Błażej *et al. BMC Bioinformatics* (2019) 20:114
https://doi.org/10.1186/s12859-019-2661-4

BMC Bioinformatics

RESEARCH ARTICLE Open Access

The influence of different types of translational inaccuracies on the genetic code structure

Paweł Błażej[*], Małgorzata Wnetrzak, Dorota Mackiewicz and Paweł Mackiewicz

Abstract

Background: The standard genetic code is a recipe for assigning unambiguously 21 labels, i.e. amino acids and stop translation signal, to 64 codons. However, at early stages of the translational machinery development, the codons did not have to be read unambiguously and the early genetic codes could have contained some ambiguous assignments of codons to amino acids. Therefore, the goal of this work was to obtain the genetic code structures which could have evolved assuming different types of inaccuracy of the translational machinery starting from unambiguous assignments of codons to amino acids.

Results: We developed a theoretical model assuming that the level of uncertainty of codon assignments can gradually decrease during the simulations. Since it is postulated that the standard code has evolved to be robust against point mutations and mistranslations, we developed three simulation scenarios assuming that such errors can influence one, two or three codon positions. The simulated codes were selected using the evolutionary algorithm methodology to decrease coding ambiguity and increase their robustness against mistranslation.

Conclusions: The results indicate that the typical codon block structure of the genetic code could have evolved to decrease the ambiguity of amino acid to codon assignments and to increase the fidelity of reading the genetic information. However, the robustness to errors was not the decisive factor that influenced the genetic code evolution because it is possible to find theoretical codes that minimize the reading errors better than the standard genetic code.

Keywords: Amino acid, Codon, Evolution, Evolutionary algorithm, Graph theory, Optimization, The standard genetic code

Background

The standard genetic code (SGC) is a template according to which the information stored in a DNA molecule is transmitted to the protein world in the process called translation. This coding system is nearly universal, with some rare exceptions, for almost all living organisms on Earth. The investigations of the unique organization and properties of this code have been carried out ever since the first encoding rules were determined [1, 2]. Many hypotheses were developed to explain the origin and evolution of the SGC (see for review: [3–7]). However, it is still unclear which factor had the decisive impact on its present structure because the results so far are inconclusive and do not allow us to formulate a final explanatory theory [8]. One of the popular hypotheses assumes that the SGC structure has evolved to minimize harmful consequences of mutations or mistranslations of coded proteins [9–24]. Originally, it was assumed that the optimality of the SGC was directly selected.

However, other models of the genetic code evolution were also proposed. In one of such simulation models both the code and the coded message (i.e. genes) could coevolve [25]. The simulations resulted in the codes that were substantially, but not optimally, error-correcting and reproduced the error-correcting patterns of the SGC. In another model, an important role was assigned to horizontal gene transfer, which made the code not only universal and compatible between translational machineries but also optimal [26]. The self-referential model for the

[*]Correspondence: pawel.blazej@uwr.edu.pl
Department of Genomics, University of Wrocław, ul. Joliot-Curie 14a, 50-383 Wrocław, Poland

formation of the SGC assumes that peptides and RNAs coevolved and were mutual stimulators for the whole system [27]. In this model, a big role was played by tRNA dimers, which directed the initial protein synthesis and showed peptidyl-transferase activity in creation of peptide bonds.

The models assuming a gradual addition of amino acids to the code postulated that this incorporation was: (i) associated with the minimization of disturbance in already synthesized proteins [28], (ii) favoured to promote the diversity of amino acids in proteins [5, 8, 28, 29], (iii) initially driven by catalytic propensity of amino acids functioning in ribozymes [30], (iv) proceeded according to biosynthetic pathways [31–40], or (v) a consequence of duplications of genes coding for tRNAs and aminoacyl-tRNA synthetases (aaRS) [6, 8, 41–47]. The latter proposition, however, was recently criticized in favour of the coevolution theory assuming that the structure of the genetic code was determined by biosynthetic relationships between amino acids [48], although other authors believe that there was a coevolution between the aaRS and the anticodon code as well as an operational code [49]. Thus, the coevolution theory does not necessarily discard the proposition that aaRS and tRNAs played a major role in the formation of the SGC [39].

Considering many factors together, the evolution of the code was probably a combination of adaptation and frozen accident, although contributions of metabolic pathways and weak affinities between amino acids and nucleotide triplets cannot be ruled out [50, 51].

The optimality of the SGC can be reformulated as an attractive problem from the computational and mathematical points of view. For example, a general method of constructing error-correcting binary group codes, represented by channels transmitting binary information, was proposed [52]. Moreover, the analysis of the structure and symmetry of the genetic code using binary dichotomy algorithms also showed its immunity to noise in terms of error-detection and error-correction [53–55]. The code can be also described as a single- or multi-objective optimization problem using the Evolutionary Algorithms (EA) technique to find optimal genetic codes under various criteria [11, 50, 56–58]. Such approach revealed that it is possible to find the theoretical codes much better optimized than the SGC.

The properties of the genetic code can be also tested using techniques borrowed from graph theory [59, 60]. The analysis of the SGC as a partition of an undirected and unweighted graph showed that the majority of codon blocks are optimal in terms of the conductance measure, which is the ratio of non-synonymous substitutions between the codons in this group to all possible single nucleotide substitutions affecting these codons [60]. Therefore, this parameter can be interpreted as a measure of robustness against the potential changes in protein-coding sequences generated by point mutations. The SGC turned out to be far from the optimum according to the conductance but many codon groups in this code reached the minimum conductance for their size [60].

The unique features of the SGC indicate that the structure of this coding system is not fully random and must have evolved under some mechanisms. It is obvious that if we assume that 64 codons encode 20 amino acids and stop coding signal in a potential genetic code then this code must be redundant, i.e. there must exist an amino acid which is encoded by more than one codon. In consequence, such code can be represented as a partition of the set of 64 codons into 21 disjoint subsets (codon groups) so that each codon group encodes unambiguously a respective amino acid or stop signal. Interestingly, these codon groups are generally characterized by a very specific structure in the SGC, namely, the codons belonging to the same group differ usually in the same codon position. Most often the third codon position is different, whereas the first and the second ones stay the same. To explain this specific pattern, Crick developed the wobble rule, which states that the first nucleotide of the tRNA anticodon can interact with one of the several possible nucleotides in the third codon position of a transcript (mRNA) [61]. This non-standard base pairing is often associated with the post-transcriptional modifications of the nucleotide at the first position of the anticodon in the tRNA [62]. The weakened specificity in the base interaction has many consequences. Particularly, it reduces the number of different tRNA molecules which have to recognize codons during the protein synthesis process. Moreover, single point mutations in the third codon position can be synonymous, i.e. do not change the coded amino acid. The wobble base pairing plays also a role in the adoption of the proper structure by tRNA and determines whether the tRNA will be aminoacylated with a specific amino acid.

Our approach to the study of the origins and the possible evolution of the specific structure of the SGC assumes that the early translational machinery was not perfect and codons could be translated ambiguously. Such assumption is in agreement with a hypothesis that protoribosomes could form spontaneously and were able to produce a variety of random peptides, whose sequences depended on the distribution of various amino acids in their vicinity, without the need of a code [63, 64]. Our model also concerns the evolvability of the genetic code as shown in the case of the alternative variants of the genetic code [5, 65–70]. The evolutionary models of these codes postulate the presence of ambiguous assignments of codons to amino acids [71, 72]. Indeed, such assignments were found in Condylostoma, Blastocrithidia and Karyorelict nuclear codes [73–75] as well as *Bacillus subtilis* and *Candida* [76–78]. For these reasons we assumed that the genetic code structure went through intermediate stages in which a particular codon could be translated into more than one amino acid. Obviously, such property of the genetic code is directly related to the level of inaccuracy of the translational machinery. Therefore the goal of our work was to learn which structures of the genetic code can evolve assuming different types of inaccuracy in codon reading in comparison to the structure of the SGC.

Using the approach based on an evolutionary algorithm [79, 80], we analysed a population of randomly generated genetic codes whose codons encoded ambiguously more than one amino acid. The population evolved under the conditions which preferred unambiguous encoding. The scenario which was run under the assumption similar to the wobble rule, produced very quickly the coding systems that are more unambiguous and robust to errors in comparison to other scenarios.

Methods

In this section we give a brief overview of the technical aspects of our work. First, we set up the notation and the terminology necessary to present the crucial steps of our simulation procedure. Then, we introduce a detailed description of the fitness function F, which was used during the selection process. Finally, we describe several measures to study the properties of the optimal genetic codes extracted from the simulations.

Evolutionary algorithm

To simulate the process of the genetic code emergence, we applied an adapted version of EA class algorithm. This technique is widely used in many optimization tasks, especially in the case when analytical solutions do not exist or they are computationally infeasible [80].

The simulation starts with a population of 1000 candidate solutions (individuals). Each candidate represents a random assignment of 64 codons c to 21 labels l corresponding to 20 amino acids and stop translation signal. For simplicity of notation, we use the following set of labels $l = 1, 2, 3, \dots, 20, 21$ and denote the codons $c = 1, 2, 3, \dots, 63, 64$. Therefore, $\mathcal{P} = (p_{cl})$ is a matrix with 64 rows and 21 columns. Each entry p_{cl} in the matrix $\mathcal{P}$ is a probability that a given codon c encodes a given label l and every row sums up to one. At the beginning of our simulations, we used the genetic code matrices whose rows were generated according to the uniform distribution. These codes create an unbiased starting population with high volatility.

The simulation process is divided into consecutive steps called generations. During each step, two important operators, i.e. mutation and selection, are applied to the population. The mutation is a classical genetic operator used in all EA algorithms because it is responsible for random modifications of selected individuals, thus creating new solutions. Here this operator is realized by changing the probability that the selected codon encodes one of 21 possible labels. All changes are introduced using random values generated from the normal distribution and normalized to obtain a probability function in each row. The selection operator requires a fitness function F which allows for assessing the quality of solutions, i.e. the fitness value. Candidate solutions with greater fitness values (scores) are more likely selected to survive and reproduce for the next generation. In this case, we applied a random process of drawing candidate solutions to the next generation with the probability proportional to their fitness. We run the simulations up to 50,000 steps and repeated them 50 times using different seeds.

Fitness function

The fitness function F plays the decisive role in the procedure of genetic codes selection. As a fitness measure, we used a modified version of the total probability function, i.e. the probability that a given genetic code encodes 20 amino acids and stop translation signal. This measure assumes some restrictions on the structure of the codon group assigned to a specific label, e.g. the size of the potential codon group. Moreover, it favours greater probability of encoding a selected label, which reduces the ambiguity in coding. Below we present a detailed description of F in three consecutive steps:

1 Let $L = l_1, l_2, \dots, l_{21}$ be a sequence of all labels and let $C = c_{r_1}, c_{r_2}, \dots, c_{r_{21}}, r_i = 1, 2, \dots, 64$ be a sequence of random codons where every codon c_{r_i} encodes a respective label l_i. Each codon $c_{r_i} \in C$ is drawn randomly from the set of all possible codons $c = c_1, c_2, \dots, c_{64}$ according to the following probability:

$$P\left(c_{r_i} = c_j\right) = P\left(c_j | l_i\right) = \frac{P\left(l_i | c_j\right)}{\sum_{j=1}^{64} P\left(l_i | c_j\right)}, \qquad (1)$$

where $p\left(l_i | c_j\right) = p_{l_i c_j}$ is an element from l_i^{th}-row and c_j^{th}-column of the matrix $\mathcal{P}$. It is evident that $\sum_{j=1}^{64} P(l_i | c_j)$ is a sum of all elements extracted from the column l_i of the matrix $\mathcal{P}$. Therefore, the Eq. (1) is clearly an application of Bayes rule under the assumption that *a priori* probability, i.e. the probability of choosing a given codon c_j, is uniformly distributed i.e. $P\left(c_j\right) = 1/64$.

2 For each codon c_{r_i} belonging to C, we define a codon neighbourhood $N\left(c_{r_i}\right)$. $N\left(c_{r_i}\right)$ is a set of codons that contains the original codon c_{r_i} and the codons c'_{r_i} differing in one nucleotide from c_{r_i}. The size of $N\left(c_{r_i}\right)$ depends on the simulation assumptions. We considered three possible scenarios:

M_1 - all codons belonging to a given $N\left(c_{r_i}\right)$ have two fixed codon positions identical and differ in exactly one nucleotide at the other position in codon;

M_2 - all codons belonging to a given $N\left(c_{r_i}\right)$ have one fixed codon position identical and differ in exactly one nucleotide in one of the other two codon positions;

M_3 - all codons belonging to a given $N\left(c_{r_i}\right)$ differ in exactly one nucleotide in any codon position.

For example, the neighbourhood for the codon GGG is:

- GGG, GGA, GGC, GGT for the scenario M_1;
- GGG, AGG, CGG, TGG, GAG, GCG, GTG for the scenario M_2;
- GGG, AGG, CGG, TGG, GAG, GCG, GTG, GGA, GGC, GGT for the scenario M_3.

Thus, the size of the neighbourhood for M_1 is $|N(c_r)| = 4$, for M_2 is $|N(c_r)| = 7$ and for M_3 is $|N(c_r)| = 10$.

3 Using the assumptions presented in step 1 and 2, we can define the fitness function F as:

$$F = \sum_{c'_{r_1},\ldots,c'_{r_{21}}:\, c'_{r_i} \in N\left(c_{r_i}\right)} P\left(l_1|c'_{r_1}\right) P\left(l_2|c'_{r_2}\right) \cdot \ldots \cdot P\left(l_{21}|c'_{r_{21}}\right) . \tag{2}$$

It is evident that assuming $P\left(c'_{r_i}\right) = \frac{1}{64}$, $c'_{r_i} = 1, 2, \ldots, 64$ and the independence of $P\left(l_n|c'_{r_i}\right)$ in the formula (2), we obtain the following equality:

$$P(l_1, l_2, \ldots, l_{21}) = F \cdot \left(\frac{1}{64}\right)^{21} ,$$

which is the total probability that a given genetic code generates a sequence of labels L. Therefore, a high value of F suggests that a given genetic code is more likely to encode 20 amino acids and stop coding signal unambiguously.

It should be noted that the computation of F, using the formula (2) directly, involves the order of $O\left(|N(c_r)|^{21}\right)$ calculations [81]. Therefore, fast calculation of the fitness values for many candidate solutions becomes a problem because the "direct" method is computationally infeasible even for small sizes of $N(c_r)$. To deal with it, we incorporated a modified version of the forward algorithm [81], which is more efficient in computing the exact fitness values than the direct approach. This procedure follows from some basic observations. Let us consider $\alpha_l(c)$ defined inductively as:

$$\alpha_l(c) = \begin{cases} \alpha_1(c) = P(l_1|c), & c \in N(c_{r_1}) \\ \alpha_k(c) = \sum_{c' \in N\left(c_{r_{k-1}}\right)} \alpha_{k-1}\left(c'\right) \cdot P(l_k|c), & 1 < k \leq 21, c \in N(c_{r_k}). \end{cases}$$

From this definition we can deduce that $F = \sum_{c \in N(c_{r_{21}})} \alpha_{21}(c)$. If we take into account the computational effort required to calculate $\alpha_l(c)$ $c \in N(c_{r_l})$ and then compute the fitness value, we need the order of $O\left(|N(c_{r_l})|^2\right)$ calculations. Thereby, assuming that $\left|N\left(c_{r_l}\right)\right| = 10$, which is the maximum size of the codon neighbourhood in the M_3 model, we need about 2100 computations for the modified forward method in comparison to about 10^{21} computations for the "direct" approach. This forward procedure allowed us to calculate the fitness values fast and effectively, which is essential in the case of many individuals constantly modified during simulations.

There is also another important feature related to the fitness function, namely, F is non-deterministic. This is because the fitness value is dependent on a randomly generated codon sequence C. Therefore, F is a random variable and in consequence, genetic codes are rated according to their randomly generated fitness values during the selection process. However, the chance to be selected to the next generation is not only a matter of luck because the selection of the sequence C prefers the codons that have relatively high probabilities to encode respective labels (see Eq. (1)). Thereby, the distribution of F prefers larger values. They are compared during the selection process and finally, the method of codon selection is crucial in terms of the convergence of genetic codes to the stable solutions. We observed such convergence of the fitness values to the stable solution during the simulations steps. An example of the variation in the fitness function values calculated for 50 independent simulations under the same parameters but different seeds is presented in the Fig. 1.

Measures of the properties of genetic codes

Because of the large amount of data to analyse, we introduced some definitions to test in details the properties of the obtained genetic codes. One of the most important questions which arose in our investigations was how to measure the level of the genetic code ambiguity at the global scale, because the fitness function delivered us only a piece of information about the probability of encoding 21 labels. To test the quality of a given genetic code, we defined the genetic code entropy.

Definition 1 *Let $\mathcal{P} = (p_{cl})$ be a matrix of a genetic code, where each row contains a discrete probability distribution, then the entropy of the genetic code $H(\mathcal{P})$ is defined as:*

$$H(\mathcal{P}) = -\sum_{c=1}^{64}\sum_{l=1}^{21} p_{cl} log(p_{cl}) . \tag{3}$$

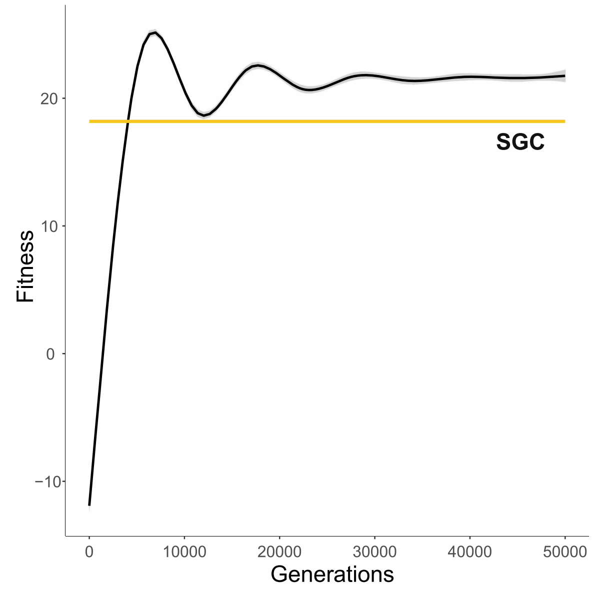

Fig. 1 Changes in the best approximation of the fitness function *F* with the number of generations (the black line). All approximations were done for 50 simulations using the Generalized Additive Models. The simulations were run under M_1 scenario with different initial seeds. The independent simulations show a very narrow confidence interval depicted by the grey strip. The results were compared with the average fitness value calculated for the standard genetic code (the orange line)

It should be noted that $H(\mathcal{P})$ is in fact the sum of Shannon entropy calculated for each row of the matrix $\mathcal{P}$, separately. Therefore, $H(\mathcal{P})$ corresponds to the multidimensional entropy of independent distributions. The definition 1 appears useful in testing the general properties of genetic codes in terms of changes in their ambiguity. Moreover, it allows us to make more detailed comparisons between the results obtained under different scenarios i.e. M_1, M_2 and M_3. In our analyses we also calculated the average genetic code entropy value $H_{av}(\mathcal{P})$, which is the arithmetic mean of the genetic code entropy $H(\mathcal{P})$ evaluated for all candidate solutions.

Furthermore, we used a graph representation of the genetic code. This approach was effectively applied by [59] and [60]. The authors considered a graph $G(V, E)$ with 64 nodes (codons) V and the set of edges E representing point mutations between codons. According to this approach, every genetic code $\mathcal{C}$ is a partition of V into 21 disjoint subsets S_l, $l = l_1, l_2, \ldots, l_{21}$, i.e. groups of codons. To investigate further the properties of a given graph clustering, [60] introduced the set conductance, which turned out a very useful measure in testing the properties of codon groups. The definition of the set conductance is as follows:

Definition 2 *For a given graph G, let S be a subset of V. The conductance of S is defined as:*

$$\phi(S) = \frac{E\left(S, \bar{S}\right)}{vol(S)} ,$$

where $E\left(S, \bar{S}\right)$ *is the number of edges of G crossing from S to its complement* $\bar{S}$ *and* $vol(S)$ *is the sum of all degrees of the vertices belonging to S.*

The set conductance has a useful interpretation from the biological point of view because for a given codon group S, $\phi(S)$ is the ratio of non-synonymous codon changes to all possible changes concerning all codons belonging to this set. Therefore, it is interesting to find the optimal codon blocks in terms of $\phi(S)$. To do so, we used the k-size-conductance $\phi_k(G)$ described as the minimal set conductance over all subsets of V with the fixed size k.

Definition 3 *The k-size-conductance of the graph G, for* $k \geq 1$*, is defined as:*

$$\phi_k(G) = min_{S \subseteq V, |S|=k} \phi(S) .$$

Moreover, the properties of a given genetic code $\mathcal{C}$ can be expressed as the average code conductance $\Phi(\mathcal{C})$, which is the arithmetic mean calculated from all set conductances of all codon groups. The detailed definition of the average code conductance is given in the following way:

Definition 4 *The average conductance of a genetic code* $\mathcal{C}$ *is defined as:*

$$\Phi(\mathcal{C}) = \frac{1}{21} \sum_{S \in \mathcal{C}} \phi(S) .$$

The relationship between matrix and graph representation of the genetic code

As mentioned in the previous section, we used two different representations of the genetic code. The first one describes the genetic codes as a matrix, whereas the other one presents the genetic code as a partition of graph nodes into 21 non-empty disjoint clusters. It is evident that for every graph representation we can construct directly a unique matrix. Then, each row c of the matrix $\mathcal{P}$ contains a degenerated probability distribution, i.e. $p_{cl} = 1$, where a codon c encodes a label l. On the other hand, without additional assumptions, it is impossible to obtain a unique graph partition from a selected matrix representation. Therefore, we have to assume that each row of the matrix $\mathcal{P}$ contains a unimodal probability distribution. Only in such case we can transform $\mathcal{P}$ unambiguously into an equivalent graph representation. To do so, we introduced the maximum likelihood graph partition (MLGP) approach.

Definition 5 *Let $\mathcal{P} = (p_{cl})$ be a matrix representation of a genetic code, where each row contains a unimodal discrete probability distribution. Assume also that for every label l there exists a codon c such that:*

$$p_{cl} = max_{1\leq l'\leq 21} p_{cl'} \ .$$

Then the maximum likelihood graph partition is a partition of the set of the graph G nodes into 21 non-empty disjoint subsets $S_1, S_2, \ldots, S_{21}$ according to the following formula:

$$c \in S_l \iff p_{cl} = max_{1\leq l'\leq 21} p_{cl'} \ .$$

To measure the quality of the selected codon block S_l, $l = 1, 2, \ldots, 21$, created according to the definition 5, we defined the coding strength of the set S_l.

Definition 6 *Let $\mathcal{P} = (p_{cl})$ be a matrix representation of a genetic code, where each row contains a unimodal discrete probability distribution and let $\mathcal{C} = \{S_1, S_2, \ldots S_{21}\}$ be its respective MLGP representation, then for every S_l we define $\psi(S_l)$, the coding strength of the set S_l, in the following way:*

$$\psi(S_l) = \frac{1}{|S_l|}\sum_{c\in S_l} p_{cl} \ .$$

Following the definition 6 of the coding strength, we can also consider the average coding strength $\Psi(\mathcal{C})$ of a genetic code $\mathcal{C}$, which is defined as the arithmetic mean of all coding strengths $\psi(S_l)$ computed for all S_l belonging to the graph representation of a genetic code $\mathcal{C}$:

$$\Psi(\mathcal{C}) = \frac{1}{21}\sum_{l=1}^{21} \psi(S_l) \ .$$

Results

The uncertainty level of simulated genetic codes

The aim of these simulations was to learn, which structures of the genetic codes can evolve assuming different inaccuracy of the translational machinery. We simulated three scenarios of the genetic code evolution that started from an ambiguous coding state. The scenarios M_1, M_2 and M_3 assumed that respectively one, two or three codon positions can be mutated or erroneously read during the translation process. We started our analysis by looking at the differences between the average entropy value $H_{av}(\mathcal{P})$ of the genetic codes calculated for the three scenarios. The high value of the entropy means that a code is characterized by a high level of coding ambiguity, i.e. a individual codon can be translated into various amino acids, while the low values indicate that the coding is more unambiguous. The code with the perfect unambiguity should be characterized by $H_{av}(\mathcal{P}) = 0$. The changes in the coding ambiguity during the simulation time are presented in the Fig. 2 for all types of scenarios. It is evident that $H_{av}(\mathcal{P})$ decreases substantially from the beginning of the simulations under all scenarios and then stabilizes around 10,000 to 30,000 simulation steps. This result indicates that the assumptions used in the optimization procedure are generally responsible for decreasing the uncertainty level of genetic codes. In addition, the level of $H_{av}(\mathcal{P})$ differs between the scenarios. The less extensive the neighbourhood, i.e. the number of similar codons in the group, the smaller the entropy. Under the M_1 scenario, where the neighbourhood size $|N(c_r)| = 4$, the entropy is the smallest, i.e. 5.48 and the equilibrium is reached much faster than in the other models. The value of $H_{av}(\mathcal{P})$ decreased about 33 times in comparison to the initially ambiguous codes with $H_{av}(\mathcal{P}) \approx 182$. On the other hand, the simulation run under the M_3 scenario, where the neighbourhood is the largest, i.e. $|N(c_r)| = 10$, reaches its minimum of the $H_{av}(\mathcal{P})$ much later. The entropy of the M_3 scenario is the largest of all scenarios and is almost six times greater than the entropy of M_1 (Fig. 2).

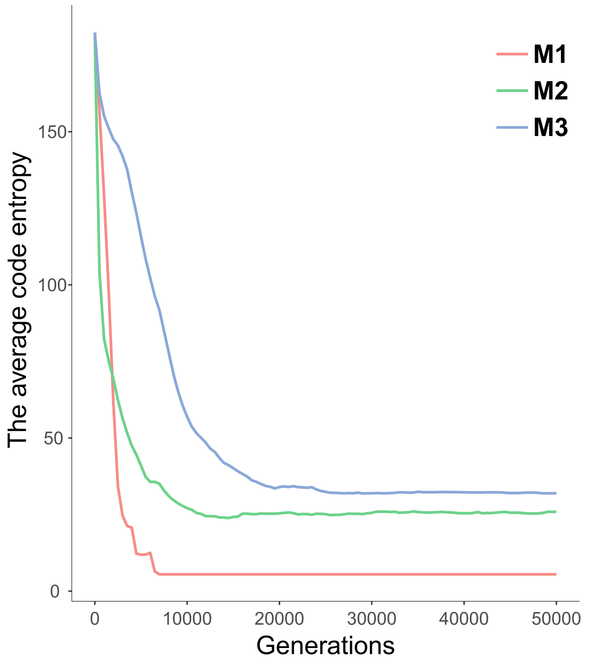

Fig. 2 Changes in the average genetic code entropy value $H_{av}(\boldsymbol{\mathcal{P}})$ during the simulation time calculated for three scenarios M_1, M_2, M_3. The average genetic code entropy is the arithmetic mean of the genetic code entropy $H(\boldsymbol{\mathcal{P}})$ evaluated for all candidate solution

In contrast to the entropy measure, which includes in the calculation the probabilities of all possible assignments of amino acids to codons, the average coding strength Ψ takes into account only the maximum probability of these assignments. Large values of Ψ indicate that the assignments are highly unambiguous in a given code, while small values mean that many amino acids can be encoded by many codons with a comparable probability. The code with no ambiguous assignment of amino acids to codons ought to have the value $\Psi = 1$. Similarly to the entropy, the highest unambiguity and the largest values of Ψ are observed in the case of M_1 but the values of Ψ do not show the relationship with the size of $N(c_r)$ as the $H_{av}(\mathcal{P})$ (Fig. 3). We could expect that a decrease in the neighbourhood would result in an increase of the coding signal. However, it is not fully fulfilled because Ψ for M_2 is slightly smaller than for M_3 (Fig. 3). This observation suggests that the MGLP graph representations of the genetic codes computed under the M_2 scenario are composed of codon blocks characterized by a weaker coding signal in comparison to the other simulation scenarios.

The robustness level of simulated genetic codes

To describe the robustness of the structure of the genetic code to mutations and mistranslations, we applied the average code conductance Φ. Its large value indicates that the code is not robust against point mutations. The Φ values were calculated following the MLGP representation of the codes obtained at the end of each simulation run. It is interesting that the Φ values for each simulation run under the M_1 assumption, are smaller than the average code conductance computed for the standard genetic code, i.e. $\Phi(SGC) = 0.8112$ (Fig. 4). Moreover, the M_1-type optimal genetic codes are closer to the best (minimum) possible value of $\Psi = 0.7724$ for any code assigning 21 labels to 64 codons. The results strongly suggest that the M_1 scenario of code evolution is able to create the genetic codes quite robust to mutation and mistranslations. In contrast to that, the genetic codes obtained under the M_2 and M_3 assumptions are characterized by much larger values of the average code conductance than SGC (Fig. 4). Thereby their structures are less robust against point mutation. The genetic codes obtained in the M_2 type of simulations show generally the worst Φ in comparison to the other simulation types.

The types of codon groups in simulated genetic codes

The genetic codes obtained under M_1, M_2 and M_3 scenarios differ in the codon group distribution (Fig. 5). In the the genetic codes produced at the end of 50 independent

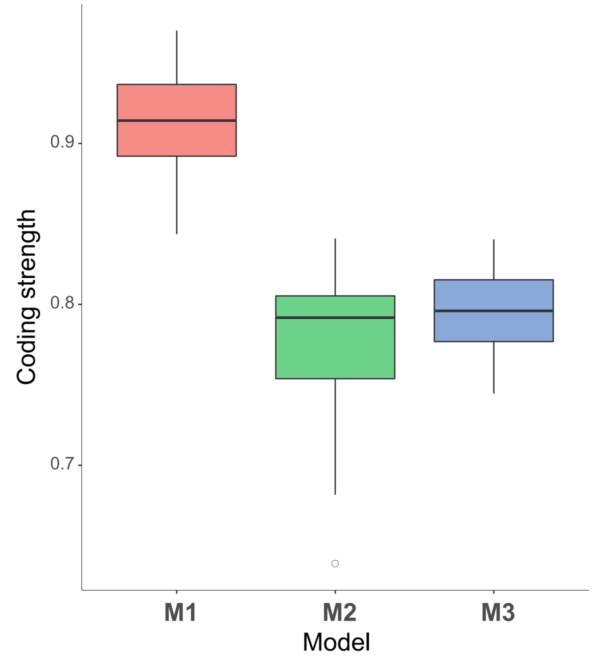

Fig. 3 Box-plots of the average coding signal strength calculated at the end of the simulations under three scenarios M_1, M_2 and M_3 for 50 independent simulation runs per scenario. The thick horizontal line indicates the median (*IQR*, the inter-quartile range), the box shows the range between the first and the third quartiles and the whiskers determine the range without outliers for the assumption 1.5 × *IQR*

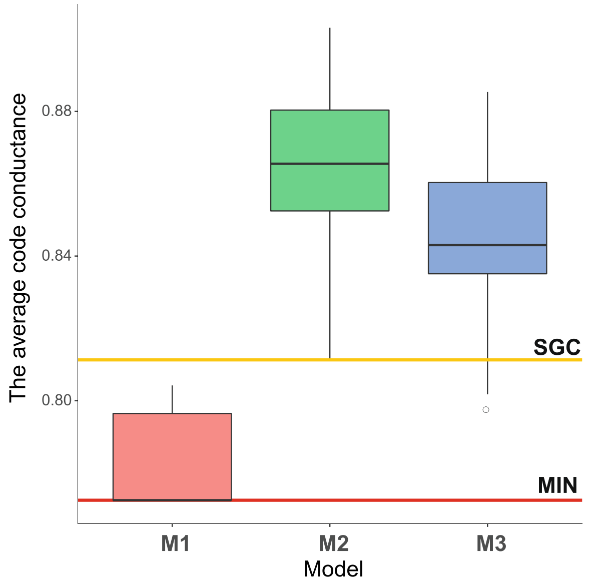

Fig. 4 Box-plots of the average code conductance calculated at the end of the simulations under three scenarios M_1, M_2 and M_3 for 50 independent simulation runs per scenario. The thick black horizontal line (inside each box) indicates the median (*IQR*, the inter-quartile range), the box shows the range between the first and the third quartiles and the whiskers determine the range without outliers for the assumption 1.5 × *IQR*. The results were compared with the average code conductance Φ calculated for the standard genetic code (the orange horizontal line) and the minimum value of the average code conductance (the red horizontal line)

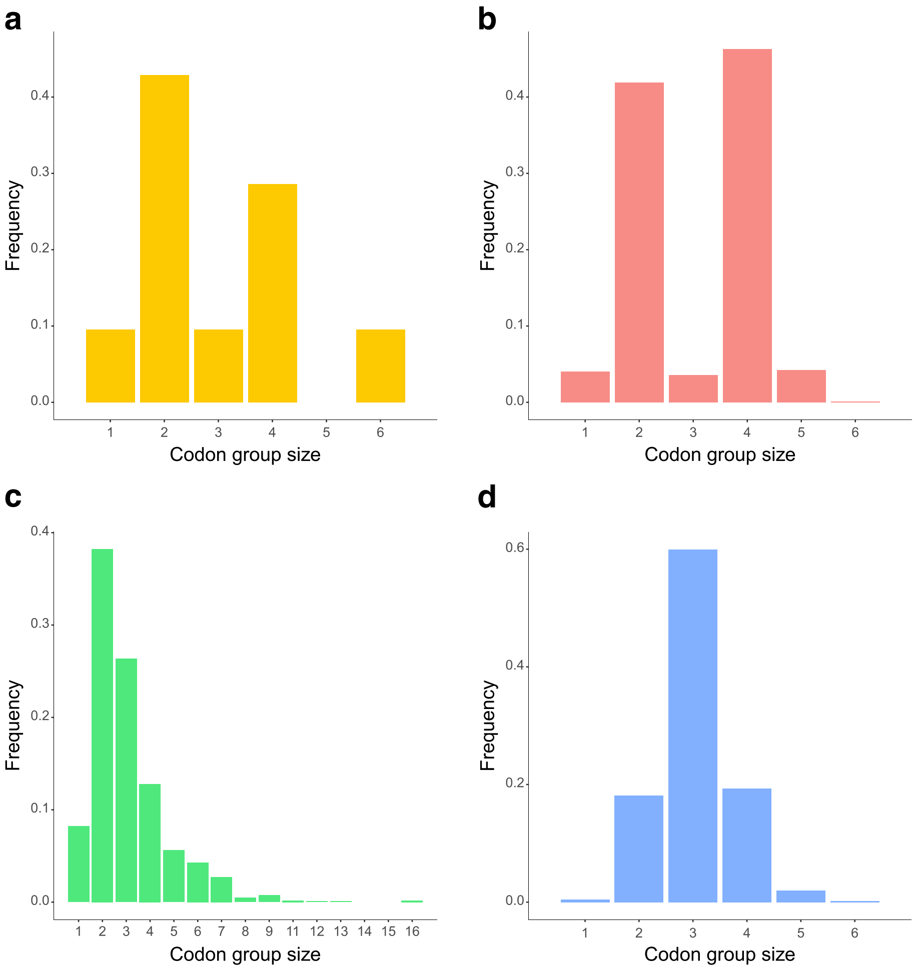

Fig. 5 The frequencies of codon group sizes observed in the standard genetic code (**a**) as well as in the MLGP representations of genetic codes at the end of 50 independent simulation runs under the M_1 (**b**), M_2 (**c**) and M_3 (**d**) scenarios

simulations in the M_1 scenario, there are two most frequent types of groups, consisting of two and four codons (Fig. 5b), similarly to the SGC (Fig. 5a). They constitute in total over 87% of all codon groups in the M_1 codes and 71% in the case of the SGC. The groups of one, three, five and six codons are in the minority, constituting in total less than 13% of the codon groups in the M_1 codes. However, there are also some differences in comparison to the SGC. In the SGC the contribution of two-codon groups is greater than the four-codon groups, while in the M_1 codes the opposite is true. Moreover, there are no groups of five codons in the SGC, which occur in the M_1 codes.

The codes produced by the M_2 model show definitely different distribution of the codon groups and are characterized by a greater variability in codon group sizes, being in the range from 1 to 16 (Fig. 5c). However, the codon groups of the size from 1 to 6 have the joint frequency over 95%. The most frequent are two-codon groups as in the SGC. They constitute 38% and 43%, respectively. What is more, an intriguing kind of symmetry is present in the distribution of codon groups in the genetic codes simulated under the M_3 scenario (Fig. 5d). The most frequently observed codon group consists of three codons and constitutes about 60% of all groups. The frequencies of other codon groups are nearly symmetrically arranged around the most frequent group. The next most common groups (about 20%) include two and four codons. This type of codes are the most different form the SGC in the distribution of the codon groups because in the SGC the three-codon groups are poorly represented.

The presence of codon groups with the number of codons different than in the SGC would seem intriguing and artificial for the simulated codes. However, such groups have actually evolved in some alternative variants of the SGC. In total in these codes, there are five pentacodonic amino acids, four heptacodonic amino acids and five octacodonic amino acids (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi). For example, in the alternative yeast nuclear code, serine is encoded additionally by the seventh codon CUG, which was taken from leucine, encoded in consequence by five codons.

The properties of the best genetic codes

In this section, we discussed the properties of the best genetic codes that were selected according to their maximum fitness values from all simulation runs for all types of scenarios. In the Fig. 6, we presented four heatmaps depicting the selected matrix representations of the genetic codes at the beginning as well as at the end of the simulations under the M_1, M_2 and M_3 scenarios.

As expected, the random code at the start of simulation is highly ambiguous (Fig. 6a), while the code emerged under the M_1 scenario is characterized by a very high unambiguity and is filled mainly with the codon blocks consisting of two and four codons (Fig. 6b). The codons in each of such groups differ in pairwise comparison in only one nucleotide (Fig. 7). The graph representation of this code following the definition 5 is also optimal in terms of the k-size conductance $\phi_k(G)$, $k = 2, 4$. All the codon groups show the minimum possible conductance for their size. Therefore, these groups are the most robust against single non-synonymous nucleotide mutations. In consequence, this genetic code reaches the minimum of the average code conductance $\Phi(\mathcal{C}) = 0.7725$, which is the minimum value of all possible genetic codes and is smaller than the conductance of the standard genetic code $\Phi(SGC) = 0.8113$. Moreover, many codon groups in the M_1-type code are characterized by a relatively large unambiguity. Fifteen groups have the maximal coding strength

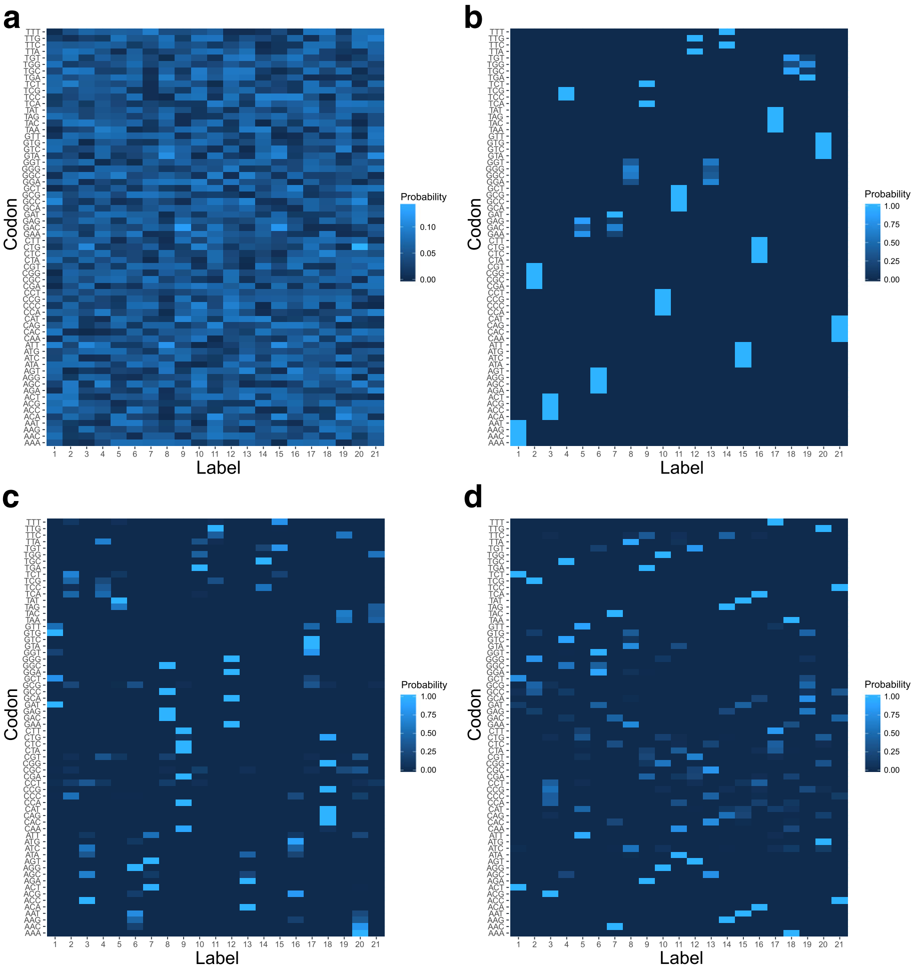

Fig. 6 The matrix representation of a genetic code at the beginning of the simulations (**a**) as well as obtained at the end of the simulations under the M_1 (**b**) , M_2 (**c**) and M_3 (**d**) scenarios. Each row contains values of the probability function represented by a respective rectangle. The colour of the rectangles indicates high (light blue) or low (dark blue) probability that a given codon (row) encodes a given label (column). It is evident that codon blocks of the size 2 and 4 show high probabilities (light blue colour) and dominate in the code under the M_1 scenario. In the case of other scenarios the codes show much greater ambiguity

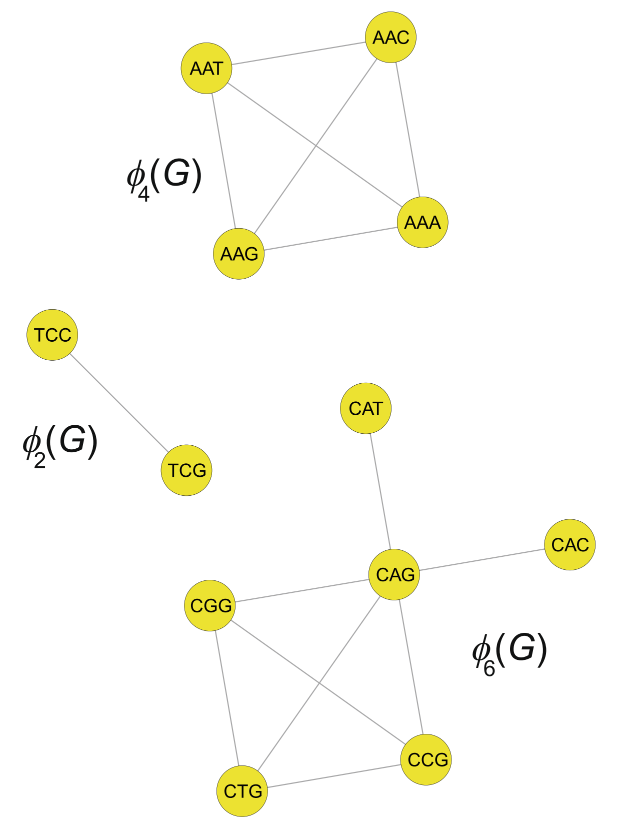

Fig. 7 The examples of graph representations of codon groups with the minimal 2, 4 and 6-size conductance: $\phi_2(G)$, $\phi_4(G)$ and $\phi_6(G)$, respectively. The first two cases dominate in the best genetic code produced under the M_1 scenario and the latter is observed in the best genetic code produced under the M_2 scenario

Table 1 The codon groups of the best genetic code in terms of the fitness function *F* extracted from 50 independent simulations under the M_1 scenario

Codon group (S)	k	$\psi(S)$	$\phi(S)$	$\phi_k(G)$
$\{AAA, AAT, AAG, AAC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{CGA, CGT, CGG, CGC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{ACA, ACT, ACG, ACC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{GTA, GTT, GTG, GTC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{CAA, CAT, CAG, CAC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{AGA, AGT, AGG, AGC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{CCA, CCT, CCG, CCC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{ATA, ATT, ATG, ATC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{CTA, CTT, CTG, CTC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{TAA, TAT, TAG, TAC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{GCA, GCT, GCG, GCC\}$	4	1.0000000	$\frac{2}{3}$	$\frac{2}{3}$
$\{TCG, TCC\}$	2	1.0000000	$\frac{8}{9}$	$\frac{8}{9}$
$\{TCA, TCT\}$	2	1.0000000	$\frac{8}{9}$	$\frac{8}{9}$
$\{TTATTG\}$	2	1.0000000	$\frac{8}{9}$	$\frac{8}{9}$
$\{TTT, TTC\}$	2	1.0000000	$\frac{8}{9}$	$\frac{8}{9}$
$\{TGT, TGC\}$	2	0.8648035	$\frac{8}{9}$	$\frac{8}{9}$
$\{TGA, TGG\}$	2	0.8648030	$\frac{8}{9}$	$\frac{8}{9}$
$\{GAT, GAC\}$	2	0.8344635	$\frac{8}{9}$	$\frac{8}{9}$
$\{GAA, GAG\}$	2	0.8344630	$\frac{8}{9}$	$\frac{8}{9}$
$\{GGG, GGC\}$	2	0.6446255	$\frac{8}{9}$	$\frac{8}{9}$
$\{GGA, GGT\}$	2	0.6446250	$\frac{8}{9}$	$\frac{8}{9}$

The groups S are characterized by: the size k, the coding strength $\psi(S)$, the conductance $\phi(S)$ and the minimal conductance of the codon group with the size k $\phi_k(G)$

$\psi(S) = 1$ and the average coding strength calculated over all 21 groups is equal to 0.9375 (Table 1).

The best codes produced under the M_3 scenario (Fig. 6d) show completely different composition of codon groups in comparison to the best code of the M_1 scenario. The M_3-type code is composed of codon groups of the size $k = 2, 3, 4$ with the domination of three-codon groups (Table 2). This code is also less robust against point mutation because its average code conductance is equal to 0.8457, which is slightly greater than the conductance of the standard genetic code $\Phi(SGC) = 0.8113$. This is caused by the presence of as many as twelve non-optimal codon groups in terms of the k-size conductance (Table 2). The code shows a higher ambiguity than that of the M_1 scenario because its average coding strength ψ is 0.8023. Only four codon groups consisting of two codons are perfectly unambiguous and robust to non-synonymous mutations.

The best genetic code evaluated under the M_2 model (Fig. 6c) is characterized by the most diversified size of codon groups in comparison to the M_1 and M_3 cases because it is composed of codon groups of the size $k = 1, 2, 3, 4, 6$ (Table 3). These groups are also characterized by generally smaller coding strength values of ψ. Therefore, the average coding strength calculated in this case is equal to 0.7996. Moreover, thirteen codon blocks are not optimal in terms of the set conductance $\phi(S)$. In consequence, the average code conductance is relatively high and equals 0.8580. Therefore, it is the least robust genetic code structure against point mutation in comparison to the M_1- and M_3-type codes. The M_2 code contains no codon groups including at least two codons that simultaneously encode unambiguously one label and are the most robust to single point mutations. On the other hand, the two largest groups of six codons in this code are optimal in terms of the k-size conductance $\phi_k(G)$ (Fig. 7) and are characterized by quite big values of coding strength, over 0.98.

Discussion

We carried out a simulation study to find out how the structure of the genetic code could have evolved

Table 2 The codon groups of the best genetic code in terms of the fitness function F extracted from 50 independent simulations under the M_3 scenario

Codon group (S)	k	$\psi(S)$	$\phi(S)$	$\phi_k(G)$
$\{AAG, TAG, TTC, CAG\}$	4	0.7453878	$\frac{5}{6}$	$\frac{2}{3}$
$\{ATC, TTA, GAA, GTA\}$	4	0.7347630	$\frac{8}{9}$	$\frac{2}{3}$
$\{ACG, CCA, CCT, CCG\}$	4	0.5928058	$\frac{7}{9}$	$\frac{2}{3}$
$\{AGC, CAC, CGC, CCC\}$	4	0.6837612	$\frac{7}{9}$	$\frac{2}{3}$
$\{GAG, GTG, GCA, GCG\}$	4	0.5734470	$\frac{7}{9}$	$\frac{2}{3}$
$\{ACT, TCT, GCT\}$	3	0.9164170	$\frac{7}{9}$	$\frac{7}{9}$
$\{AGG, TGG, CGG\}$	3	0.8623860	$\frac{7}{9}$	$\frac{7}{9}$
$\{TGC, GTC, GGC\}$	3	0.8267687	$\frac{23}{27}$	$\frac{7}{9}$
$\{AGT, TGT, CGT\}$	3	0.8261313	$\frac{7}{9}$	$\frac{7}{9}$
$\{ACC, TCC, GAC\}$	3	0.8157710	$\frac{25}{27}$	$\frac{7}{9}$
$\{ATG, TTG, CTG\}$	3	0.7968670	$\frac{7}{9}$	$\frac{7}{9}$
$\{GAT, GGA, GGT\}$	3	0.7812357	$\frac{23}{27}$	$\frac{7}{9}$
$\{ATT, GTT, CAT\}$	3	0.7741430	$\frac{25}{27}$	$\frac{7}{9}$
$\{TTT, CTT, CTC\}$	3	0.7475287	$\frac{23}{27}$	$\frac{7}{9}$
$\{AGA, TGA, CGA\}$	3	0.7347630	$\frac{7}{9}$	$\frac{7}{9}$
$\{ATA, CAA, CTA\}$	3	0.7112670	$\frac{23}{27}$	$\frac{7}{9}$
$\{TCG, GGG, GCC\}$	3	0.7241587	1	$\frac{7}{9}$
$\{AAC, TAC\}$	2	1.0000000	$\frac{8}{9}$	$\frac{8}{9}$
$\{AAT, TAT\}$	2	1.0000000	$\frac{8}{9}$	$\frac{8}{9}$
$\{ACA, TCA\}$	2	1.0000000	$\frac{8}{9}$	$\frac{8}{9}$
$\{AAA, TAA\}$	2	1.0000000	$\frac{8}{9}$	$\frac{8}{9}$

The groups S are characterized by: the size k, the coding strength $\psi(S)$, the conductance $\phi(S)$ and the minimal conductance of the codon group with the size k $\phi_k(G)$

Table 3 The codon groups of the best genetic code in terms of the fitness function F extracted from 50 independent simulations under the M_2 scenario

Codon group (S)	k	$\psi(S)$	$\phi(S)$	$\phi_k(G)$
$\{CAT, CAG, CAC, CTG, CGG, CCG\}$	6	0.9867038	$\frac{36}{54}$	$\frac{36}{54}$
$\{CAA, CTA, CTT, CTC, CGA, CCA\}$	6	0.9866648	$\frac{35}{54}$	$\frac{36}{54}$
$\{GAG, GAC, GGC, GCC\}$	4	1.0000000	$\frac{7}{9}$	$\frac{2}{3}$
$\{GAA, GGA, GGG, GCA\}$	4	1.0000000	$\frac{7}{9}$	$\frac{2}{3}$
$\{GAT, GTT, GTG, GCT\}$	4	0.8262542	$\frac{7}{9}$	$\frac{2}{3}$
$\{GTA, GTC, GGT, GCG\}$	4	0.7837840	$\frac{34}{36}$	$\frac{2}{3}$
$\{ATC, AGC, ACC, CCT\}$	4	0.6426162	$\frac{5}{6}$	$\frac{2}{3}$
$\{ATT, AGT, ACT\}$	3	0.8457703	$\frac{7}{9}$	$\frac{7}{9}$
$\{ATA, AGA, ACA\}$	3	0.8136870	$\frac{7}{9}$	$\frac{7}{9}$
$\{AAT, AAG, AGG\}$	3	0.7359857	$\frac{23}{27}$	$\frac{7}{9}$
$\{AAA, AAC, CGC\}$	3	0.7170777	$\frac{25}{27}$	$\frac{7}{9}$
$\{TAA, TAC, TTC\}$	3	0.5927813	$\frac{23}{27}$	$\frac{7}{9}$
$\{TCT, TCG, CCC\}$	3	0.5598730	$\frac{25}{27}$	$\frac{7}{9}$
$\{TTA, TCA, CGT\}$	3	0.4932700	$\frac{25}{27}$	$\frac{7}{9}$
$\{ATG, ACG\}$	2	0.8650405	$\frac{8}{9}$	$\frac{8}{9}$
$\{TAT, TAG\}$	2	0.8027995	$\frac{8}{9}$	$\frac{8}{9}$
$\{TTT, TGT\}$	2	0.7850055	$\frac{8}{9}$	$\frac{8}{9}$
$\{TGC, TCC\}$	2	0.7838025	$\frac{8}{9}$	$\frac{8}{9}$
$\{TGA\}$	1	1.0000000	1	1
$\{TTG\}$	1	0.9967190	1	1
$\{TGG\}$	1	0.5734410	1	1

The groups S are characterized by: the size k, the coding strength $\psi(S)$, the conductance $\phi(S)$ and the minimal conductance of the codon group with the size k $\phi_k(G)$

under various types of inaccurate translation of codons to amino acids. The simulations started from the set of ambiguous assignments of amino acids to codons, which evolved into patterns with lower levels of uncertainty. The reduction of ambiguity was driven by a fitness function, which preferred the codes that are characterized by the robustness to incorrect amino acid translation due to point mutations in codons. We developed three theoretical models of the genetic code evolution, M_1, M_2, and M_3, which corresponded to various types and levels of inaccuracy of a primordial translation apparatus.

All the models are in agreement with the ambiguous intermediate mechanism acting in the evolution of the alternative genetic codes [71, 72]. In this case, the codon is translated ambiguously to two different amino acids during the period of reassignment. Such cases of ambiguous translation were reported in different organisms [73–78]. What is more, such ambiguous state can also promote phenotypic diversity and adaptability, e.g. it helps yeasts to cope with more stressful environments [82, 83].

Moreover, the models M_1 and M_2 match the stages of the genetic code evolution postulated by the 2-1-3 model [44, 84] and the four-column theory [28]. They assume that in the beginning of the genetic code evolution the second codon position decided about the encoded amino acids, whereas other positions were not important. Next, the coding specificity occurred in the first codon position and then, to some extent, in the third position.

The initial ambiguity in the assignments of amino acids to codons disappeared the fastest under the M_1 model. The codes generated under this scenario are characterized by the biggest unambiguity of coding and the most effective minimization of mutations changing encoded amino acids or stop translation signal. On the other hand, the genetic codes simulated under the M_3 assumptions maintained the highest level of ambiguity and the M_2-type codes produced the biggest number of amino acid changes due to point mutations in codons.

It is interesting to consider, which of the simulated codes is the most similar to the SGC based on unambiguity, minimization of point mutations and the structure.

According to the unambiguity measured by entropy or coding strength, the most similar are the codes obtained under the M_1 scenario. They show almost unambiguous assignments of amino acids to codons. However, they are not perfect. Similarly, the SGC is usually presented as a table with the unambiguous assignments but the translation process is not ideal and some errors can occur. It was estimated that one mistranslation occurs with the rate of 10^{-3} to 10^{-6} per codon [85] or 10^{-3} to 10^{-5} per amino acid [86]. Moreover, errors associated with replication and transcription processes can also change the encoded amino acid. If the initial genetic codes had been characterized by much bigger ambiguity of assignments of amino acids to codons, they would have been quickly eliminated by selection, which resembles the rapid decrease in entropy during the simulation of the M_1 codes. The entropy in other models was also reduced but stabilized at the much larger level. It indicates that the assumption on a imprecise recognition of only one fixed codon position is necessary to reduce the initial ambiguity, which corresponds to the wobble rule characterizing the current process of translation.

In terms of minimization of amino acid replacements resulting from point mutations in codons, measured here by the average conductance, the SGC is placed between the M_1 codes, characterized by the lowest conductance, and the codes from the M_2 and M_3 models. In agreement with our simulation study, other analyses also showed that the SGC is not perfectly optimized in this respect and better codes can be found [11, 44, 50, 57, 58, 87–89]. Therefore, it is possible that the assignments of amino acids to codons occurred in accordance with other mechanisms, while the minimization of mutation errors was adjusted by the direct optimization of the mutational pressure around the established genetic code [90–94]. Moreover, some minimization properties of the SGC could have evolved as a by-product of the duplication of genes for tRNAs and aminoacyl-tRNA synthetases charging similar amino acids [6, 8, 41–47]. It is also possible that new amino acids were added into the code in an order that ensured the minimal disturbance of already synthesized proteins but the code itself was not directly optimized [28].

When we compare the structure of the SGC with the structure of the codes produced by the three models, the standard code is the most similar to the M_1 codes because they are also characterized by the domination of amino acids encoded by the groups of two and four codons. All these codon groups are also optimal in terms of the conductance in both the simulated and the SGC. However, four-codon groups are the most numerous in the M_1 codes, while in the SGC the most frequent are two-codon groups, which dominate also in the M_2 codes. The degeneracy of the SGC is usually associated with the presence of codons encoding the same amino acid and differing in the third codon position. It corresponds to the M_1 model, in which the codons for a given amino acid have two fixed positions identical and one different. However, the SGC contains also two codon groups encoding arginine and leucine, which resemble the codon groups in the M_2 model, where all the codons in the groups have one fixed codon position identical and differ in one nucleotide in one of the other two codon positions. Three codons recognized as the stop translation signal also show this property in the SGC. Therefore, the SGC is a mixture of the M_1 and M_2 models in this respect.

The models M_3 and M_2 can represent the initial stages of evolution when the translational apparatus did not read codons perfectly. Therefore, there was a selection to improve the translation process and to develop a stable form of the genetic code. The fixing of two codon positions, represented by the M_1 model, was crucial and enough to unambiguously encode 20 amino acids and the stop translation signal by 64 codons. The wobble base pairing could be a relic of the initial ambiguity.

Since the SGC turned out to be most similar to the codes evolved under the M_1 and M_2 models, we may assume that at certain stage it could have evolved according to the theories proposed by [44, 84] and [28], which means that in the beginning the translation machinery could have recognised only the second codon position, then the first and the third positions. It would be interesting to combine our models with others or enrich it with other biological assumptions to obtain a more accurate model of the evolution of the standard genetic code.

Conclusions

The initial evolution of the standard genetic code could have started from imperfect reading of the genetic information associated with ambiguous assignments of amino acids to codons. Then the selection favoured codes that improved the fidelity of the translation process. An important step was the fixation of two codon positions, which generated the typical codon block structure of the genetic code. According to this hypothesis, the wobble base pairing in the third codon position could be a relic of an early ambiguity. However, the selection for the minimization of translational errors could not have been the only factor influencing the genetic code evolution because its current level of optimization is not perfect. The simulated codes outperformed the standard genetic code in the robustness against mistranslations.

Abbreviations

aaRS: Aminoacyl-tRNA synthetases; EA: Evolutionary Algorithms; mRNA: Messenger RNA; SGC: Standard genetic code; tRNA: Transfer RNA

Acknowledgements

We are very grateful for the insightful comments and remarks of two anonymous reviewers, which significantly improved the manuscript.

Funding
This work was supported by the National Science Centre, Poland (Narodowe Centrum Nauki, Polska) under Grant number 2017/27/N/NZ2/00403.

Availability of data and materials
All the data generated or analysed during this study are included in this published article.

Authors' contributions
Conceived and designed the study: PB, PM. Performed analyses: PB, PM, MW, DM. Analysed and interpreted the data: PB, PM, MW, DM. Wrote the paper and responded to reviewers: PM, PB, MW, DM. All authors participated in the improvement of the manuscript and approved the final version.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 2 October 2018 Accepted: 29 January 2019
Published online: 06 March 2019

References

1. Khorana HG, Buchi H, Ghosh H, Gupta N, Jacob TM, Kossel H, Morgan R, Narang SA, Ohtsuka E, Wells RD. Polynucleotide synthesis and the genetic code. Cold Spring Harb Symp Quant Biol. 1966;31:39–49.
2. Nirenberg M, Caskey T, Marshall R, Brimacombe R, Kellogg D, Doctor B, Hatfield D, Levin J, Rottman F, Pestka S, Wilcox M, Anderson F. The RNA code and protein synthesis. Cold Spring Harb Symp Quant Biol. 1966;31: 11–24.
3. Knight RD, Freeland SJ, Landweber LF. Selection, history and chemistry: the three faces of the genetic code. Trends Biochem Sci. 1999;24(6):241–7.
4. Di Giulio M. The origin of the genetic code: theories and their relationships, a review. Biosystems. 2005;80(2):175–84.
5. Sengupta S, Higgs PG. Pathways of genetic code evolution in ancient and modern organisms. J Mol Evol. 2015;80(5–6):229–43.
6. Koonin EV. Frozen accident pushing 50: Stereochemistry, expansion, and chance in the evolution of the genetic code. Life (Basel). 2017;7(2):22.
7. Kun Á, Radványi Á. The evolution of the genetic code: Impasses and challenges. Biosystems. 2017;164:217–25.
8. Koonin EV, Novozhilov AS. Origin and evolution of the universal genetic code. Annual Review of Genetics. 2017;51:45–62.
9. Ardell DH. On error minimization in a sequential origin of the standard genetic code. J Mol Evol. 1998;47(1):1–13.
10. Ardell DH, Sella G. On the evolution of redundancy in genetic codes. J Mol Evol. 2001;53(4–5):269–81.
11. Błażej P, Wnetrzak M, Mackiewicz P. The role of crossover operator in evolutionary-based approach to the problem of genetic code optimization. Biosystems. 2016;150:61–72.
12. Di Giulio M. The extension reached by the minimization of the polarity distances during the evolution of the genetic code. J Mol Evol. 1989;29(4): 288–93.
13. Di Giulio M, Medugno M. Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. J Mol Evol. 1999;49(1):1–10.
14. Epstein CJ. Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. Nature. 1966;210(5031):25–8.
15. Freeland SJ, Hurst LD. Load minimization of the genetic code: history does not explain the pattern. Proc R Soc B Biol Sci. 1998;265(1410):2111–9.
16. Freeland SJ, Hurst LD. The genetic code is one in a million. J Mol Evol. 1998;47(3):238–48.
17. Freeland SJ, Wu T, Keulmann N. The case for an error minimizing standard genetic code. Orig Life Evol Biosph. 2003;33(4–5):457–477.
18. Freeland SJ, Knight RD, Landweber LF, Hurst LD. Early fixation of an optimal genetic code. Mol Biol Evol. 2000;17(4):511–8.
19. Gilis D, Massar S, Cerf NJ, Rooman M. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biol. 2001;2(11):0049.
20. Goldberg AL, Wittes RE. Genetic code: aspects of organization. Science. 1966;153(3734):420–4.
21. Goodarzi H, Najafabadi HS, Torabi N. Designing a neural network for the constraint optimization of the fitness functions devised based on the load minimization of the genetic code. Biosystems. 2005;81(2):91–100.
22. Haig D, Hurst LD. A quantitative measure of error minimization in the genetic-code. J Mol Evol. 1991;33(5):412–7.
23. Sella G, Ardell DH. The coevolution of genes and genetic codes: Crick's frozen accident revisited. J Mol Evol. 2006;63(3):297–313.
24. Woese CR. On the evolution of the genetic code. Proc Natl Acad Sci U S A. 1965;54(6):1546–52.
25. Ardell DH, Sella G. No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. Philos Trans R Soc B Biol Sci. 2002;357(1427):1625–42.
26. Vetsigian K, Woese C, Goldenfeld N. Collective evolution and the genetic code. Proc Natl Acad Sci U S A. 2006;103(28):10696–701.
27. Guimarães RC, Moreira CHC, de Farias ST. A self-referential model for the formation of the genetic code. Theory Biosci. 2008;127(3):249–70.
28. Higgs PG. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biol Direct. 2009;4:16.
29. Weberndorfer G, Hofacker IL, Stadler PF. On the evolution of primitive genetic codes. Orig Life Evol Biosph. 2003;33(4–5):491–514.
30. Kun Á, Pongor S, Jordán F, Szathmáry E. Catalytic propensity of amino acids and the origins of the genetic code and proteins. Codes Life. 2008;1: 39–58. https://doi.org/10.1007/978-1-4020-6340-4_3.
31. Di Giulio M. The origin of the genetic code. Trends Biochem Sci. 1997;22(2):49–50.
32. Di Giulio M. The coevolution theory of the origin of the genetic code. J Mol Evol. 1999;48(3):253–5.
33. Di Giulio M. The coevolution theory of the origin of the genetic code. Phys Life Rev. 2004;1(2):128–37.
34. Di Giulio M. An extension of the coevolution theory of the origin of the genetic code. Biol Direct. 2008;3:37.
35. Di Giulio M. The lack of foundation in the mechanism on which are based the physico-chemical theories for the origin of the genetic code is counterposed to the credible and natural mechanism suggested by the coevolution theory. J Theor Biol. 2016;399:134–40.
36. Di Giulio M. Some pungent arguments against the physico-chemical theories of the origin of the genetic code and corroborating the coevolution theory. J Theor Biol. 2017;414:1–4.
37. Guimarães RC. Metabolic basis for the self-referential genetic code. Orig Life Evol Biosph. 2011;41(4):357–71.
38. Wong JT. A co-evolution theory of the genetic code. Proc Natl Acad Sci U S A. 1975;72(5):1909–12.
39. Wong JT, Ng SK, Mat WK, Hu T, Xue H. Coevolution theory of the genetic code at age forty: Pathway to translation and synthetic life. Life (Basel). 2016;6(1):12.
40. Wong JTF. Coevolution theory of the genetic code: A proven theory. Orig Life Evol Biosph. 2007;37(4–5):403–8.
41. Ribas de Pouplana L, Schimmel P. Aminoacyl-tRNA synthetases: potential markers of genetic code development. Trends Biochem Sci. 2001;26: 591–6.
42. Cavalcanti AR, Leite ES, Neto BB, Ferreira R. On the classes of aminoacyl-tRNA synthetases, amino acids and the genetic code. Orig Life Evol Biosph. 2004;34(4):407–20.
43. Cavalcanti AR, Neto BD, Ferreira R. On the classes of aminoacyl-trna synthetases and the error minimization in the genetic code. J Theor Biol. 2000;204(1):15–20.
44. Massey SE. A neutral origin for error minimization in the genetic code. J Mol Evol. 2008;67(5):510–6.
45. Massey SE. Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. Life (Basel). 2015;5(2):1301–32.

46. Carter CW, et al. The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed,. Biol Direct. 2014;9:11.
47. Massey SE. The neutral emergence of error minimized genetic codes superior to the standard genetic code. J Theor Biol. 2016;408:237–42.
48. Di Giulio M. The aminoacyl-tRNA synthetases had only a marginal role in the origin of the organization of the genetic code: Evidence in favor of the coevolution theory. J Theor Biol. 2017;432:14–24.
49. de Farias ST, Antonino D, Rêgo TG, José MV. Structural evolution of Glycyl-tRNA synthetases alpha subunit and its implication in the initial organization of the decoding system; 2018. p. 1–8.
50. Novozhilov AS, Wolf YI, Koonin EV. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. Biol Direct. 2007;2:24.
51. Koonin EV, Novozhilov AS. Origin and evolution of the genetic code: The universal enigma. Iubmb Life. 2009;61(2):99–111.
52. Bose RC, Ray-Chaudhuri DK. On a class of error correcting binary group codes. Inf Control. 1960;3:68–79.
53. Fimmel E, Strüngmann L. On the hierarchy of trinucleotide n-circular codes and their corresponding amino acids. J Theor Biol. 2015;364: 113–120.
54. Gumbel M, Fimmel E, Danielli A, Strüngmann L. On models of the genetic code generated by binary dichotomic algorithms. Biosystems. 2015;128:9–18.
55. Fimmel E, Strüngmann L. Mathematical fundamentals for the noise immunity of the genetic code. Biosystems. 2018;164:186–98.
56. Monteagudo A, Santos J. Simulated evolution of the adaptability of the genetic code using genetic algorithms. Bio-Inspired Model Cogn Tasks Pt 1 Proc. 2007;4527:478–87.
57. Santos J, Monteagudo A. Simulated evolution applied to study the genetic code optimality using a model of codon reassignments. BMC Bioinforma. 2011;12:56.
58. Wnetrzak M, Błażej P, Mackiewicz D, Mackiewicz P. The optimality of the standard genetic code assessed by an eight-objective evolutionary algorithm. BMC Evol Biol. 2018;18:192.
59. Tlusty T. A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes. Phys Life Rev. 2010;7(3):362–76.
60. Błażej P, Kowalski D, Mackiewicz D, Wnetrzak M, Aloqalaa D, Mackiewicz P. The structure of the genetic code as an optimal graph clustering problem. bioRxiv. 2018. https://doi.org/10.1101/332478. http://arxiv.org/abs/https://www.biorxiv.org/content/early/2018/05/28/332478.full.pdf.
61. Crick FH. Codon–anticodon pairing: the wobble hypothesis. J Mol Biol. 1966;19(2):548–55.
62. Murphy FVt, Ramakrishnan V. Structure of a purine-purine wobble base pair in the decoding center of the ribosome. Nat Struct Mol Biol. 2004;11(12):1251–2.
63. Agmon I. The dimeric proto-ribosome: structural details and possible implications on the origin of life. Int J Mol Sci. 2009;30:2921–34.
64. Belousoff MJ, Davidovich C, Bashan A, Yonath A. On the development towards the modern world: a plausible role of uncoded peptides in the RNA world. In: Origins of life and evolution of biospheres. vol 40. Berlin: Springer; 2010. p. 415–9.
65. Knight RD, Freeland SJ, Landweber LF. Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet. 2001;2:49–58.
66. Yokobori S, Suzuki T, Watanabe K. Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. J Mol Evol. 2001;53(4–5):314–26.
67. Sengupta S, Higgs PG. A unified model of codon reassignment in alternative genetic codes. Genetics. 2005;170(2):831–40.
68. Swire J, Judson OP, Burt A. Mitochondrial genetic codes evolve to match amino acid requirements of proteins. J Mol Evol. 2005;60(1):128–39.
69. Sengupta S, Yang X, Higgs PG. The mechanisms of codon reassignments in mitochondrial genetic codes. J Mol Evol. 2007;64(6):662–88.
70. Błażej P, Wnetrzak M, Mackiewicz P. The importance of changes observed in the alternative genetic codes; 2018. p. 154–9.
71. Schultz DW, Yarus M. Transfer-RNA mutation and the malleability of the genetic code. J Mol Biol. 1994;235(5):1377–80.
72. Schultz DW, Yarus M. On malleability in the genetic code. J Mol Evol. 1996;42(5):597–601.
73. Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. Novel Ciliate Genetic Code Variants Including the Reassignment of All Three Stop Codons to Sense Codons in Condylostoma magnum,. Mol Biol Evol. 2016;33:2885–9. https://doi.org/10.1093/molbev/msw166.
74. Swart EC, Serra V, Petroni G, Nowacki M. Genetic codes with no dedicated stop codon: Context-dependent translation termination. Cell. 2016;166:691–702. https://doi.org/10.1016/j.cell.2016.06.020.
75. Zahonova K, Kostygov AY, Sevcikova T, Yurchenko V, Elias M. An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. Curr Biol. 2016;26(17): 2364–9.
76. Lovett PS, Ambulos NP, Mulbry W, Noguchi N, Rogers EJ. UGA can be decoded as tryptophan at low efficiency in Bacillus subtilis. J Bacteriol. 1991;173(5):1810–2.
77. Matsugi J, Murao K, Ishikura H. Effect of B. subtilis tRNA(Trp) on readthrough rate at an opal UGA codon. J Biochem. 1998;123(5):853–8.
78. Santos MAS, Ueda T, Watanabe K, Tuite MF. The non-standard genetic code of Candida spp.: an evolving genetic code or a novel mechanism for adaptation? Mol Microbiol. 1997;26(3):423–31.
79. Mitchell M. An Introduction to Genetic Algorithms. Cambridge, MA, USA: MIT Press; 1998.
80. Sivanandam SN, Deepa SN. Introduction to Genetic Algorithms. Berlin, Heidelberg, New York: Springer; 2008.
81. Rabiner L, Juang B. An introduction to hidden markov models. IEEE ASSP Mag. 1986;3(1):4–16. https://doi.org/10.1109/MASSP.1986.1165342.
82. Santos MAS, Cheesman C, Costa V, Moradas-Ferreira P, Tuite MF. Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp,. Mol Microbiol. 1999;31(3):937–47.
83. Gomes AC, Miranda I, Silva RM, Moura GR, Thomas B, Akoulitchev A, Santos MA. A genetic code alteration generates a proteome of high diversity in the human pathogen Candida albicans. Genome Biol. 2007;8(10):R206. https://doi.org/10.1186/gb-2007-8-10-r206.
84. Massey SE. A sequential "2-1-3" model of genetic code evolution that explains codon constraints. J Mol Evol. 2006;62(6):809–10.
85. Ribas de Pouplana L, Santos M, Zhu J, Farabaugh P, Javid B. Protein mistranslation: friend or foe?. Trends Biochem Sci. 2014;39:355–62.
86. Schwartz MH, Pan T. Function and origin of mistranslation in distinct cellular contexts. Crit Rev Biochem Mol Biol. 2017;52(2):205–19.
87. Błażej P, Wnetrzak M, Mackiewicz D, Mackiewicz P. Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. PLoS ONE. 2018;13(8):0201715.
88. Santos J, Monteagudo Á. Inclusion of the fitness sharing technique in an evolutionary algorithm to analyze the fitness landscape of the genetic code adaptability. BMC Bioinforma. 2017;18(1):195. https://doi.org/10.1186/s12859-017-1608-x.
89. Błażej P, Wnetrzak M, Mackiewicz D, Gagat P, Mackiewicz P. Many alternative and theoretical genetic codes are more robust to amino acid replacements than the standard genetic code. J Theor Biol. 2019;464: 21–32.
90. Dudkiewicz A, Mackiewicz P, Nowicka A, Kowalczuk M, Mackiewicz D, Polak N, Smolarczyk K, Banaszak J, Dudek MR, Cebrat S. Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution. Futur Gener Comput Syst. 2005;21(7): 1033–9.
91. Mackiewicz P, Biecek P, Mackiewicz D, Kiraga J, Baczkowski K, Sobczynski M, Cebrat S. Optimisation of asymmetric mutational pressure and selection pressure around the universal genetic code. Comput Sci ICCS 2008 Pt 3 Ser Lect Notes Comput Sci. 2008;5103:100–9.
92. Błażej P, Mackiewicz P, Cebrat S, Wanczyk M. Using evolutionary algorithms in finding of optimized nucleotide substitution matrices. In: Genetic and Evolutionary Computation Conference, GECCO '13, Amsterdam, The Netherlands, July 6-10, 2013, Companion Material Proceedings; 2013. p. 41–2. https://doi.org/10.1145/2464576.2464598. https://doi.org/10.1145/2464576.2464598.
93. Błażej P, Mackiewicz D, Grabinska M, Wnetrzak M, Mackiewicz P. Optimization of amino acid replacement costs by mutational pressure in bacterial genomes. Sci Rep. 2017;7:1061. https://doi.org/10.1038/s41598-017-01130-7.
94. Błażej P, Miasojedow B, Grabinska M, Mackiewicz P. Optimization of mutation pressure in relation to properties of protein-coding sequences in bacterial genomes. PLoS ONE. 2015;10:0130411. https://doi.org/10.1371/journal.pone.0130411.