BioSystems 210 (2021) 104528

Contents lists available at ScienceDirect

BioSystems

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

Models of genetic code structure evolution with variable number of coded labels

Konrad Pawlak, Małgorzata Wnetrzak, Dorota Mackiewicz, Paweł Mackiewicz, Paweł Błażej [*]

Department of Bioinformatics and Genomics, Faculty of Biotechnology, University of Wrocław, ul. Joliot-Curie 14a, Wrocław, Poland

ARTICLE INFO

Keywords:
Genetic code
Codon
Amino acid

ABSTRACT

It is assumed that at the early stage of cell evolution its translation machinery was characterized by high noise, i.e. ambiguous assignment of codons to amino acids in the genetic code, which initially encoded only few amino acids. Next, during its evolution new amino acids were added to this code. Taking into account this facts, we investigated theoretical models of genetic code's structure, which evolved from a set of ambiguous codons assignments into a coding system with a low level of uncertainty. We considered three types of translational inaccuracies assuming a different number of fixed codon positions. We applied a modified version of evolutionary algorithm for finding the genetic codes that the most effectively reduced the initial uncertainty in the assignment of codons to encoded labels, i.e. amino acids and a stop translation signal. We examined codes with the number of labels from four to 22. Our results indicated that the quality of genetic code structure is strongly dependent on the number of encoded labels as well as the type of translational mechanism. The more strict assignments of codon to the labels was preferred by the codes encoding more number of labels. The results showed that a smaller degeneracy of codes evolved from a more tolerant coding with the stepwise addition of coded amino acids to the genetic code. The distribution of codon groups in the standard genetic code corresponds well to the translation model assuming two fixed codon positions, whereas the six-codon groups can be relics form previous stages of evolution when the code characterized by a greater uncertainty.

1. Introduction

The standard genetic code (SGC) is a template according to which 64 codons are assigned to 20 amino acids and the stop coding signal. Because the number of codons is greater then the number of encoded labels this coding system is redundant, which means that there exist codons that encode the same genetic information. These codons are arranged in codon groups called blocks. They consist of two, three, four or six codons. Generally, the codons in these groups differ in the third position. In order to explain this phenomenon, Crick put forward the wobble hypothesis, which assumes specific interactions between the first base in a tRNA anticodon and the third base of translated codon in a transcript (mRNA) (Crick, 1966). He proposed that the base pairing between two nucleotides in RNAs do not have to follow Watson–Crick base pair rules, i.e. cytosine–guanine and adenine–uracil, but other interactions are also possible, i.e. guanine–uracil, hypoxanthine–uracil, hypoxanthine–adenine and hypoxanthine–cytosine. It was also found that other modified bases can pair with the typical ones (Murphy and Ramakrishnan, 2004). This fact has many interesting consequences. For example, it reduces the number of different tRNA molecules, which are necessary in protein synthesis. What is more, single point mutations that occur in the third codon positions are synonymous, i.e. do not change the encoded information. It causes that the SGC is to some extent robust against consequences of nucleotide substitutions.

It should be noted that the wobble rule is just one out of many attempts to explain the characteristic structure of the SGC. This issue and the properties of the SGC have been hotly debated since the first codon assignment was deciphered in the sixties of the twentieth century (Khorana et al., 1966; Nirenberg et al., 1966). Nowadays in the academic world, there exist several hypotheses trying to explain the evolution of genetic code (Knight and Landweber, 1999; Di Giulio, 2005; Barbieri, 2015; Sengupta and Higgs, 2015; Koonin, 2017; Koonin and Novozhilov, 2017; Kun and Radvanyi, 2018). They focus on different features, which would be a driving force of the genetic code emergence. A popular adaptive hypothesis assumes that the present structure of the SGC has evolved to minimize the harmful effects of mutations and mistranslations (Sonneborn, 1965; Woese, 1965; Epstein, 1966; Goldberg and Wittes, 1966; Haig and Hurst, 1991; Freeland and Hurst, 1998; Di Giulio, 1999; Gilis et al., 2001; Freeland et al., 2003; Goodarzi et al., 2005). The mathematical analysis of the structure and symmetry

[*] Corresponding author.
E-mail address: pawel.blazej@uwr.edu.pl (P. Błażej).

https://doi.org/10.1016/j.biosystems.2021.104528
Received 17 May 2021; Received in revised form 26 August 2021; Accepted 27 August 2021
Available online 4 September 2021

of the genetic code confirmed its immunity to noise in terms of error-detection and error-correction (Fimmel et al., 2015; Gumbel et al., 2015; Fimmel et al., 2018). The computer simulations and optimization analyzes using genetic algorithms also showed a general tendency of the SGC to minimize errors, but it did not appear perfectly optimized in comparison to theoretical codes (Novozhilov et al., 2007; Massey, 2008; Santos et al., 2011; Błażej et al., 2016; Santos and Monteagudo, 2017; Wnetrzak et al., 2018; Błażej et al., 2018c, 2019b; Wnetrzak et al., 2019). What is more, alternative versions of the SGC turned out to be better at mitigating mutations and translational errors (Błażej et al., 2018b, 2019a). The structural properties of codon blocks in the SGC were also studied on the basis of graph theory (Błażej et al., 2018a; Aloqalaa et al., 2020; Błażej et al., 2020). The analyzes showed that the majority of codon blocks present in the SGC are optimal according to the conductance measure. However, the SGC turned out to be far from the optimum according to this measure. Another approach points out that the current codon assignments represent relationships between respective amino acids in biosynthetic pathways, i.e. a newly added amino acids took over some codons from a respective groups encoding their precursors (Wong, 1975; Di Giulio, 1997; Di Giulio and Medugno, 1999; Di Giulio, 2008, 2016; Guimaraes, 2011; Wong et al., 2016).

Nevertheless, it should be noted that it is still unclear which factor played a decisive role in the origin and evolution leading to the present SGC. It is not inconceivable that its evolution could a combination of many factors (Koonin and Novozhilov, 2009). This fact opens the field for improvements of existing models as well as construction of new ones.

The early genetic code most likely characterized by a high translational noise, which was further reduced during its evolution (Fitch and Upper, 1987; Barbieri, 2015; Błażej et al., 2019b). Such a state remained likely a long time because the last universal common ancestor of three domains of life, bacteria, archaea and eukaryotes was still a progenote, with not fully developed translational apparatus (Di Giulio, 2001; Giulio, 2014; Di Giulio, 2020a,b). Likewise, amino acids were gradually added to the evolving code. This incorporation was driven by catalytic properties of amino acids functioning in ribozymes (Kun et al., 2008) and was beneficial because it increased the diversity of synthesized proteins (Higgs, 2009; Koonin and Novozhilov, 2017; Sengupta and Higgs, 2015; Weberndorfer et al., 2003).

The order of the amino acid addition into the code was determined by the minimizing disorders in already synthesized proteins (Higgs, 2009), dependence between the amino acids in terms of precursor-product in metabolic pathways (Wong, 1975; Di Giulio, 1997; Di Giulio and Medugno, 1999; Di Giulio, 2008, 2016; Guimaraes, 2011; Wong et al., 2016) or duplications of genes coding for tRNAs and aminoacyl-tRNA synthetases (Cavalcanti et al., 2000, 2004; Massey, 2015, 2016; Koonin, 2017; Koonin and Novozhilov, 2017).

Here, we propose a model of the genetic code evolution including these two aspects, i.e. the reduction of translational noise and the stepwise addition of amino acid into the code. We considered three scenarios of translation inaccuracies in coding systems, namely $M1, M2$ and $M3$, assuming a different number of encoded labels. M_1 assumes that a given amino acid is coded by codons that have two fixed positions identical and differ in one position. In M_2, a given amino acid is coded by codons with one fixed position identical and differ in exactly one of other two codon positions from the reference codon. M_3 codons coded a given amino acid differ in exactly one any codon position from the reference codon. Therefore, $M1$ is a special case of the wobble rule, whereas $M2$ and $M3$ are its generalizations.

2. Methods

2.1. Overview

We investigated the evolution of genetic coding systems, which encode a different number of amino acids in comparison to the SGC. The evolution of these genetic codes started from a set of ambiguous assignments of 64 codons to a fixed number of labels, i.e. amino acids and stop translation signal, and evolved towards coding systems characterized by a low uncertainty of assignments of these labels to the codons. Similar to the previous approach (Błażej et al., 2019b), we run our simulations using a modified version of evolutionary algorithm, in which the genetic codes were represented by a population of candidate solutions (individuals).

The simulation procedure was divided into consecutive steps called generations. During each generation two operators, mutation and selection were applied to the population of evolving genetic codes. These operators were responsible for diversity of this population and guaranteed that generally better solutions took part in reproduction for the next generation. The codes with higher probabilities of encoding unambiguous genetic information were preferred in reproduction to the next generation.

2.2. Representation of genetic codes

As in Błażej et al. (2019b), the evolving coding system were represented by a matrix $\mathcal{P} = (p_{cl})$ consisting of 64 rows and the number of columns equal to the number of considered labels L. Values in each row $c = 1, 2, 3, \ldots, 64$ describe a probability distribution function of labels $l = 1, 2, \ldots, L$ by a given codon. Therefore, the value of the element p_{cl} in the matrix $\mathcal{P}$ is the probability that a codon c encodes a label l.

All simulations started with a population of individuals, which were represented by their respective randomly generated matrices $\mathcal{P}$. In Fig. 1, we depicted a genetic code that encodes 15 labels at the beginning of simulations. In the heatmap plot, each element of the matrix is represented by a cell and its brightness corresponds to the probability that a given codon (in a row) encodes a respective label (in a column).

2.3. Fitness function

The quality of a given coding system was measured by the fitness function F, which corresponded to the probability that a genetic code encodes a given set of labels. More specifically, this function was the sum of the products of probabilities that a given label l was encoded by a codon c^l (Błażej et al., 2019b). For each label l, a codon c^l with the highest probability to encode l was chosen to describe the most probable coding path $C = c^1 c^2 \ldots c^l$ in a given code.

In this calculation for every c^l, we also included a codon neighborhood $N(c^l)$, which was a set of codons containing not only the codon c^l but also several codons that differed from c^l in one position. These codons were chosen according to three rules called M_1, M_2 and M_3. They can be described in the following way:

- M_1 states that codons belonging to a given $N(c^l)$ have two fixed codon positions identical and differ in any codon position, but exactly one position is changed;
- M_2 states that codons belonging to a given $N(c^l)$ have one fixed codon position identical and differ in exactly one of other two codon positions from the reference codon;
- M_3 states that codons belonging to a given $N(c^l)$ differ in exactly one any codon position from the reference codon.

For example, the neighborhood for the codon GGG includes:

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

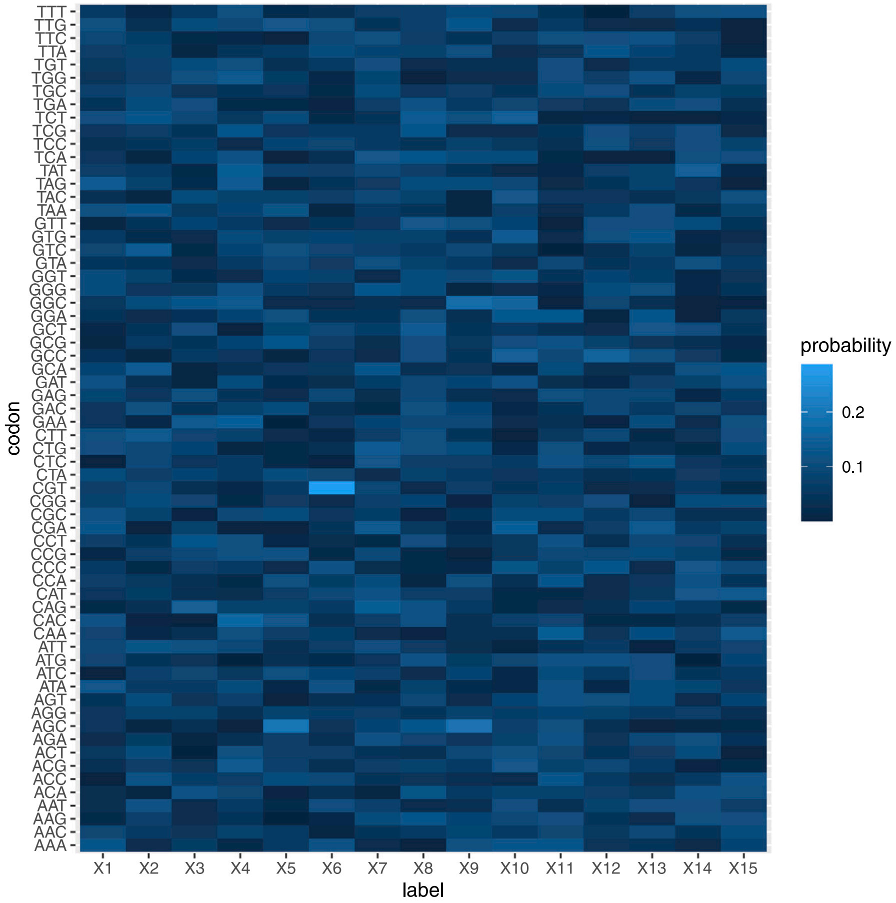

Fig. 1. The heatmap of a genetic code encoding four labels at the beginning of the simulation run under scenario M_1. This is in fact a graphical representation of the matrix $\mathcal{P} = (p_{cl})$, in which each element p_{cl} has ascribed a probability that a codon c in a row encodes a label l in a column. The probability value of this encoding is represented by brightness.

Therefore, for a given codon, its neighborhood under the rule M_3 may include the neighborhood M_1 and M_2. These rules represent different mechanisms of reading transcripts by the translational machinery, which induces various types of genetic code redundancy.

In contrast to Błażej et al. (2019b), who analyzed the codes with 21 labels as in the SGC, we considered here coding systems with a different number of encoded labels starting from the genetic codes with only four labels and ending with those consisting of 22 labels. For each type of rules, i.e. M_1, M_2 and M_3, we run simulations with the fixed number of labels.

2.4. Measure of the quality of genetic codes

In order to describe the structural properties of the genetic codes represented as the matrix $\mathcal{P} = (p_{cl})$, we applied the entropy

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

which is the sum of Shannon entropy calculated for each row of the matrix $\mathcal{P}$ over the probabilities that a codon c encodes a label l. Thus, $H(\mathcal{P})$ corresponds to the multidimensional entropy of independent distributions. In consequence, we obtained a measure of genetic code uncertainty because higher values of the entropy indicate that a given coding system is composed of ambiguous assignments of codons to labels. Conversely, unambiguous genetic codes are characterized by lower values of $H(\mathcal{P})$.

3. Results

We started our investigations with finding the best theoretical genetic codes that encode the selected number of labels $L = 4, 5, \dots, 22$. These coding systems emerged under three assumptions M_1, M_2 and M_3 for imprecise translation of transcripts to proteins. All simulations were run over 40,000 generations and were repeated for different 40 seeds. The conducted simulations provided the best genetic codes in terms of the fitness function F, which describes the ability of a given coding system to encode a fixed number of labels. After getting the convergence of the fitness function, we got the optimal coding structures for each number of labels.

3.1. Structure of genetic codes

Fig. 2 presents the optimal genetic code computed under the M_1 scenario. This coding system encodes 15 labels. As you can see, this code is characterized by low ambiguity because in their structure dominate codons with a high probability of encoding a given label. In this case, each label is encoded with a very high probability by four codons differing in one position. A weak ambiguity remained only for four codons, which can encode two, three or four labels.

In Fig. 3, we gathered information about the structure of all genetic codes produced in the simulations under three types of translational inaccuracies, i.e. M_1, M_2 and M_3, and regarding the number of encoded labels L from four to 22. Generally, the codes encoding a smaller number of labels consist of more numerous codons groups for these labels. For $L = 4$, the most frequent are groups including 14–17 codons.

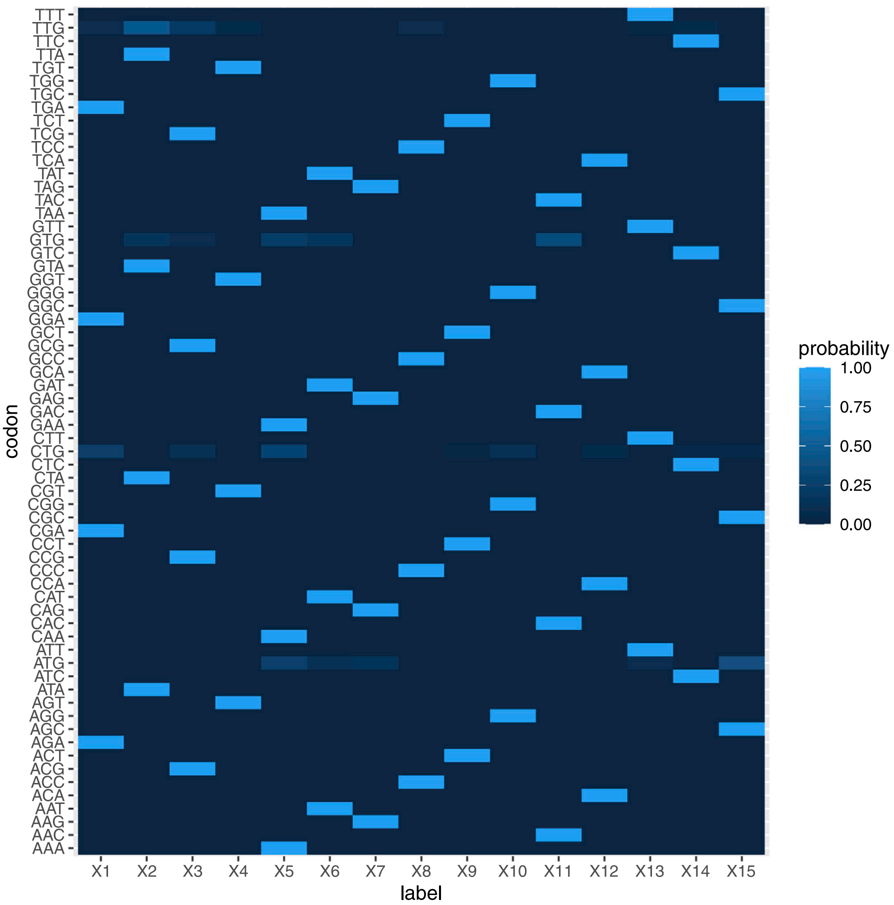

Fig. 2. The heatmap of the optimal genetic code with 15 labels, obtained after the simulation run under scenario M_1. This is in fact a graphical representation of the matrix $\mathcal{P} = (p_{cl})$, in which each element p_{cl} has ascribed a probability that a codon c in a row encodes a label l in a column. The probability value of this encoding is represented by brightness.

Next, with the increasing number of coded labels, the size of these groups decreases, e.g. for $L = 6$ and $L = 7$ the most frequent are groups with 9–11 codons, for $L = 8$ and $L = 9$ those with 6–8 codons, for $L = 12$ and $L = 14$ those with five codons, and for $L = 15$-17 those with four codons. Interestingly, the groups consisting of two codons begin to obtain a significant contribution only in the case of codes for the rule M_1 and encoding more than 9 labels, which makes the distribution bimodal. For $L = 22$, the two-codon groups exceed the number of four-codon groups. There is also an additional difference between the codes optimized under different translational inaccuracies. When $L >= 18$, the codes for the rule M_2 and M_3 have the most frequent groups including three codons, which are very poorly represented for the codes under the restriction M_1, in which the groups with four and two codons are dominated.

The changes in the size of codon groups are well visible in Fig. 4. For the genetic codes optimized under the type M_1 of translational inaccuracy, the contribution of groups with four and two codon increases gradually with the number of coded labels, whereas the groups comprising other number of codons C show the largest frequency for the specific number of labels L, e.g. $C = 5$ for $L = 13$, $C = 6$ for $L = 11$, $C = 7$ for $L = 9$ and $C = 8$ for $L = 8$. More numerous groups, i.e. those with 14 or more codons are the most frequent in the codes encoding only four labels and become rare in the codes encoding more labels. The one-codon and three-codon groups are poorly represented in any codes independently from the number of encoded labels.

For the codes simulated with M_2 and M_3 rules, the gradual increase with the number of coded labels is also demonstrated by the groups comprising two codons as for the codes under the M_1 rule but their contribution is much smaller (Fig. 4). In contrast to the M_1 code, the positive relationship between the codon group frequency and the coded labels is revealed by the groups with three codons, whereas four-codon groups do not show this trend and are the most frequent for $L = 15$ and $L = 16$. Like the codes under M_1 restrictions, other codon groups in the M_2 and M_3 codes has the highest frequency for the specific number of coded labels, e.g. the groups with the size $C = 5$ for $L = 12$ and $L = 13$, $C = 6$ for $L = 10$, $C = 7$ for $L = 9$ and $C = 8$ for $L = 8$. Similarly, the groups with $C >= 14$ dominate in the codes with four labels and the groups with one codon are marginally used.

The most frequent codon groups in the codes optimized under M_2 and M_3 scenarios well correspond to the expected values obtained after division of the number of all 64 codons by the respective number of coded labels L (Fig. 3). The size of such codon groups concerns the case when the individual labels are uniformly coded by the same or similar number of codons. Interestingly, this correspondence deviates for the codes obtained for the rule M_1 when $L >= 18$. It should be also noted that the codes under M_1 and $L = 21$, are much more similar to the codon size distribution in the SGC than the codes fulfilling the other rules of transnational ambiguity.

3.2. Unambiguity level of genetic codes

In order to compare the studied coding systems, which differed in the type of translational inaccuracies M_1, M_2 and M_3 and the number of encoded genetic information, i.e. labels L, we applied the entropy $H(\mathcal{P})$, which is a good measure of coding ambiguity. In Fig. 5, we presented the relationship between the entropy and the number of

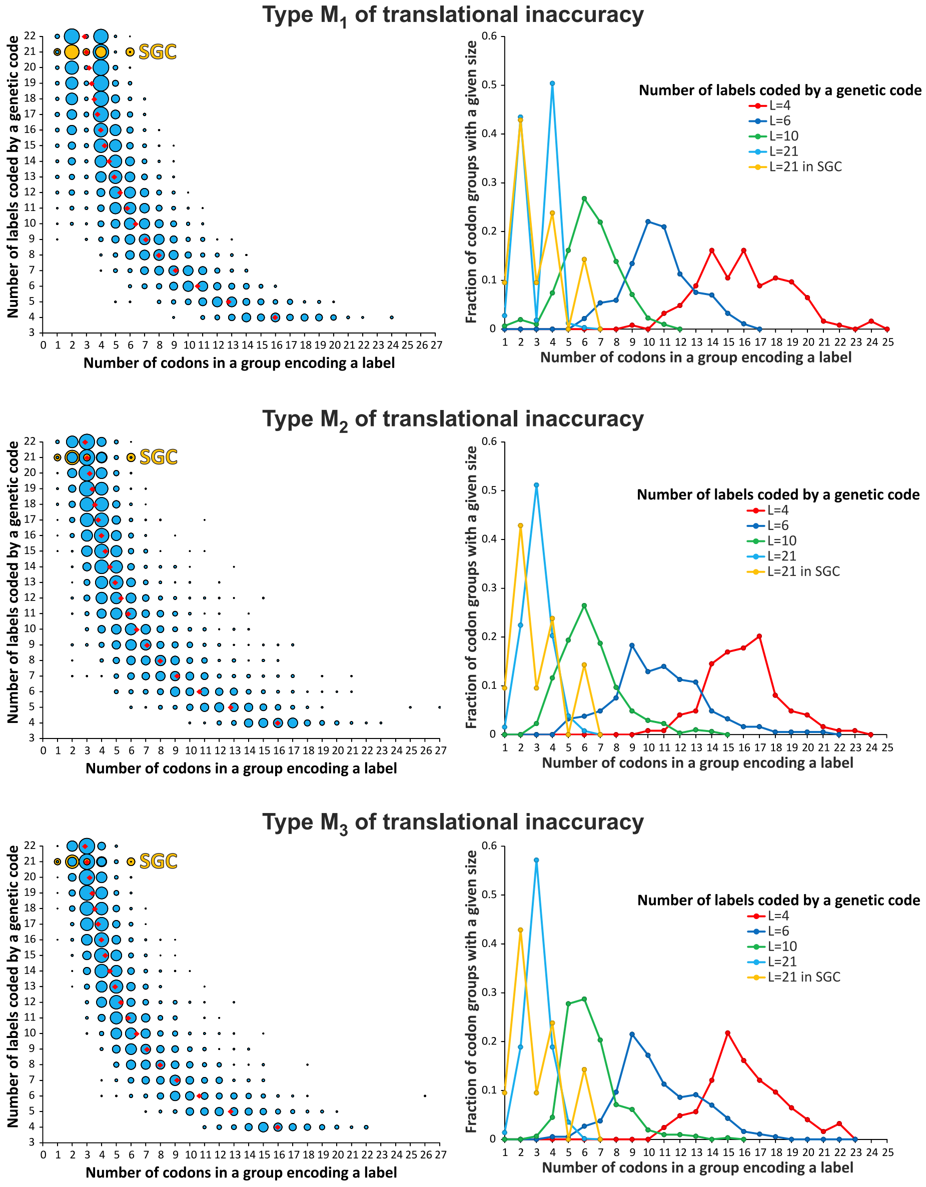

Fig. 3. The structure of genetic codes obtained under three types of translational inaccuracies, i.e. M_1, M_2 and M_3, and taking into account the various number of encoded labels L, from four to 22. The plots in the left panel show the fraction of codon groups with a given size encoding a specific label for the codes with a fixed number of encoded labels. This fraction is reflected by the area of circles. The plots in the right panel show selected distributions of codon group sizes. SGC means the standard genetic code. The red diamonds indicate expected values obtained after division of the number of 64 codons by the number of coded labels L.

encoded labels, calculated under the M_1, M_2 and M_3 assumptions. In fact, the curves show the average of genetic coding entropy calculated from all respective simulations. Interestingly, the relationships are not linear, and the least ambiguity in the assignment of codons to genetic information is reached for different combinations of labels and the assumptions on the translational inaccuracy.

The genetic codes optimized under rule M_1 are characterized by very high ambiguity in the assignment of small number of labels to codons and reach the maximum entropy for $L = 6$ (Fig. 5). Next, its value drops rapidly and reaches the minimum for the codes with 22 labels. In turn, the entropy of codes simulated with M_2 assumption is the smallest for four labels. Next it gradually increases slightly drooping for 11 and 12 labels and again rises to the maximum for $L = 22$. The entropy of the codes assuming the translational inaccuracy M_3 is very high, when these codes encode four labels. Then, it decreases to the minimum at $L = 15$ and grows up for the greater number of labels in these codes.

These complex relationships cause that, for a given number of labels, the codes with lowest entropy are for different type of translational inaccuracies M_1, M_2 and M_3 (Fig. 5). The genetic codes optimized under the M_3 restriction are generally the best possible solutions in terms of the most unambiguous assignment for very small number of encoded labels, i.e. four and five, but they turn out to be worse in comparison to the genetic codes computed under the M_2 and M_3 assumptions for a larger number of labels. For six to 14 labels, the codes with the lowest entropy are characterized by the translational inaccuracy of type M_2, whereas for the more labels encoded, the codes under the rule M_1 show the lowest entropy in comparison to the others. It should be noted, that the codes with rule M_1 are characterized by the largest range of $H(\mathcal{P})$, reaching the largest and the smallest entropy values of all simulated conditions.

4. Discussion

In this study, we performed simulations of genetic code structures assuming various number of labels encoded by these codes. These labels

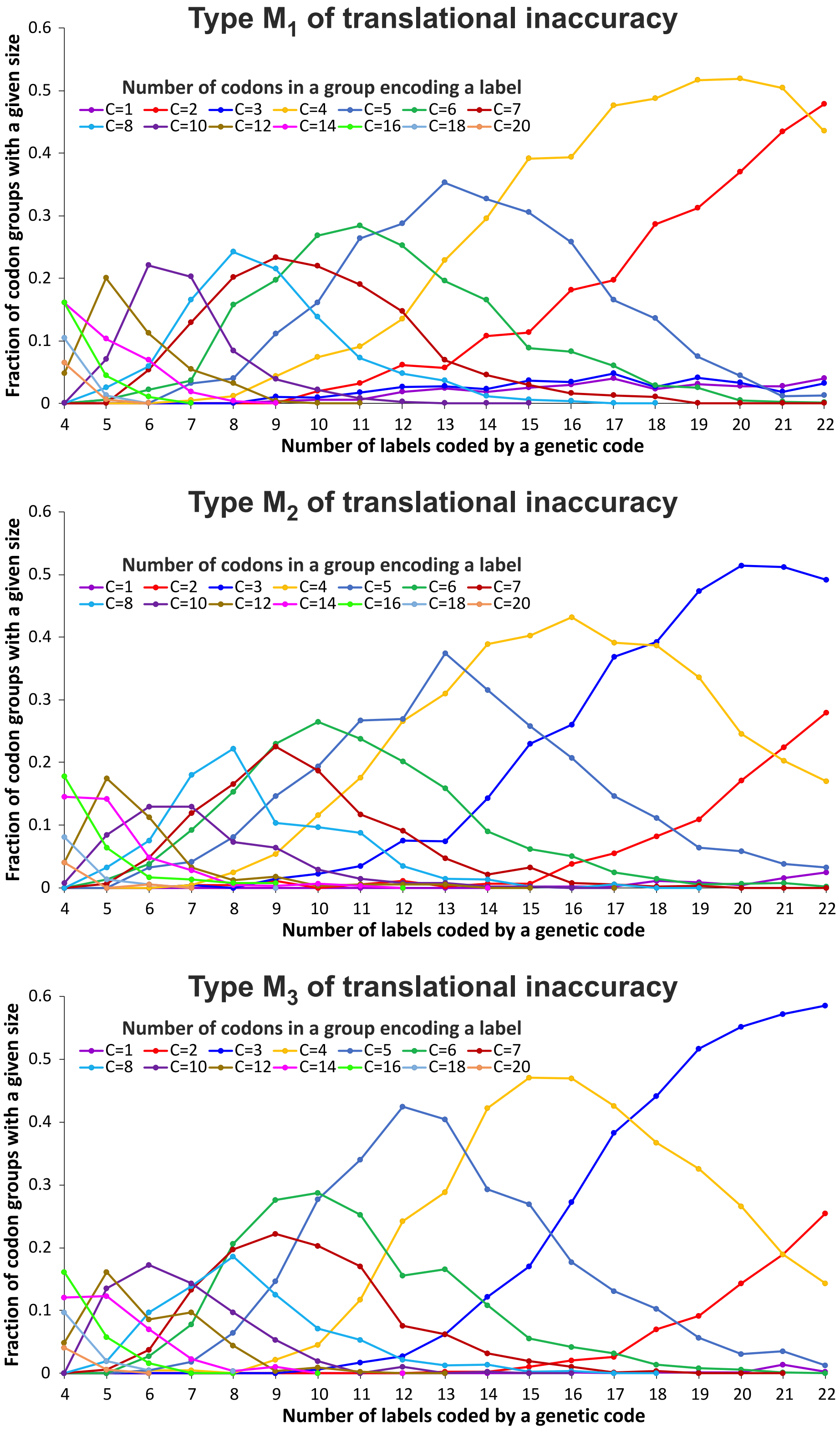

Fig. 4. Relationship between the fraction of codon groups with a given size and the number of labels coded by genetic codes obtained under three types of translational inaccuracies, i.e. M_1, M_2 and M_3. The data only for selected sizes of codon groups were shown for clarity.

represent amino acids and the stop translation signal. We considered codes with four to 22 labels. The assumption on the minimum, i.e. four labels, corresponds to the four-column model for the origin of the genetic code, which stared its evolution just from such a number of amino acids (Higgs, 2009). The comparison of the simulated codes ranked according to the number of encoded labels may represent a gradual addition of amino acids into the evolving code. We studied codes up to 22 labels because we took into account not only the classical 20 amino acids and stop translation signal but also the presence of selenocysteine or pyrrolysine in some genetic codes (Böck et al., 1991; Srinivasan et al., 2002).

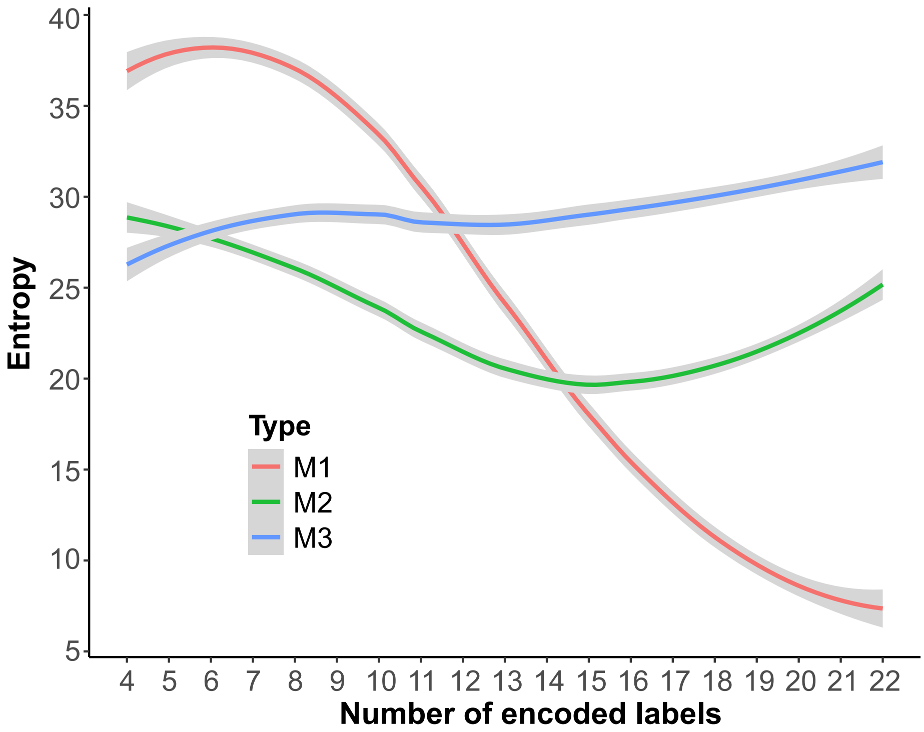

Fig. 5. Relationships between the genetic code entropy and the number of labels coded by the genetic codes simulated under three types of translational inaccuracies. The curves are the result of fitting obtained from Generalized Additive Model for 40 simulations run under different seeds. The gray band represents the 95%-confidence interval.

The simulation started with codes characterized by ambiguous assignment of codons to the labels and evolved to reduce this uncertainty (measured by the fitness function) under three models of translational inaccuracies M_1, M_2 and M_3. These rules assume various levels of translation imprecision and are defined by a neighborhood to a reference codon. The model M_3 is the most tolerant and assumes that nine other codons can have the same meaning as the reference codon. In rule M_2, there are six such codons and in the case of M_1, three additional codons can encode the same label. The codons differ in one position but the number of fixed positions depends on the model.

The models of translational inaccuracies assume that similar codons encode a specific amino acid, which is in agreement with the mechanisms of amino acid addition to the genetic code. One of them assumes that newly added amino acids captured codons due to duplications of genes coding for tRNAs and aminoacyl-tRNA synthetases (Cavalcanti et al., 2000, 2004; Massey, 2015, 2016; Koonin, 2017; Koonin and Novozhilov, 2017). The products of the duplicated genes most likely recognized initially groups of similar codons and there existed an ambiguity in the coding of amino acids until the whole system became more precise (Fitch and Upper, 1987; Barbieri, 2015).

Results of our simulations indicate that the structure of coding systems optimal in terms of unambiguous translation strongly depends on the number of coded labels. We found an interesting succession of code types with gradual addition of labels into the codes. When a code had to encode a few labels, four and five, the best reduction of translational uncertainty was under the M_3 rule. Next, the M_2 model was preferred by the codes encoding six to 14 labels, whereas for 15 and more labels, the least tolerant assumption M_1 was the best solution. Therefore, a smaller degeneracy of codes evolved from a greater ambiguity with the addition of coded items into the genetic code. The results suggest that these three different systems of translational inaccuracy can be optimal at different stages of the genetic code evolution. The replacement of these systems well corresponds to the 2-1-3 model (Massey, 2006, 2008) and the four-column theory (Higgs, 2009), which also postulate a subsequent reduction of degeneracy. In the initial genetic code, the second codon position determined the coded amino acids, whereas other positions were less crucial. Next, the first codon position became more specific and finally the third position differentiated some amino acids. Our finding show that the system of reading genetic information must have changed when new genetic information was added into the repertoire of coded amino acids.

The structure of the present standard genetic code is very similar to the codes optimized under the M_1 assumption for 21 labels. The most frequently used groups consists of two and four codons. In the SGC, there are also three six-codon groups, which disappeared in the optimized codes but are quite frequent in the codes encoding 10 or 11 amino acids. It can suggest that these groups are relics from previous stages of the genetic code evolution.

Our analyzes showed that fixation of two codon positions could be enough to reduce ambiguity in the assignment of codons to amino acids and produce the structure similar to that in the SGC. However, the current translational machinery is still characterized by mistranslation with the rate of 10^{-3} to 10^{-6} per codon (Ribas de Pouplana et al., 2014) or 10^{-3} to 10^{-5} per incorporated amino acid (Kramer and Farabaugh, 2007; Schwartz and Pan, 2017; Allan Drummond and Wilke, 2009; Mordret et al., 2019), which is much higher than DNA replication errors, i.e. 10^{-9} to 10^{-10} per residue (Lee et al., 2012; Zhu et al., 2014). Thereby, 15% of average-length protein molecules can contain at least one misincorporated amino acid (Allan Drummond and Wilke, 2009). Apparently, the minimization of mutation errors associated with replication was more important and was optimized around the established genetic code (Dudkiewicz et al., 2005; Mackiewicz et al., 2008; Błażej et al., 2015, 2017). The relatively high mistranslation rate could be in some cases profitable, e.g. in adaptation to oxidative and environmental stresses as well as in host invasion and evasion of immunity by parasites (Santos et al., 1999; Gomes et al., 2007; Netzer et al., 2009; Wiltrout et al., 2012; Miranda et al., 2013) although proteins that are synthesized with errors can incorrectly fold and lose their functions, which generally decreases fitness of organisms (Allan Drummond and Wilke, 2009).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Maja Orłowska and Kuba Nowak for their insightful remarks that significantly improved the paper.

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

References

Allan Drummond, D., Wilke, C.O., 2009. The evolutionary consequences of erroneous protein synthesis. Nature Rev. Genet. 10 (10), 715–724. http://dx.doi.org/10.1038/nrg2662.

Aloqalaa, D.A., Kowalski, D.R., Błażej, P., Wnetrzak, M., Mackiewicz, D., Mackiewicz, P., 2020. The properties of the standard genetic code and its selected alternatives in terms of the optimal graph partition. In: Roque, A., Tomczyk, A., De Maria, E., Putze, F., Moucek, R., Fred, A., Gamboa, H. (Eds.), Biomedical Engineering Systems and Technologies. Springer International Publishing, Cham, pp. 170–191.

Barbieri, M., 2015. Evolution of the genetic code: The ribosome-oriented model. Biol. Theory 10 (4), 301–310.

Błażej, P., Kowalski, D., Mackiewicz, D., Wnetrzak, M., Aloqalaa, D., Mackiewicz, P., 2018a. The structure of the genetic code as an optimal graph clustering problem. https://www.Biorxiv.Org/Content/Early/2018/05/28/332478.

Błażej, P., Mackiewicz, D., Grabinska, M., Wnetrzak, M., Mackiewicz, P., 2017. Optimization of amino acid replacement costs by mutational pressure in bacterial genomes. Sci. Rep. 7, 1061. http://dx.doi.org/10.1038/s41598-017-01130-7.

Błażej, P., Miasojedow, B., Grabinska, M., Mackiewicz, P., 2015. Optimization of mutation pressure in relation to properties of protein-coding sequences in bacterial genomes. PLoS One 10, e0130411. http://dx.doi.org/10.1371/journal.pone.0130411.

Błażej, P., Wnetrzak, M., Mackiewicz, P., 2016. The role of crossover operator in evolutionary-based approach to the problem of genetic code optimization. BioSystems 150, 61–72.

Błażej, P., Wnetrzak, M., Mackiewicz, P., 2018b. The importance of changes observed in the alternative genetic codes. In: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 4: BIOINFORMATICS, pp. 154–159.

Błażej, P., Wnetrzak, M., Mackiewicz, D., Gagat, P., Mackiewicz, P., 2019a. Many alternative and theoretical genetic codes are more robust to amino acid replacements than the standard genetic code. J. Theoret. Biol. 464, 21–32.

Błażej, P., Wnetrzak, M., Mackiewicz, D., Mackiewicz, P., 2018c. Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. PLoS One 13 (8), e0201715.

Błażej, P., Wnetrzak, M., Mackiewicz, D., Mackiewicz, P., 2019b. The influence of different types of translational inaccuracies on the genetic code structure. BMC Bioinformatics 20 (1), 114.

Błażej, P., Wnetrzak, M., Mackiewicz, D., Mackiewicz, P., 2020. Basic principles of the genetic code extension. R. Soc. Open Sci. 7 (2), 191384. http://dx.doi.org/10.1098/rsos.191384, arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rsos.191384, https://royalsocietypublishing.org/doi/abs/10.1098/rsos.191384.

Böck, A., Forchhammer, K., Heider, J., Leinfelder, W., Sawers, G., Veprek, B., Zinoni, F., 1991. Selenocysteine: the 21st amino acid. Mol. Microbiol. 5, 515–520.

Cavalcanti, A.R.O., Leite, E.S., Neto, B.B., Ferreira, R., 2004. On the classes of aminoacyl-tRNA synthetases, amino acids and the genetic code. Origins Life Evol. Biosph. 34 (4), 407–420.

Cavalcanti, A.R., Neto, B.D., Ferreira, R., 2000. On the classes of aminoacyl-tRNA synthetases and the error minimization in the genetic code. J. Theoret. Biol. 204 (1), 15–20.

Crick, F.H., 1966. Codon - anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19 (2), 548–555.

Di Giulio, M., 1997. The origin of the genetic code. Trends Biochem. Sci. 22 (2), 49–50.

Di Giulio, M., 1999. The coevolution theory of the origin of the genetic code. J. Mol. Evol. 48 (3), 253–255.

Di Giulio, M., 2001. The non-universality of the genetic code: the universal ancestor was a progenote. J. Theoret. Biol. 209 (3), 345–349.

Di Giulio, M., 2005. The origin of the genetic code: theories and their relationships, a review. BioSystems 80 (2), 175–184.

Di Giulio, M., 2008. An extension of the coevolution theory of the origin of the genetic code. Biol. Direct. 3, 37.

Di Giulio, M., 2016. 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. Theoret. Biol. 399, 134–140.

Di Giulio, M., 2020a. LUCA as well as the ancestors of archaea, bacteria and eukaryotes were progenotes: Inference from the distribution and diversity of the reading mechanism of the AUA and AUG codons in the domains of life. Bio Syst. 198, 104239.

Di Giulio, M., 2020b. The phylogenetic distribution of the glutaminyl-tRNA synthetase and Glu-tRNA(Gln) amidotransferase in the fundamental lineages would imply that the ancestor of archaea, that of eukaryotes and LUCA were progenotes. Bio Syst. 196, 104174.

Di Giulio, M., Medugno, M., 1999. Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. J. Mol. Evol. 49 (1), 1–10.

Dudkiewicz, A., Mackiewicz, P., Nowicka, A., Kowalezuk, M., Mackiewicz, D., Polak, N., Smolarczyk, K., Banaszak, J., Dudek, M.R., Cebrat, S., 2005. Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution. Future Gener. Comput. Syst. 21 (7), 1033–1039.

Epstein, C.J., 1966. Role of the amino-acid ''code'' and of selection for conformation in the evolution of proteins. Nature 210 (5031), 25–28.

Fimmel, E., Giannerini, S., Gonzalez, D.L., Strungmann, L., 2015. Circular codes, symmetries and transformations. J. Math. Biol. 70 (7), 1623–1644.

Fimmel, E., Michel, C.J., Starman, M., Strungmann, L., 2018. Self-complementary circular codes in coding theory. Theory Biosci. 137 (1), 51–65.

Fitch, W.M., Upper, K., 1987. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harbor. Symp. Quant. Biol. 52, 759–767.

Freeland, S.J., Hurst, L.D., 1998. The genetic code is one in a million. J. Mol. Evol. 47 (3), 238–248.

Freeland, S.J., Wu, T., Keulmann, N., 2003. The case for an error minimizing standard genetic code. Origins Life Evol. Biosphere 33 (4–5), 457–477.

Gilis, D., Massar, S., Cerf, N.J., Rooman, M., 2001. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biol. 2 (11), research0049.1–0049.12.

Giulio, M., 2014. On how many fundamental kinds of cells are present on earth: Looking for phylogenetic traits that would allow the identification of the primary lines of descent. J. Mol. Evol. 78, 313–320.

Goldberg, A.L., Wittes, R.E., 1966. Genetic code: aspects of organization. Science 153 (3734), 420–424.

Gomes, A.C., Miranda, I., Silva, R.M., Moura, G.R., Thomas, B., Akoulitchev, A., Santos, M.A., 2007. A genetic code alteration generates a proteome of high diversity in the human pathogen Candida albicans. Genome Biol. 8, R206.

Goodarzi, H., Najafabadi, H.S., Hassani, K., Nejad, H.A., Torabi, N., 2005. On the optimality of the genetic code, with the consideration of coevolution theory by comparison of prominent cost measure matrices. J. Theoret. Biol. 235 (3), 318–325.

Guimaraes, R.C., 2011. Metabolic basis for the self-referential genetic code. Origins Life Evol. Biosph. 41 (4), 357–371.

Gumbel, M., Fimmel, E., Danielli, A., Strungmann, L., 2015. On models of the genetic code generated by binary dichotomic algorithms. BioSystems 128, 9–18.

Haig, D., Hurst, L.D., 1991. A quantitative measure of error minimization in the genetic code. J. Mol. Evol. 33 (5), 412–417.

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.

Khorana, H.G., Buchi, H., Ghosh, H., Gupta, N., Jacob, T.M., Kossel, H., Morgan, R., Narang, S.A., Ohtsuka, E., Wells, R.D., 1966. Polynucleotide synthesis and the genetic code. Cold Spring Harb. Symp. Quant. Biol. 31, 39–49.

Knight, R.D., Landweber, L.F., 1999. Is the genetic code really a frozen accident? New evidence from in vitro selection. Mol. Strateg. Biol. Evol. 870, 408–410.

Koonin, E.V., 2017. Frozen accident pushing 50: stereochemistry, expansion, and chance in the evolution of the genetic code. Life (Basel) 7 (2), 22.

Koonin, E.V., Novozhilov, A.S., 2009. Origin and evolution of the genetic code: the universal enigma. Iubmb Life 61 (2), 99–111.

Koonin, E.V., Novozhilov, A.S., 2017. Origin and evolution of the universal genetic code. Annu. Rev. Genet. 51, 45–62.

Kramer, E.B., Farabaugh, P.J., 2007. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA 13 (1), 87–96.

Kun, A., Pongor, S., Jordan, F., Szathmary, E., 2008. Catalytic Propensity of Amino Acids and the Origins of the Genetic Code and Proteins. vol. 1, pp. 39–58.

Kun, A., Radvanyi, A., 2018. The evolution of the genetic code: Impasses and challenges. BioSystems 164, 217–225.

Lee, H., Popodi, E., Tang, H., Foster, P.L., 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proc. Natl. Acad. Sci. 109 (41), E2774–E2783. http://dx.doi.org/10.1073/pnas.1210309109, arXiv:https://www.pnas.org/content/109/41/E2774.full.pdf, https://www.pnas.org/content/109/41/E2774.

Mackiewicz, P., Biecek, P., Mackiewicz, D., Kiraga, J., Baczkowski, K., Sobczynski, M., Cebrat, S., 2008. Optimisation of asymmetric mutational pressure and selection pressure around the universal genetic code. In: Computational Science - ICCS 2008, Pt 3. In: Lecture Notes in Computer Science, vol. 5103, pp. 100–109.

Massey, S.E., 2006. A sequential ''2-1-3" model of genetic code evolution that explains codon constraints. J. Mol. Evol. 62 (6), 809–810.

Massey, S.E., 2008. A neutral origin for error minimization in the genetic code. J. Mol. Evol. 67 (5), 510–516.

Massey, S.E., 2015. Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. Life (Basel) 5 (2), 1301–1332.

Massey, S.E., 2016. The neutral emergence of error minimized genetic codes superior to the standard genetic code. J. Theoret. Biol. 408, 237–242.

Miranda, I., Silva-Dias, A., Rocha, R., Teixeira-Santos, R., Coelho, C., Gonçalves, T., Santos, M.A.S., Pina-Vaz, C., Solis, N.V., Filler, S.G., Rodrigues, A.G., 2013. Candida albicans CUG mistranslation is a mechanism to create cell surface variation. MBio 4.

Mordret, E., Dahan, O., Asraf, O., Rak, R., Yehonadav, A., Barnabas, G.D., Cox, J., Geiger, T., Lindner, A.B., Pilpel, Y., 2019. Systematic detection of amino acid substitutions in proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity. Mol. Cell 75 (3), 427–441.e5. http://dx.doi.org/10.1016/j.molcel.2019.06.041, https://www.sciencedirect.com/science/article/pii/S1097276519304988.

Murphy, F.V.t., Ramakrishnan, V., 2004. Structure of a purine-purine wobble base pair in the decoding center of the ribosome. Nat. Struct. Mol. Biol. 11 (12), 1251–1252.

Netzer, N., Goodenbour, J.M., David, A., Dittmar, K.A., Jones, R.B., Schneider, J.R., Boone, D., Eves, E.M., Rosner, M.R., Gibbs, J.S., Embry, A., Dolan, B., Das, S., Hickman, H.D., Berglund, P., Bennink, J.R., Yewdell, J.W., Pan, T., 2009. Innate immune and chemically triggered oxidative stress modifies translational fidelity. Nature 462 (7272), 522–526. http://dx.doi.org/10.1038/nature08576.

Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M., Anderson, F., 1966. The RNA code and protein synthesis. Cold Spring Harb. Symp. Quant. Biol. 31, 11–24.

Novozhilov, A.S., Wolf, Y.I., 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.

Ribas de Pouplana, L., Santos, M.A., Zhu, J.H., Farabaugh, P.J., Javid, B., 2014. Protein mistranslation: friend or foe? Trends Biochem. Sci. 39 (8), 355–362.

Santos, M.A.S., Cheesman, C., Costa, V., Moradas-Ferreira, P., Tuite, M.F., 1999. Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp. Mol. Microbiol. 31 (3), 937–947.

Santos, M.A.S., Gomes, A.C., Santos, M.C., Carreto, L.C., Moura, G.R., 2011. The genetic code of the fungal CTG clade. C. R. Biol. 334 (8–9), 607–611.

Santos, J., Monteagudo, A., 2017. Inclusion of the fitness sharing technique in an evolutionary algorithm to analyze the fitness landscape of the genetic code adaptability. BMC Bioinformatics 18 (1), 195.

Schwartz, M.H., Pan, T., 2017. Function and origin of mistranslation in distinct cellular contexts. Critic. Rev. Biochem. Mol. Biol. 52 (2), 205–219.

Sengupta, S., Higgs, P.G., 2015. Pathways of genetic code evolution in ancient and modern organisms. J. Mol. Evol. 80 (5–6), 229–243.

Sonneborn, T.M., 1965. Degeneracy of the genetic code: extent, nature, and genetic implications.. In: Evolving Genes and Proteins. Academic Press, New York, pp. 377–397.

Srinivasan, G., James, C.M., Krzycki, J.A., 2002. Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science 296, 1459–1462.

Weberndorfer, G., Hofacker, I.L., Stadler, P.F., 2003. On the evolution of primitive genetic codes. Origins Life Evol. Biosph. 33 (4–5), 491–514.

Wiltrout, E., Goodenbour, J.M., Fréchin, M., Pan, T., 2012. Misacylation of tRNA with methionine in saccharomyces cerevisiae. Nucleic Acids Res. 40, 10494–10506.

Wnetrzak, M., Błażej, P., Mackiewicz, P., 2019. Optimization of the standard genetic code in terms of two mutation types: Point mutations and frameshifts. Bio Syst. 181, 44–50.

Wnetrzak, M., Błażej, P., Mackiewicz, D., Mackiewicz, P., 2018. The optimality of the standard genetic code assessed by an eight-objective evolutionary algorithm. BMC Evol. Biol. 18, 192.

Woese, C.R., 1965. On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA 54 (6), 1546–1552.

Wong, J.T., 1975. A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA 72 (5), 1909–1912.

Wong, J.T., Ng, S.K., Mat, W.K., Hu, T., Xue, H., 2016. Coevolution theory of the genetic code at age forty: pathway to translation and synthetic life. Life (Basel) 6 (1), 12.

Zhu, Y.O., Siegal, M.L., Hall, D.W., Petrov, D.A., 2014. Precise estimates of mutation rate and spectrum in yeast. Proc. Natl. Acad. Sci. 111 (22), E2310–E2318. http://dx.doi.org/10.1073/pnas.1323011111, arXiv:https://www.pnas.org/content/111/22/E2310.full.pdf, https://www.pnas.org/content/111/22/E2310.