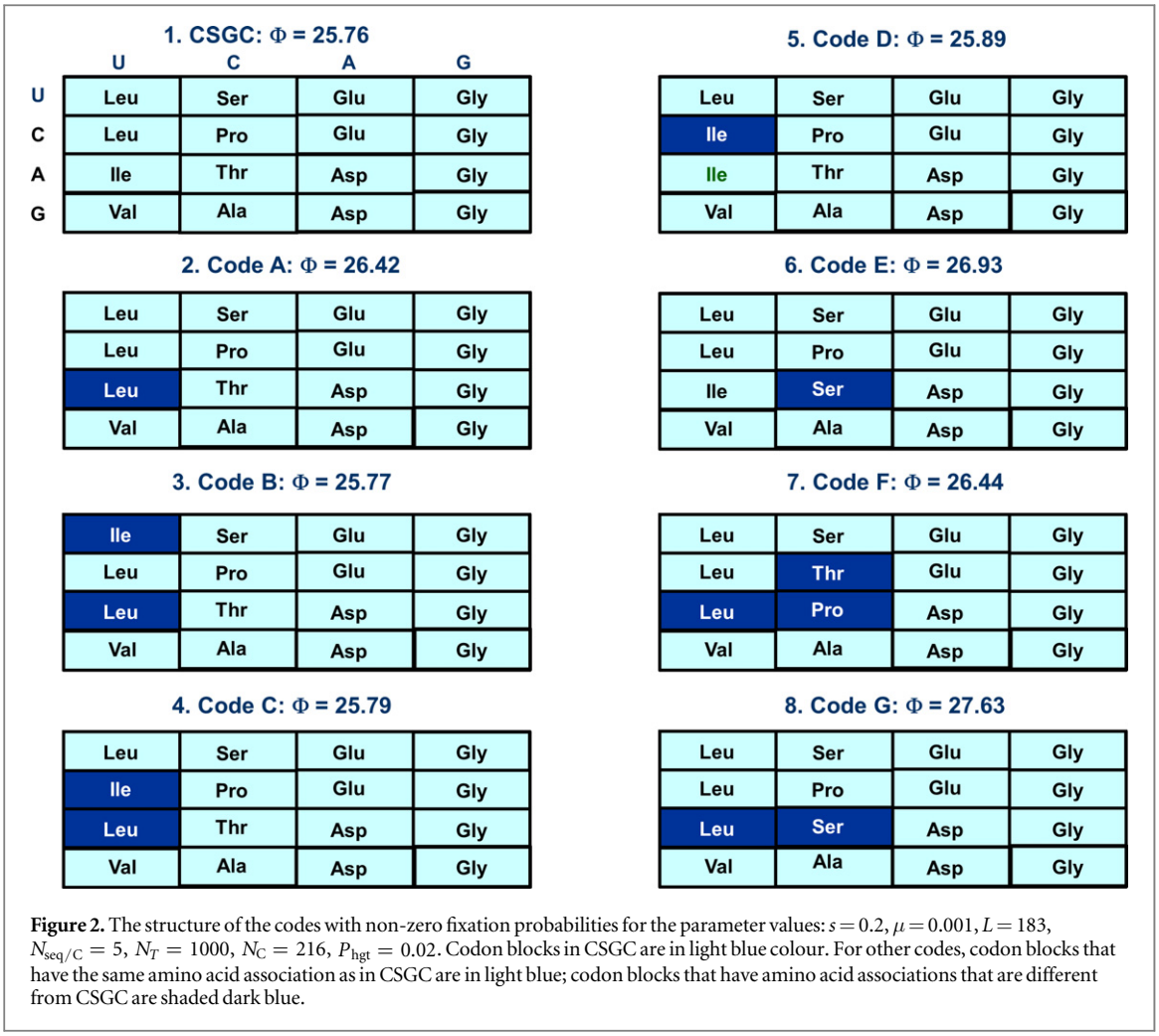

Figure 2. The structure of the codes with non-zero fixation probabilities for the parameter values: $s = 0.2$, $\mu = 0.001$, $L = 183$, $N_{seq/C} = 5$, $N_T = 1000$, $N_C = 216$, $P_{hgt} = 0.02$. Codon blocks in CSGC are in light blue colour. For other codes, codon blocks that have the same amino acid association as in CSGC are in light blue; codon blocks that have amino acid associations that are different from CSGC are shaded dark blue.

Table 2. The code cost versus fixation probabilities for the codes with six highest fixation probabilities for the parameter values: $s = 0.2$, $\mu = 0.001$, $L = 183$, $N_{seq/C} = 100$, $N_T = 1000$, $N_C = 216$, $P_{hgt} = 0.001$. Code labels correspond to code structures shown in figures 2 and S3 (see supplementary material file).

Code label	Code cost	Fixation probability
CSGC	25.76	0.856
A	26.42	0.100
B	25.77	0.038
C	28.79	0.004
J	25.80	0.002

When the selection coefficient is lowered, the diversity of codes in the population increases since lower fitness sequences associated with some codes are retained in the population for a longer time (see figure S2 in 'supplementary material'). Eventually, sequences associated with only a few codes survive with neither one of those possessing a sufficient selective advantage over the others to get fixed in the population. This can be attributed to the close similarity in the structures of those surviving codes.

Increasing the number of sequences per code ($N_{seq/C}$) in the population does not alter the nature of results or the structure of the codes with the highest fixation probabilities but facilitates the process for a lower probability of HGT as can be seen in table 2 and figure S3 of 'supplementary material'. Figures 3(a) and (b) shows the time variation of frequency of codes for a single trial for two different values of P_{hgt}. Increase in $N_{seq/C}$ largely reduces the fluctuations in the time variation of the mean fitness and optimum code fitness (figures 3(c) and (d)) due to the increase in the total population size.

For a fixed mutation rate per site per generation, longer sequences are prone to accumulate more errors and a smaller selection coefficient ensures that sequences with a greater diversity in fitness are tolerated in the population. In such a situation, HGT is limited in its ability to ensure that the population efficiently converges on the code that optimizes the fitness of the associated sequences relative to the benchmark sequence. This is evident from the results of our simulations using sequences of length $L = 732$. Table 3 shows the cost and associated fixation probabilities of codes for two different values of selection coefficient. Even though CSGC has the highest fixation probability when $s = 0.05$, its value is comparable to a few other alternative codes (figure S4 in the 'supplementary material' gives the structures of those codes).

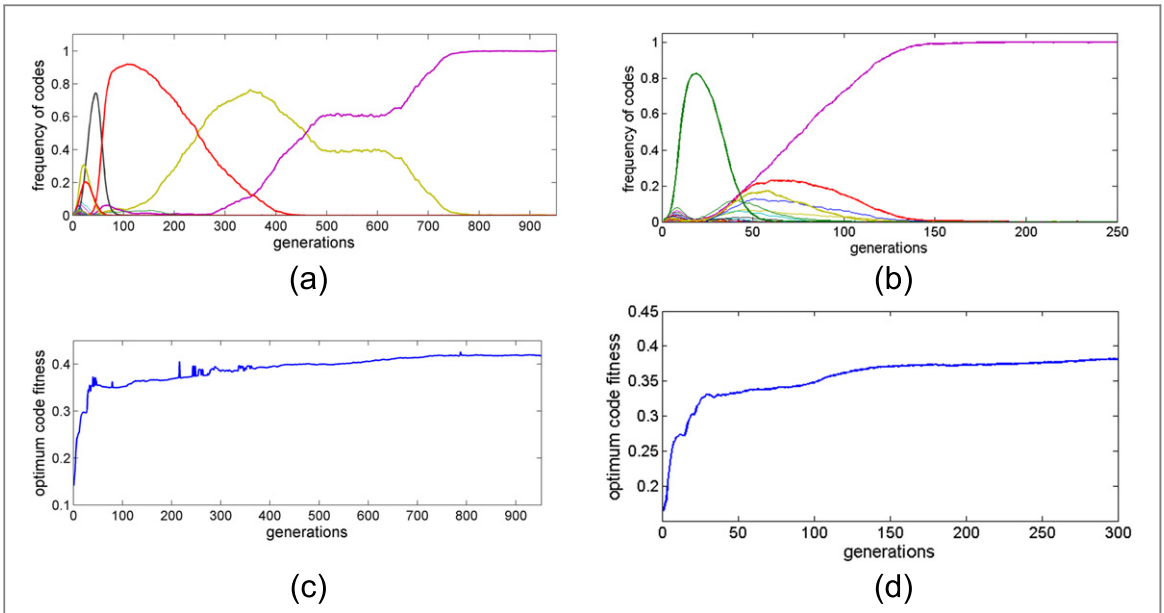

Figure 3. Frequency of the different codes (indicated by different coloured lines) in the population; for a single trial for (a) $P_{hgt} = 0.001$ (b) $P_{hgt} = 0.02$. Though the other codes remain for a long time the one indicated by the purple line in (a) and (b) eventually gets fixed in the population. (c) Variation of the optimum code fitness with time for $P_{hgt} = 0.001$ and (d) $P_{hgt} = 0.02$. Other parameters used: $s = 0.2$, $\mu = 0.001$, $L = 183$, $N_{seq/C} = 100$, $N_C = 216$.

Table 3. The code cost versus fixation probabilities for the codes with six highest fixation probabilities for the parameter values: $\mu = 0.001$, $L = 732$, $N_{seq/C} = 5$, $N_T = 1000$, $N_C = 216$, $P_{hgt} = 0.01$ (a) $s = 0.05$ (b) $s = 0.02$. For case (b) the non-zero fixation probabilities are given while for case (a) six highest fixation probabilities are shown. Code labels correspond to code structures shown in figures 2 and S4, S5 (see supplementary material file).

Code label	Code cost	Fixation probability
CSGC	25.76	0.434
A	26.42	0.403
K	28.42	0.099
L	28.44	0.012
M	28.52	0.012
N	28.54	0.010

(a)

Code label	Code cost	Fixation probability
O	29.13	0.918
CSGC	25.76	0.053
A	26.42	0.012
P	28.42	0.008
Q	27.13	0.008
R	28.42	0.001

(b)

When s is reduced to 0.02, the effect of HGT is no longer evident and consequently CSGC has a very low fixation probability compared to a seven amino-acid code that has the highest fixation probability (figure S5 in the 'supplementary material' gives the structures of these codes). Under such circumstances, the stochastic nature of the evolutionary dynamics can occasionally allow a sub-optimal code to get fixed with a significantly high probability since a low HGT rate reduces the likelihood of fitter code variants from emerging and competing with the existing codes.

The benchmark code affects the fitness of a sequence since it is used to obtain the target proteins (that are adapted to the benchmark code) against which the evolved amino acid sequences are compared to ascertain their fitness. We therefore decided to explore the effect of changing the benchmark code on the structure of the emergent universal code. We used another benchmark code called the AGC that was obtained by reshuffling the amino acid associated with codons within the same column of the SGC. In this case, we found that for values of P_{hgt} above a threshold, the code fixed with the highest probability is the one most consistent with AGC, labelled as CAGC (see figure S6 in 'supplementary material'). The codes with the five highest fixation probabilities are given in figure S7 of 'supplementary material'. Table S1 of 'supplementary material' shows how the fixation probability varies with code cost in this case. For values of P_{hgt} lower than the threshold, CAGC has a fixation probability that is much less than the code with the highest fixation probability.

The high frequency of horizontal transfer of translational components across leaky protocells is manifest in the update rules (see 'Methods' section and 'appendix A' for details) which allows a sequence to sample different codes in the population before selecting one which retains or increases its fitness relative to its current fitness that was obtained by translating the sequence with the code originally associated with it. For the results described above, we followed the protocol adopted in [45] for code update after an HGT event

(see 'Methods' section for details). A smaller frequency of horizontal transfer of translational components can be implemented by changing the update rule. We did so by enforcing that the code update process of the acceptor sequence is allowed only *once*, after an HGT event. The original code translating the sequence is replaced by the new code only if the fitness of the sequence translated using the new code exceeds or equals the fitness of the sequence translated using the original code. In this case, CSGC gets fixed with a higher probability than other codes in the pool provided the probability of HGT events is higher than before. For example, when $P_{\rm hgt} = 0.1$, the above update criterion ensures that CSGC is no longer fixed with the highest probability. However, when $P_{\rm hgt}$ is increased to 0.2, CSGC has the highest fixation probability but even then its fixation probability is comparable to other codes having similar structure. Table S2 in the 'supplementary material' file gives the fixation probability of CSGC and six other codes that get fixed with the highest fixation probabilities for three different values of $P_{\rm hgt}$.

4. Discussions and conclusions

We developed the first finite population code-sequence co-evolution model in the presence of HGT to ascertain the effects of HGT on primordial code evolution under various circumstances. Our results highlight the extent to which HGT affects the outcome of competition between primordial codes and the structure of the emergent universal code. We show that the efficacy of HGT on shaping the outcome of the code-sequence coevolutionary dynamics depends on the initial number of sequences associated with each code in the population, length of the sequences, the selection coefficient, and the frequency of transfer of translational components like tRNA's. HGT alone cannot always guarantee the emergence of a code with a structure that is consistent with the SGC. There is a critical probability of HGT beyond which the presence of HGT can lead to the emergence of an optimized and universal code. This critical threshold varies with $N_{\rm seq/C}$, sequence length (L) and selection coefficient. While increasing $N_{\rm seq/C}$ ensures the effectiveness of HGT leading to the emergence of CSGC for a lower threshold of $P_{\rm hgt}$, an increase in sequence length L makes it more difficult for CSGC to emerge unless the selection coefficient is appropriately increased to prevent the sustained presence of less optimum codes in the population. Lowering the selection coefficient also makes the fixation of CSGC more difficult unless $P_{\rm hgt}$ is enhanced.

As noted in [45], a crucial role is played by HGT of translational components which can result in the same sequence producing two distinct proteins as a result of translating the sequence before and after HGT. Since the nature of translational components transferred between the leaky protocells is random, such a process gives rise to the so-called statistical proteins. A high frequency of horizontal transfer of translational components across genomes of primordial organisms leads to a more efficient search for codes that minimize the fitness difference between the sequences in the population and the benchmark sequences. This process eventually allows the population to converge to the CSGC code resulting in its emergence as the universal and optimized code with a high fixation probability. If the horizontal transfer of translational components occurs less frequently, $P_{\rm hgt}$ needs to be increased in order for the coevolutionary dynamics to efficiently select CSGC over other similar codes.

It is standard practice [51] in population genetic simulations of sequence evolution to calculate the fitness of a sequence by comparing the translated sequence with a previously specified target protein. The target protein will carry the imprint of the code which is used to translate it. But this is not what causes convergence of the code to one which is similar to the SGC. In the absence of HGT, the code getting fixed with the highest probability is not the one most consistent with the SGC. It is clear that HGT plays a crucial role in determining the code that gets fixed in the population. The consequences of changing the benchmark code (used to translate the equilibrated sequences to produce the target protein sequences) depend on the pool of competing codes. When AGC was used as the benchmark code, the 10-amino acid code that got fixed with the highest probability was not CSGC but CAGC as long as the probability of HGT was above the threshold value. These results further reinforce the efficiency with which the process of HGT can identify the code which optimizes the fitness of the sequences it translates relative to the target protein that carries signatures of the benchmark code. Once such a code is identified, the code-sequence co-evolutionary dynamics in the presence of HGT facilitates its fixation in the population.

Several interesting directions suggested by our work can be explored further. HGT is expected to be more effective between sequences that are translated by similar genetic codes. As pointed out in Vetsigian *et al* [45], when a HGT event occurs, the transferred sequence segment undergoes mutations to better adapt to the original code used by the host (acceptor) and the host simultaneously attempts to adjust its original code to better utilize the transferred sequence segment. HGT between significantly different codes makes this process of code-sequence co-adaptation generally more difficult. It therefore remains to be seen how effective a role HGT plays on code-sequence coevolution when it is no longer unconstrained with intra community HGT between sequences translated by similar codes occurring at higher rates than inter community HGT occurring between sequences associated with widely different codes. This would enable us to explore the effect of competition between two or more distinct pools of codes characterized by

markedly different patterns of association between codons and amino acids. The results presented here clarifies the conditions under which HGT can be effective in facilitating the emergence of the SGC from a population of competing code-sequence sets with similar levels of physico-chemical optimization.

Acknowledgments

We thank Devapriya Choudhury for valuable discussions. NA is supported by a DBT-BINC fellowship provided by the Department of Biotechnology (DBT), India.

References

[1] Haig D and Hurst L D 1991 A quantitative measure of error minimization in the genetic code *J. Mol. Evol.* **33** 412–7
[2] Freeland S J and Hurst L D 1998 The genetic code is one in a million *J. Mol. Evol.* **47** 238–48
[3] Pelc S 1965 Correlation between coding-triplets and amino-acids *Nature* **207** 597–9
[4] Pelc S and Welton M 1966 Correlation between coding-triplets and amino-acids *Nature* **209** 868–70
[5] Dunnill P 1966 Triplet nucleotide-amino-acid pairing; a stereochemical basis for the division between protein and non-protein amino-acids *Nature* **210** 1265–7
[6] Yarus M 2000 RNA-ligand chemistry: a testable source for the genetic code *RNA* **6** 475–84
[7] Farias S T D, Moreira C H C and Guimarães R C 2007 Structure of the genetic code suggested by the hydropathy correlation between anticodons and amino acid residues *Orig. Life Evol. Biosph.* **37** 83–103
[8] Sonneborn T M 1965 *Evolving Genes and Proteins* (Amsterdam: Elsevier)
[9] Epstein C J 1966 Role of the amino-acid 'code' and of selection for conformation in the evolution of proteins *Nature* **210** 25–8
[10] Woese C R 1965 On the evolution of the genetic code *Proc. Natl Acad. Sci. USA* **54** 1546–52
[11] Goldberg A L and Wittes R E 1966 Genetic code: aspects of organization *Science* **153** 420–4
[12] Woese C 1967 *The Genetic Code* (New York: Harper & Row) pp 179–95
[13] Di Giulio M 1989 The extension reached by the minimization of the polarity distances during the evolution of the genetic code *J. Mol. Evol.* **29** 288–93
[14] Ardell D H 1998 On error minimization in a sequential origin of the standard genetic code *J. Mol. Evol.* **47** 1–13
[15] Woese C R 1965 Order in the genetic code *Proc. Natl Acad. Sci. USA* **54** 71–5
[16] Gilis D, Massar S, Cerf N J and Rooman M 2001 Optimality of the genetic code with respect to protein stability and amino-acid frequencies *Genome Biol.* **2** 1–12
[17] Freeland S J, Knight R D, Landweber L F and Hurst L D 2000 Early fixation of an optimal genetic code *Mol. Biol. Evol.* **17** 511–8
[18] Novozhilov A S, Wolf Y I and Koonin E V 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
[19] Higgs P G 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
[20] Novozhilov A S and Koonin E V 2009 Exceptional error minimization in putative primordial genetic codes *Biol. Direct* **4** 44
[21] Goodarzi H, Najafabadi H S, Hassani K, Nejad H A and Torabi N 2005 On the optimality of the genetic code, with the consideration of coevolution theory by comparison of prominent cost measure matrices *J. Theor. Biol.* **235** 318–25
[22] Goodarzi H, Nejad H A and Torabi N 2004 On the optimality of the genetic code, with the consideration of termination codons *Biosystems* **77** 163–73
[23] Judson O P and Haydon D 1999 The genetic code: what is it good for? An analysis of the effects of selection pressures on genetic codes *J. Mol. Evol.* **49** 539–50
[24] Morgens D W and Cavalcanti A R O 2013 An alternative look at code evolution: using non-canonical codes to evaluate adaptive and historic models for the origin of the genetic code *J. Mol. Evol.* **76** 71–80
[25] Bandhu A V, Aggarwal N and 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** 465–89
[26] Wong J T 1975 A co-evolution theory of the genetic code *Proc. Natl Acad. Sci. USA* **72** 1909–12
[27] Wong J T-F 1976 The evolution of a universal genetic code *Proc. Natl Acad. Sci.* **73** 2336–40
[28] Wong J T 1980 Role of minimization of chemical distances between amino acids in the evolution of the genetic code *Proc. Natl Acad. Sci. USA* **77** 1083–6
[29] Wong J T-F 2005 Coevolution theory of the genetic code at age thirty *Bioessays* **27** 416–25
[30] Di Giulio M 2002 Genetic code origin: are the pathways of type Glu-tRNA(Gln) –> Gln-tRNA(Gln) molecular fossils or not? *J. Mol. Evol.* **55** 616–22
[31] Di Giulio M 2008 An extension of the coevolution theory of the origin of the genetic code *Biol. Direct* **3** 37
[32] Di Giulio M 1996 The β-sheets of proteins, the biosynthetic relationships between amino acids, and the origin of the genetic code *Orig. Life Evol. Biosph.* **26** 589–609
[33] Di Giulio M and Medugno M 1998 The historical factor: the biosynthetic relationships between amino acids and their physicochemical properties in the origin of the genetic code *J. Mol. Evol.* **46** 615–21
[34] Di Giulio M and Amato U 2009 The close relationship between the biosynthetic families of amino acids and the organisation of the genetic code *Gene* **435** 9–12
[35] Taylor F J and Coates D 1989 The code within the codons *Biosystems* **22** 177–87
[36] Guimarães R C 2011 Metabolic basis for the self-referential genetic code *Orig. Life Evol. Biosph.* **41** 357–71
[37] Di Giulio M 2005 The origin of the genetic code: theories and their relationships, a review *Biosystems* **80** 175–84
[38] Knight R D, Freeland S J and Landweber L F 2001 Rewiring the keyboard: evolvability of the genetic code *Nat. Rev. Genet.* **2** 49–58
[39] Sengupta S, Yang X and Higgs P G 2007 The mechanisms of codon reassignments in mitochondrial genetic codes *J. Mol. Evol.* **64** 662–88
[40] Sengupta S and Higgs P G 2005 A unified model of codon reassignment in alternative genetic codes *Genetics* **170** 831–40
[41] Sengupta S and Higgs P G 2015 Pathways of genetic code evolution in ancient and modern organisms *J. Mol. Evol.* **80** 229–43
[42] Woese C R 2002 On the evolution of cells *Proc. Natl Acad. Sci. USA* **99** 8742–7
[43] Poole A M 2009 Horizontal gene transfer and the earliest stages of the evolution of life *Res. Microbiol.* **160** 473–80
[44] Vogan A A and Higgs P G 2011 The advantages and disadvantages of horizontal gene transfer and the emergence of the first species *Biol. Direct* **6** 1
[45] Vetsigian K, Woese C and Goldenfeld N 2006 Collective evolution and the genetic code *Proc. Natl Acad. Sci. USA* **103** 10696–701

[46] Ardell D H and Sella G 2001 On the evolution of redundancy in genetic codes *J. Mol. Evol.* **53** 269–81
[47] Ardell D H and Sella G 2002 No accident: genetic codes freeze in error-correcting patterns of the standard genetic code *Phil. Trans. R. Soc.* B **357** 1625–42
[48] Linz S, Radtke A and von Haeseler A 2007 A likelihood framework to measure horizontal gene transfer *Mol. Biol. Evol.* **24** 1312–9
[49] Trifonov E N 2000 Consensus temporal order of amino acids and evolution of the triplet code *Gene* **261** 139–51
[50] Higgs P G and Pudritz R E 2009 A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code *Astrobiology* **9** 483–90
[51] Zhu W and Freeland S 2006 The standard genetic code enhances adaptive evolution of proteins *J. Theor. Biol.* **239** 63–70