OXFORD GENETICS

GENETICS, 2021, **218(1)**, iyab040
DOI: 10.1093/genetics/iyab040
Advance Access Publication Date: 26 March 2021
Investigation

Some theoretical aspects of reprogramming the standard genetic code

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. F. Joliot-Curie 15, 50-383 Wrocław, Poland
[2]Department of Bioinformatics and Genomics, Faculty of Biotechnology, University of Wrocław, ul F. Joliot-Curie 14a, 50-383 Wrocław, Poland

*Corresponding author: Department of Bioinformatics and Genomics, Faculty of Biotechnology, University of Wrocław, ul. Joliot-Curie 14a, 50-383 Wrocław, Poland. pawel.blazej@uwr.edu.pl

Abstract

Reprogramming of the standard genetic code to include non-canonical amino acids (ncAAs) opens new prospects for medicine, industry, and biotechnology. There are several methods of code engineering, which allow us for storing new genetic information in DNA sequences and producing proteins with new properties. Here, we provided a theoretical background for the optimal genetic code expansion, which may find application in the experimental design of the genetic code. We assumed that the expanded genetic code includes both canonical and non-canonical information stored in 64 classical codons. What is more, the new coding system is robust to point mutations and minimizes the possibility of reversion from the new to old information. In order to find such codes, we applied graph theory to analyze the properties of optimal codon sets. We presented the formal procedure in finding the optimal codes with various number of vacant codons that could be assigned to new amino acids. Finally, we discussed the optimal number of the newly incorporated ncAAs and also the optimal size of codon groups that can be assigned to ncAAs.

Keywords: genetic code; codon; code reprogramming; code expansion

Introduction

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. Thanks to that, genetic information can be stored in DNA and transmitted into the protein world. It is clear that the SGC is redundant because there are 18 amino acids encoded by more than one codon, *i.e.* 2, 3, 4, or 6 codons. Such codons encoding the same amino acid are named synonymous and are organized in groups called blocks or boxes.

The redundancy is a consequence of necessity of coding 21 items by non-overlapping words with a constant length and having at their disposal four nucleotides. Codons, *i.e.* words composed of three nucleotides, are enough to encode all these 21 elements. Shorter words, *e.g.* with the length of two nucleotides would encode up to 16 items. According to the adaptation hypothesis, the redundancy of the SGC could evolve to minimize adverse effects of mutations or translational errors 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). This property causes that the coded information is more resistant to changes. Mutations in codons encoding the same amino acid are neutral in terms of the coded amino acid. Therefore, the SGC appears to be a good buffer to the mutations. The redundancy may also result from the necessity to fill in as many as possible codons with sense information, otherwise the unassigned codons could pause or break protein synthesis. It would result in the production of shorter products without function or with disabled activity. Moreover, the redundancy enables coding of selected amino acids by more than one codon, which may increase the number of this amino acid in coded proteins. The encoding a given amino acid by several codons differently used enables regulation of efficiency and speed of translation, which can be important in correct protein folding (Orešič and Shalloway 1998; Xia 1998; Kanaya *et al.* 1999; Akashi 2003; Rocha 2004; Zhou *et al.* 2009; Plotkin and Kudla 2011; D'Onofrio and Abel 2014).

Despite the additional roles, the presence of the surplus codons suggests to reduce their redundancy and exploit it in expanding the genetic code. Thanks to that, we could use the extra codons for introducing new genetic information into the canonical coding system. The inclusion of non-canonical amino acids (ncAAs) in the code can allow us for producing new artificial proteins with novel functions and properties. This approach is very promising for synthetic biology and may find many applications in medicine, industry and biotechnology.

There are several approaches to the expansion of the SGC (Chin 2014). The first one is stop-codon suppression (Noren *et al.* 1989; Chin 2017; Italia *et al.* 2017; Young and Schultz 2018). In this method, stop translation codons, especially those that are very rarely used, *e.g.* UAG, are applied to encode new ncAAs. This technique needs a modified aminoacyl-tRNA synthetase that charges a tRNA molecule with the ncAA. However, this approach has several drawbacks. For example, we can expand the SGC by

Received: September 24, 2020. **Accepted:** February 11, 2021

only up two new amino acids, because one of the three stop codons must be left to function as a termination signal of translation (Ozer *et al.* 2017). What is more, the newly added ncAAs could compete with translation release factors, which may have an impact on the speed and efficiency of the protein synthesis.

The second method is related to programmed frameshift suppression. In this approach, four-base codons, called quadruplets, are used to encode new ncAAs (Hohsaka *et al.* 1996; Anderson *et al.* 2004; Neumann *et al.* 2010). Generally, these quadruplets are composed of rarely used classical codons with an additional base. They are decoded by a modified tRNAs containing complementary four-base anticodons. It should be noted, that the competition between tRNAs reading classical codons and the respective quadruplets can decrease the efficiency of the whole procedure.

The third method postulates the expansion of the SGC by using selected synonymous codons, whose corresponding tRNAs are pre-charged with ncAAs (Iwane *et al.* 2016). In this approach, up to 1, 2, 3, and 5 codons from corresponding synonymous codon blocks can be used to encode ncAAs leaving at least one codon for the canonical amino acid. This method can significantly increase the number of ncAAs by using many codon boxes. However, changes in the using of synonymous codons can disturb the translation and protein folding process, because the codon usage is associated with the speed of protein synthesis (Plotkin and Kudla 2011).

Another approach is based on addition one pair of unnatural nucleotides to the canonical four bases (Ishikawa *et al.* 2000; Ohtsuki *et al.* 2001; Yang *et al.* 2007; Kimoto *et al.* 2009; Malyshev *et al.* 2009; Dien *et al.* 2018; Hamashima *et al.* 2018). It is thereby possible to generate up to $6^3 - 4^3 = 216 - 64 = 152$ new codons to which ncAAs can be assigned. Since the new genetic information does not involve the canonical codons, it does not interfere with the natural system. For example, the new unnatural codons may not compete with tRNAs charged with canonical amino acids. However, this method must deal with some molecular problems: the pairing efficiency of unnatural bases and their recognition by polymerases during DNA replication. Hopefully, these technical problems can be solved with the development of molecular biology and biological chemistry, so it is interesting to consider the expansion of the SGC from theoretical point of view, which may be used in the experimental solutions. The theoretical approach to the expansion of the SGC has been recently proposed by Błażej *et al.* (2020). The authors analyzed how to expand the SGC up to 216 codons generated by a six-letter nucleotide alphabet, including besides four canonical bases also one pair of new bases. The model of the code assumed the gradual addition of the codons to minimize the consequences of point mutations.

In this paper, we investigated other theoretical aspects of the SGC expansion using 64 canonical codons and the code redundancy. We focused on finding the rules of the code expansion via optimal partition of codon boxes into two parts coding canonical and new information. In the first step, we found the minimal set of codons that encodes the complete canonical information, whereas the vacant codons can be used to encode non-canonical items. At the same time, this code was supposed to be the most robust to point mutations, which could change the information between the canonical and non-canonical parts. We considered codes with various number of the codons in the canonical set and studied the robustness of the codes to lose encoded information including physicochemical properties of amino acids.

Methods

Representation of the genetic code as a graph

We described properties of the SGC using the methodology of graph theory, which studies graphs, *i.e.* mathematical structures, consisting of objects that are related to each other in some way. According to this approach, the objects are represented by vertices (nodes), which are connected by edges (links). This representation is suitable to describe relationships between all possible 64 codons of the SGC in terms of point mutations. In this case, vertices are codons, whereas edges are all possible single point mutations, which may occur between codons in protein coding sequences. Assuming that, each codon has nine connections with others, which results from three possible point mutations in each of three codon positions (Figure 1). For example, codon *UUU* can mutate into: *AUU*, *CUU*, and *GUU* due to mutations in the first codon position, *UAU*, *UCU*, and *UGU* because of mutations in the second codon position, as well as *UUA*, *UUC*, and *UUG* on account of mutations in the third codon position. This code representation was successfully used in many problems related to the optimality of the SGC in terms of point mutations (Aloqalaa *et al.* 2019; Błażej *et al.* 2019b; Aloqalaa *et al.* 2020). Moreover, some rules of the optimal genetic code expansion using an additional pair of unnatural bases were investigated by Błażej *et al.* (2020).

In order to describe the SGC as a graph in a more formal way, let us assume that $G(V, E)$ is 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. In other words, these codons can mutate one into another with one base substitution. Thus, this graph is undirected, because its edges are bidirectional. The graph is also regular, because each vertex has the same number of neighbors, *i.e.* the same degree. In this case, it is nine. Moreover, the graph presented in Figure 1 is unweighted, because its edges do not have assigned different numerical values, *i.e.* weights.

Codons can be clustered in different groups encoding 20 amino acids and the translation termination signal, as well as ncAAs in the case of expanded versions. Thus, following graph theory, we can say that every possible genetic code induces a partition $\mathcal{P}$ of the codon set V into $l \geq 21$ disjoint non-empty subsets S, *i.e.* codon groups encoding at least 21 items. This can be written in the formal way as:

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

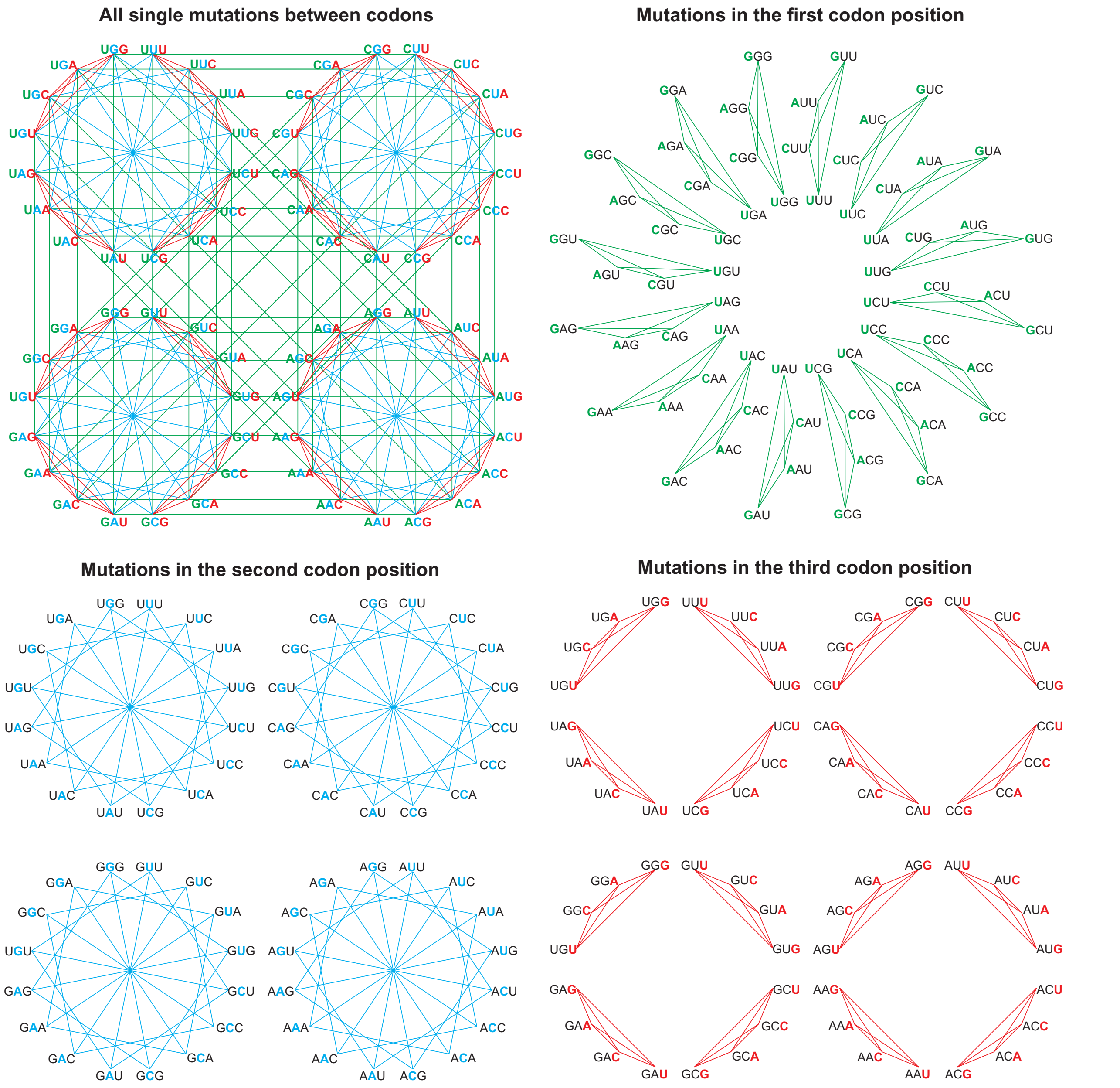
Measures of genetic code robustness to point mutations

A good measure, which describes the amount of information lost due to mutations of codons, is related to the set conductance and its modifications (Aloqalaa *et al.* 2019; Błażej *et al.* 2019b; Aloqalaa *et al.* 2020; Błażej *et al.* 2020). They were defined below.

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

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

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


All single mutations between codons

Mutations in the first codon position

Mutations in the second codon position

Mutations in the third codon position

Figure 1 The representation of the standard genetic code as a graph, in which vertices are represented by 64 codons, whereas edges are all possible single mutations that may occur between these codons. For clarity, the connections between the codons were presented also separately for mutations in three codon positions. In fact, each codon can mutate into nine others with one substitution. The codons can be clustered into four groups differing in one codon position. The codons connected by edges representing mutations in the first codon positions were shown as concentric circles.

The set conductance has an interesting interpretation. Let us assume that S is a codon group encoding a selected amino acid. Then, $\phi(S)$ is the ratio of non-synonymous substitutions to all possible point mutations for the given codon group S. Therefore, this measure gives us an information about the robustness level of this codon group to single mutations that can change an amino acid coded by this group to other amino acids coded by other codon groups $\overline{S}$. For example, the group of six codons encoding arginine and serine have $\phi(S)$ equal to 36/54 and 40/54, respectively. This observation rises immediately a question about the minimum value of the set conductance for a codon group with a given number of codons, *i.e.* its size denoted here as k. It is particularly interesting in the context of the optimal encoding of genetic information by a codon block. In order to study this property, we used other measure called k-size conductance.

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

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

where $|S| = k$ is the number of codons in subset S and $\phi(S)$ is the set conductance.

In other words, it is the minimum possible value of set conductance for a group consisting of k codons. It is helpful in finding the codon groups that minimize consequences of changing genetic information due to point mutations. For example, in the case of codon groups encoding arginine and serine, only the former has the minimal set conductance for their size, *i.e.* $\phi_6(G) = (36/54) = (2/3)$ but the latter is greater. To evaluate the general quality of genetic code, we introduced the third characteristic, called the average conductance of set collection.

Definition 3. *Let $\mathcal{P}$ be a set collection that fulfills the following property:*

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

The average conductance of $\mathcal{P}$ is then defined as:

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

where $|\mathcal{P}|$ is the number of subsets S in a code partition $\mathcal{P}$ and $\phi(S)$ is the set conductance of an individual subset.

This measure is an average robustness of a code $\mathcal{P}$ to consequences of point mutations, which can occur between codon blocks S of this code. $\Phi(\mathcal{P})$ is a generalization of the average code conductance calculated for the SGC (Aloqalaa *et al.* 2019, 2020). In fact, if the number of codon groups $|\mathcal{P}| = 21$ and $S_1 \cup S_2 \cup \ldots \cup S_l = V$ are codon blocks arranged as in the SGC, then $\Phi(\mathcal{P})$ is the average conductance for the SGC.

Finding the optimal codon groups in terms of changing genetic information

The characteristics presented above appeared useful in studying the structural properties of genetic codes. However, they require a fast and effective method for determining the optimal conductance $\phi_k(G)$ for groups with k, *i.e.* a specific number of codons. Fortunately, the graph G including 64 codons possesses many interesting properties, which are helpful in the solution of the optimality problem. First of all, this graph can be represented as a Cartesian graph product, *i.e.* the set of all ordered combinations of three cliques:

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

where K_4 is a 4-vertex clique, *i.e.* the set of four vertices corresponding to nucleotides $\{A, U, G, C\}$, such that every two distinct vertices are adjacent. This property allows us to characterize the set of codons reaching the minimal set conductance from all possible subsets with a given codon number k. The following proposition presented in Aloqalaa *et al.* (2019, 2020) is a natural consequence of the Theorem 1 given by Bezrukov and Elsässer (2003).

Proposition 1.*Let us consider a linear order of the set of vertices of 4-clique K_4, e.g. $A < C < G < U$, and let C_k be a collection of the first k vertices of a graph $G = K_4 \times K_4 \times K_4$ in the lexicographic order. Then we get:*

$$\phi(C_k) \le \phi(S)\,,$$

where $S \subseteq K_4 \times K_4 \times K_4$, $|S| = k$, for any $k \ge 1$. Therefore, the following equations hold for any $k \ge 1$:

$$\phi(C_k) = \phi_k(G)\,.$$

As a result, each sequence of k, $1 < k < 64$ codons of graph G sorted according to a given lexicographic order can reach the minimum of the set conductance over all possible set of codons with the number k. We used the notation C_k in the whole paper to denote a general set of k-codons in the lexicographic order. Briefly, sorting codons according to the lexicographic order enabled us to find the codon groups that minimize changes in coded genetic information, *e.g.* substitutions between codons encoding different information. It should be noted that there are exactly 144 different lexicographic codon orders, which can be used to build a genetic code as a graph G. It results from all possible linear orders of the four nucleotides and all possible orders of three codon positions: $4! * 3! = 144$.

Inclusion of physicochemical properties of amino acids

The model of the genetic code described above assumed equal consequences of substitutions between all amino acids. This general assumption allowed us for analytical finding of the optimal codon groups minimizing the amino acid changes due to point codon mutations. It would not be possible after the inclusion of amino acid properties, although the model would contain an important information. Nevertheless, we also calculated the costs of amino acid substitutions assuming physicochemical properties of these amino acids for the optimal codes found according to the lexicographic approach in the procedure described in the previous section. These calculations were done for the codon groups that encoded canonical amino acids but not for those encoding newly incorporated ncAAs, because we do not know *a priori* their properties.

We included eight amino acid indices describing their physicochemical properties: BLAM930101, BIOV880101, MAXF760101, TSAJ990101, NAKH920108, CEDJ970104, LIFS790101, and MIYS990104. They represent diverse features, such as: electric properties (isoelectric point and polarity), hydrophobicity, alpha-helix and turn propensities, general physicochemical properties, residue propensity (molecular weight, average accessible surface area, and mutability), composition, beta-strand propensity, and intrinsic propensities (hydration potential, refractivity, optical activity, and flexibility). The indices were selected as representatives out of more than 500 amino acid indices present in the AAindex database (Kawashima *et al.* 2008) using a consensus fuzzy clustering method (Saha *et al.* 2012).

Based on these indices, we calculated F function for the canonical part of a genetic code, which is defined as:

$$F = \sum_{n=1}^{8} \sum_{<i,j> \in D} [f_n(i) - f_n(j)]^2\,,$$

where D is the set of all possible pairs of codons i and j from canonical part of genetic code that differ by a single-point mutation, whereas $f_n(i)$ and $f_n(j)$ are values of amino acid index n for the amino acids encoded by these codons, respectively. Simply speaking, this function is the sum of squared differences in eight amino acid properties. In the case of mutations involving stop translation codons, we assumed the maximum possible squared difference over all possible pairs of amino acids for the given amino acid index. The values of corresponding amino acid indices were standardized by dividing by the maximum squared difference of the given index. The final F values were additionally normalized by the total number of codons k belonging to the canonical subset of the considered code.

Data availability

The computations were conducted using Python 3.9.1 programming language. All source codes and raw data relevant to our investigations were included in supplementary material at figshare: https://figshare.com/s/10.25386/genetics.14079452.

Results and Discussion

We began our investigation with finding the smallest set of codons encoding all 20 amino acids and stop coding signal, which still preserves the canonical codon assignments and is simultaneously optimal in terms of changing genetic information between the set of canonical codons and the vacant codons, which can encode ncAAs. We discussed different scenarios of reprogramming the SGC assuming various number of vacant codons. We used the average conductance as a measure of the quality of given genetic code structures, *i.e.* codon blocks. We also discussed structural features of the canonical codes in terms of its robustness against changes in encoded amino acids.

In the construction of the codes, we assumed their robustness to changes causing the loss of genetic information due to mutations. This assumption follows the adaptation hypothesis, which claims that the SGC evolved to minimize harmful consequences 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). Although this code did not turn out perfectly optimized in this respect (Błażej *et al.* 2016; Massey 2008; Novozhilov *et al.* 2007; Santos *et al.* 2011; Santos and Monteagudo 2017; Wnętrzak *et al.* 2018; Błażej *et al.* 2018a, 2019b; Wnętrzak *et al.* 2019), it shows a general tendency to error minimization in the global scale. This property is better exhibited by its alternative versions (Błażej *et al.* 2018b, 2019a), which occurred later in the evolution. Therefore, the analysis of the genetic code expansion in this context seems to be a natural consequence of its evolution.

The smallest set of codons encoding canonical information

It is well known that the SGC is redundant, which means that a smaller number of codons is enough to encode all 20 canonical amino acids and one stop translation signal. Therefore, the set encoding the canonical information can be reduced to a smaller number of codons, allowing for encoding new genetic information by the set of vacant codons. It seems reasonable to postulate some conditions that must be met to obtain minimalistic genetic codes encoding the canonical genetic information, which can be a potential starting point for further analysis of genetic code expansion. We assumed that this codon set must be optimal in terms of the set conductance ϕ, which means that, for a given number k of codons in the set, the number of connections between canonical information and the set of vacant codons is as small as possible. In other words, the number of mutations between the canonical and the non-canonical codes should be minimal. This assumption has a sensible biological meaning because it reduces a possibility of unwanted changes between the new and the old genetic information.

Table 1 The example of the codon set C_k for $k=8$, which is a sequence of the first eight codons taken in a selected lexicographic order

Codon	Amino acid
AAA	Lys
AAC	Asn
AAG	Lys
AAU	Asn
ACA	Thr
ACC	Thr
ACG	Thr
ACU	Thr

According to Proposition 1, this set is characterized by the minimal set conductance over all sets with the size of $k=8$. The codons have assigned encoded amino acids as in the standard genetic code.

Following Proposition 1, we get that the first k-codons ordered in lexicographic order C_k constitute the set with the minimum set conductance ϕ over all possible sets with k codons. In consequence, this codon set is the most resistant against loosing information. An example of such set is shown in Table 1 for eight codons. This property poses a question about the minimum number of codons k such that there exists a set C_k composed of codons that encode 20 amino acids and stop translation signal. In order to deal with this problem, we denote $\mathcal{P}_k$ as a partition of the set C_k of k lexicographically ordered codons that encode 21 canonical items creating a code, *i.e.* $\mathcal{P}_k$ is a set collection of codons encoding canonical information:

$$\mathcal{P}_k = \{S_1, S_2, \ldots, S_{21}, S_i \cap S_j = \varnothing, S_1 \cup S_2 \cup \ldots \cup S_{21} = C_k\},$$

where S_l, $l = 1, 2, \ldots, 21$ is a non-empty set of codons encoding 21 items according to the SGC rules.

We tested all possible sets C_k encoding 21 canonical items induced by all 144 codon orders and SGC assignments. Using this method, we obtained that $k=28$ is the minimal number of codons in the set C_k, which induces partition $\mathcal{P}_{28}$, and encodes 20 amino acids and stop coding signal. In fact, there are two lexicographic orders, which produce such a code. The first is induced by a linear order between nucleotides $U < G < A < C$ and an order between codon positions $1<2<3$. The second is generated by a linear order $G < U < A < C$ between nucleotides and an order between codon positions $1<2<3$.

Tables 2 and 3 include representations of 64 codons in the classical SGC table showing the structure of the optimal C_{28} codon set. The codons C_{28} belonging to the canonical part of this code $\mathcal{P}_{28}$ are marked in red, whereas the vacant codons are in blue. In the first, third and fourth column of the tables, two codons in the block comprising four codons differing in the third codon position encode a classical amino acid or stop signal, whereas in the second column, only one codon in the block encodes an amino acid. Interestingly, only codons ending with G and U are involved in the coding of the canonical information. These two codes differ only in the assignment of four amino acids in the second column of the tables. In the first code, these amino acids are coded by the codons ending with G, whereas in the second code, the codons end with U. This way of codon selection by the algorithm causes that the number of mutations changing the canonical information to the non-canonical one coded by the vacant codons is minimized. At the same time, all 20 canonical amino acids and at least one stop translation signal are included in the code.

Table 2 The smallest set of 28 codons encoding canonical information and minimizing changes of information between the canonical (labeled according to canonical assignments) and non-canonical (unassigned codons) partition of the code

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

These codons were chosen according to a lexicographic order induced by the linear order of nucleotides $U < C < A < G$ and the order of codon positions $1 < 2 < 3$.

Table 3 The smallest set of 28 codons encoding canonical information and minimizing changes of information between the canonical (labeled according to canonical assignments) and non-canonical (unassigned codons) partition of the code

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

These codons were chosen according to a lexicographic order induced by the linear order of nucleotides $G < U < A < C$ and the order of codon positions $1 < 2 < 3$.

Properties of the codes in terms of robustness to point mutations

We compared the quality of the obtained codes in terms of the average conductance Φ, to find out to what extent these codes minimize consequences of point mutations between codon blocks encoding the canonical information $\mathcal{P}_k$. We considered codes with the increasing number of these codons $k = 28, \ldots, 64$ at the expense of the codons for non-canonical information. We also present $\mathcal{P}_k$ which were generated for two lexicographic orders C_{28}, for which we found the smallest set of codons encoding the canonical information.

Figure 2 presents a relationship between $\Phi(\mathcal{P}_k)$ and the codon number k calculated for the two lexicographic orders, for which we found the smallest coding set C_{28} (the blue and orange lines). The lower bound of $\Phi(\mathcal{P}_k)$ calculated over all possible 144 orders is also shown for comparison (the green line). This line corresponds to the codes whose structure allows for the best possible minimization of substitutions between the coded canonical items. As we can observe, in all considered cases the average conductance Φ decreases with the number of codons k involved in the code. This trend is related with an increasing redundancy of the code for the same number of coded items. The maximum is reached at $k = 28$ and is equal $\Phi(\mathcal{P}_{28}) = 0.91$, whereas the minimum equals to $\Phi(\mathcal{P}_{64}) = 0.81$ for all set collections at $k = 64$. It should be noted that $\Phi(\mathcal{P}_{64})$ corresponds to the average code code conductance calculated for the SGC and was discussed in Aloqalaa *et al.* (2019, 2020). However, the presented relationships are not strictly linear, because local changes in the course of this trend occur. What is more, the two lexicographic orders that generate the smallest codon sets C_{28}, generally do not induce the optimal collections of sets $\mathcal{P}_k$ for $k > 28$ in terms of Φ. In other words, it is not possible to generate a set collection $\mathcal{P}_k$ for each $k = 29, \ldots, 64$ using lexicographic orders shown in Tables 2 and 3 that would be minimal in terms of Φ.

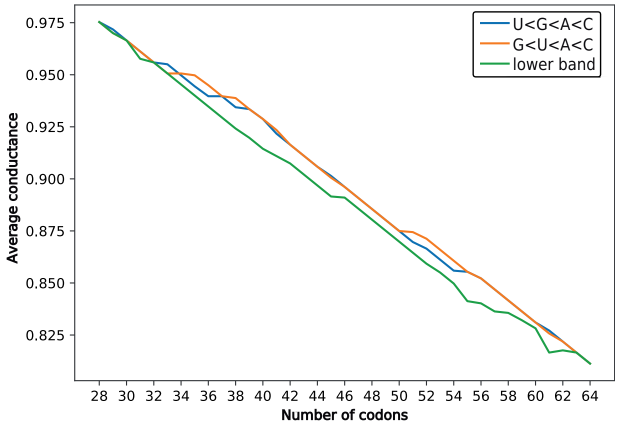


Figure 2 The relationship between the average conductance $\Phi(\mathcal{P}_k)$ and the number of codons in the code $k = 28, \ldots, 64$ calculated for two lexicographic orders, for which we found the smallest set coding canonical information C_{28} (blue and orange lines). The lower bound calculated over 144 orders is shown for comparison (green line).

We also analyzed these codes in terms of consequences of amino acid substitutions considering their physicochemical properties. Figure 3 presents the relationship between the smallest possible average costs of amino acid replacements and the number of codons encoding these amino acids in the codes that minimize changes between the canonical and non-canonical information. The costs were normalized by the number of codons for the canonical information in the corresponding code. Interestingly, the maximum of this normalized cost is taken by the code with all 64 codons, *i.e.* the SGC and the minimum is for the code including 55 codons for the canonical information (Table 4). This code has nine codons released, which can be used to code ncAAs. Interestingly, these codons have U in the first codon position and among them are two encoded stop translation signal in the SGC. Moreover, these codons encode two amino acids, which are very rarely used, *i.e.* cysteine and tyrosine, as well as those that are abundant in proteins and can be coded by six codons, *i.e.* leucine and serine. Therefore, reprogramming of these codons seems sensible.

Table 5 presents how many times the individual codons were selected as vacant in 37 codes that minimized changes of information between the canonical and non-canonical partition and showed the smallest possible average costs of amino acid replacements considering their physicochemical properties. Interestingly, codons encoding stop translation signal in the SGC were most often released. However, among them there is not

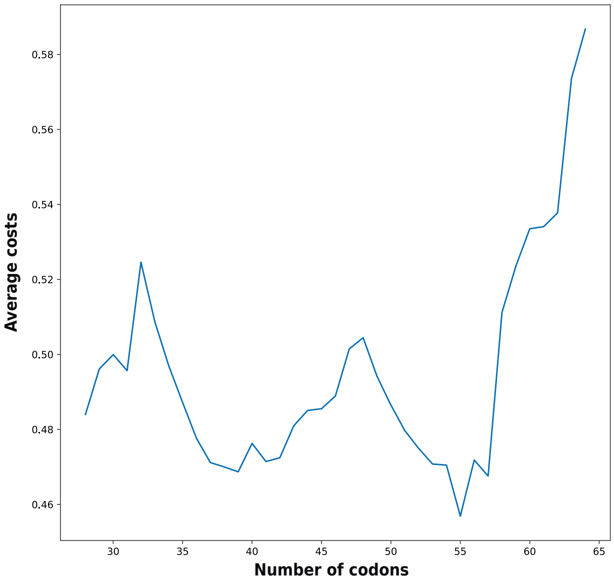


Figure 3 The relationship between the smallest possible average costs of amino acid substitutions regarding their physicochemical properties based on *F* function and the number of codons in the canonical codes that minimize changes between the canonical and non-canonical information. The costs were normalized by the number of codons for the canonical information in the corresponding code.

Table 4 The set of 55 codons (labeled according to canonical assignments) encoding canonical information and minimizing changes of information between the canonical and non-canonical (unassigned codons) partition of the code

UUU	UCU	UAU	UGU
UUC Phe	UCC Ser	UAC Tyr	UGC Cys
UUA	UCA	UAA	UGA
UUG	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 Gly
GUC Val	GCC Ala	GAC Asp	GGC Gly
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly

These codons were chosen according to a lexicographic order induced by the linear order of nucleotides $C < G < A < U$ and the order of codon positions $1 < 3 < 2$. This code shows also the smallest possible average costs of amino acid replacements considering their physicochemical properties normalized by the codon number.

UAG codon, which is often used in experimental approaches due to its low usage in protein coding sequences (Noren *et al.* 1989; Chin 2017; Italia *et al.* 2017). Our algorithm preferred other stop codons because it applies different criteria, *i.e.* the minimization of changing the canonical and non-canonical information. Next frequently used codons in the non-canonical partition are those with A in the second and third codon position encoding lysine, glutamic acid and glutamate. On the other hand, codons with G in the third codon position were released very rarely or not at all. Two of these codons are the only ones that encode methionine and tryptophan.

Table 5 The number of times when the individual codons were selected as vacant in the codes minimizing changes of information between the canonical and non-canonical partition and showing the smallest possible average costs of amino acid replacements considering their physicochemical properties

UUU Phe 8	UCU Ser 8	UAU Tyr 9	UGU Cys 7
UUC Phe 7	UCC Ser 11	UAC Tyr 13	UGC Cys 8
UUA Leu 27	UCA Ser 25	UAA Stop 36	UGA Stop 31
UUG Leu 1	UCG Ser 1	UAG Stop 0	UGG Trp 0
CUU Leu 5	CCU Pro 5	CAU His 5	CGU Arg 4
CUC Leu 7	CCC Pro 13	CAC His 13	CGC Arg 10
CUA Leu 22	CCA Pro 23	CAA Gln 29	CGA Arg 25
CUG Leu 0	CCG Pro 3	CAG Gln 0	CGG Arg 0
AUU Ile 5	ACU Thr 5	AAU Asn 7	AGU Ser 4
AUC Ile 8	ACC Thr 12	AAC Asn 14	AGC Ser 10
AUA Ile 24	ACA Thr 24	AAA Lys 31	AGA Arg 27
AUG Met 0	ACG Thr 2	AAG Lys 0	AGG Arg 0
GUU Val 5	GCU Ala 4	GAU Asp 5	GGU Gly 4
GUC Val 7	GCC Ala 10	GAC Asp 13	GGC Gly 9
GUA Val 22	GCA Ala 22	GAA Glu 30	GGA Gly 22
GUG Val 0	GCG Ala 1	GAG Glu 0	GGG Gly 0

The results were obtained from the set of 37 codes in which 28–63 codons encoded the canonical information.

Properties of expanded genetic code

The reducing number of codons encoded canonical amino acids and stop translation signal, as shown in The smallest set of codons encoding canonical information section, implies that the rest codons can be used to encode ncAAs. Mathematically speaking, the codon set C_k encoding the canonical information by the $k \geq 28$ codons, induces its own complement, *i.e.* the set C'_k of vacant codons for ncAAs. Thus, the new genetic information would be encoded by $1 \leq n \leq (64 - k)$ codon blocks, which constitute a partition of the set C'_k. In consequence, we introduced a set collection of n codon blocks for the new genetic information:

$$\mathcal{P}'_k(n) = \{S_1, S_2, \ldots, S_n,\ S_i \cap S_j = \emptyset,\ S_1 \cup S_2 \cup \ldots \cup S_n = C'_k\}, \quad (1)$$

where each $S_i, i = 1, \ldots, n$ is a non-empty set of codons that encodes the same genetic information, *e.g.* a specific ncAA. The new set $\mathcal{P}'_k(n)$ together with the set $\mathcal{P}_k$ encoding canonical information constitutes an expanded genetic code denoted by $\mathcal{P}(n, k)$, which encodes exactly n new ncAAs and $k \geq 28$ items of canonical genetic information:

$$\mathcal{P}(n, k) = \mathcal{P}_k \cup \mathcal{P}'_k(n). \quad (2)$$

Please note that according to the definition of $\mathcal{P}_k$, we get that the number of connections between the canonical and non-canonical parts of the expanded code is as small as possible, which may causes a low probability of potential reversion between the new and old information. It is very useful from experimental point of view, when we want to keep the information about the canonical amino acids and the stop translation, and simultaneously not lose the new information encoded in the vacant codons.

Similar to the average conductance of the canonical code $\Phi(\mathcal{P}_k)$, it is theoretically possible to calculate this measure also for the codon set encoding ncAAs, denoted as $\Phi(\mathcal{P}'_k(n))$. Finally, the average conductance of the whole expanded genetic code $\Phi(\mathcal{P}(n, k))$ can be derived to assess its optimality in terms of point mutations. However, these measures can be obtained only when the assignments of individual ncAAs to the vacant codons are

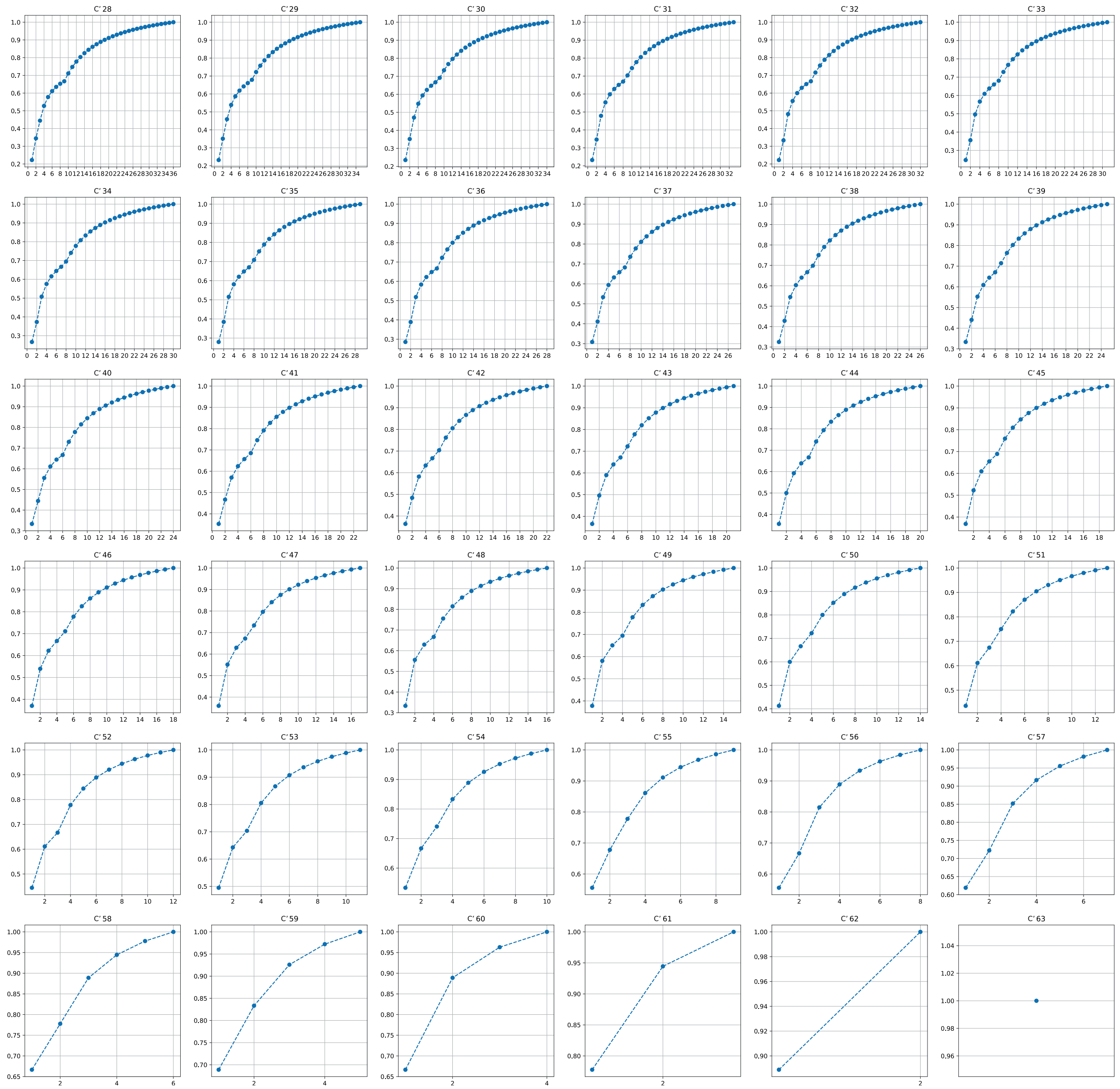


Figure 4 The lower bound of the average conductance $\Phi(\mathcal{P}'_k(n))$ calculated for the set collection $\mathcal{P}'_k(n)$ encoding ncAAs in relation to the number n of coded ncAAs. The relationship was presented for all possible partitions of the set C'_k containing vacant codons, which encode $n = 1, \ldots, (64-k)$ ncAAs, for $k \geq 28$ being the number of codons in the canonical set.

known, because there are many possible set collections for the fixed number of codons in the canonical set k and the number of ncAAs coded in the non-canonical set n, which differ in Φ values. Therefore, we decided to find a lower bound on the values of $\Phi(\mathcal{P}'_k(n))$ for the fixed $k \geq 28$ and $n = 1, \ldots, (64-k)$. It could be done using the representation 1 of $\mathcal{P}'_k(n)$ as well as the definition 2 of the k-size conductance and a simple observation that for every $\mathcal{P}'_k(n)$, we have:

$$\Phi(\mathcal{P}'_k(n)) = \frac{1}{n}\sum_{i=1}^{n} \phi(S_i) \geq \frac{1}{n}\sum_{i=1}^{n} \phi_{|S_i|}(G)\,. \tag{3}$$

It means that the average conductance of the non-canonical part of the code is greater or at most equal to the average of respective $|S_i|$-size conductances of the codon blocks encoding n ncAAs and having the optimal structure in terms of the set conductance.

Therefore, for every $\mathcal{P}'_k(n)$, there exists a lower bound on the average conductance of the set collection imposed on its codon set C'_k. What is more, these optimal collections are composed of the best codon blocks in terms of the k-size conductance, *i.e.* the minimum possible value of set conductance for a group consisting of k codons. This feature gives us a general overview on the optimal structures of the genetic code expansions including the selected number n of ncAAs.

Following the property 3, we found all possible lower bounds for every $k \geq 28$ and $n = 1, 2, \ldots, (64-k)$. Figure 4 presents their graphical representations. As we can see, the lower bound on $\Phi(\mathcal{P}'_k)$ increases with the number n of coded ncAAs.

This relationship shows an interesting course, *e.g.* for $k = 28$ (Figure 3), the curve of the lower bound increases with n but slows down for n close to $n' = 9$ and then blows up again for $n > n'$. This fact results from that there is no set with the number of codons lower than four for $n \leq n'$, whereas these sets appear

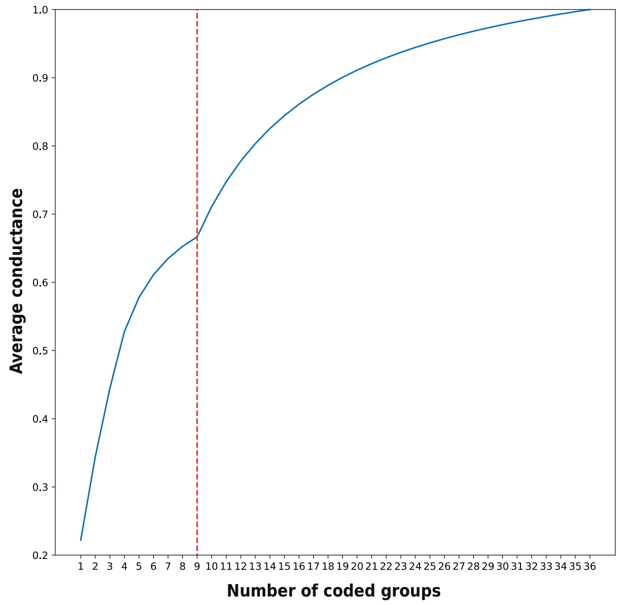


Figure 5 The minimum of the average set conductance $\Phi(C'_{28})$ (blue line) in relation to the number n of coded ncAAs. The minimum of Φ was found over all possible partitions of the set C'_{28} containing 28 codons for canonical information and 36 vacant codons for n ncAAs. The red dashed line shows the minimum of the average set conductance obtained for $n=9$. As we can see, $n=9$ is a deflection point, in which the rate of the curve increase is changing.

for $n > n'$. Since the k-size conductance $\phi_k(G)$ for groups of codons in the number $k=1,\ 2,\ 3$ is ≥ 0.778 and larger than for more numerous groups with $\phi_k \leq 0.689$, a set collection containing the codon group of size lower than four have the average conductance of the set collection generally higher in comparison to the collections that are composed of codon blocks with the size greater or equal than four. This fact could explain the presence of the Φ minimum for $n > n'$. This phenomenon is also observed for $28 \leq k \leq 52$ for respective changing points n' (Figures 4 and 5).

The balance of expanded genetic code

Using equation (2), we can compare the structural differences between the canonical $\mathcal{P}_k$ and the non-canonical $\mathcal{P}'_k(n)$ parts of the expanded code. In order to do so, we introduced a balance measure Ψ defined as:

$$\Psi(\mathcal{P}(n,k)) = \frac{\Phi(\mathcal{P}'_k(n))}{\Phi(\mathcal{P}_k)}\,,\ 28 \leq k \leq 63\,, \tag{4}$$

where $\Phi(\mathcal{P}_k)$ is the average conductance of the canonical code partition and $\Phi(\mathcal{P}'_k(n))$ is the average conductance of the non-canonical code partition.

The balance function $\Psi < 1$ indicates that the non-canonical partition $\mathcal{P}'_k(n)$ possesses better structural properties in terms of the average conductance, *i.e.* minimization of non-synonymous substitutions than the canonical partition $\mathcal{P}_k$, whereas $\Psi > 1$ means that the canonical genetic information is better optimized in this respect. From our point of view, the value of Ψ around one is the most interesting because it suggests a similar robustness of codon blocks to point mutations in both types of the expanded genetic code. Therefore, the balance measure Ψ appears to be useful in studying properties of codon groups belonging to $\mathcal{P}_k$ and $\mathcal{P}'_k(n)$. Thanks to that, we can compare the quality of coding system for the new and old information.

We tested the balance under the assumption that the non-canonical set $\mathcal{P}'_k(n)$ attains lower bound of the average conductance value $\Phi(\mathcal{P}'_k(n))$. Figure 6 presents the balance values Ψ calculated for various number of codons in the non-canonical set C'_k in relationship with the number n of coded ncAAs. It is visible that the expanded genetic code is extremely unbalanced for small n, *i.e.* when $\Psi < 1$, which indicates that the non-canonical partition $\mathcal{P}'_k(n)$ have in general codon block structure that minimize non-synonymous substitutions better than the canonical partition $\mathcal{P}_k$. In all considered cases Ψ increases with the number of newly incorporated ncAAs. However, it is possible to find a balanced code for which Ψ are around one.

After comparison of codes with the balance in the range of 0.991–1.010, we noticed a biased distribution of codons in the canonical and non-canonical partitions. Codons with the G in the third codon position dominated in the canonical partition. Each of them was present in more than 2.3% among codons belonging to this set. In total, these codons constituted 40%. In the case of uniform distribution of all codons, we should expect the 1.56% usage for each codon and 25% for the group having one nucleotide type in one codon position. On the other hand, in the non-canonical partition, codons ending with A were most frequent, each in more than 3.2% and 60% in total. Interestingly, such codons were also very often selected to the non-canonical partition in the expanded codes showing the smallest possible average costs of amino acid replacements in terms of their physicochemical properties (Table 5). The biased usage of the codon groups in the canonical and non-canonical sets is significantly different in comparison to the uniform distribution (P-value $< 10^{-10}$ in the proportion test).

What is more, the number of newly included ncAAs required to obtain the balanced code is in some cases quite large. For example, in the case of $k=28$, possible balanced genetic codes are obtained for the number of ncAAs $n=28$, 29, and 30. This result shows in fact a huge redundancy level of the SGC.

Concluding remarks

The redundancy of the SGC suggests that this coding system can be expanded. In literature, we can find several approaches to this problem. These findings encouraged us to start studying the issue of the optimal expansion of the SGC from theoretical perspective. In this paper, we proposed a method of genetic code expansion using graph theory. Following this methodology, we described the smallest set of codons still encoding 21 canonical items (20 amino acids with one stop translation signal) and characterizing by the minimal set conductance for its size. This property provides the smallest number of connections between codons in this minimalistic canonical code and the set of vacant codons, which can be assigned to new genetic information. Thanks to that, such a code is characterized by the minimized possibility of reversions between these two parts of the expanded code, the canonical and non-canonical one. What is more, we investigated the optimal structure of many expanded codes with various number of codons released for encoding potential ncAAs. Among these codes, we found those that minimized average costs of amino acid replacements considering their physicochemical properties. In addition, the introduced balance measure, *i.e.* the ratio of the average conductance of the non-canonical to canonical code, allows us for finding the expanded genetic codes whose canonical and non-canonical sets show a similar robustness to point

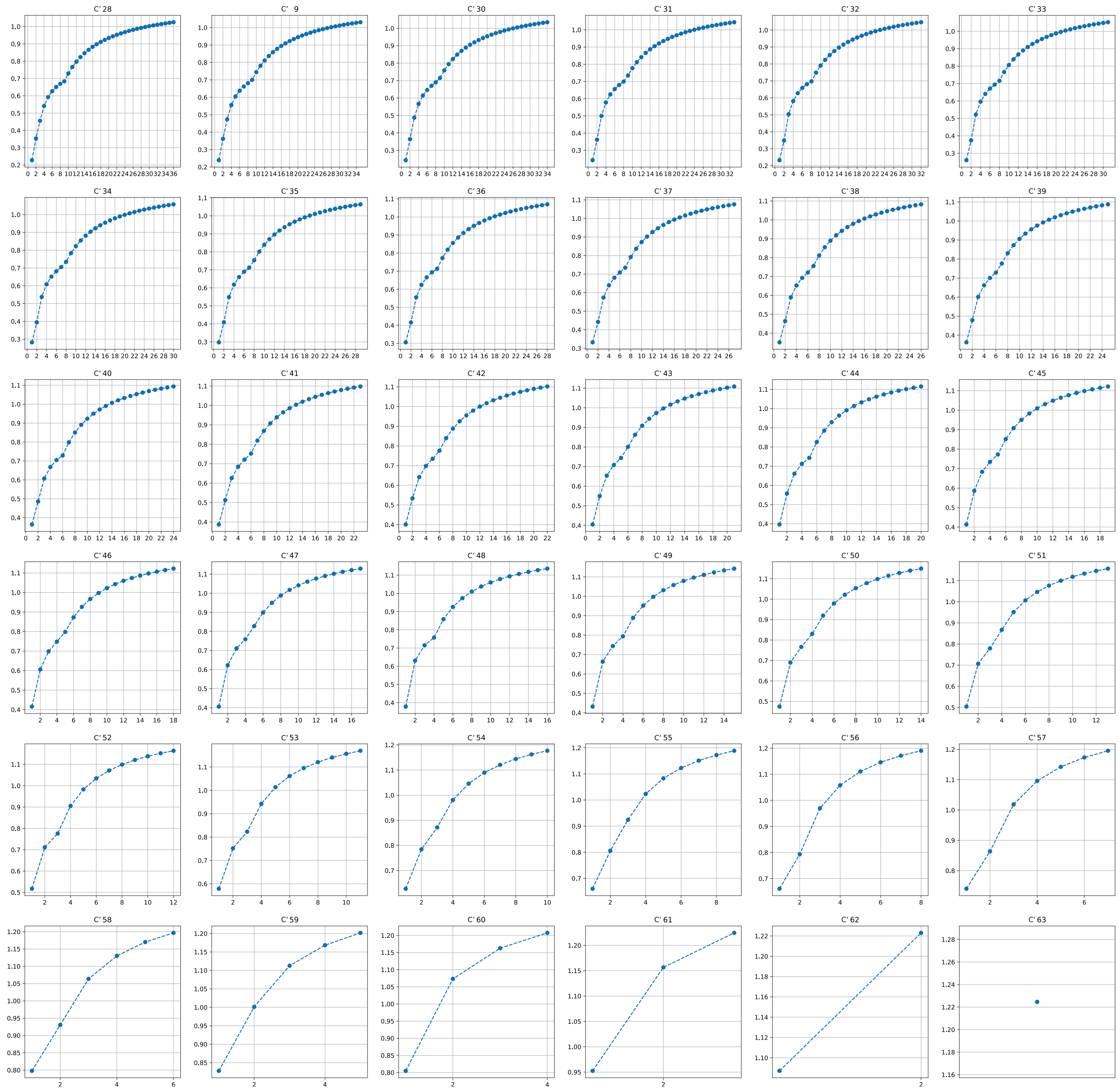


Figure 6 The balance Ψ, *i.e.* the ratio of the average conductance of the non-canonical to canonical code, in relation to the number n of coded ncAAs. The relationship was presented for all possible partitions of the set C'_k containing vacant codons, which encode $n = 1, \ldots, (64 - k)$ ncAAs, for $k \geq 28$ being the number of codons in the canonical set. It was assumed that $\mathcal{P}'_k(n)$ attains the lower bound of $\Phi(\mathcal{P}'_k(n))$.

mutations. Using these approaches, we identified the codons that can be used for reprogramming to encode new ncAAs.

It should be noted that the results presented here are based on some theoretical assumptions, which were necessary to conduct the analytical calculations and reasoning as well as make general conclusions about the expansion of the SGC. First of all, we proposed an universal approach, which does not take into account the different probabilities of nucleotide mutations and the codon usage. These features are much diversified and specific not only between various species but even within the same genome. Therefore, it is not possible to construct a general model of the genetic code expansion including the huge diversity of the mutations and codon frequency. Secondly, we did not regard the number and types of tRNAs, which can be used to decode unambiguously respective codons. Nevertheless, it seems reasonable to investigate the problem of the SGC expansions starting from the general foundations. Interestingly, using these assumptions, we found several interesting limitations on the number of codons required to encode canonical information and also on the codon blocks that would encode new information. Our approach can be considered a null model and a starting point to other more complex models, most probably heuristic and genome-specific, including the different mutation rate between nucleotides and codon usage.

Funding

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

Conflicts of interest

None declared.

Literature cited

Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics. 164:1291–1303.

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

Aloqalaa DA, Kowalski DR, Błażej P, Wnetrzak M, Mackiewicz D. 2020. The properties of the standard genetic code and its selected alternatives in terms of the optimal graph partition. In: Roque A, Tomczyk A, De Maria E, Putze F, Moucek R, et al. editors. Communications in Computer and Information Science, Vol. 1211. p. 170–191. Springer.

Anderson JC, Wu, N Santoro, SW Lakshman, V King, DS, *et al.* 2004. An expanded genetic code with a functional quadruplet codon. Proc Natl Acad Sci U S A. 101:7566–7571.

Bezrukov SL, Elsässer R. 2003. Edge-isoperimetric problems for Cartesian powers of regular graphs. Theor Comput Sci. 307: 473–492.

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

Błażej P, Wnętrzak M, Mackiewicz D, Mackiewicz P. 2018a. Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. PLoS One. 13: e0201715.

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

Błażej P, Wnetrzak, M Mackiewicz, D Mackiewicz P. 2020. Basic principles of the genetic code extension. R Soc Open Sci. 7:191384.

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

Błażej P, Wnetrzak M. and Mackiewicz P. 2018b. The importance of changes observed in the alternative genetic codes. Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 3:BIOINFORMATICS. p. 154–159.

Chin JW. 2014. Expanding and reprogramming the genetic code of cells and animals. Annu Rev Biochem. 83:379–408.

Chin JW. 2017. Expanding and reprogramming the genetic code. Nature. 550:53–60.

Dien VT, Morris, SE Karadeema, RJ Romesberg FE. 2018. Expansion of the genetic code via expansion of the genetic alphabet. Curr Opin Chem Biol. 46:196–202.

D'Onofrio DJ, Abel DL. 2014. Redundancy of the genetic code enables translational pausing. Front Genet. 5:140.

Epstein CJ. 1966. Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. Nature. 210:25–28.

Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. J Mol Evol. 47:238–248.

Freeland SJ, Knight, RD Landweber LF. 2000. Measuring adaptation within the genetic code. Trends Biochem Sci. 25:44–45.

Gilis D, Massar, S Cerf, NJ Rooman M. 2001. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biol. 2:research0049.1–0049–12.

Goldberg AL, Wittes RE. 1966. Genetic code: aspects of organization. Science. 153:420–424.

Haig D, Hurst LD. 1991. A quantitative measure of error minimization in the genetic code. J Mol Evol. 33:412–417.

Hamashima K, Kimoto, M Hirao I. 2018. Creation of unnatural base pairs for genetic alphabet expansion toward synthetic xenobiology. Curr Opin Chem Biol. 46:108–114.

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

Ishikawa M, Hirao, I Yokoyama S. 2000. Synthesis of 3-(2-deoxy-beta-d-ribofuranosyl)pyridin-2-one and 2-amino-6-(n,n-dimethylamino)-9-(2-deoxy-beta-d-ribofuranosyl)purine derivatives for an unnatural base pair. Tetrahedron Lett. 41:3931–3934.

Italia JS, Addy PS, Wrobel CJJ, Crawford LA, Lajoie MJ, *et al.*, 2017. An orthogonalized platform for genetic code expansion in both bacteria and eukaryotes. Nat Chem Biol. 13:446–450.

Iwane Y, Hitomi A, Murakami H, Katoh T, Goto Y, *et al.*, 2016. Expanding the amino acid repertoire of ribosomal polypeptide synthesis via the artificial division of codon boxes. Nat Chem. 8: 317–325.

Kanaya S, Yamada, Y Kudo, Y Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene. 238:143–155.

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, *et al.*, 2008. Aaindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36:D202–D205.

Kimoto M, Kawai, R Mitsui, T Yokoyama, S Hirao I. 2009. An unnatural base pair system for efficient PCR amplification and functionalization of DNA molecules. Nucleic Acids Res. 37:e14–e14.

Malyshev DA, Seo, YJ Ordoukhanian, P Romesberg FE. 2009. PCR with an expanded genetic alphabet. J Am Chem Soc. 131:14620–14621.

Massey SE. 2008. A neutral origin for error minimization in the genetic code. J Mol Evol. 67:510–516.

Neumann H, Wang, K Davis, L Garcia-Alai, M Chin JW. 2010. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. Nature. 464:441–444.

Noren CJ, Anthony-Cahill, SJ Griffith, MC Schultz PG. 1989. A general method for site-specific incorporation of unnatural amino acids into proteins. Science. 244:182–188.

Novozhilov AS, Wolf, YI Koonin EV. 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.

Ohtsuki T, Kimoto, M Ishikawa, M Mitsui, T Hirao, I, *et al.* 2001. Unnatural base pairs for specific transcription. Proc Natl Acad Sci U S A. 98:4922–4925.

Orešič M, Shalloway D. 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol. 281:31–48.

Ozer E, Chemla, Y Schlesinger, O Aviram, HY Riven, I, *et al.* 2017. In vitro suppression of two different stop codons. Biotechnol Bioeng. 114:1065–1073.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 12:32–42.

Rocha EP. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 14:2279–2286.

Saha I, Maulik, U Bandyopadhyay, S Plewczynski D. 2012. Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino Acids. 43:583–594.

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

Santos MAS, Gomes, AC Santos, MC Carreto, LC Moura GR. 2011. The genetic code of the fungal CTG clade. C R Biol. 334:607–611.

Sonneborn T. 1965. Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications. New York, NY: Academic Press. p. 377–397.

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

Wnętrzak M, Błażej P, Mackiewicz P. 2019. Optimization of the standard genetic code in terms of two mutation types: point mutations and frameshifts. BioSystems. 181: 44–50.

Woese CR. 1965. On the evolution of the genetic code. Proc Natl Acad Sci U S A. 54:1546–1552.

Xia X. 1998. How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae?* Genetics. 149:37–44.

Yang Z, Sismour, AM Sheng, P Puskar, NL Benner SA. 2007. Enzymatic incorporation of a third nucleobase pair. Nucleic Acids Res. 35:4238–4249.

Young DD, Schultz PG. 2018. Playing with the molecules of life. ACS Chem Biol. 13:854–870.

Zhou T, Weems, M Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol. 26:1571–1580.

Communicating editor: J. Dai