Article

Model of Genetic Code Structure Evolution under Various Types of Codon Reading

Paweł Błażej *, Konrad Pawlak, Dorota Mackiewicz and Paweł Mackiewicz

Department of Bioinformatics and Genomics, Faculty of Biotechnology, University of Wrocław,
ul. Joliot-Curie 14a, 50-383 Wrocław, Poland; konrad.pawlak@uwr.edu.pl (K.P.);
dorota.mackiewicz@uwr.edu.pl (D.M.); pamac@smorfland.uni.wroc.pl (P.M.)
* Correspondence: pawel.blazej@uwr.edu.pl

Abstract: The standard genetic code (SGC) is a set of rules according to which 64 codons are assigned to 20 canonical amino acids and stop coding signal. As a consequence, the SGC is redundant because there is a greater number of codons than the number of encoded labels. This redundancy implies the existence of codons that encode the same genetic information. The size and organization of such synonymous codon blocks are important characteristics of the SGC structure whose evolution is still unclear. Therefore, we studied possible evolutionary mechanisms of the codon block structure. We conducted computer simulations assuming that coding systems at early stages of the SGC evolution were sets of ambiguous codon assignments with high entropy. We included three types of reading systems characterized by different inaccuracy and pattern of codon recognition. In contrast to the previous study, we allowed for evolution of the reading systems and their competition. The simulations performed under minimization of translational errors and reduction of coding ambiguity produced the coding system resistant to these errors. The reading system similar to that present in the SGC dominated the others very quickly. The survived system was also characterized by low entropy and possessed properties similar to that in the SGC. Our simulation show that the unambiguous SGC could emerged from a code with a lower level of ambiguity and the number of tRNAs increased during the evolution.

Keywords: amino acid; codon; evolution; genetic code

1. Introduction

The structure and properties of the standard genetic (SGC) code have intrigued scientists since the first ascriptions of codons to amino acids were discovered in the sixties of the last century [1,2]. From that time, many hypotheses concerning the origin and evolution of SGC have been proposed (see for review: [3–8]). However, it is still unclear which factors played a decisive role in the process of the genetic code emergence and evolution.

These investigations been started with two of Crick's seminal papers, i.e., [9,10]. In the first work, the author discussed the general codon block structure of the SGC. Briefly, the codons in the SGC are arranged in groups, which is an immediate consequence of the fact that the number of possible codons, i.e., 64, is greater than the number of encoded items, i.e., 20 amino acids and a stop translation signal. The codon groups encoding the same item are called synonymous and are composed of two, three, four, or six codons. Codons in a given group differ mostly in the third codon position. In order to explain this fact, Crick proposed the wobble rule, which refers to interactions between the first base in a tRNA anticodon and the third base of translated codon in a transcript (mRNA). This rule assumes that the base pairing between these two RNAs does not have to follow Watson–Crick base pair rules, i.e., cytosine–guanine and adenine–uracil, but that other interactions are also possible, i.e., guanine–uracil, hypoxanthine–uracil, hypoxanthine–adenine and hypoxanthine–cytosine. Additional experiments revealed that other modified bases can also pair with the typical ones [11].

This phenomenon has fundamental consequences. It reduces the number of necessary tRNA molecules for protein synthesis. Moreover, it determines the structure of codon blocks and introduces specific robustness of the SGC against single nucleotide substitutions occurred in the third position in codons belonging to a group encoding the same genetic information. Therefore, such mutations do not change coded amino acids.

In the second paper [10], Crick considered the evolution of genetic code on a general level. He put forward "the frozen accident" hypothesis. According to this scenario, it is not inconceivable that, in the past, there were various genetic codes that were used by different organisms. They coexisted and evolved simultaneously. Finally, the present structure of the SGC won this competition by accident and stayed universal among all domains of life, because any change in the codon assignment would be highly deleterious at later stages of SGC evolution.

The existence of many different coding systems at the early evolution of genetic code seems very probable. This assumption appears to be a good starting point for further investigations of the SGC evolution. It was proposed that these codes evolved together from a set of ambiguous codon assignments towards a coding system with a low level of ambiguity to reduce an initial high translational noise [5,12–14]. During the genetic code evolution, amino acids were gradually added to the code, which was profitable because it increased the diversity of synthesized proteins [6,15–17].

These assumptions were included in a simulation model to show the origin of coding system structure [13,14]. This model assumed that every code was a set of rules, described by selected random variables, according to which 64 codons were assigned to specific genetic information, i.e., labels. What is more, this model did not take into account any properties of amino acids in order to avoid additional assumptions, which made the model general. The model considered three types of reading mechanisms, called M_1, M_2, and M_3, which induced different codon groups assigned to respective encoded labels. In M_1, a given amino acid was coded by codons that had two fixed identical positions and differed in one position from the reference codon. M_2 assumed that a given amino acid was coded by codons with one fixed identical position and differed in exactly one of the other two codon positions from the reference codon. Codons of M_3 for a given amino acid differed in exactly one of any codon positions from the reference codon. Thereby, the codon groups defined by the M_1 rule had a structure similar to that in the SGC, whereas M_2 and M_3 were potential generalizations of M_1. The codes were evolved to a state characterized by a low level of uncertainty.

The results based on these simulations showed that the structure of the SGC would emerge from highly ambiguous coding systems under relatively simple criteria, i.e., the reduction of translational noise and a stepwise addition of amino acids into the code [13,14]. The codon block structure observed in the best surviving code was characterized by low translational noise as well as a high robustness against point mutations. However, the authors assumed that the type of reading, i.e., M_1, M_2, and M_3, stayed fixed and constant during simulations. Therefore, to make the simulation model more realistic, we assumed here that these three mechanisms of reading genetic information could coexisted at the same time and evolved simultaneously according to Crick's scenario. The computer simulations performed including this possibility showed that the initial coding systems evolved very quickly to one characterized by unambiguous reading of genetic information.

2. Results

2.1. Changes of Genetic Codes during Simulations

We performed computer simulations to study the process of coding system emergence. Each simulation was run up to 100,000 steps. All simulations started with a population of 1000 genetic codes. They were described by initially randomly generated probability distribution functions for codon assignments and reading system types. During the simulations, we collected several descriptive characteristics for a given coding system, i.e., values of the fitness function, the expected number of genetic labels encoded by a given reading

mechanism, and the structure of codon blocks. They together allowed us to describe tendencies in the evolution of coding systems. All simulations were run under different seeds because we wanted to find out if the observed tendencies in the genetic code structure have a general character, i.e., is they are independent of starting constraints.

The simulations started from non-optimized genetic systems. However, the fitness values showed a tendency to increase substantially. During the code evolution, the fitness function increased from a low value to higher one after 15,000 steps. After some fluctuations it stabilized after 50,000 steps above the value of -20 in the logarithmic scale (Figure 1).
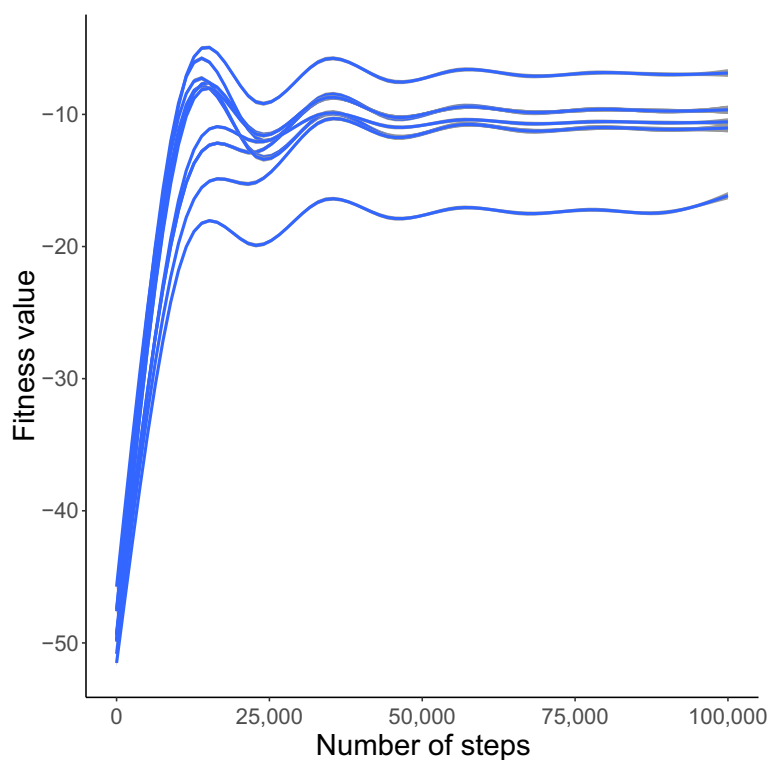
Figure 1. Changes in the best approximation of the fitness function *F* (y-axis) with the number of generations (x-axis) based on the GAM model and 10 simulation runs with different initial seeds. The y-axis is shown in a logarithmic scale.

The increase in the fitness function corresponds to the decrease in entropy values calculated for probability distributions of codon assignments and reading types. The average genetic code entropy decreased with the simulation time and remained rather constant at a low level after 30,000 steps (Figure 2A). However, the drop and stabilization of the average entropy for reading systems occurred much earlier, i.e., after 5000 steps (Figure 2B). The results show that the studied coding systems changed their properties during simulation runs. However, the reading system was selected by the genetic codes much faster than the translational noise was reduced.
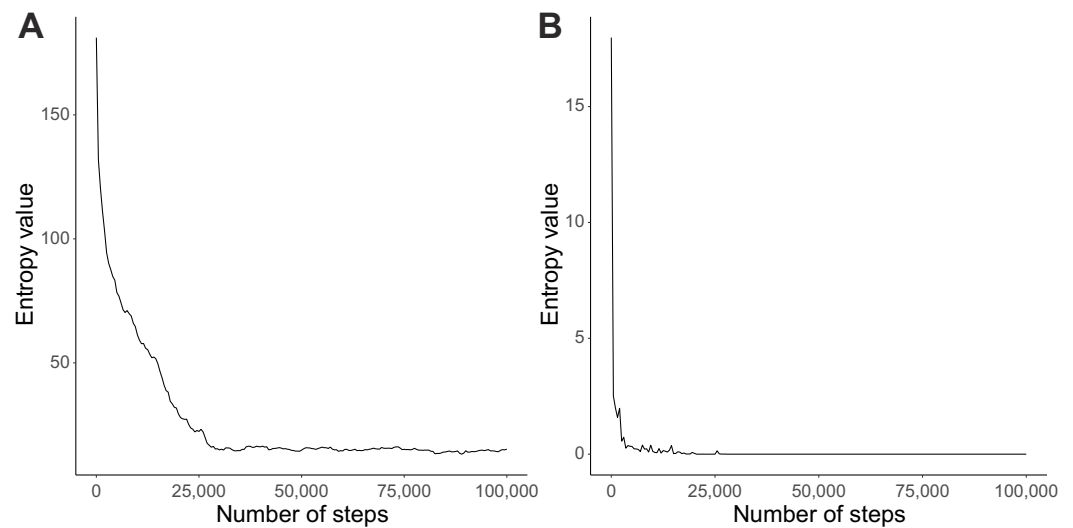
Figure 2. Changes in the average entropy values (y-axis) calculated from distributions of codon assignments $H_c(\mathcal{P})$ (**A**), and reading system types $H_r(\mathcal{P})$ (**B**), with the number of generations (x-axis).

2.2. Structure of Genetic Codes

Interestingly, we observed a relatively fast emergence of nearly homogeneous reading systems, i.e., one type out of M_1, M_2, and M_3 was preferred for all possible labels $l = 1, 2, 3, \ldots 21$. The model M_1 started to dominate very quickly over the two others (Figure 3).
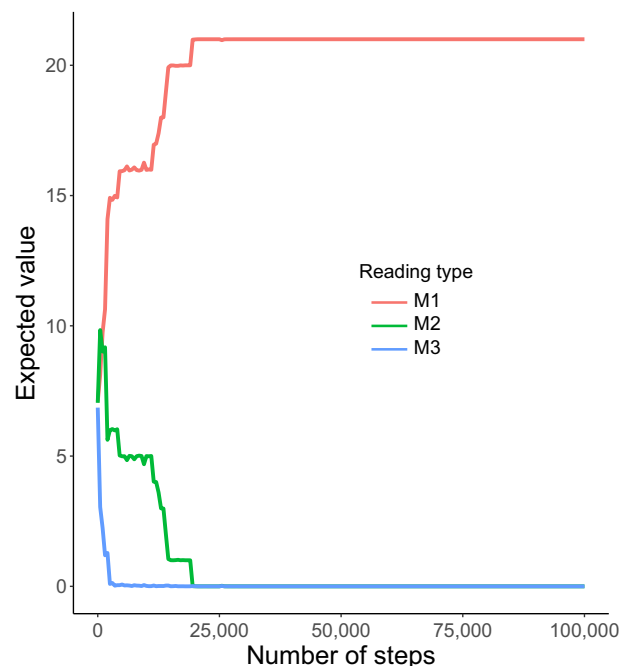
**Figure 3.** The expected value of the total number of encoded labels using various reading types M_1, M_2, and M_3 (y-axis) with the number of generations (x-axis). Notice that the expected number of labels read by the M_1 system started dominated among all considered types of reading very quickly.

This tendency was detected from the beginning of simulations. The M_3 system was the first one eliminated after about 2500 steps and M_2 coexisted with M_1 to about 20,000 steps. After that, only M_1 was used by the genetic codes to read encoded information, whereas the other systems, i.e., M_2 and M_3, disappeared and the coding system stayed homogeneous. The homogeneity is visualized by the heatmap, which shows that M_1 was selected as the only reading type receiving the probability one for all encoded labels (Figure 4).
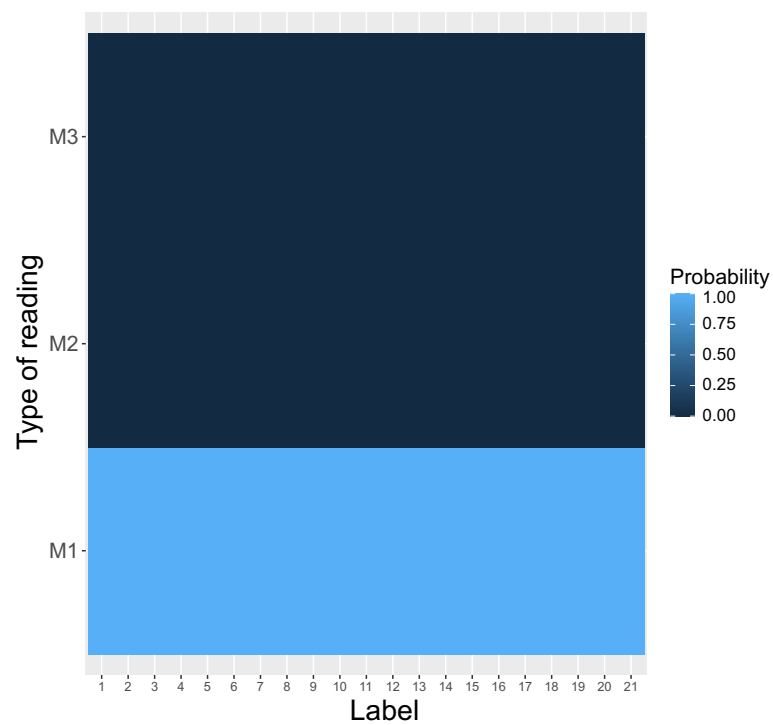
Figure 4. The heatmap of different types of reading systems (y-axis) at the end of the simulations. This is a graphical representation of a matrix in which each genetic label (x-axis) ascribes a probability that is read by a given reading type. Please compare with Figure 5 at the beginning of the simulations.

In contrast to that, no reading system was preferred at the beginning of simulations (Figure 5).
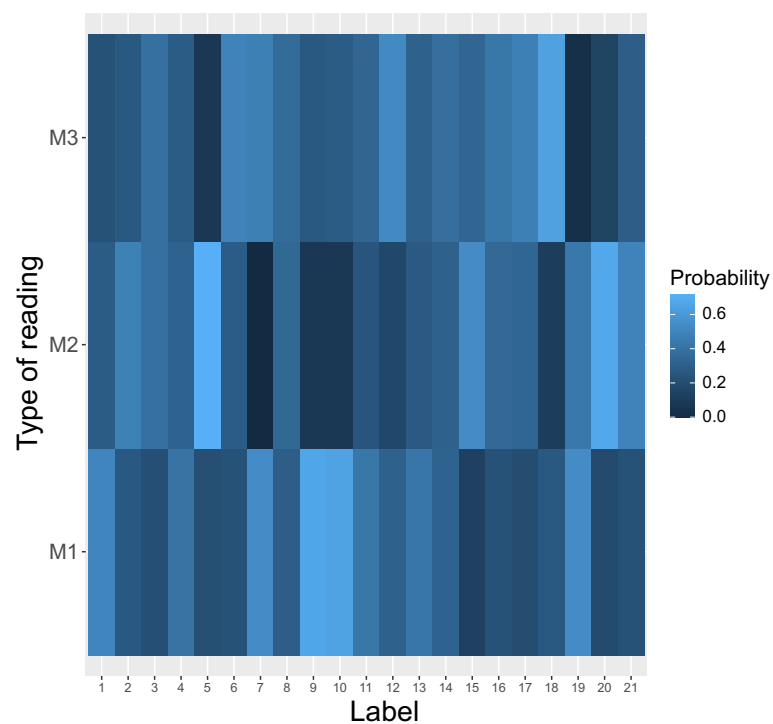
Figure 5. The heatmap of different types of reading systems (y-axis) at the beginning of simulation. This is a graphical representation of a matrix in which each genetic label ascribes a probability that is read by a given reading type.

It is also interesting to investigate the codon block structure of the computed coding system at the end of simulations. The results presented in in the previous paragraph indicate that the M_1 reading system was the most preferred, which suggests that codon blocks induced by this rule should be present in the structure of the genetic codes selected after simulations. In Figure 6, we present the coding system that was characterized by the highest fitness value among all simulation runs. In contrast to the beginning of simulations (Figure 7), we can distinguish for each label a group of codons with a high coding probability, i.e., over 0.8.
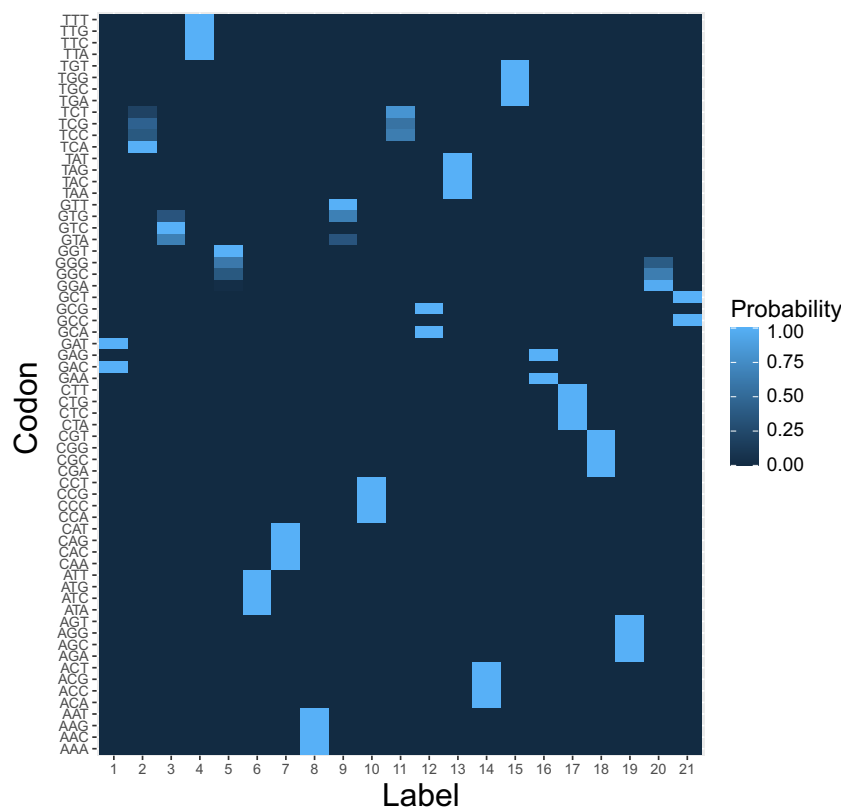

Figure 6. The heatmap of genetic code encoding 21 labels by 64 codons at the end of simulation. This is a graphical representation of the matrix $\mathcal{P} = (p_{cl})$, in which each element p_{cl} ascribes a probability that a codon c in a row encodes a label l in a column. Please compare with Figure 7 at the beginning of the simulations.

Many of these groups were composed of codons that differed in one codon position as assumed by the M_1 rule and resemble the structure of the standard genetic code.

The distribution of size of codon groups encoding a given label is also very interesting (Figure 8). Eleven labels were encoded by blocks consisting of four codons, eight labels by two-codon blocks, one label by three-codon blocks, and one label by one codon. This is very similar to the number of codon blocks in the SGC, in which five amino acids are encoded by four-codon blocks, nine amino acids by two-codon blocks, one amino acid and the stop signal by three-codon blocks, and two amino acids by one codon each. Moreover, there are three amino acids, each of which is ascribed to six codons. These codons are, however, organized into the subgroup of four codons and the subgroup of two codons. These tendencies were observed in all genetic codes evolved at the end of all simulation runs.
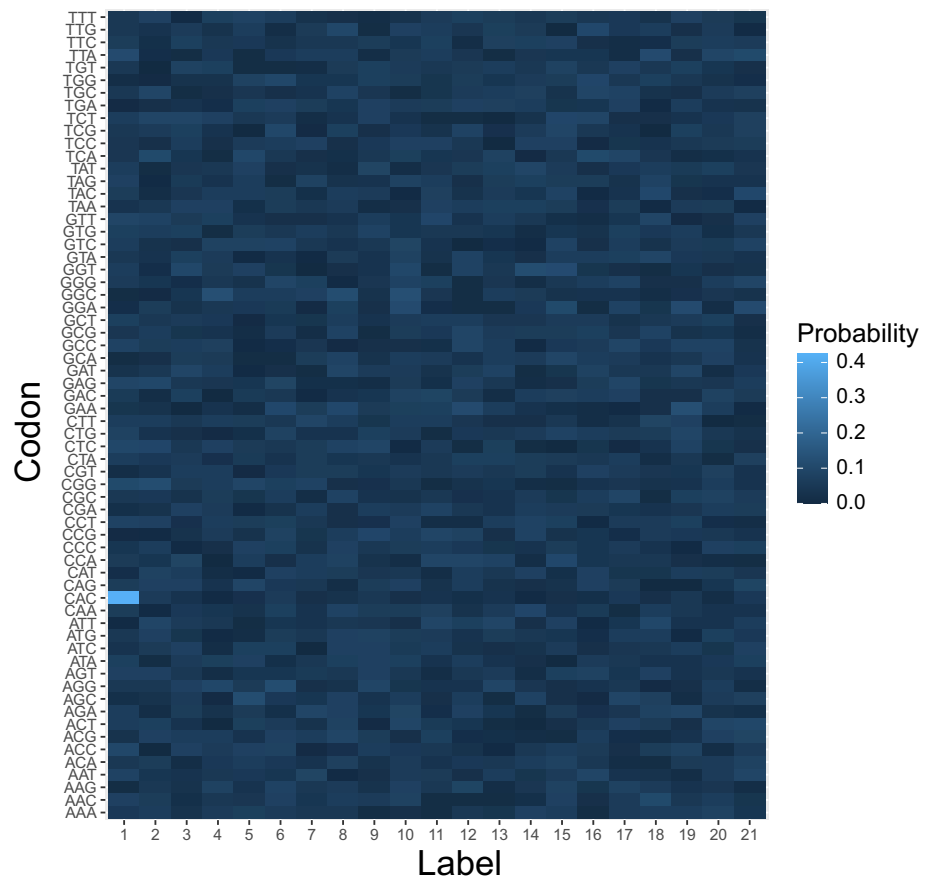
Figure 7. The heatmap of genetic code encoding 21 labels by 64 codons at the beginning of the simulation. This is a graphical representation of the matrix $\mathcal{P} = (p_{cl})$, in which each element p_{cl} ascribes a probability that a codon c in a row encodes a label l in a column.
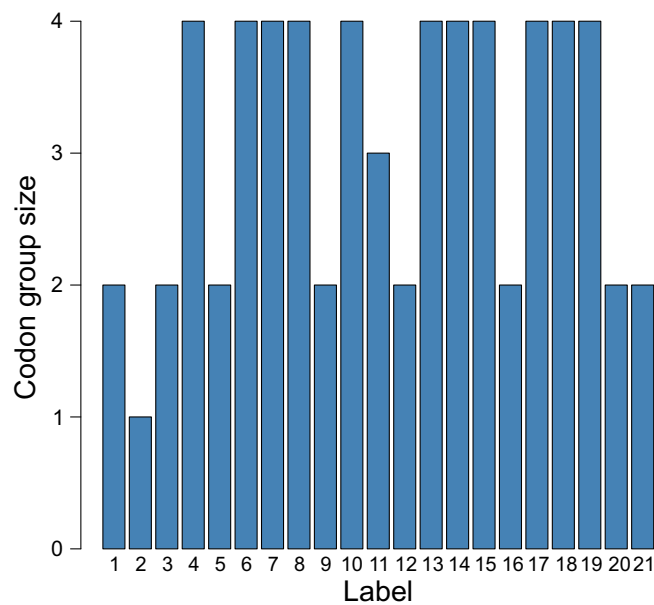
Figure 8. The distribution of the size of codon groups encoding a given label in the genetic code at the end of the simulation.

3. Discussion

In this paper, we investigated a potential emergence of genetic codes from a set of highly ambiguous codon assignments to the codes characterized by an unequivocal

assignment of codons to encoded labels. A similar problem was discussed in our previous papers [13,14], where we showed that the structure of genetic codes similar to that in the SGC could evolve under error minimization restrictions and the stepwise addition of amino acids to the genetic code. However, in the previous studies, the method of reading genetic information and its degeneracy, or more precisely, the way of recognizing not only a codon but also its codon neighborhood encoding the same label, was fixed and unchanged from the start of genetic code evolution. Therefore, in this approach, we not only included the evolution of codon assignments, but also took into account the evolution of the reading systems, which are responsible for codon block structures in the genetic code.

This assumption seems reasonable because the genetic code at very early stages of evolution was likely a mixture of different reading systems for the set of encoded items and characterized by ambiguity in the coding of amino acids until the whole system became more precise [5,12]. The genetic code could have evolved from a highly ambiguous and non-homogeneous system to an unambiguous system with nearly identical reading mechanisms for each encoded label. The results of our simulations are in agreement with the 2-1-3 model [18,19] and the four-column theory [15], which also assume that the degeneracy of the genetic code was subsequently reduced with time. Initially, the second codon position was used to determine encoded amino acids, whereas other codon positions were meaningless. Only later the additional codon position positions were used for distinguishing of coded amino acids.

In our simulations, we tested three different models of reading systems, namely, M_1, M_2, and M_3, which from the beginning of simulations could be used to read genetic information simultaneously with different probabilities. The model M_3 was the most tolerant in codon reading because it assumed that nine other codons can code for the same label as the reference codon. In M_2, six codons could have the same meaning, whereas in M_1, three codons besides the reference codon could code for the same label.

Although the simulations produced coding systems with high unambiguity in reading genetic information, there are some codons that can still code more than one label, one with a very high and the other with a very lower probability (Figure 6). This resembles the situation in natural biological systems because the present translational machinery also shows errors with rates of 10^{-3} to 10^{-6} per codon [20] or 10^{-3} to 10^{-5} per incorporated amino acid [21–24]. Mistranslation errors can be beneficial for parasites, for example, in adaptation to oxidative and environmental stresses as well as in host invasion and evasion of host immunity [25–29].

The evolution of ambiguous codes to those showing unequivocal assignment of codons implies that the number of tRNAs matching appropriate codons in protein synthesis increased during evolution. In the initial state of high ambiguity, a small fraction of tRNA could be enough to recognize codons due to the tolerant codon reading. When the codes became more unequivocal, a larger number of tRNAs were required to distinguish the codons and assign an appropriate amino acid. This scenario corresponds well to the view that the number of genes coding for tRNAs and aminoacyl-tRNA synthetases charging amino acids to the tRNA molecules increased via duplication during the genetic code evolution [7,16,30–33]. The increase in the number of tRNAs is associated with the more precise reading of genetic information.

The population of genetic codes evolved to reduce coding errors, which was manifested by a higher probability that a given codon encodes a fixed label. What is more, we also observed that the model M_1 started to dominate in coding systems among all possible types of reading very quickly. Therefore, at the end of simulations. we obtained genetic codes with low entropy values and a nearly homogeneous reading system. It is worth mentioning that the M_1 rule generates codon blocks with similar properties to those observed in the standard genetic code. Moreover, the codes at the end of the simulations were similar to the SGC in the number and structure of codon blocks encoding the same label. Such organization means that the codes are robust to point mutations that could

change the encoded information, because changing a nucleotide in one codon position does not result in changing the encoded label.

This tendency to minimize errors in the coded amino acid replacements is also present in the SGC and has been reported in many analytical and statistical studies [34–43]. However, the minimization of the errors turned out to be imperfect when genetic algorithms were applied and the SGC was compared with optimized codes [13,19,44–50]. Interestingly, alternative codes appeared to better minimize the harmful effects of mutations than the SGC [51,52]. This means that there are possibilities for improving the SGC. However, substantial changes in the code would be deleterious due to the universality of this code associated with beneficial exchange of genetic information between organisms, e.g., via horizontal gene transfer [16]. Therefore, other mechanisms, e.g., optimization of mutational pressure, evolved to mitigate the mutation errors associated with replication [53–56].

4. Materials and Methods

4.1. Representation of Genetic Codes

In order to investigate the problem of genetic structure evolution, we consider a population of 1000 theoretical genetic codes, i.e., candidate solutions, where each coding system is represented as a matrix $\mathcal{P} = (p_{cl})$, $1 \leq c \leq 64$ and $1 \leq l \leq 21$ with 64 rows, i.e., codons c, and 21 columns, i.e., amino acids and the stop translation signal, or more generally, encoded labels or items l. Each row in the matrix $\mathcal{P}$ denotes a probability distribution function of codon assignments to one out of 21 possible labels. Therefore, p_{cl} is the probability that a codon c encodes a label l. As a result, $\mathcal{P}$ describes a potential ambiguity in codon ascription. Figure 6 is a graphical representation of the coding system at the beginning of the simulations. The color gradient represents the probability of coding a given label in a column by a given codon in a row.

This approach has been used in our previous works [13,14], where we studied the process of genetic code evolution from a set of ambiguous codon assignments characterized by a large value of entropy to nearly unambiguous coding systems. However, it was assumed that the type of reading system was fixed during the whole simulation run. By a reading system, we mean the type of inaccuracy in codon recognition. In order to explain this important feature, let us consider that a codon c has the highest probability to encode a given label l. We also assume that the reading system allows that several codons belonging to the neighborhood of the reference codon c can encode the same label l. We considered in simulations the following different ways of reading:

M_1 all codons belonging to a given group encoding a fixed label have two fixed codon positions identical with codon c and differ in exactly one nucleotide at other codon positions;

M_2 all codons belonging to a given group encoding a fixed label have one fixed codon position with codon c and differ in exactly one nucleotide in one of the other two codon positions;

M_3 all codons belonging to a given group encoding a fixed label differ in exactly one nucleotide with codon c in any codon position.

For example, let us assume that codon GGG encodes a fixed label l. Then the neighborhood of GGG is:

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

4.2. Simulation Procedure

In contrast to [13,14], we took into account the fact that the reading systems can evolve at the same time and compete with each other. In other words, the reading mechanisms M_1, M_2, and M_3 can coexist in one coding system to encode the same labels with different probabilities. Thus, we considered that the respective reading mechanism for a given

encoded label is a random variable. This variable was generated at the beginning of simulated evolution. Figure 5 shows probability distribution functions of these three reading systems (y-axis) for each encoded label (x-axis). The probability values are depicted by respective color gradients. As we can see, the reading system is highly ambiguous, because a given label can be read by different mechanisms with various probabilities.

To sum up, we had two parameters responsible for shaping the structure of genetic code in our simulations. The first one is described by the probability that a given codon encodes a selected label. The second parameter is described by the probability that a given genetic label is read by a mechanism M_1, M_2, and M_3.

We ran simulations using the methodology of evolutionary algorithms. This was applied to find solutions in optimization problems where it is not possible to use classical analytical methods because their respective assumptions do not hold. According to this approach, the simulation process is divided into steps called generations. During each step, a population of individuals (solutions) is subjected to two important operators, namely, mutation and selection. They both act on the potential solutions, but in different way. The mutation is responsible for maintaining diversity of population, so we can test many different solutions during one generation. Thanks to the selection, it is possible to choose better solutions in terms of fitness. Such individuals have a greater probability to reproduce.

In the case considered here, the mutation operator randomly modifies the probability distribution functions for the assignment of codons and the type of reading system. The selection operator chooses better solutions with a greater probability for the next generation. Therefore, the central role in the simulations is the fitness function F, which allowed us to measure the quality of evolving genetic codes.

4.3. The Fitness Function F

Simply speaking, F is a modified version of the fitness function used in [13,14]. Similarly to [13], for each label $l = 1, 2, \ldots, 21$, we choose a respective codon c_i according to the Bayes rule. Therefore, we obtain the sequence of codons $C = c_1, c_2, \ldots c_{21}$, which is the most probable path to encode all 21 labels. What is more, each codon c_i belonging to C has its own neighborhood $N_j(c_i)$, describing a potential j way of reading. We considered $j = 1, 2, 3$ three types of neighborhoods defined according to the M_1, M_2, and M_3 reading models. In contrast to the previous works, the type of codon neighborhood was assigned at random to each codon $c_i \in C$. This fact was manifested in the form of the fitness function F, which is defined in the following way:

$$F = \sum_{c_1' \in N_{j_1}(c_1), \ldots, c_{21}' \in N_{j_{21}}(c_{21})} \frac{P(l=1|c_1')P(j_1)}{|N_{j_1}(c_1)|} \frac{P(l=2|c_2')P(j_2)}{|N_{j_2}(c_2)|} \cdot \ldots \cdot \frac{P(l=21|c_{21}')P(j_{21})}{|N_{j_{21}}(c_{21})|} ,$$

where $P(j_i)$, $j_i = 1, 2, 3$ is a probability that a label i is encoded by a codon group $N_{j_i}(c_i)$, which is defined according to reading the M_{j_i} constraint.

For a fixed label l, the formula

$$\frac{P(l|c')P(j)}{|N_j(c)|}, \ c' \in N_j(c)$$

has a simple interpretation. In the denominator, there is a joint probability that the codon neighborhood $N_j(c)$ of reading type M_j is chosen and the codon $c' \in N_j(c)$ encodes the label l. This probability is normalized by the size of codon block $N_j(c)$. Thus, the fitness function F gives information about the coding strength of genetic code. It is calculated over all possible combinations of codons $c_1', c_2', \ldots c_{21}'$ belonging to randomly selected codon neighborhoods.

In Figure 1, we presented the general tendency of the fitness function observed during simulations. The plot presents 10 simulation runs with different initial seeds. The values of the fitness function are shown in the logarithmic scale. We applied the general additive

model (GAM) approximation method to illustrate the changes in this function during the simulations. As we can see, the fitness values increase from very low values to very high ones after 20,000 generations and then stabilize. This indicates that the simulations converged to very similar final results independently of the initial state. Moreover, the course of the curves means that the studied genetic code evolved from that characterized by a high ambiguity to that showing a low ambiguity in the assignment of codons to labels and the selection of reading mechanisms.

4.4. 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 used the entropy

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

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. Higher values of entropy indicate that a given coding system is composed of ambiguous assignments of codons to labels, whereas lower values mean that the coding is unambiguous.

Similarly, for the entropy of codon to label assignments, we calculated the entropy for the distributions of three reading systems.

$$H_r(\mathcal{P}) = -\sum_{l=1}^{21}\sum_{j=1}^{3} p(j)\,log(p(j)),$$

where l is an encoded label, j is a type of reading, and $p(j)$ is a probability that j-type was chosen for reading a given label.

References

1. Khorana, H.G.; Buchi, H.; Ghosh, H.; Gupta, N.; Jacob, T.M.; Kossel, H.; Morgan, R.; Narang, S.A.; Ohtsuka, E.; Wells, R.D. Polynucleotide synthesis and the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **1966**, *31*, 39–49. [CrossRef]
2. Nirenberg, M.; Caskey, T.; Marshall, R.; Brimacombe, R.; Kellogg, D.; Doctor, B.; Hatfield, D.; Levin, J.; Rottman, F.; Pestka, S.; et al. The RNA code and protein synthesis. *Cold Spring Harb. Symp. Quant. Biol.* **1966**, *31*, 11–24. [CrossRef] [PubMed]
3. Knight, R.D.; Landweber, L.F. Is the genetic code really a frozen accident? New evidence from in vitro selection. *Mol. Strateg. Biol. Evol.* **1999**, *870*, 408–410. [CrossRef] [PubMed]
4. Di Giulio, M. The origin of the genetic code: Theories and their relationships, a review. *BioSystems* **2005**, *80*, 175–184. [CrossRef] [PubMed]
5. Barbieri, M. Evolution of the Genetic Code: The Ribosome-Oriented Model. *Biol. Theory* **2015**, *10*, 301–310. [CrossRef]
6. Sengupta, S.; Higgs, P.G. Pathways of Genetic Code Evolution in Ancient and Modern Organisms. *J. Mol. Evol.* **2015**, *80*, 229–243. [CrossRef]

7. Koonin, E.V. Frozen accident pushing 50: Stereochemistry, expansion, and chance in the evolution of the genetic code. *Life* **2017**, *7*, 22. [CrossRef]
8. Kun, Á.; Radványi, Á. The evolution of the genetic code: Impasses and challenges. *Biosystems* **2017**, *164*, 217–225. [CrossRef]
9. Crick, F.H. Codon—Anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* **1966**, *19*, 548–555. [CrossRef]
10. Crick, F.H. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38*, 367–379. [CrossRef]
11. Murphy, F.V.T.; Ramakrishnan, V. Structure of a purine-purine wobble base pair in the decoding center of the ribosome. *Nat. Struct. Mol. Biol.* **2004**, *11*, 1251–1252. [CrossRef] [PubMed]
12. Fitch, W.M.; Upper, K. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **1987**, *52*, 759–767. [CrossRef] [PubMed]
13. Błażej, P.; Wnetrzak, M.; Mackiewicz, D.; Mackiewicz, P. The influence of different types of translational inaccuracies on the genetic code structure. *BMC Bioinform.* **2019**, *20*, 114. [CrossRef] [PubMed]
14. Pawlak, K.; Wnetrzak, M.; Mackiewicz, D.; Mackiewicz, P.; Błażej, P. Models of genetic code structure evolution with variable number of coded labels. *Biosystems* **2021**, *210*, 104528. [CrossRef]
15. Higgs, P.G. 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. [CrossRef] [PubMed]
16. Koonin, E.V.; Novozhilov, A.S. Origin and evolution of the universal genetic code. *Annu Rev Genet* **2017**, *51*, 45–62. [CrossRef] [PubMed]
17. Weberndorfer, G.; Hofacker, I.L.; Stadler, P.F. On the evolution of primitive genetic codes. *Orig. Life Evol. Biosph.* **2003**, *33*, 491–514. [CrossRef]
18. Massey, S.E. A sequential "2-1-3" model of genetic code evolution that explains codon constraints. *J. Mol. Evol.* **2006**, *62*, 809–810. [CrossRef]
19. Massey, S.E. A Neutral Origin for Error Minimization in the Genetic Code. *J. Mol. Evol.* **2008**, *67*, 510–516. [CrossRef]
20. Ribas de Pouplana, L.; Santos, M.A.; Zhu, J.H.; Farabaugh, P.J.; Javid, B. Protein mistranslation: Friend or foe? *Trends Biochem. Sci.* **2014**, *39*, 355–362. [CrossRef]
21. Kramer, E.B.; Farabaugh, P.J. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* **2007**, *13*, 87–96. [CrossRef] [PubMed]
22. Schwartz, M.H.; Pan, T. Function and origin of mistranslation in distinct cellular contexts. *Crit. Rev. Biochem. Mol. Biol.* **2017**, *52*, 205–219. [CrossRef] [PubMed]
23. Allan Drummond, D.; Wilke, C.O. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* **2009**, *10*, 715–724. [CrossRef] [PubMed]
24. Mordret, E.; Dahan, O.; Asraf, O.; Rak, R.; Yehonadav, A.; Barnabas, G.D.; Cox, J.; Geiger, T.; Lindner, A.B.; Pilpel, Y. Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. *Mol. Cell* **2019**, *75*, 427–441.e5. [CrossRef] [PubMed]
25. Santos, M.A.S.; Cheesman, C.; Costa, V.; Moradas-Ferreira, P.; Tuite, M.F. Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp. *Mol. Microbiol.* **1999**, *31*, 937–947. [CrossRef] [PubMed]
26. Gomes, A.C.; Miranda, I.; Silva, R.M.; Moura, G.R.; Thomas, B.; Akoulitchev, A.; Santos, M.A. A genetic code alteration generates a proteome of high diversity in the human pathogen Candida albicans. *Genome Biol* **2007**, *8*, R206. [CrossRef]
27. 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.; et al. Innate immune and chemically triggered oxidative stress modifies translational fidelity. *Nature* **2009**, *462*, 522–526. [CrossRef]
28. Wiltrout, E.; Goodenbour, J.M.; Fréchin, M.; Pan, T. Misacylation of tRNA with methionine in Saccharomyces cerevisiae. *Nucleic Acids Res.* **2012**, *40*, 10494–10506. [CrossRef]
29. 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.; et al. Candida albicans CUG mistranslation is a mechanism to create cell surface variation. *mBio* **2013**, *4*, e00285-13. [CrossRef]
30. Cavalcanti, A.R.; Neto, B.D.; Ferreira, R. On the classes of aminoacyl-tRNA synthetases and the error minimization in the genetic code. *J. Theor. Biol.* **2000**, *204*, 15–20. [CrossRef]
31. Cavalcanti, A.R.O.; Leite, E.S.; Neto, B.B.; Ferreira, R. On the classes of aminoacyl-tRNA synthetases, amino acids and the genetic code. *Orig. Life Evol. Biosph.* **2004**, *34*, 407–420. [CrossRef] [PubMed]
32. Massey, S.E. Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. *Life* **2015**, *5*, 1301–1332. [CrossRef] [PubMed]
33. Massey, S.E. The neutral emergence of error minimized genetic codes superior to the standard genetic code. *J. Theor. Biol.* **2016**, *408*, 237–242. [CrossRef]
34. Sonneborn, T., Degeneracy of the genetic code: Extent, nature, and genetic implications. In *Evolving Genes and Proteins*; Academic Press: New York, NY, USA, 1965; pp. 377–397.
35. Woese, C.R. On the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* **1965**, *54*, 1546–1552. [CrossRef] [PubMed]
36. Epstein, C.J. Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. *Nature* **1966**, *210*, 25–28. [CrossRef] [PubMed]
37. Goldberg, A.L.; Wittes, R.E. Genetic code: Aspects of organization. *Science* **1966**, *153*, 420–424. [CrossRef]

38. Haig, D.; Hurst, L.D. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **1991**, *33*, 412–417. [CrossRef]

39. Freeland, S.J.; Hurst, L.D. The genetic code is one in a million. *J. Mol. Evol.* **1998**, *47*, 238–248. [CrossRef]

40. Di Giulio, M. The coevolution theory of the origin of the genetic code. *J. Mol. Evol.* **1999**, *48*, 253–255. [CrossRef]

41. Gilis, D.; Massar, S.; Cerf, N.J.; Rooman, M. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol* **2001**, *2*, research0049.1. [CrossRef]

42. Freeland, S.J.; Wu, T.; Keulmann, N. The case for an error minimizing standard genetic code. *Orig. Life Evol. Biosph.* **2003**, *33*, 457–477. [CrossRef] [PubMed]

43. Goodarzi, H.; Najafabadi, H.S.; Hassani, K.; Nejad, H.A.; Torabi, N. On the optimality of the genetic code, with the consideration of coevolution theory by comparison of prominent cost measure matrices. *J. Theor. Biol.* **2005**, *235*, 318–325. [CrossRef] [PubMed]

44. Novozhilov, A.S.; Wolf, Y.I.; Koonin, E.V. 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. [CrossRef] [PubMed]

45. Santos, M.A.S.; Gomes, A.C.; Santos, M.C.; Carreto, L.C.; Moura, G.R. The genetic code of the fungal CTG clade. *C. R. Biol.* **2011**, *334*, 607–611. [CrossRef] [PubMed]

46. 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. [CrossRef] [PubMed]

47. Santos, J.; Monteagudo, A. Inclusion of the fitness sharing technique in an evolutionary algorithm to analyze the fitness landscape of the genetic code adaptability. *BMC Bioinform.* **2017**, *18*, 195. [CrossRef]

48. 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. [CrossRef]

49. 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*, e0201715.

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

51. Błażej, P.; Wnetrzak, M.; Mackiewicz, P. The Importance of Changes Observed in the Alternative Genetic Codes. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018)—Volume 3: BIOINFORMATICS*; SciTePress: Funchal, Portugal, 2018; pp. 154–159.

52. 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. [CrossRef]

53. Dudkiewicz, A.; Mackiewicz, P.; Nowicka, A.; Kowalczuk, M.; Mackiewicz, D.; Polak, N.; Smolarczyk, K.; Banaszak, J.; Dudek, M.R.; Cebrat, S. Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution. *Future Gener. Comput. Syst.* **2005**, *21*, 1033–1039. [CrossRef]

54. 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. In *International Conference on Computational Science—ICCS 2008*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5103, pp.100–109.

55. 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*, e0130411. [CrossRef] [PubMed]

56. 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. [CrossRef] [PubMed]