The Importance of Changes Observed in the Alternative Genetic Codes

Paweł Błażej, Małgorzata Wnetrzak and Paweł Mackiewicz
Department of Genomics, Faculty of Biotechnology, University of Wrocław, F. Joliot-Curie 14a, 50-383 Wrocław, Poland

Keywords: Alternative Genetic Code, Error Minimization, Genetic Code, Mutation.

Abstract: The standard genetic code is a way of transmitting genetic information from DNA into protein world. The code is universal for almost all living organisms on Earth but small deviations have been observed for many cellular organelles and some specific groups of microorganisms with highly reduced genomes. Such modifications are called alternative genetic codes. There is no consensus about the factors that caused or allowed these changes. A popular concept assumes that the codes evolved under neutral evolution without adaptive constraints. In this paper we present findings that argue with such view. We examined the level of error minimization in amino acid replacements generated by the standard genetic code and its alternatives. We found that only 3 out of 23 tested alternative codes have worse quality than the standard genetic code. In agreement with that, many single codon reassignments observed in the variants of the standard genetic code are generally responsible for improving the quality of the codes under the studied criteria. These results indicate that the codon reassignments observed in the existing alternative genetic codes could play an adaptive role in their evolution to minimize translational and mutational errors. The study can help in designing alternative genetic codes for artificially modified organisms in the framework of synthetic biology.

1 INTRODUCTION

The assumption about the universality of the standard genetic code (SGC) was challenged by the discoveries of the genetic code variants (Osawa et al., 1992; Jukes, 1996) especially because the SGC was initially considered a 'frozen accident' (Crick, 1968). These deviations are mainly observed in the codes operating in organelles, especially in mitochondria (Osawa et al., 1989; Crozier and Crozier, 1993; Boore and Brown, 1994). The alternative codes are also involved in translation of proteins coded in nuclear genomes of various unicellular eukryotes (Schneider et al., 1989; Sanchez-Silva et al., 2003) and some bacteria, especially in parasites and symbionts with highly reduced genomes (Lim and Sears, 1992; McCutcheon et al., 2009; Campbell et al., 2013). In recent years, the number of newly discovered alternative genetic codes has significantly increased (Heaphy et al., 2016; Zahonova et al., 2016). This quite large set of alternative codes is a good starting point to analyze the properties and the potential evolutionary tendencies of these codes and the SGC.

There are three main types of deviations from the standard genetic code found in its alternatives: (i) reassignments of codons coding for typical 20 amino acids and stop translation signal, (ii) loss of codon assignments resulting in the occurrence of unused codons, (iii) incorporation of new amino acids, e.g. selenocysteine and pyrrolysine. These three types of changes were discussed and modeled by (Sengupta et al., 2007). They result usually from mutations and the editing of tRNA genes or the posttranscriptional modifications of bases in tRNA molecules (Santos et al., 2004). However, it is not clear how and why these changes occurred and were fixed. It is generally believed that they evolved neutrally without any adaptive pressure through genetic drift and mutational pressure that drove small populations and their tiny genomes toward the high AT-content (Freeland et al., 2000; Koonin, 2017). Such changes can influence codon usage and, in extreme cases, can lead to disappearance of GC-containing codons (Santos et al., 2004). Alternatively, the variants of the SGC could have evolved to reduce protein synthesis costs (Swire et al., 2005) or to minimize effects of point mutations; such properties were observed in some of these codes (Kurnaz et al., 2010; Morgens and Cavalcanti, 2013). Here we focused on the latter aspects of the genetic code evolution and analyzed the optimality of many genetic code variants to assess their robustness in terms of amino acid replacements.

Błażej, P., Wnetrzak, M. and Mackiewicz, P.
The Importance of Changes Observed in the Alternative Genetic Codes.
DOI: 10.5220/0006642001540159
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 3: BIOINFORMATICS*, pages 154-159
ISBN: 978-989-758-280-6

2 METHODS

The examined alternative genetic codes were downloaded from the NCBI taxonomy web page: www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi. From the whole set of 30 available genetic code variants we chose those which differed in codon assignments from the standard genetic code, including also the codes with ambiguous codon assignments. In total, we studied 23 alternative codes (Table 1).

Table 1: The genetic code variants studied in this work with the number of codon reassignments. In the case of the genetic codes with ambiguous reassignments, we included two different versions of their structures: with all possible codon assignments (all) and with only unambiguous assignments (unamb).

Genetic code name	Num. of reassig.
Alternative Flatworm Mitochondrial Code	5
Trematode Mitochondrial Code	5
Invertebrate Mitochondrial Code	4
Flatworm Mitochondrial Code	4
Ascidian Mitochondrial Code	4
Vetebrate Mitochondrial Code	4
Candylostoma Nuclear Code (all)	3
Karyorelict Nuclear Code (all)	3
Blastocrithidia Nuclear Code (all)	3
Pterobranchia Mitochondrial Code	3
Hexamita Nuclear Code	2
Karyorelict Nuclear Code (unamb)	2
Mesodium Nuclear Code	2
Peritrich Nuclear Code	2
Scenedesmus Mitochondrial Code	2
Gracilibacteria Code	1
Euploid Nuclear Code	1
Blastocrithidia Nuclear Code (unamb)	1
Protozoan Mitochondrial Code	1
Chlorophycean Mitochondrial Code	1
Tannophilus Nuclear Code	1
Alternative Yeast Nuclear Code	1
Thraustochytrium Mitochondrial Code	1

To test the optimality of a given code, we had to use a specific measure describing costs of amino acid replacements. In this work, for a given genetic code $(code)$ we used the following cost measure:

$$F(code) = \sum_{<i,j>\in D} [f(i) - f(j)]^2, \qquad (1)$$

where D is the set of pairs of codons that differ in one nucleotide substitution, whereas $f(i)$ and $f(j)$ are the polarity values of the amino acids (Woese, 1973) coded by the codons i and j, respectively. Therefore, the measure $F(code)$ represents the total sum of the squared differences between polarity properties of amino acids for the codon pairs differing in one substitution. The main reason for using the polarity property to evaluate the cost of a genetic code follows from the fact that this characteristic is independent of the specificity of the SGC structure and was commonly applied in testing the optimality of the standard genetic code (Di Giulio, 1989; Haig and Hurst, 1991; Freeland and Hurst, 1998; Santos and Monteagudo, 2010; Błażej et al., 2016).

All the single nucleotide substitutions that lead to nonsense mutations, i.e. to the replacement of an amino acid by a stop translation codon, were included in the calculation as the maximum of squared differences computed for any possible pair of amino acids. One could argue with this assumption. However, it is known that the nonsense mutations are very deleterious because they result in incomplete and usually nonfunctional proteins. Therefore, it seems reasonable to assume such large costs for this type of substitution.

Furthermore, we calculated three characteristics in the case of the genetic codes with ambiguous codon assignments. $F(code\_all)$ included all possible codon assignments and $F(code\_unamb)$ included only unambiguous codon assignments. Additionally, we calculated the arithmetic mean $F(code\_mean)$ from $F(code\_all)$ and $F(code\_unamb)$. As a result, $F(code\_mean)$ assumes that ambiguous codon assignments occur with equal probability.

To validate the properties of the genetic codes we compare them with all possible 1240 theoretical codes that differed from the SGC by one codon assignment.

3 RESULTS

3.1 Optimality of the Standard and Alternative Genetic Codes

The comparison of the cost function F calculated for the standard genetic code and its alternatives showed that the SGC is not the best optimized code, in terms of the polarity property, similarly to the results obtained by (Morgens and Cavalcanti, 2013). Only 3 out of 23 considered alternative genetic codes have greater (i.e. worse) F value than the SGC, i.e. $F(SGC) = 5641.46$ (Figure 1):

1. Vetebrate Mitochondrial Code: $F(code) = 6716.48$;
2. Alternative Yeast Nuclear Code: $F(code) = 5651.86$;
3. Thraustochytrium Mitochondrial Code: $F(code) = 6283.02$.

The best found code according to the polarity costs is Karyorelict Nuclear Code including all am-

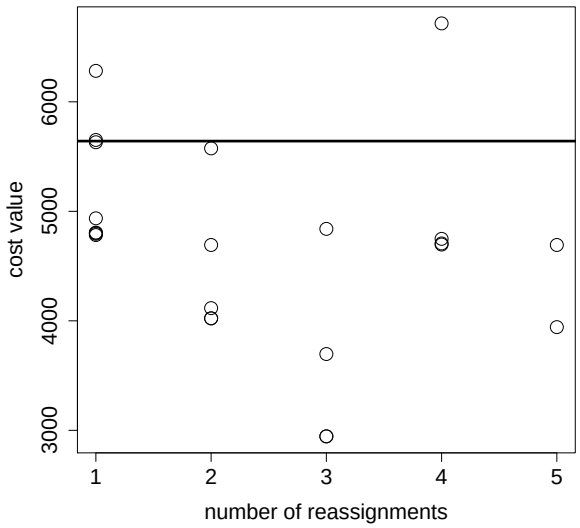

Figure 1: The cost values (black circles) calculated for alternative genetic codes. They are compared with the number of codon reassignments (x-axis) and also with the cost of the standard genetic code (the bold horizontal line). It is visible that many alternative codes lie below the horizontal line and are better optimized to minimize the costs of amino acid replacements than the SGC.

biguous reassignments of codons. It reaches the minimum cost value $F(code) = 2945.12$ which is nearly two times lower than $F(SGC)$. Similarly, there are theoretical genetic codes with one codon reassignment that have smaller cost values than the SGC. The best one of them has the cost value $F(code) = 4766.28$. These results suggest that the standard genetic code can be significantly improved even by a small number of codon reassignments.

On the other hand, there is a quite large probability to deteriorate the SGC structure in terms of the F value just by one random reassignment because more than 330 theoretical codes out of all 1240 considered, i.e. 27% have lower cost values than the SGC (Figure 2).

3.2 The Properties of Codon Assignments in Alternative Genetic Codes

It is interesting to examine the features of the alternative genetic codes which are better optimized than the SGC. First we calculated the number of occurrences of all individual codon reassignments observed in the alternative genetic codes under study (Table 3). It is evident that these changes can be classified into three groups (Figure 3). The first one (A) contains all codon reassignments from stop translation signal

Table 2: The cost values F calculated for all genetic code variants studied in this work in comparison to the the standard genetic code (SGC) and the best theoretical genetic code with single codon reassignment. We included three different characteristics of the genetic code variants: with all possible codon assignments (all) and with only unambiguous assignment (unamb) and the average value (mean) of the cost value.

genetic code name	cost
Candylostoma Nuclear Code (all)	2945.12
Karyorelict Nuclear Code (all)	2945.12
Karyorelict Nuclear Code (mean)	3484
Blastocrithidia Nuclear Code (all)	3697.04
Alternative Flatworm Mitochondrial Code	3942.66
Hexamita Nuclear Code	4022.88
Karyorelict Nuclear Code (unamb)	4022.88
Mesodium Nuclear Code	4116.26
Blastocrithidia Nuclear Code (mean)	4250.7
Candylostoma Nuclear Code (mean)	4293.29
Peritrich Nuclear Code	4692.9
Trematode Mitochondrial Code	4692.92
Invertebrate Mitochondrial Code	4696.84
Flatworm Mitochondrial Code	4706.84
Ascidian Mitochondrial Code	4748.84
the best theoret. code with one reassign.	**4766.28**
Gracilibacteria Code	4783.56
Euploid Nuclear Code	4795.08
Blastocrithidia Nuclear Code (unamb)	4804.36
Protozoan Mitochondrial Code	4804.36
Pterobranchia Mitochondrial Code	4839.88
Chlorophycean Mitochondrial Code	4936.3
Scenedesmus Mitochondrial Code	5575.14
Tannophilus Nuclear Code	5630.96
the standard genetic code	**5641.46**
Alternative Yeast Nuclear Code	5651.86
Thraustochytrium Mitochondrial Code	6283.02
Vetebrate Mitochondrial Code	6716.48

to one of the 20 standard amino acids. There are 31 such changes which make over 55% of all possible 56 codon reassignments found in the studied alternative genetic codes. The second group (B) includes in total 21 codon reassignments which are involved in changing encoded amino acids. Finally, we have only 4 cases in which codons originally encoding amino acids change their meaning to the stop translation signal (the group C).

It should be noted that the codon reassignments belonging to the first and, at the same time, the largest group (Table 3) are the most desired in terms of minimizing the F value because each of these changes decreases the cost value in comparison with the $F(SGC)$ (Figure 3). For example, the single assignment of stop codon TGA to tryptophane, which is the most frequently observed change, i.e. 12 times, diminishes the cost of the $F(SGC) = 5641.46$ to $F(TGA \rightarrow Trp) = 4804.36$. This reassignment is one of the best ones because the resulting cost for the code differs by less

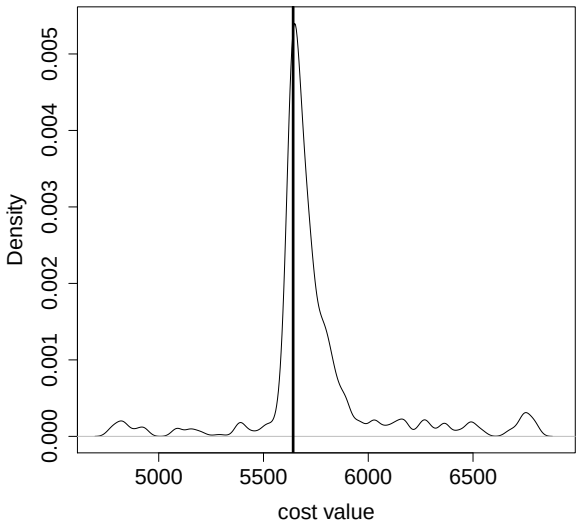

Figure 2: The density plot of the frequency of cost values F calculated for the theoretical genetic codes that differ from the standard genetic code in one codon assignment. The cost value calculated for the standard genetic code $F(SGC)$ is marked by the vertical bold line. It is evident that $F(SGC)$ value is situated closer to smaller values. Thereby, it is less probable to generate at random a code better than the SGC just by one codon reassignment.

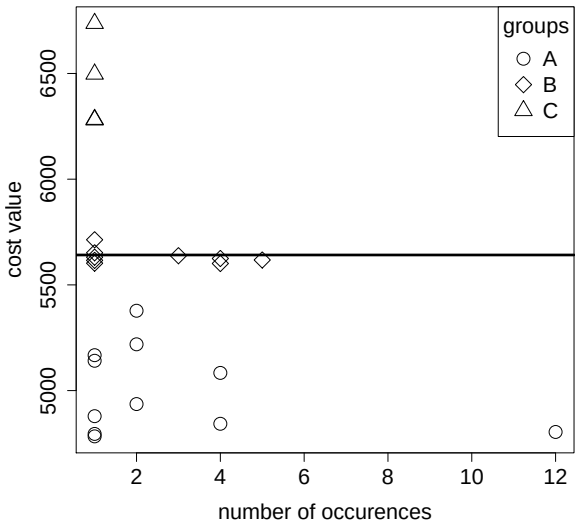

Figure 3: The relationship between the cost values and the number of occurrences of individual codon reassignments in the studied alternative genetic codes. The reassignments were classified into three groups where the codon's meaning is changed: from a stop translation signal to an amino acid (A), from one amino acid to another (B) and from an amino acid to a stop translation signal (C). The results are compared with the cost value calculated for the standard genetic code (the horizontal bold line).

than 1% from the best possible cost obtained by the single change $F(TGA \rightarrow Ala) = 4766.28$.

The reassignments in the second group have a rather small influence on the cost values in comparison with the change of codon's meaning from a stop translation signal to an amino acid (Figure 3). However, many of these missense reassignments can also improve, although slightly, the F value in comparison with $F(SGC)$, i.e. $F(AGG \rightarrow Ser) = 5601.14$. The reassignments that decrease the F value are more frequent in the studied alternative genetic codes than the reassignments increasing the costs of amino acid replacement, e.g. $F(AGG \rightarrow Lys) = 5713.46$. The former were found in 21 cases, whereas the latter in only four cases.

The third group of reassignments contain codons that formerly coded for an amino acid but then changed their meaning into the stop translation signal. Such changes have the most dramatic impact on the cost value and encoded proteins. Generally, they are responsible for the significant increase (over 10%) of the F value in comparison with $F(SGC)$ (Figure 3). However, they were observed only in four alternative genetic codes (Table 3).

4 DISCUSSION

Our study on the optimality of the alternative genetic codes in comparison to the standard genetic code showed that many of these variants contain codon reassignments that decrease the costs of amino acid replacements described by the polarity values. It implies that the alternatives did not necessarily originate as a result of the neutral evolution but they could have evolved under adaptational factors and at least some of their codon reassignments were favored by the selection (Kurnaz et al., 2010). Such reorganizations of the code could have occured in small populations with tiny genomes, in which the changes did not influence the large number of encoded proteins.

The SGC is, however, less optimal in comparison to most of its alternatives. This finding does not fully support the adaptive hypothesis postulating that the code structure evolved to minimize the effects of amino acid replacements and errors during translation of proteins (Epstein, 1966; Haig and Hurst, 1991; Freeland et al., 2003; Goodarzi et al., 2005). This concept is attractive but the deleterious effects of mutations on protein properties can be minimized by other mechanisms, i.e. the direct optimization of the

Table 3: The number of occurrences of single codon reassignments observed in the studied alternative genetic codes. The table includes also the costs calculated for genetic codes with exactly one of such substitutions and the type of reassignment used in Figure 3.

Codon	In SGC	In altern.	Cost	Occur.	Type
TGA	Stp	Trp	4804.36	12	A
AGA	Arg	Ser	5617.78	5	B
TAA	Stp	Gln	5083.72	4	A
TAG	Stp	Gln	4843.06	4	A
ATA	Ile	Met	5624.82	4	B
AGG	Arg	Ser	5601.14	4	B
AAA	Lys	Asn	5638.02	3	B
TAA	Stp	Glu	5377.78	2	A
TAG	Stp	Glu	5219.02	2	A
TAG	Stp	Leu	4936.3	2	A
CTG	Leu	Ala	5630.96	1	B
TGA	Stp	Cys	4795.08	1	A
AGA	Arg	Gly	5616.02	1	B
AGG	Arg	Gly	5602.58	1	B
TGA	Stp	Gly	4783.56	1	A
AGG	Arg	Lys	5713.46	1	B
CTG	Leu	Ser	5651.86	1	B
AGA	Arg	Stp	6737.06	1	C
AGG	Arg	Stp	6497.1	1	C
TTA	Leu	Stp	6283.02	1	C
TCA	Ser	Stp	6280.3	1	C
TAA	Stp	Trp	5167.54	1	A
TAA	Stp	Tyr	5141.14	1	A
TAG	Stp	Tyr	4879.02	1	A

mutational rate and pattern on the fixed genetic code (Błażej et al., 2015; Błażej et al., 2017). Therefore, the main role of the assignments of amino acids to codons in the SGC could have been played by the expansion of biosynthetic pathways and the stepwise addition of newly synthesized amino acids into the code, according to the co-evolution hypothesis (Wong, 1975; Di Giulio, 1999; Wong et al., 2016; Di Giulio, 2017). Under this scenario, the present structure of SGC evolved from an ancestral version including a smaller number of simple amino acids that were at the beginning of the biosynthetic pathways. Next, other amino acids were incorporated into the code when more complex metabolic networks evolved. The newly synthesized amino acids took over the codons of their precursors.

The idea of studying the properties of the genetic code variants seems very promising in the light of designing alternative versions of the code for artificially modified organisms (Xie and Schultz, 2006; Chin, 2014). Such modifications can lead to production of peptides or proteins including unnatural amino acids and showing enhanced or novel properties. The introduced codon reassignments can also help to test the protein structure and function in global scale. Moreover, the knowledge about the optimality of the genetic codes may enable us to construct new artificial organisms in the framework of synthetic biology. Such organisms could be characterized by for example a higher fidelity of the protein synthesis and a higher resistance to the mutations causing amino acid replacements.

ACKNOWLEDGMENTS

This work was supported by the National Science Centre Poland (Narodowe Centrum Nauki, Polska) under Grant Miniatura no. 2017/01/X/NZ2/00608.

REFERENCES

Błażej, P., Mackiewicz, D., Grabinska, M., Wnetrzak, M., and Mackiewicz, P. (2017). Optimization of amino acid replacement costs by mutational pressure in bacterial genomes. *Scientific Reports*, 7:1061.

Błażej, P., Miasojedow, B., Grabinska, M., and Mackiewicz, P. (2015). Optimization of mutation pressure in relation to properties of protein-coding sequences in bacterial genomes. *PloS One*, 10:e0130411.

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

Boore, J. L. and Brown, W. M. (1994). Complete dna sequence of the mitochondrial genome of the black chiton, katharina tunicata. *Genetics*, 138(2):423–43.

Campbell, J. H., O'Donoghue, P., Campbell, A. G., Schwientek, P., Sczyrba, A., Woyke, T., Sll, D., and Podar, M. (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci U S A*, 110:5540–5545.

Chin, J. W. (2014). Expanding and reprogramming the genetic code of cells and animals. *Annu Rev Biochem*, 83:379–408.

Crick, F. H. (1968). The origin of the genetic code. *J Mol Biol*, 38(3):367–79.

Crozier, R. H. and Crozier, Y. C. (1993). The mitochondrial genome of the honeybee Apis mellifera: complete sequence and genome organization. *Genetics*, 133(1):97–117.

Di Giulio, M. (1989). The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J Mol Evol*, 29(4):288–93.

Di Giulio, M. (1999). The coevolution theory of the origin of the genetic code. *J Mol Evol*, 48(3):253–5.

Di Giulio, M. (2017). Some pungent arguments against the physico-chemical theories of the origin of the genetic code and corroborating the coevolution theory. *J Theor Biol*, 414:1–4.

Epstein, C. J. (1966). Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. *Nature*, 210(5031):25–8.

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

Freeland, S. J., Knight, R. D., Landweber, L. F., and Hurst, L. D. (2000). Early fixation of an optimal genetic code. *Mol Biol Evol*, 17(4):511–8.

Freeland, S. J., Wu, T., and Keulmann, N. (2003). The case for an error minimizing standard genetic code. *Origins of Life and Evolution of the Biosphere*, 33(4-5):457–477.

Goodarzi, H., Najafabadi, H. S., Hassani, K., Nejad, H. A., and Torabi, N. (2005). On the optimality of the genetic code, with the consideration of coevolution theory by comparison of prominent cost measure matrices. *J Theor Biol*, 235(3):318–25.

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

Heaphy, S. M., Mariotti, M., Gladyshev, V. N., Atkins, J. F., and Baranov, P. V. (2016). Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in Condylostoma magnum. *Mol Biol Evol*, 33:2885–2889.

Jukes, T. H. (1996). Neutral changes and modifications of the genetic code. *Theoretical Population Biology*, 49(2):143–145.

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

Kurnaz, M. L., Bilgin, T., and Kurnaz, I. A. (2010). Certain non-standard coding tables appear to be more robust to error than the standard genetic code. *J Mol Evol*, 70(1):13–28.

Lim, P. O. and Sears, B. B. (1992). Evolutionary relationships of a plant-pathogenic mycoplasmalike organism and Acholeplasma laidlawii deduced from two ribosomal protein gene sequences. *J. Bacteriol*, 174:2606–2611.

McCutcheon, J. P., McDonald, B. R., and Moran, N. A. (2009). Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *Plos Genetics*, 5(7).

Morgens, D. W. and Cavalcanti, A. R. (2013). An alternative look at code evolution: using non-canonical codes to evaluate adaptive and historic models for the origin of the genetic code. *J Mol Evol*, 76(1-2):71–80.

Osawa, S., Jukes, T. H., Watanabe, K., and Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiol Rev*, 56(1):229–64.

Osawa, S., Ohama, T., Jukes, T. H., and Watanabe, K. (1989). Evolution of the mitochondrial genetic code. I. origin of AGR serine and stop codons in metazoan mitochondria. *J Mol Evol*, 29(3):202–7.

Sanchez-Silva, R., Villalobo, E., Morin, L., and Torres, A. (2003). A new noncanonical nuclear genetic code: translation of UAA into glutamate. *Curr Biol*, 13:442–447.

Santos, J. and Monteagudo, A. (2010). Study of the genetic code adaptability by means of a genetic algorithm. *J Theor Biol*, 264(3):854–865.

Santos, M. A., Moura, G., Massey, S. E., and Tuite, M. F. (2004). Driving change: the evolution of alternative genetic codes. *Trends Genet*, 20(2):95–102.

Schneider, S. U., Leible, M. B., and Yang, X. P. (1989). Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of Acetabularia and the occurrence of unusual codon usage. *Mol Gen Genet*, 218:445–452.

Sengupta, S., Yang, X., and Higgs, P. G. (2007). The mechanisms of codon reassignments in mitochondrial genetic codes. *J Mol Evol*, 64(6):662–88.

Swire, J., Judson, O. P., and Burt, A. (2005). Mitochondrial genetic codes evolve to match amino acid requirements of proteins. *J Mol Evol*, 60(1):128–39.

Woese, C. R. (1973). Evolution of the genetic code. *Naturwissenschaften*, 60(10):447–59.

Wong, J. T. (1975). A co-evolution theory of the genetic code. *Proc Natl Acad Sci U S A*, 72(5):1909–12.

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

Xie, J. M. and Schultz, P. G. (2006). Innovation: A chemical toolkit for proteins - an expanded genetic code. *Nat Rev Mol Cell Biol*, 7(10):775–782.

Zahonova, K., Kostygov, A. Y., Sevcikova, T., Yurchenko, V., and Elias, M. (2016). An unprecedented noncanonical nuclear genetic code with all three termination codons reassigned as sense codons. *Curr Biol*, 26:2364–2369.