The Extension of the Standard Genetic Code via Optimal Codon Blocks Division

Kuba Nowak[1], Paweł Błażej[2], Małgorzata Wnetrzak[2], Dorota Mackiewicz[2] and Paweł Mackiewicz[2]
[1]*Faculty of Mathematics and Computer Science, University of Wrocław, ul. Joliot-Curie 15, Wrocław, Poland*
[2]*Department of Genomics, Faculty of Biotechnology, University of Wrocław, F. Joliot-Curie 14a, 50-383 Wrocław, Poland*

Keywords: Code Extension, Codon Reassignment, Genetic Code, Graph Theory, Mutation, Optimisation.

Abstract: The standard genetic code (SGC) is a crucial biological system, which allows to transmit genetic information from DNA sequences to the protein world. The idea of the optimal extension of the SGC with new information appears especially interesting in the context of successful experimental achievements in reprogramming of this code. The aim of this code engineering is incorporating non-canonical amino acids (ncAAs) into synthesised artificial proteins with novel functions. Such molecules open new perspectives in medicine, chemistry and biotechnology. Several methods extending the canonical coding system were proposed. Here, we would like to investigate a problem of the optimal genetic code extension using graph theory methodology. We measured the quality of considered coding systems applying the set conductance, which determines the robustness level against point mutations of individual codon blocks encoding the same information. Thanks to that, we were able to find several possible optimal extensions of the SGC based on utilization of the codons redundant in the original code. We found codes that could encode up to 16 ncAAs and simultaneously code for 20 canonical amino acids and one stop translation signal. One of these codes was the most balanced, i.e. it consisted of the canonical set and the extended set that were characterized by the same level of robustness against point mutations. The proposed codes could be helpful in experimental construction of artificial genetic codes, which can encode new amino acid with new useful properties.

1 INTRODUCTION

The standard genetic code (SGC) is a template according to which genetic information stored in DNA sequences is decoded into the protein world. This coding system is nearly universal in all domains of life, with some rare exceptions (Santos et al., 2004; Sengupta and Higgs, 2005; Błażej et al., 2019a). The SGC is redundant because its 64 codons are assigned to 20 amino acids and stop translation signal. Therefore, 18 amino acids as well the stop signal are coded by more than one codon called synonymous. This property poses a question about potential using of the redundant codons in coding new information.

Recently, several methods for reducing the redundancy of the SGC and using it in the genetic code extension were proposed (Chin, 2014). Thanks to them, new non-canonical amino acids (ncAAs) with new desired functions can be introduced into the coding system. One of the most commonly used method to extend the SGC is based on the stop-codon suppression. In this approach, a rarely used stop codon, for example amber UAG codon, is assigned to a new amino acid (Noren et al., 1989; Chin, 2017; Italia et al., 2017; Young and Schultz, 2018). However, this method has a strong limitation, i.e. the number of newly added ncAAs cannot be greater than two because one of three stop codons must be left in the code.

Other method is based on a programmed frameshift suppression. Here, four-base codons (quadruplets) are used to encode new genetic information (Hohsaka et al., 1996; Anderson et al., 2004; Neumann et al., 2010). Generally, these extended codons are constructed by an addition of one base to rare classical codons in a given coding system. Additionally, modified tRNAs with four-base anticodons recognising the corresponding quadruplets should be constructed. However, the reading of such codons is not always perfect, because tRNAs reading classical codons can compete with those reading the quadruplets.

Another interesting approach uses synonymous codons of the SGC to encode various ncAAs (Iwane

Nowak, K., Błażej, P., Wnetrzak, M., Mackiewicz, D. and Mackiewicz, P.
The Extension of the Standard Genetic Code via Optimal Codon Blocks Division.
DOI: 10.5220/0010215500300037
In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 3: BIOINFORMATICS*, pages 30-37
ISBN: 978-989-758-490-9

et al., 2016). This is obtained by depletion of their corresponding tRNAs and addition of tRNAs precharged with the desired ncAAs. Using this methodology, the authors expanded the repertoire of canonical amino acids by three new ncAAs via division of multiple codon boxes. Despite methodological difficulties of these methods, there is no doubt that they allow us to synthesis of new proteins with demanding new properties, which can revolutionize drug discovery and chemical biology in the future.

However, it should be noted that all the studies that have been carried out so far do not investigate the problem of potential effectiveness of newly created coding systems. The optimality of extended SGC seems to be desired feature in terms of the quality of the preservation, storing and decoding of the genetic information. Recently, Nowak and co-authors have opened a discussion about the optimal extension of the SGC (Nowak et al., 2020). Basing on methodology borrowed from graph theory, the authors examined the minimum set of codons encoding the canonical genetic information as well as the set of vacant codons that potentially might encode ncAAs.

Here, we would like to continue the previous study about the genetic code extension. However, in contrast to the previous work, we would like to investigate in details the way of the SGC extension assuming the specific codon blocks structure observed in the SGC. These blocks consists of four codons which encode the same information and differ only in the third codon positions. We propose that new ncAAs can be encoded by codon groups fulfilling the same rule. Following this approach, we can extend the canonical coding system to encode up to 16 new ncAAs.

2 PRELIMINARIES

We used the methodology based on graph theory to describe the properties of individual codon blocks as well as the whole coding system represented by a graph partition. This approach was applied successfully to studies on the structure of the SGC and its robustness against point mutations (Błażej et al., 2018a; Błażej et al., 2019b; Aloqalaa et al., 2019). In addition, a possible extension of this coding system using six-letter alphabet (Błażej et al., 2020) was considered.

Following (Błażej et al., 2018a; Błażej et al., 2019b; Aloqalaa et al., 2019; Błażej et al., 2020), let $G(V,E)$ be a graph in which V is the set of vertices representing all possible 64 codons, whereas E is the set of edges between these vertices. We say that two codons $u,v \in V$ are connected by the edge $e(u,v) \in E$ if and only if the codon u differs from the codon v in exactly one position. Set E defined in this way describe all possible point mutations, which may occur between codons. So G is by definition unweighted and regular. According to this representation, every coding system, which uses 64 codons, is a partition P of the graph G into a selected number $l \geq 21$ of codon sets S_i. Therefore, we used the following definition:

$$P = \{S_1, S_2, \ldots, S_l : S_i \cap S_j = \emptyset,\ S_1 \cup S_2 \cup \ldots \cup S_l = V\}.$$

As a consequence of that, the problem of studying genetic code structure can be reformulated as a question about properties of its respective graph partition P. Similarly to the previous studies (Błażej et al., 2018a; Błażej et al., 2019b; Aloqalaa et al., 2019; Błażej et al., 2020), we used the set conductance measure in order to describe the level of robustness of a given codon group against point mutations.

Definition 1. *For a given graph G, let S be a subset of V. Then the conductance of S is defined as:*

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

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

Clearly, ϕ describes the robustness of a given codon group S against point mutations that may change the encoded genetic information. Using the definition of ϕ, it is also possible to define the k-size conductance as a lower bound on the set conductance for the sets of a given size k. A smaller value of ϕ indicates that there are relatively few point mutations that can change an amino acid or stop signal encoded by a given codon group.

Definition 2. *The k-size-conductance of the graph G, for $k \geq 1$, is defined as:*

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

The definition 2 appears to be especially useful in describing the most optimal codon structure. Particularly, if we take into account that the graph G has the representation as a Cartesian product of graphs (Błażej et al., 2018a), i.e.

$$G = K_4 \times K_4 \times K_4,$$

where K_4 is a 4-clique with the set of vertices representing four bases $\{A,U,G,C\}$, then using the Theorem 1 from (Bezrukov, 1999), we get that codon set S composed of the first k codons according to a selected lexicographic order fulfils the property $\phi(S) = \phi_{|S|}(G)$. It should be noted that including all

possible linear orders on the set $\{A,U,G,C\}$ and also all possible orders on the three codon positions, there are exactly 144 different lexicographic orders that can be introduced in G. Consequently, it is possible to characterise the most robust codon groups in terms of minimising the effect of point mutations.

The definitions 1 and 2 both characterise the properties of a selected set or group of sets with the same size. In order to obtain the characteristics of the whole coding system, we introduced the average conductance:

Definition 3. *The average conductance of a set collection P is defined as:*

$$\Phi(P) = \frac{1}{|P|} \sum_{S \in P} \phi(S) \ .$$

Thanks to that it is possible to measure the quality of a given group of sets P. A smaller Φ value indicates that the given code is built of codon groups on average better resistant to change of coded information due to point mutations.

3 RESULTS AND DISCUSSION

Similarly to (Nowak et al., 2020), we defined the set C_k composed of exactly k codons which encode the whole canonical genetic information, i.e. 20 amino acids and stop translation signal. Moreover, we claim that $\phi(C_k) = \phi_k(G)$, hence C_k is a set with the minimum possible number of connections to its complement, for fixed k. In order to obtain this set, we chose an lexicographic order as presented in (Nowak et al., 2020) and found exactly k-first codons that at the same time encode 20 amino acids and stop signal (Tab. 1)

. Consequently, the set of all 64 codons V, i.e. vertices of the graph G, has a representation:

$$V = C_k \cup C'_k \ ,$$

where C'_k is a set of vacant codons, which could be assigned to ncAAs. Moreover, the standard codon assignments induces the partition P_k of the set C_k into 21 disjoint codon groups encoding canonical genetic information. In addition, the new genetic information can creates also a partition P'_k, which is the set of vacant codons C'_k. As a result, the extended genetic code P has a representation as a union:

$$P = P_k \cup P'_k \ .$$

Table 1 presents a partition of the set of 64 possible codons into two sets, namely, the canonical set C_{28} (in red) consisting of 28 codons encoding the canonical information and the extended C'_{28} of 36 vacant codons (in blue). Clearly, P_{28} is in this case a partition of C_{28} induced by the canonical codon assignments, whereas codons belonging to C'_{28} may encode new ncAAs. It was shown in (Nowak et al., 2020) that $k = 28$ is the smallest number of codons selected in a lexicographic order, which is able to encode all 21 canonical items. Thus, the presented code is simultaneously the most 'minimalistic' and optimal. It still preserves the canonical information of 21 elements, is the most robust against losing this information due to point mutations and contains the largest possible number of vacant codons for these conditions. In this code, 15 amino acid and stop signal are encoded by single codons, three amino acids by two codons and two amino acids by three codons.

Table 1: The partition of the set of 64 possible codons into two sets. The set C_{28} (red) contains exactly 28 codons which encode canonical genetic information. They were chosen according to a lexicographic order induced by the linear order of bases $U < C < A < G$ and the order of codon positions $1 < 2 < 3$. The set C'_{28} of 36 vacant codons (blue) can be split into 16 disjoint codon groups with the same first two codon positions.

UUU Phe	UCU	UAU Tyr	UGU Cys
UUC	UCC	UAC	UGC
UUA	UCA	UAA	UGA
UUG Leu	UCG Ser	UAG Stop	UGG Trp
CUU Leu	CCU	CAU His	CGU Arg
CUC	CCC	CAC	CGC
CUA	CCA	CAA	CGA
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUU Ile	ACU	AAU Asn	AGU Ser
AUC	ACC	AAC	AGC
AUA	ACA	AAA	AGA
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU	GAU Asp	GGU Gly
GUC	GCC	GAC	GGC
GUA	GCA	GAA	GGA
GUG Val	GCG Ala	GAG Glu	GGG Gly

3.1 The Extended Genetic Codes

In contrast to the previous work (Nowak et al., 2020), we considered here another aspect of the SGC structure. We started with the observation that the set of all 64 codons in the SGC can be split into 16 disjoint codon blocks $B_i, i = 1,2,\ldots,16$ in such a way that codons belonging to a given group B_i differ only in the third codon position. Moreover, it is clearly visible in the SGC that a group of codons encoding the same amino acid is generally composed of codons also with only the third positions different. Using this property, we investigated a potential extension of the genetic code according to a similar rule. Here,

we claim that any new non-canonical amino acid can be encoded by vacant codons belonging only to fixed blocks $B_i, i = 1,2,\ldots,16$. In consequence, the codons encoding the same amino acid differ only in the third position. Mathematically speaking, partition P'_k is composed of codon blocks S_i, $1 \leq i \leq 16$ belonging to the set of vacant codons C'_k. These codon blocks fulfil the following property:

$$S_i = B_i \cap C'_k \ , \ i = 1,2,\ldots,16 \ .$$

Tab. 2, presents an example of codon block of B_i type. In this case, $S_i = B_i \cap C'_k$ is composed of exactly two vacant codons (in blue) and belongs to the partition of the set C'_{28} from the 'minimalistic' code. All codons in this group have only the third codon positions different.

Table 2: The example of codon block B_i, extracted from the top left-hand corner of the code shown in Table 1. This block is divided into two sets. The codons in red encode canonical genetic information, whereas those in blue are vacant and follow the property $S_i = C'_k \cap B_i$.

UUU	Phe
UUC	
UUA	
UUG	Leu

Interestingly, the family of vacant codons in the set P'_{28} from the code shown in Table 1 has a better robustness in terms of Φ than P_{28} because $\Phi(P'_{28}) = 0.861$, whereas $\Phi(P_{28}) = 0.975$. This property follows from the fact that the canonical amino acids and stop are in most cases encoded by a single codon. This causes an imbalance between the 'minimalistic' canonical set P_{28} and the non-canonical set P'_{28}.

In order to improve the robustness of the canonical set, we have to increase the number of codons by inclusion of consecutive codons in the fixed lexicographic order. However, this procedure causes simultaneously deterioration of conductance Φ in the non-canonical set P'_k. It also causes that the number of newly amino acids incorporated into the code decreases. To measure and control the properties of these two sets simultaneously, we investigated their conductance in relation to k, i.e. the number of codons building the set C_k. Thus, we introduced a balance measure

$$\Psi(P'_k, P_k) = \frac{\Phi(P'_k)}{\Phi(P_k)} \ ,$$

which is defined as a ratio of the average conductance of non-canonical set $\Phi(P'_k)$ to that of canonical sets $\Phi(P_k)$. This measure allows us to compare the conductance properties of the non-canonical set in relation to the canonical set. It is evident that the value of Ψ near to one means that both P'_k and P_k achieve the same level of robustness against point mutations, which could change the coded information. Therefore, it seems reasonable to find for every k the most balanced codes that minimize $|1 - \Psi(P'_k, P_k)|$ over all possible P'_k and P_k for fixed k.

Table 3: The values of selected measures for the most balanced genetic codes calculated for the fixed number of codons in the canonical set $k = 28,29,\ldots,63$; Ψ - the balance; $\Phi(P'_k)$ and $\Phi(P_k)$ - the average conductance for non-canonical and canonical sets, respectively; $|P|$ - the number of potentially encoded ncAAs.

k	Ψ	$\Phi(P'_k)$	$\Phi(P_k)$	$\lvert P\rvert$
28	0.8829	0.8611	0.9753	16
29	0.8949	0.8681	0.9700	16
30	0.9053	0.8750	0.9665	16
31	0.9209	0.8819	0.9577	16
32	0.9299	0.8889	0.9559	16
33	0.9424	0.8958	0.9506	16
34	0.9550	0.9028	0.9453	16
35	0.9678	0.9097	0.9400	16
36	0.9807	0.9167	0.9347	16
37	0.9937	0.9236	0.9295	16
38	1.0012	0.9306	0.9295	16
39	1.0043	0.9375	0.9335	16
40	1.0169	0.9444	0.9287	16
41	1.0302	0.9514	0.9235	16
42	1.0458	0.9583	0.9164	16
43	1.0595	0.9653	0.9111	16
44	1.0733	0.9722	0.9058	16
45	1.0863	0.9792	0.9014	16
46	1.1004	0.9861	0.8961	16
47	1.1148	0.9931	0.8908	16
48	1.1293	1.0000	0.8855	16
49	1.1358	1.0000	0.8804	15
50	1.1358	1.0000	0.8804	14
51	1.1408	1.0000	0.8765	13
52	0.8963	0.7778	0.8677	4
53	0.9418	0.8056	0.8554	4
54	0.9752	0.8333	0.8545	4
55	1.0026	0.8611	0.8589	4
56	1.0392	0.8889	0.8554	4
57	1.0783	0.9167	0.8501	4
58	1.1180	0.9444	0.8448	4
59	1.1581	0.9722	0.8395	4
60	1.1987	1.0000	0.8342	4
61	0.9525	0.7778	0.8166	1
62	1.0815	0.8889	0.8219	1
63	1.2246	1.0000	0.8166	1

Figure 1 shows the relationship between the Ψ values and the size of the set C_k calculated for the most balanced codes for fixed k. In the plot, we also

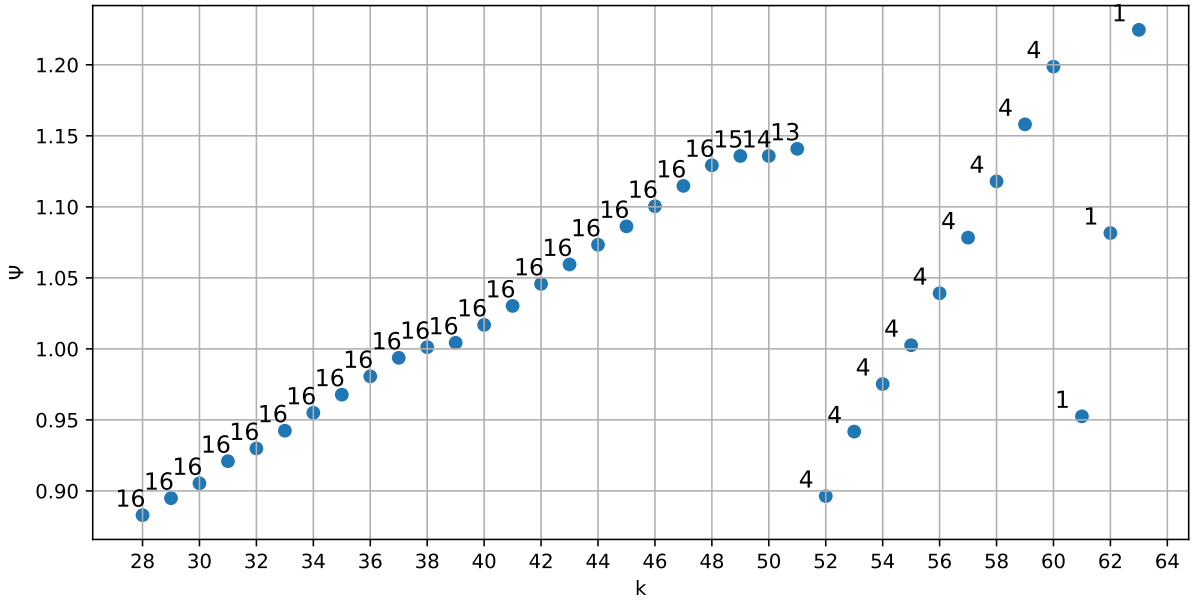

Figure 1: The values of Ψ, calculated for the most balanced codes, in relation to the size k, i.e. the number of codons encoding canonical genetic information. The number of newly encoded ncAAs is presented above dots.

included the number of potential ncAAs for the given code. The numerical details are presented in Table 3.

Interestingly, we can distinguish three increasing trends in this plot. In one of them, the balance Ψ is constantly increasing with $k \in [28,51]$, which results from the fact that the set conductance calculated for codon groups of the canonical set P_k is generally decreasing, whereas that for codon groups of the non-canonical set P'_k is increasing (Table 3). For $k < 38$, the set conductance Φ of the canonical set is generally higher than that calculated for respective codon groups belonging to the non-canonical set. For $k = 38$, the Φ values of these sets are the most similar and the Ψ value is most close to one. Thereby, it is possible to extend the SGC by 16 amino acids and achieve a good balance between the canonical and non-canonical set. Ψ is still growing till $k = 51$ and in the range $k = [38,51]$, the set conductance of the canonical set is smaller than that of the non-canonical set (Table 3).

The value of Ψ calculated for the most balanced extended genetic codes declines sharply for $k = 52$, in comparison to Ψ for $k \leq 51$ (Figure 1). What is more, the number of possibly encoded ncAAs also decreases dramatically till only four and persists for the range of $k = [52,60]$. It is mainly related with a much smaller number of vacant codons, which were left for $k > 51$. The increasing trend of the balance in this range results from a gradual dropping of Φ calculated for the canonical set and its rise for the non-canonical set (Table 3). Within the range of $k = [52,60]$, we can also find the most balanced code for $k = 55$. In this case, we can extend the SGC with four ncAAS. The third growing trend in this plot, for $k = 61,62,63$, is mainly associated with an increase in the set conductance of the non-canonical set because this measure for the canonical set is rather stable (Table 3). These codes can be extended with only one ncAA.

3.2 The Most Balanced Extended Genetic Codes

Here, we present two examples of the most balanced extended genetic codes. These coding systems achieve the values of Ψ most close to 1 among all genetic codes generated in the lexicographic order method.

Table 4 shows the most balanced code that contains $k = 38$ codons in the canonical set (in red) and encodes up to 16 non-canonical amino acids. The potential ncAAs can be coded by also 16 codon blocks (in blue), which are induced by codons with the same the first and the second positions. In the case of the non-canonical set, the size of the codon groups is one or two, whereas in the canonical set, the codon groups encoding an amino acid or stop signal has one, two, three or four codons. The average conductance for the canonical set P_k and the non-canonical set P'_k are nearly the same, i.e. 0.93, which causes that the balance $\Psi(P'_k, P_k)$ is 1.001 in this case.

Another example of the most balanced code that is able to encode up to four ncAAs is given in Table 5. As we can see, the canonical set contains 55

Table 4: The extended genetic code that encodes at most 16 ncAAs by new blocks with 26 codons (blue) and is the most balanced among all extended genetic codes achieving $\Psi = 1.001$. The set encoding canonical information contains $k = 38$ codons (red).

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC	UAC	UGC Cys
UUA	UCA	UAA	UGA
UUG Leu	UCG Ser	UAG Stop	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC	CCC	CAC	CGC Arg
CUA	CCA	CAA	CGA
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC	ACC	AAC	AGC Ser
AUA	ACA	AAA	AGA
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC Val	GCC	GAC	GGC Gly
GUA	GCA	GAA	GGA
GUG Val	GCG Ala	GAG Glu	GGG Gly

codons (in red) and nine vacant codons belonging to the non-canonical set (in blue). The non-canonical set is composed of four groups, three contain two codons and one has three codons, whereas in the canonical set there are groups with one, two, three, four or six codons. The average conductance Φ calculated for the canonical set $\Phi(P_{55})$ and extended set $\Phi(P'_{55})$ are nearly similar and equal 0.86. Therefore, the balance measure Ψ is 1.003 for this code.

Table 5: The extended genetic code that encodes at most four ncAAs by new blocks with nine codons (blue) and is the most balanced among all extended genetic codes achieving $\Psi = 1.003$. The set encoding canonical information contains $k = 55$ codons (red).

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA Leu	UCA Ser	UAA Stop	UGA Stop
UUG Leu	UCG Ser	UAG Stop	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC Leu	CCC Pro	CAC His	CGC Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC Ile	ACC Thr	AAC Asn	AGC Ser
AUA Ile	ACA Thr	AAA Lys	AGA Arg
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUU Val	GCU Ala	GAU Asp	GGU
GUC	GCC	GAC	GGC
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG	GCG	GAG	GGG

4 CONCLUSIONS

We presented a potential method of genetic code extension. Using it, we searched for codes that encode all 21 elements, i.e. 20 amino acids and one stop translation signal, but can simultaneously release as many as possible vacant codons, which can be assign to non-canonical amino acids. We also imposed conditions for the codes to minimize consequences of the point mutations, which could change the coded information.

This assumption is in good agreement with the hypothesis, that the SGC evolved to minimize the damaging effects of mutations or mistranslations of coded proteins (Woese, 1965; Sonneborn, 1965; Epstein, 1966; Goldberg and Wittes, 1966; Haig and Hurst, 1991; Freeland and Hurst, 1998; Freeland et al., 2000; Gilis et al., 2001). Detailed studies revealed that the code is not ideally optimized in this respect but shows a global tendency to minimization this harmful consequences (Błażej et al., 2016; Massey, 2008; Novozhilov et al., 2007; Santos et al., 2011; Santos and Monteagudo, 2017; Wnetrzak et al., 2018; Błażej et al., 2018b; Błażej et al., 2019b; Wnetrzak et al., 2019). Thus, our assumption seems reasonable because the biological systems require the optimization in terms of mutations (Radman et al., 1999; Sniegowski et al., 2000; Dudkiewicz et al., 2005; Mackiewicz et al., 2008; Błażej et al., 2015; Błażej et al., 2017).

Therefore, we used the set conductance, i.e. the fraction of mutations changing genetic information, as a measure of the code robustness against point mutations of individual codon blocks encoding the same information. Accordingly, the applied algorithm found the new codon groups for non-canonical amino acids consisted of codons that differ only in the third codon position for a given group. This is in line with the general structure of canonical codon blocks observed in the standard genetic code and very good optimization of this code in this codon position in terms of minimization of point mutations (Santos and Monteagudo, 2010; Błażej et al., 2018b).

Following these assumptions and using the methodology borrowed from graph theory, we showed several examples of the extended genetic codes, which potentially could encode up to 16 non-canonical amino acids. We also searched for the extended genetic code in which both the canonical part and the extended part possess the same level of robustness against point mutations. We found two such the most balanced codes. One includes 38 codons for the canonical information and 26 codons for 16 ncAAs, and another contains 55 codons for the canon-

ical elements and nine codons for encoding four non-canonical amino acids. The proposed extended codes can be useful in designing artificial codes encoding new amino acid with specific properties, which can find application in various branches of biotechnology and synthetic biology.

5 FUNDING STATEMENT

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

REFERENCES

Aloqalaa, D. A., Kowalski, D. R., Blazej, P., Wnetrzak, M., Mackiewicz, D., and Mackiewicz, P. (2019). The impact of the transversion/transition ratio on the optimal genetic code graph partition. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019) - Volume 3: BIOINFORMATICS*, pages 55–65.

Anderson, J. C., Wu, N., Santoro, S. W., Lakshman, V., King, D. S., and Schultz, P. G. (2004). An expanded genetic code with a functional quadruplet codon. *Proc Natl Acad Sci U S A*, 101(20):7566–7571.

Bezrukov, S. L. (1999). *Edge isoperimetic problems on graphs.*, volume 7, pages 157–197. Akademia Kiado, Budapest.

Błażej, P., Kowalski, D., Mackiewicz, D., Wnetrzak, M., Aloqalaa, D., and Mackiewicz, P. (2018a). The structure of the genetic code as an optimal graph clustering problem. *www.biorxiv.org/content/early/2018/05/28/332478*.

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., Mackiewicz, D., Gagat, P., and Mackiewicz, P. (2019a). Many alternative and theoretical genetic codes are more robust to amino acid replacements than the standard genetic code. *Journal of Theoretical Biology*, 464:21–32.

Błażej, P., Wnetrzak, M., Mackiewicz, D., and Mackiewicz, P. (2018b). 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., and 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., and Mackiewicz, P. (2020). Basic principles of the genetic code extension. *Royal Society Open Science*, 7(2):191384.

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.

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

Chin, J. W. (2017). Expanding and reprogramming the genetic code. *Nature*, 550(7674):53–60.

Dudkiewicz, A., Mackiewicz, P., Nowicka, A., Kowalezuk, M., Mackiewicz, D., Polak, N., Smolarczyk, K., Banaszak, J., Dudek, M. R., and Cebrat, S. (2005). Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution. *Future Generation Computer Systems*, 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):25–8.

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

Freeland, S. J., Knight, R. D., and Landweber, L. F. (2000). Measuring adaptation within the genetic code. *Trends Biochem Sci*, 25(2):44–5.

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

Goldberg, A. L. and Wittes, R. E. (1966). Genetic code: aspects of organization. *Science*, 153(3734):420–4.

Haig, D. and Hurst, L. D. (1991). A quantitative measure of error minimization in the genetic code. *Journal of Molecular Evolution*, 33(5):412–417.

Hohsaka, T., Ashizuka, Y., Murakami, H., and Sisido, M. (1996). Incorporation of nonnatural amino acids into streptavidin through in vitro frame-shift suppression. *J Am Chem Soc*, 118(40):9778–9779.

Italia, J. S., Addy, P. S., Wrobel, C. J., Crawford, L. A., Lajoie, M. J., Zheng, Y., and Chatterjee, A. (2017). An orthogonalized platform for genetic code expansion in both bacteria and eukaryotes. *Nat Chem Biol*, 13(4):446–450.

Iwane, Y., Hitomi, A., Murakami, H., Katoh, T., Goto, Y., and Suga, H. (2016). Expanding the amino acid repertoire of ribosomal polypeptide synthesis via the artificial division of codon boxes. *Nature Chemistry*, 8(4):317–325.

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

Massey, S. E. (2008). A neutral origin for error minimization in the genetic code. *Journal of Molecular Evolution*, 67(5):510–516.

Neumann, H., Wang, K., Davis, L., Garcia-Alai, M., and Chin, J. W. (2010). Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature*, 464(7287):441–4.

Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C., and Schultz, P. G. (1989). A general method for site-specific incorporation of unnatural amino acids into proteins. *Science*, 244(4901):182–8.

Novozhilov, A. S., Wolf, Y. I., and Koonin, E. V. (2007). Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol Direct*, 2:24.

Nowak, K., Błażej, P., Wnetrzak, M., Mackiewicz, D., and Mackiewicz, P. (2020). Some theoretical aspects of reprogramming the standard genetic code. *www.biorxiv.org/content/10.1101/2020.09.12.294553v1*.

Radman, M., Matic, I., and Taddei, F. (1999). Evolution of evolvability a. *Annals of the New York Academy of Sciences*, 870(1):146–155.

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

Santos, J. and 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.

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

Santos, M. A. S., Gomes, A. C., Santos, M. C., Carreto, L. C., and Moura, G. R. (2011). The genetic code of the fungal ctg clade. *Comptes Rendus Biologies*, 334(8-9):607–611.

Sengupta, S. and Higgs, P. G. (2005). A unified model of codon reassignment in alternative genetic codes. *Genetics*, 170(2):831–40.

Sniegowski, P. D., Gerrish, P. J., Johnson, T., and Shaver, A. (2000). The evolution of mutation rates: separating causes from consequences. *Bioessays*, 22(12):1057–1066.

Sonneborn, T. (1965). *Degeneracy of the genetic code: extent, nature, and genetic implications.*, pages 377–397. Academic Press, New York.

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

Wnetrzak, M., Błażej, P., Mackiewicz, D., and Mackiewicz, P. (2018). The optimality of the standard genetic code assessed by an eight-objective evolutionary algorithm. *BMC Evolutionary Biology*, 18:192.

Woese, C. R. (1965). On the evolution of the genetic code. *Proc Natl Acad Sci U S A*, 54(6):1546–52.

Young, D. D. and Schultz, P. G. (2018). Playing with the molecules of life. *ACS Chem Biol*, 13(4):854–870.