Bulletin of Mathematical Biology (2019) 81:2239–2257
https://doi.org/10.1007/s11538-019-00603-2

Society for Mathematical Biology

The Quality of Genetic Code Models in Terms of Their Robustness Against Point Mutations

P. Błażej[2] · E. Fimmel[1] · M. Gumbel[1]

Received: 16 November 2018 / Accepted: 25 March 2019 / Published online: 5 April 2019

Abstract

In this paper, we investigate the quality of selected models of theoretical genetic codes in terms of their robustness against point mutations. To deal with this problem, we used a graph representation including all possible single nucleotide point mutations occurring in codons, which are building blocks of every protein-coding sequence. Following graph theory, the quality of a given code model is measured using the set conductance property which has a useful biological interpretation. Taking this approach, we found the most robust genetic code structures for a given number of coding blocks. In addition, we tested several properties of genetic code models generated by the binary dichotomic algorithms (BDA) and compared them with randomly generated genetic code models. The results indicate that BDA-generated models possess better properties in terms of the conductance measure than the majority of randomly generated genetic code models and, even more, that BDA-models can achieve the best possible conductance values. Therefore, BDA-generated models are very robust towards changes in encoded information generated by single nucleotide substitutions.

Keywords Genetic code · Dichotomy classes · Point mutations

✉ E. Fimmel
e.fimmel@hs-mannheim.de

P. Błażej
pawel.blazej@uwr.edu.pl

M. Gumbel
m.gumbel@hs-mannheim.de

[1] Competence Center for Algorithmic and Mathematical Methods in Biology, Biotechnology and Medicine, Mannheim University of Applied Sciences, 68163 Mannheim, Germany

[2] Department of Genomics, University of Wroclaw, Wrocław, Poland

1 Introduction

The standard genetic code is a template according to which 64 codons are mapped into 20 amino acids and a stop coding signal. This set of rules, with rare exceptions, is nearly universal for all domains of life and is responsible for transmitting genetic information stored in the DNA molecule into proteins. The questions about the origin and also the present structure of the standard genetic code have been puzzling biologists since the first codon assignments were discovered in the sixties of the last century (Khorana et al. 1966; Nirenberg et al. 1966). In particular, the question about the way of the standard genetic code's degeneracy appears to be intriguing because if we assume that potential theoretical genetic code must encode 20 amino acids and stop coding signals then we get around 10^{84} possible variations (Schönauer and Clote 1997). Therefore, the standard genetic code is just one potential solution out of extremely many different scenarios. This fact gives a motivation for studying which features are playing a decisive role in the process of genetic code emerging.

There are three main hypotheses concerning origin of the standard genetic code (Di Giulio 2005). These are stereochemical, coevolution, and adaptive. The first claims that the genetic code evolves as a result of a high affinity between amino acids and respective codons/anti-codons or other aptamers and oligomers (Dunnill 1966; Pelc and Welton 1966; Yarus et al. 2005). However, some evidences to support this evolutionary scenario were found only in very few cases. The coevolution hypothesis posits that the present structure of the standard genetic code evolved from its ancestral version including small number of simple amino acids. This code has been changed simultaneously with the development of metabolic pathways, i.e. the newly synthesized amino acids took over the codons of their precursors (Wong 1975; Di Giulio 2017). This process required the simultaneous evolution of the genetic code and biosynthetic pathways. In the framework of this hypothesis, the physicochemical properties of amino acid played only a subsidiary role in the standard genetic code evolution. The adaptive hypothesis postulates that the code was created to minimize the effect of amino acid replacements and errors which occur during the translation process (Epstein 1966; Freeland and Hurst 1998a, b). The main argument for this explanations follows from the present structure of the standard genetic code where there is observed a tendency to group similar amino acids in the same column of the code table. However, using several optimization methods it was shown that the standard genetic code can be significantly improved under some criteria to minimize genetic errors and is rather a suboptimal solution in the vast space of possible genetic codes (Di Giulio 1989; Santos and Monteagudo 2010; Blazej et al. 2016, 2018a). It should be noted that proposed explanations are still not satisfactory, and none of these hypotheses became a comprehensive explanatory theory. On the other hand, they are not to be mutually exclusive because the main drivers of the standard genetic code evolution postulated by these hypotheses could have a significant impact on the evolution at different stages.

It is interesting that many questions concerning the properties of the standard genetic code can be formulated as an interesting problem from mathematical and also computational point of view. Many authors used some optimization methods, such as

single or multi-objective evolutionary algorithms, to test the quality of the standard genetic code under selected criteria. Moreover, they used many different and—at the same time—interesting mathematical approaches to describe properties or more generally to develop some rules for generating theoretical genetic codes. These techniques mainly follow from graph theory, coding theory, and group theory (Fimmel et al. 2016, 2017, 2018, 2014; José et al. 2017; Tlusty 2010).

A broader class of models of the genetic code, the so-called BDA-generated models (binary dichotomic algorithms) (Gumbel et al. 2015), is based on overlappings of the so-called dichotomic partitions of the set of codons. The most known dichotomic partition is the Rumer's degeneracy dichotomy (Fimmel and Strüngmann 2016; Rumer 2016a, b, c) which decomposes codons into two disjoint equal-sized classes: the first Rumer class identifies amino acids with degeneracy 4 (for which the first two bases of the triplet are sufficient to define unambiguously the amino acid), while the second one specifies amino acids with degeneracy non-4 (i.e. 1, 2 or 3). A generalization of Rumer's algorithm (Fimmel et al. 2013) has led to a family of binary dichotomic algorithms (BDAs) which derive their decision for classifying a codon from biochemical properties of the bases involved. These algorithms distinguish whether a base is of type

- purine (denoted as R = {A, G}) or pyrimidine (Y = {C, T})
- keto (K = {T, G} or amino (Am = (C, A)
- strong (S = {C, G}) or weak (W = {A, T}).

and classify the codons into two disjoint classes of equal size. Dichotomic partitions seem to contribute to frame retrieval and error detection properties, sustaining a robustness of the code (Giannerini et al. 2012).

The other approach involved to investigate the properties of the standard genetic code follows on graph theory. Many authors (Tlusty 2010) used this methodology to develop several genetic code representations. They applied this approach to their studies about the structure and evolution of the standard genetic code. Among them, the set conductance approach presented in Blazej et al. (2018b) appears to be especially interesting in testing the quality of codon blocks structure generated by BDA-algorithms. Moreover, this methodology has an interesting biological interpretation as the level of robustness of the genetic code structure against single point mutation.

In the present work, we are developing the conductance approach for studying the genetic code further, applying it for a measurement of the quality of BDA-generated models and comparing them with randomly generated models. This enables us to construct variants of the genetic code with a good set conductance.

2 Methods

2.1 BDA-Models

In the sequel $\mathcal{B} = \{A, C, G, T(U)\}$ will denote the set of four nucleotide bases Uracil (Thymine), Cytosine, Adenine, and Guanine, in short $T(U)$, C, A, G. A codon is an element of $\mathcal{B}^3$, e.g. ACU.

The alphabet $\mathcal{B}$ can be decomposed in three different ways into two disjoint equal-sized subsets. Each of these decompositions has a biochemical meaning:

$$\begin{aligned}\mathcal{B} &= \{C, G\} \cup \{A, T\} \quad \text{(strong/weak)},\\ \mathcal{B} &= \{C, A\} \cup \{G, T\} \quad \text{(amino/keto)},\\ \mathcal{B} &= \{C, T\} \cup \{A, G\} \quad \text{(pyrimidine/purine)}.\end{aligned}$$

Based on these biochemical properties of nucleotides, we can classify the set of codons into two disjoint equal-sized subsets, establishing a dichotomic partition of $\mathcal{B}^3$. Let us first give a precise definition of how we understand a dichotomic partition:

Definition 2.1 An ordered pair (H_0, H_1) of subsets $H_0, H_1 \subseteq \mathcal{B}^3$ is called a **dichotomic partition** of $\mathcal{B}^3$ if $H_0 \cap H_1 = \emptyset$, $H_0 \cup H_1 = \mathcal{B}^3$ and $\mid H_0 \mid = \mid H_1 \mid$.

In other words: the set of 64 codons is divided into two disjoint subsets of equal size as, for instance, the so-called Rumer partition does, which separates the codons, where two first bases are enough to determine the encoded amino acid, from the codons, where the third base for the decision is needed.

In Fimmel et al. (2013), the notion of binary dichotomic algorithms was introduced for sequences of nucleotide bases of arbitrary length, i.e. for classification of n-nucleotides $c \in \mathcal{B}^n$, $n \in \mathbb{N}$. For the purposes of the present work, it is sufficient to consider only the set of codons, i.e. $c \in \mathcal{B}^3$. Let us recall the definition from Fimmel et al. (2013) in this special case:

Definition 2.2 Let (H_0, H_1) be a dichotomic partition of $\mathcal{B}^3$. We call an algorithm $\mathcal{A}$ a **binary dichotomic algorithm (BDA) with dichotomic partition** (H_0, H_1) if it follows the following scheme: $\mathcal{A}$ chooses two indices $i_1, i_2 \in \{1, 2, 3\}$ with $i_1 \neq i_2$, an ordered pair of different nucleotide bases $Q_1 = (B_1, B_2)$ and a subset $Q_2 \subset \mathcal{B}$ with $|Q_2| = 2$. Now $\mathcal{A}$ classifies $c = (b_1, b_2, b_3) \in \mathcal{B}^3$ as follows:

(A) if $b_{i_1} \in \{B_1, B_2\}$, then

$$(c \in H_0 \quad \text{if} \quad b_{i_1} = B_1,) \quad \text{and} \quad (c \in H_1 \quad \text{if} \quad b_{i_1} = B_2,)$$

(B) if $b_{i_1} \notin \{B_1, B_2\}$, then

$$(c \in H_0 \quad \text{if} \quad b_{i_2} \in Q_2,) \quad \text{and} \quad (c \in H_1 \quad \text{if} \quad b_{i_2} \notin Q_2.)$$

We will call Q_1 and Q_2 the **questions** of $\mathcal{A}$, $i_1, i_2 \in \{1, 2, 3\}$ the **indices** of $\mathcal{A}$, and the pair (H_0, H_1) a **dichotomic partition of $\mathcal{B}^3$ generated by the binary dichotomic algorithm** $\mathcal{A}$.

Remark 2.3 We will call in short a binary dichotomic algorithm BDA and will write for all $c \in H_0$ $\mathcal{A}(c) = 0$ and for all $c \in H_1$ $\mathcal{A}(c) = 1$.

Figure 1 depicts an example of how a BDA works in order to get Rumer's degeneracy dichotomy:

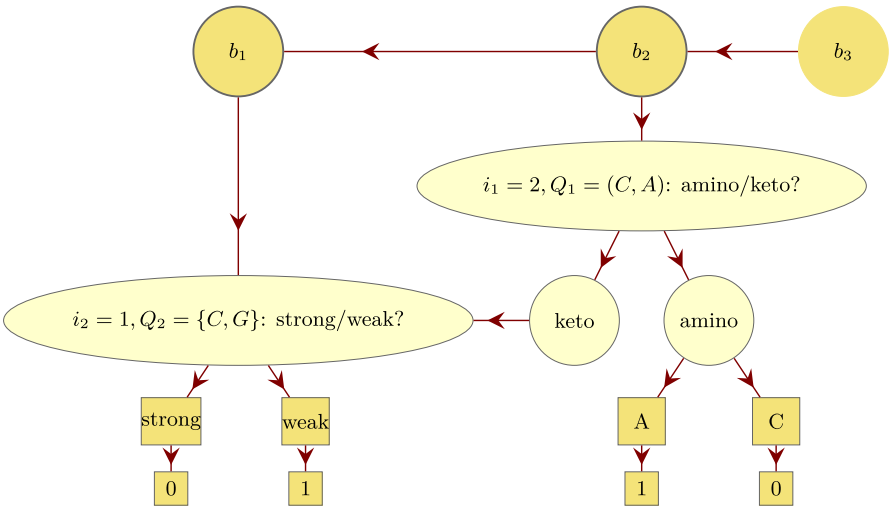

Fig. 1 Algorithmic way to define Rumer's dichotomy (Colour figure online)

If we apply several BDAs successively, we 'cut' the set of codons $\mathcal{B}^3$ into disjoint subsets. For example, we obtain four subsets labelled as $(0, 0)$, $(1, 0)$, $(0, 1)$, $(1, 1)$ (or in short 00, 10, 01, 11) when two different BDAs are applied. In the subset $(1, 0)$ we have, for instance, codons which are classified by the first BDA into the class 1 and by the second BDA into the class 0. In Gumbel et al. (2015), based on this notion, a class of models of the genetic code was introduced:

Definition 2.4 Let $k \in \mathbb{N}$. We will call a bijective mapping

$$M : \mathcal{B}^3 \to \{0, 1\}^k$$

a BDA-generated model of the genetic code of grade k if there exist k different BDAs

$$\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_k$$

such that for all $c \in \mathcal{B}^3$ the following equation holds:

$$M(c) = (\mathcal{A}_1(c), \mathcal{A}_2(c), ..., \mathcal{A}_k(c)).$$

Remark 2.5 It means for a codon $c \in \mathcal{B}^3$ that the mapping M assigns c in the jth coordinate 1 if c is classified by $\mathcal{A}_j$ for the dichotomic class H_1 and 0 if c is classified by $\mathcal{A}_j$ for the dichotomic class H_0.

Table 1 shows the standard genetic code and the Rumer classification as an example for a BDA.

Table 1 Standard genetic code where ! represents a stop codon

	U		C		A		G		
U	**Phe**	1	Ser	0	**Tyr**	1	**Cys**	1	U
U	**Phe**	1	Ser	0	**Tyr**	1	**Cys**	1	C
U	**Leu**	1	Ser	0	**!**	1	**!**	1	A
U	**Leu**	1	Ser	0	**!**	1	**Trp**	1	G
C	Leu	0	Pro	0	**His**	1	Arg	0	U
C	Leu	0	Pro	0	**His**	1	Arg	0	C
C	Leu	0	Pro	0	**Gln**	1	Arg	0	A
C	Leu	0	Pro	0	**Gln**	1	Arg	0	G
A	**Ile**	1	Thr	0	**Asn**	1	**Ser**	1	U
A	**Ile**	1	Thr	0	**Asn**	1	**Ser**	1	C
A	**Ile**	1	Thr	0	**Lys**	1	**Arg**	1	A
A	**Met**	1	Thr	0	**Lys**	1	**Arg**	1	G
G	Val	0	Ala	0	**Asp**	1	Gly	0	U
G	Val	0	Ala	0	**Asp**	1	Gly	0	C
G	Val	0	Ala	0	**Glu**	1	Gly	0	A
G	Val	0	Ala	0	**Glu**	1	Gly	0	G

The Rumer classification is labelled as 0 and 1 (bold)

2.2 Conductance

In this section, we introduce a methodology which comes from graph theory. Using these characteristics, we can describe some new features of BDA-algorithms in terms of the graph partition quality. We start our investigation by giving a definition of specific graph, which includes all information about single point mutations occurred in protein-coding sequences.

Definition 2.6 Let $G(V; E)$ be a graph in which V is the set of vertices (nodes) representing all possible 64 codons, whereas E is the set of edges connecting these vertices. All connections are defined in such a way that two vertices $c, c' \in V$, i.e. respective codons, are connected by the edge $e(c; c') \in E$ if and only if the codon c differs from c' in exactly one position.

According to Definition 2.6, the graph G is unweighted, undirected and also regular because the degree of each node is equal to nine (compare Fig. 2).

Furthermore, G has a nice interpretation from biological point of view because the set of edges E represents all possible single point mutations, i.e. single nucleotide substitutions, which can occur between codons in protein-coding sequences. In this work, we would like to investigate properties of the selected partitions of the vertices of the graph represented in Fig. 3 into fixed number $1 < k \leq 64$ of disjoint and non-empty subsets $\mathcal{C}_k$, i.e. k codon groups. The $\mathcal{C}_k$ partition is defined in the following way:

$$\mathcal{C}_k = \{S_1, S_2, \ldots, S_k : S_i \cap S_j = \emptyset,\ S_1 \cup S_2 \cup \ldots \cup S_k = V\}.$$

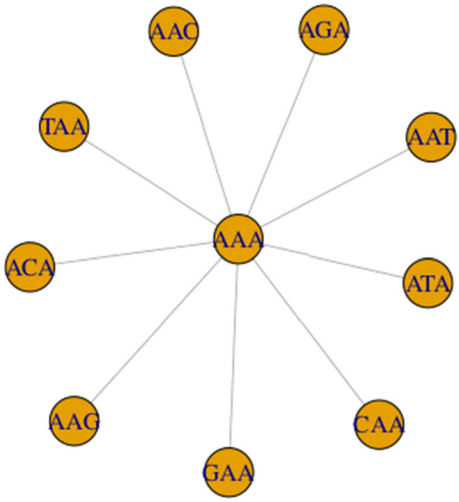

Fig. 2 The example of AAA codon neighbourhood generated by single nucleotide substitution. There are exactly nine codons which differ from AAA in exactly one nucleotide (Colour figure online)

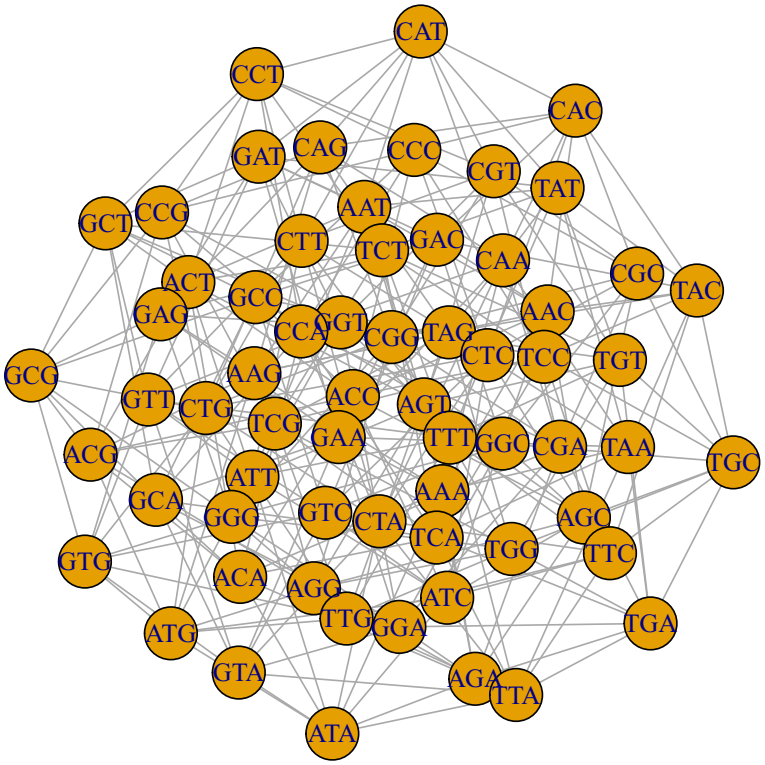

Fig. 3 Graphical representation of the graph defined in 2.6 (Colour figure online)

It is easy to see that for $k = 21$, we get $\mathcal{C}_{21}$ which is a representation of the genetic code as a partition of the set of vertices V into 21 disjoint and non-empty subsets. Therefore, it is interesting to test some characteristics of the $\mathcal{C}_k$ following graph theory. Particularly, we considered properties of $\mathcal{C}_k$ in terms of the optimal graph partitioning. Generally, the problem of finding optimal, in some sense, partition of G can be investigated from different perspectives. However, the idea presented in Blazej et al. (2018b), using the conductance property, appears to be promising for further research around the standard genetic code. The central role in this approach plays the set conductance measure which is used to calculate the quality of a given genetic code but clearly this method is used in more general clustering problem. This characteristic is defined in the following way:

Definition 2.7 For a given graph $G = (V, E)$ let S be a subset of V . The conductance of S is defined as:

$$\phi(S) = \frac{E(S, \overline{S})}{9|S|}$$

where $E(S, \bar{S})$ is the number of edges of G crossing from S to its complement $\bar{S}$.

The set conductance has many applications, for example, in the theory of random walks, theory of Markov processes (Levin et al. 2009) and also in social networks (Bollobàs 1998). Moreover, $\phi(S)$ has also a very interesting biological interpretation. Assuming that all codons belonging to S encode the same label, i.e. amino acids or stop coding signal, $\phi(S)$ is the ratio of the total number of non-synonymous single nucleotide substitution to all possible nucleotide substitution generated by all codons from S. Moreover, the Definition 2.7 allowed us as to define the conductance of the partition $\mathcal{C}_k$.

Definition 2.8 The conductance of a partition $\mathcal{C}_k$ is defined as

$$\Phi(\mathcal{C}_k) = \max_{S \in \mathcal{C}_k} \phi(S).$$

Therefore, the Φ measure gives us a characterization of the quality of a given partition $\mathcal{C}_k$ as the set conductance of the worst, in this term, codon group. What is more, Φ measure involves a question about the structure of the optimal graph G partition $\mathcal{C}_k$ for a fixed k. In this context, the best partition $\mathcal{C}_k$ of the graph G in terms of Φ follows in a natural way and is given by the formula:

$$\Phi_{\min}(k) = \min_{\mathcal{C}_k} \Phi(\mathcal{C}_k).$$

The definition of $\Phi_{\min}$ is identical with the definition of the kth-order graph conductance presented in Lee et al. (2014) and has an interpretation as lower boundary of a set robustness against changes which cause transitions between codon groups. It should be noted that in the case of $k = 2$, the minimum partition conductance $\Phi_{\min}$ is equivalent to the definition of the graph conductance (Lee et al. 2014).

3 Results and Discussion

3.1 Conductance of BDA-Partitions

We consider in this paper BDA-generated models of the genetic code from the viewpoint of their conductance. The next proposition shows that the conductance of only one BDA-partition is independent on the algorithm applied:

Proposition 3.1 *Let $\mathcal{A}$ be a BDA with the indices $i_1, i_2 \in \{1, 2, 3\}, i_1 \neq i_2$ and the questions $Q_1 = (B_1, B_2)(B_1 \neq B_2)$, and $Q_2 = \{B_3, B_4\}$ with $B_3 \neq B_4$, $\mathcal{C} = (H_0, H_1)$ the corresponding BDA-partition of $\mathcal{B}^3$. Then, the following formula holds:*

$$\phi(H_0) = \phi(H_1) = \Phi(\mathcal{C}) = \frac{80}{9 \cdot 32} = 0.2(7).$$

Proof Since $|H_0| = |H_1|$ and $E(H_1, H_2) = E(H_2, H_1)$, we get immediately $\phi(H_0) = \phi(H_1) = \Phi(\mathcal{C}_2)$. Therefore, it is sufficient to show $\phi(H_0) = 0.2(7)$. Let us consider $c = (b_1, b_2, b_3) \in H_0$ and assume also without loss of generality that $i_1 = 1, i_2 = 2$. Since all codons of the form $c = (B_2, b_2, b_3)$ are in H_1, we have to take into account the following two cases:

Case 1: Let $b_1 = B_1$. There are 16 codons fulfilling this condition in total. We have two cases in which the edge could go outside the set H_0, namely if the base in the first position in codon is substituted, i.e. $b_1 \rightarrow B_2$, we have one possible edge going outside H_0. Moreover, when codon c fulfils additional condition, i.e. for 8 codons out of 16: $b_2 \notin Q_2$, we obtain two additional edges, i.e. $b_1 \rightarrow \overline{\{B_1, B_2\}}$. Therefore, the total number of crossing edges calculated for all (B_1, b_2, b_3) type codons is equal to 32.

Case 2: Let $b_1 \notin \{B_1, B_2\}$. In this case, all codons c belonging to the set H_0, 16 in total, have the following form: $c = (b_1, [B_3|B_4], b_3)$. In this case, each c has one crossing edge generated by substitution in the first position in codon i.e. $b_1 \rightarrow B_2$. Moreover, they all have two possible crossing edges generated by nucleotide substitution in the second position in codon, i.e. $b_2 \rightarrow \overline{Q_2}$. As a result, the total number of crossing edges is equal, in this case, to 48.

To sum up, we have $48 + 32 = 80$ edges crossing from H_0 to H_1 and, thus,

$$\phi(H_0) = \frac{80}{9 \cdot 32}$$

what completes the proof. □

The following Theorem helps to calculate the minimal possible conductances for subsets of $\mathcal{B}^3$ of arbitrary size:

Theorem 3.2 *Let $G = (V; E)$ be the graph according to the Definition 2.6, $N_1 < N_2 < N_3 < N_4$ a linear order over the alphabet $\mathcal{B} = \{A; C; G; T(U)\}$, e.g. $C < G < A < T$, and $S_k \subseteq V$ the collection of the first $k = 1, 2, \ldots, 64$ vertices of*

the graph G in the lexicographic order. Then, the following recursive formula for the number of edges of G crossing from S_k to its complement $\bar{S}_k$ holds:

$$E(S_{k+1}, \overline{S_{k+1}}) = E(S_k, \overline{S_k}) + 9 - 2 \cdot (k_1 + k_2 + k_3), \quad E(S_1, \overline{S_1}) = 9$$

where $(k_1, k_2, k_3)_4$, $k_i \in \{0, 1, 2, 3\}$[1] *is the representation of k to base 4, i.e.*

$$k = k_1 \cdot 4^2 + k_2 \cdot 4^1 + k_3 \cdot 4^0.$$

The conductance of S_k is accordingly equal to

$$\phi(S_k) = \frac{E(S_k, \overline{S_k})}{9 \cdot k}.$$

Proof It is clear that $E(S_1, \overline{S_1}) = 9$ since the graph 2.6 is 9-regular and for only one codon in S_1 all edges are crossing edges between S_1 and $\bar{S}_1$.

Let us assume now that we already have calculated $E(S_k, \overline{S_k})$ for $k \geq 1$ and we are inserting now the next codon $c \in \mathcal{B}^3$ in the lexicographic order. It is easy to see that all codons ordered in lexicographic order can be rewritten as a sequence of consecutive three-digit numbers to the base 4 if we assign, for example, $N_1 \to 0, N_2 \to 1, N_3 \to 2, N_4 \to 3$. Therefore, newly included codon c has a numeric representation $c = (k_1, k_2, k_3)_4$, where $k_i = 0, 1, 2, 3$. What is more, k_i, $i = 1, 2, 3$ is the number of codons that differ from c at the position i which are smaller than c in the lexicographic order and the total number of edges crossing S_k and c, i.e. $E(S_k, \{c\})$, is, consequently, equal to $k_1 + k_2 + k_3$. As a result, the total number of edges between S_{k+1} to its complement fulfils the equation:

$$\begin{aligned} E(S_{k+1}, \overline{S_{k+1}}) &= E(S_k, \overline{S_k}) - E(S_k, \{c\}) + 9 - E(S_k, \{c\}) \\ &= E(S_k, \overline{S_k}) + 9 - 2 \cdot (k_1 + k_2 + k_3). \end{aligned}$$

That completes the proof. □

With Table 2, we calculate conductances for all S_k from Theorem 3.2 for $1 \leq k \leq 32$. It suffices if we calculate it for $1 \leq k \leq 32$ since in the case of at least one partitioning of $\mathcal{B}^3$ into at least 2 subsets, the size of one of them will be at most 32:

Following the approach from Proposition 3.2, we can calculate the minimal conductances for arbitrary partitions:

Proposition 3.3 *Let $\mathcal{C}_n$ be a partition of graph G where $n \in \mathbb{N}$, i.e. with n classes. Then, we have the following lower boundary for the conductance of the partition $\mathcal{C}_n$:*

(1) *For $n \neq 12, n \neq 3$*

$$\Phi(\mathcal{C}_n) \geq \phi(S_k) \text{ with } k = \left\lfloor \frac{64}{n} \right\rfloor,$$

[1] In the formula, we need for calculations always the 'previous' k. For instance, for calculation of $E(S_{64}, \bar{S}_{64})$ we need $k = 63$. This is why we can always represent k as a three-digit number to base 4.

Table 2 The Table shows exact values for conductances of all S_k from Theorem 3.2 and $1 \leq k \leq 32$

$k = \lvert S_k \rvert$	$E(S_k, \bar{S}_k)$	$\phi(S_k)$	$k = \lvert S_k \rvert$	$E(S_k, \bar{S}_k)$	$\phi(S_k)$
1	9	$\frac{9}{9} = 1$	17	55	$\frac{55}{9 \cdot 17} = \frac{55}{153} = 0.(359477..)$
2	16	$\frac{8 \cdot 2}{9 \cdot 2} = \frac{8}{9} = 0.(8)$	18	60	$\frac{60}{9 \cdot 18} = \frac{10}{27} = 0.(370)$
3	21	$\frac{7 \cdot 3}{9 \cdot 3} = \frac{7}{9} = 0.(7)$	19	63	$\frac{63}{9 \cdot 19} = \frac{7}{19} = 0.(368421053..)$
4	24	$\frac{6 \cdot 4}{9 \cdot 4} = \frac{6}{9} = 0.(6)$	20	64	$\frac{64}{9 \cdot 20} = \frac{16}{45} = 0.3(5)$
5	31	$\frac{31}{9 \cdot 5} = \frac{31}{45} = 0.6(8)$	21	69	$\frac{69}{9 \cdot 21} = \frac{23}{63} = 0.(365079)$
6	36	$\frac{36}{9 \cdot 6} = \frac{6}{9} = 0.(6)$	22	72	$\frac{72}{9 \cdot 22} = \frac{4}{11} = 0.(36)$
7	39	$\frac{39}{9 \cdot 7} = \frac{13}{21} = 0.(6190407)$	23	73	$\frac{73}{9 \cdot 23} = \frac{73}{207} = 0.(352657005...)$
8	40	$\frac{5 \cdot 8}{9 \cdot 8} = \frac{5}{9} = 0.(5)$	24	72	$\frac{72}{9 \cdot 24} = \frac{1}{3} = 0.(3)$
9	45	$\frac{45}{9 \cdot 9} = \frac{5}{9} = 0.(5)$	25	75	$\frac{75}{9 \cdot 25} = \frac{1}{3} = 0.(3)$
10	48	$\frac{48}{9 \cdot 10} = \frac{8}{15} = 0.5(3)$	26	76	$\frac{76}{9 \cdot 26} = \frac{38}{117} = 0.(324786)$
11	49	$\frac{49}{9 \cdot 11} = \frac{49}{99} = 0.(49)$	27	75	$\frac{75}{9 \cdot 27} = \frac{25}{81} = 0.(308641975...)$
12	48	$\frac{4 \cdot 12}{9 \cdot 12} = \frac{4}{9} = 0.(4)$	28	72	$\frac{72}{9 \cdot 28} = \frac{2}{7} = 0.(285714)$
13	51	$\frac{51}{9 \cdot 13} = \frac{51}{117} = 0.(435897)$	29	73	$\frac{73}{9 \cdot 29} = \frac{73}{261} = 0.(279693487...)$
14	52	$\frac{52}{9 \cdot 14} = \frac{26}{63} = 0.(412698)$	30	72	$\frac{72}{9 \cdot 30} = \frac{4}{15} = 0.2(6)$
15	51	$\frac{51}{9 \cdot 15} = \frac{17}{45} = 0.3(7)$	31	69	$\frac{69}{9 \cdot 31} = \frac{23}{93} = 0.(247311828...)$
16	48	$\frac{3 \cdot 16}{9 \cdot 16} = \frac{1}{3} = 0.(3)$	32	64	$\frac{64}{9 \cdot 32} = \frac{2}{9} = 0.(2)$

The notation, for example, 0.2(6) means the periodical fraction 0.2666666... where the repetend is taken in parentheses

(2) *For* $n = 12$

$$\Phi(\mathcal{C}_{12}) \geq \phi(S_6),$$

(3) *For* $n = 3$

$$\Phi(\mathcal{C}_3) \geq \phi(S_{23}),$$

where $\phi(S_k)$ *is the entry corresponding to* k *from Table* 2.

Proof According to Proposition 3 from Blazej et al. (2018b), the collection of the first k vertices taken in lexicographic order of a graph as defined in 2.6 has the minimal conductance among all subsets of the same size k.

(1) For $n = 2$ we have two subsets, the size of one of them has to be at most 32. Since

$$\phi(S_{64}) < \phi(S_k) \quad \text{for all} \quad k < 32$$

takes place, a partition with the minimal conductance has to consist of two equal-sized subsets. Thus, we get a partition with the minimum conductance if we classify in the lexicographic order the first 32 codons into one class and the remaining 32

into the other one. Corresponding to Table 2, the conductance of such partition is equal to

$$\Phi(\mathcal{C}) = \frac{64}{9 \cdot 32} = 0.(2)$$

Let $n \geq 4, n \neq 12$. In this case, $k = \lfloor \frac{64}{n} \rfloor \leq 16$ represents the average size of a subset in a partition and $\phi(S_k)$ is decreasing with increasing of k with only one exception $\phi(S_4) < \phi(S_5)$.

Since $\Phi(\mathcal{C}_n)$ is defined as the maximal value of conductances of all subsets from $\mathcal{C}$, it is equal to the conductance of the 'worst', in this sense, subset. Since, on the one hand, the following inequality takes place

$$\phi(S_i) > \phi(S_{15}) \geq \phi(S_j), \quad 1 \leq i \leq 14, \quad 17 \leq j \leq 32 \quad \text{(compare table 2)}$$

and, on the other hand, increasing the size of one subset leads to decreasing it for another subset of the partition, we have that it is not favourable to have bigger than 16 codons subsets in the partition.

Let us now assume that

$$\Phi(\mathcal{C}_n) < \phi(S_k).$$

It means that for all subsets $C_i \in \mathcal{C}_n$, we have

$$\phi(C_i) < \phi(S_k).$$

According to the behaviour of the function $\phi(S_k)$ (compare Table 2), it means that for all $i = 1, ..., n \quad |C_i| > k$ or $k = 5$. In the first case, we obtain immediately a contradiction since then we have

$$\sum_{i=1}^{n} |C_i| \geq n \cdot (k+1) > 64.$$

In the second case ($k = 5$), we have $n = 12$ what is excluded.

(2) Let $n = 12$ and $\Phi(\mathcal{C}_{12}) < \phi(S_6)$. It means that for all subsets $C_i \in \mathcal{C}_{12}$, we have

$$\phi(C_i) < \phi(S_6)$$

and, thus, for all $i = 1, ..., 12 \quad |C_i| > 6$. Consequently,

$$\sum_{i=1}^{12} |C_i| \geq 12 \cdot 7 = 84 > 64.$$

This is a contradiction.

(3) Let $n = 3$ and $\Phi(\mathcal{C}_3) < \phi(S_{23})$. It means that for all subsets $C_i \in \mathcal{C}_3$, we have

$$\phi(C_i) < \phi(S_{23})$$

and, thus, for all $i = 1, 2, 3$ $|C_i| > 23$. Consequently,

$$\sum_{i=1}^{3} |C_i| \geq 3 \cdot 23 = 69 > 64.$$

This is a contradiction.

□

Applying the Proposition 3.3 in the special case of BDA-partitions, we obtain:

Corollary 3.4 *Let M be a BDA-model of $\mathcal{B}^3$ with $n \in \mathbb{N}$ classes and $\mathcal{C}$ the corresponding BDA-partition. Then, we have for the conductance of $\mathcal{C}$ the following lower boundary:*

(1) *For $n \neq 12$*

$$\Phi(\mathcal{C}) \geq \phi(S_k) \text{ with } k = \left\lfloor \frac{64}{n} \right\rfloor$$

and $\phi(S_k)$ the entry corresponding k from Table 2.

(2) *For $n = 12$*

$$\Phi(\mathcal{C}) \geq 0.(6).$$

Proof According to Proposition 6 in Gumbel et al. (2015), it is not possible for a BDA-generated model that $n = 3$. The remaining part follows immediately from 3.3.

□

3.2 Best Conductance BDAs

In the previous section, we found lower boundaries for BDA-generated partitions. However, it is not clear yet whether these boundaries are sharp. We have adapted the algorithm described in Gumbel et al. (2015) and searched for models of the genetic code based on BDAs with the best conductance $\Phi_{\min}$. The following examples in Tables 3, 4, 5 and 6 show that we can indeed obtain partitions generated with BDAs with the best possible conductance for the class sizes 8, 12, 16, and 24. All these code tables contain the Rumer-BDA, the code table for 24 classes additionally uses the Complementary-BDA. These partitions are achieved with the minimum number of BDAs required, i.e. 3 for 8 classes, 4 for 12 and 16 classes and 5 for 24 classes.

According to the corollary 3.3, we have an exception if the number of generated classes is equal to 12. Table 4 shows that we can also obtain in this case a partition with the best possible conductance with a BDA-model.

Table 3 $|M| = 8$ classes generated by three BDAs including Rumer. (A) Code table, (B) list of BDAs. Conductance of $\mathcal{C}_8$ is $\Phi(\mathcal{C}_8) = 5/9 = \Phi_{\min}(8)$. ! represents a stop codon (Colour figure online)

(A)

	T		C		A		G		
T	Phe	100	Ser	000	Tyr	110	Cys	110	T
T	Phe	100	Ser	000	Tyr	110	Cys	110	C
T	Leu	100	Ser	000	!	110	!	110	A
T	Leu	100	Ser	000	!	110	Trp	110	G
C	Leu	010	Pro	010	His	111	Arg	011	T
C	Leu	010	Pro	010	His	111	Arg	011	C
C	Leu	010	Pro	010	Gln	111	Arg	011	A
C	Leu	010	Pro	010	Gln	111	Arg	011	G
A	Ile	100	Thr	000	Asn	101	Ser	101	T
A	Ile	100	Thr	000	Asn	101	Ser	101	C
A	Ile	100	Thr	000	Lys	101	Arg	101	A
A	Met	100	Thr	000	Lys	101	Arg	101	G
G	Val	001	Ala	001	Asp	111	Gly	011	T
G	Val	001	Ala	001	Asp	111	Gly	011	C
G	Val	001	Ala	001	Glu	111	Gly	011	A
G	Val	001	Ala	001	Glu	111	Gly	011	G

(B)

BDA	(i_1, i_2)	Q_1	Q_2
Rumer	(2, 1)	(C, A)	$\{C, G\}$
$\mathcal{A}_2$	(1, 2)	(A, C)	$\{T, C\}$
$\mathcal{A}_3$	(1, 2)	(T, G)	$\{T, C\}$

Table 4 $|M| = 12$ classes generated by three BDAs including Rumer. (A) Code table, (B) list of BDAs. Conduction of $\mathcal{C}_{12}$ is $\Phi(\mathcal{C}_{12}) = 2/3 = 0.(6) = \Phi_{\min}(12)$ (Colour figure online)

(A)

	T		C		A		G		
T	Phe	1100	Ser	0101	Tyr	1101	Cys	1101	T
T	Phe	1100	Ser	0101	Tyr	1101	Cys	1101	C
T	Leu	1100	Ser	0101	!	1101	!	1101	A
T	Leu	1100	Ser	0101	!	1101	Trp	1101	G
C	Leu	0010	Pro	0111	His	1010	Arg	0110	T
C	Leu	0010	Pro	0111	His	1010	Arg	0110	C
C	Leu	0010	Pro	0111	Gln	1010	Arg	0110	A
C	Leu	0010	Pro	0111	Gln	1010	Arg	0110	G
A	Ile	1000	Thr	0011	Asn	1001	Ser	1011	T
A	Ile	1000	Thr	0011	Asn	1001	Ser	1011	C
A	Ile	1000	Thr	0011	Lys	1001	Arg	1011	A
A	Met	1000	Thr	0011	Lys	1001	Arg	1011	G
G	Val	0000	Ala	0111	Asp	1000	Gly	0110	T
G	Val	0000	Ala	0111	Asp	1000	Gly	0110	C
G	Val	0000	Ala	0111	Glu	1000	Gly	0110	A
G	Val	0000	Ala	0111	Glu	1000	Gly	0110	G

(B)

BDA	(i_1, i_2)	Q_1	Q_2
Rumer	(2, 1)	(C, A)	$\{C, G\}$
$\mathcal{A}_2$	(1, 2)	(A, T)	$\{A, T\}$
$\mathcal{A}_3$	(1, 2)	(T, C)	$\{A, T\}$
$\mathcal{A}_4$	(2, 1)	(T, C)	$\{C, G\}$

Next, it was analysed whether a code table for the standard genetic code could be created by means of BDAs, i.e. if we can classify 21 classes (20 for amino acids plus 1 for stop codons). If we consider codons of degeneracy 6 (like those for Serine) as codons belonging to two groups of size 4 and 2 each—like in the Rumer transformation—we get three extra classes, and thus 24 classes in total. Table 6 shows such a model of the genetic code with 24 classes and optimal conductance ($\Phi(\mathcal{C}_{24}) = 8/9$). It is striking that again the Rumer-BDA but this time also the Complementary-BDA is included. Moreover, the code with optimal conductance is also to some extend compatible with the standard universal code. In Gumbel et al. (2015), an error metric E, $(0 \leq E \leq 1)$ was introduced to indicate how "good" a code is compatible with the standard genetic code, where an error of $E = 0$ represents a perfect match. It is known that the standard genetic code does not have an optimal conductance as there are codons coding only

Table 5 $|M| = 16$ classes generated by four BDAs including Rumer. (A) Code table, (B) list of BDAs. Conductance of $\mathcal{C}_{16}$ is $\Phi(\mathcal{C}_{16}) = 2/3 = \Phi_{\min}(16)$ (Colour figure online)

(A)

	T		C		A		G		
T	Phe	1000	Ser	0000	Tyr	1100	Cys	1101	T
T	Phe	1000	Ser	0000	Tyr	1100	Cys	1101	C
T	Leu	1000	Ser	0000	!	1100	!	1101	A
T	Leu	1000	Ser	0000	!	1100	Trp	1101	G
C	Leu	0110	Pro	0100	His	1110	Arg	0101	T
C	Leu	0110	Pro	0100	His	1110	Arg	0101	C
C	Leu	0110	Pro	0100	Gln	1110	Arg	0101	A
C	Leu	0110	Pro	0100	Gln	1110	Arg	0101	G
A	Ile	1010	Thr	0001	Asn	1011	Ser	1001	T
A	Ile	1010	Thr	0001	Asn	1011	Ser	1001	C
A	Ile	1010	Thr	0001	Lys	1011	Arg	1001	A
A	Met	1010	Thr	0001	Lys	1011	Arg	1001	G
G	Val	0010	Ala	0011	Asp	1111	Gly	0111	T
G	Val	0010	Ala	0011	Asp	1111	Gly	0111	C
G	Val	0010	Ala	0011	Glu	1111	Gly	0111	A
G	Val	0010	Ala	0011	Glu	1111	Gly	0111	G

(B)

BDA	(i_1, i_2)	Q_1	Q_2
Rumer	(2, 1)	(C, A)	$\{C, G\}$
$\mathcal{A}_2$	(1, 2)	(A, C)	$\{T, C\}$
$\mathcal{A}_3$	(1, 2)	(T, G)	$\{G, C\}$
$\mathcal{A}_4$	(2, 1)	(T, G)	$\{T, C\}$

for one amino acid, e.g. AUG for Methionine (Blazej et al. 2018b). However, the code in Table 6 with optimal conductance has only a compatibility error of $E = 12/64$. That is, only 12 changes in the assignments of codons to amino acids are required to get a perfect standard universal code. When the mitochondrial vertebrate code is considered (table not shown), this comparability error could even be reduced to $E = 10/64 = 0.15625$.

For the sake of completeness and to ensure that the best conductance BDAs are not a multiple of four for the number of classes, other class sizes ranging from 4 to 23 have been tested. BDA-generated models with best possible conductance values could be indeed obtained for

$$|M| = 4, 7, 8, 10, 11, 12, 13, 14, 15, 16, 22, 23, 24.$$

However, it is not possible, for example, for $|M| = 21$ as it was proven with a comprehensive search as explained in Gumbel et al. (2015). In this case, the best possible partitioning has the conductance value equal to 7/9 (Blazej et al. 2018b) and cannot be obtained using a BDA-generated model of the genetic code. For the remaining class sizes ($|M| = 5, 6, 9, 17, 18, 19, 20$), a sample of 10,000 partitions for each class size (compare Fig. 4) did not show any best BDA-model and it remains to be shown, whether there are any BDA-models; however, this is very unlikely.

3.3 Distribution of Conductance

Lower boundaries for the optimal conductance of BDA-generated models were derived, and it was proven that there are BDA-models which are optimal. This section shows that those models have a much better conductance compared to random partitions with the same number of classes. Figure 4a–c shows the distribution of the set conductance for code tables (1) generated by random BDAs and (2) random partitions

Table 6 $|M| = 24$ classes generated by six BDAs including Rumer and Complementary. (A) Code table, (B) list of BDAs. Conductance of $\mathcal{C}_{24}$ is $\Phi(\mathcal{C}_{24}) = 8/9 = \Phi_{\min}(24)$. The compatibility error is $E = 0.1875 = 12/64$ (Colour figure online)

(A)

	T		C		A		G		
T	Phe	11110	Ser	01111	Tyr	11101	Cys	11111	T
T	Phe	11110	Ser	01111	Tyr	11101	Cys	11111	C
T	Leu	10110	Ser	00111	!	10101	!	10111	A
T	Leu	10110	Ser	00111	!	10101	Trp	10111	G
C	Leu	00010	Pro	00101	His	10000	Arg	00100	T
C	Leu	00010	Pro	00101	His	10000	Arg	00100	C
C	Leu	00010	Pro	00101	Gln	10000	Arg	00100	A
C	Leu	00010	Pro	00101	Gln	10000	Arg	00100	G
A	Ile	11010	Thr	01011	Asn	11001	Ser	11011	T
A	Ile	11010	Thr	01011	Asn	11001	Ser	11011	C
A	Ile	10010	Thr	00011	Lys	10001	Arg	10011	A
A	Met	10010	Thr	00011	Lys	10001	Arg	10011	G
G	Val	01010	Ala	01101	Asp	11000	Gly	01100	T
G	Val	01010	Ala	01101	Asp	11000	Gly	01100	C
G	Val	01010	Ala	01101	Glu	11000	Gly	01100	A
G	Val	01010	Ala	01101	Glu	11000	Gly	01100	G

(B)

BDA	(i_1, i_2)	Q_1	Q_2
Rumer	(2, 1)	(C, A)	$\{C, G\}$
Complementary	(1, 3)	(C, G)	$\{G, A\}$
$\mathcal{A}_3$	(1, 2)	(A, T)	$\{A, T\}$
$\mathcal{A}_4$	(2, 1)	(A, T)	$\{C, G\}$
$\mathcal{A}_5$	(2, 1)	(T, C)	$\{C, G\}$

with 8, 16, and 24 classes. The number of BDAs of a random model ranges from the minimum number required, e.g. 3 for 8 classes to four extra BDAs, i.e. 7. for 8 classes. Those redundant BDAs were included because it was shown in Gumbel et al. (2015) that some partitions can only be achieved with more than the minimum number of BDAs.

In any case, random BDAs create code tables with a better conductance and some of them are optimal. All random partitions for 24 classes have the worst conductance of 1. BDA-generated partitions, however, have either the best conductance ($\Phi(\mathcal{C}_{24}) = 8/9$) or the worst, too.

Even if there exist no BDA-generated models with the best possible conductance for some class sizes (compare the previous section), the BDA-generated models perform in terms of average conductance significantly better than the randomly generated ones.

4 Conclusions

In this work, we are discussing more deeply the properties of BDA-generated models of the genetic code. To do so, we are incorporating a new methodology following on graph theory. According to this approach, each BDA algorithm and, more generally, genetic code induces its own partition of graph vertices into disjoint and non-empty subsets, corresponding to amino acids to be encoded. The quality of a given partition was calculated by using the maximum partition conductance measure. This measure

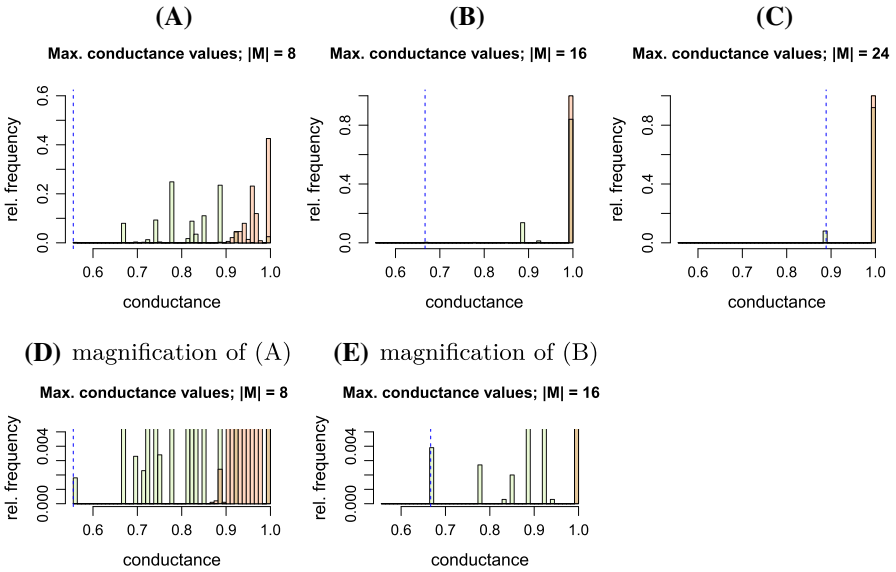

Fig. 4 Distribution of conduction for different partitions $\mathcal{C}_k$. Green bars show partitions generated by BDAs, and red bars show random partitions. Blue dashed line indicates the best conductance of $\mathcal{C}_k$. No random partition has an optimal conductance. **d**, **e** zoom in and show only a fraction of the y-axis as the scale in **a** and **b** is not sufficient to see the details. This is not required in **c** as all frequencies are visible. Sample size is 10,000. (**a**, **d**) Eight classes: there are about 0.2 % BDA-partitions with optimal conductance ($\Phi(\mathcal{C}_8) = 5/9$). Numbers of BDAs per model ($\mathcal{A}_1, \ldots, \mathcal{A}_l$) range from $3 \leq l \leq 7$. **b**, **e** 16 classes: again there are BDA-partitions (about 0.4 %) with optimal conductance ($\Phi(\mathcal{C}_{16}) = 2/3$). Numbers of BDAs per model range from 4 to 8. **c** 24 classes: also BDA-partitions (8 %) with optimal conductance ($\Phi(\mathcal{C}_{24}) = 8/9$). Numbers of BDAs per model range from 5 to 9 (Color figure online)

gives us a general overview of the quality of each set of codons belonging to its partition because it is based on calculating a proportion of a number of all possible non-synonymous substitution to all nucleotide changes for every codon group. Therefore, the maximum conductance has a biological interpretation as a measure of robustness of partition sets against point mutations. Moreover, the maximum partition conductance can be used in general for evaluating quality of theoretical genetic codes which encode different number of amino acids. In this context, we found a formula for the lower boundary of the maximum conductance for graph partitions with a given number of classes corresponding to amino acids. We also compared it to a large number of randomly generated partitions. It should be noted that, taking a single BDA, none of the dichotomic partitions obtained has the minimal conductance; however, applying overlappings of BDA-partitions, i.e. BDA-generated models, we can reach the minimal possible conductance values, i.e. create the most robust against point mutations models of the genetic code. Moreover, the BDA-models have a substantially better quality in comparison with randomly generated partitions. This result indicates that the quality of models generated by BDA-algorithms can not easily be overcome by just a random process of amino acids' assignment to codons. The results obtained can, for instance, be useful for understanding the evolution of the genetic code.

Acknowledgements We would like to thank Lutz Strüngmann for stimulating discussions.

References

Blazej 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

Blazej P, Wnetrzak M, Mackiewicz D, Mackiewicz P (2018a) Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. PLoS ONE. https://doi.org/10.1371/journal.pone.0201715

Blazej P, Kowalski D, Mackiewicz D, Wnetrzak M, Aloqalaa D, Mackiewicz P (2018b) The structure of the genetic code as an optimal graph clustering problem. https://doi.org/10.1101/332478

Bollobàs B (1998) Modern graph theory. Springer, New York

Di Giulio M (1989) The extension reached by the minimization of the polarity distances during the evolution of the genetic code. J Mol Evol 29(4):288–293

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

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

Dunnill P (1966) Triplet nucleotide-amino-acid pairing—a stereochemical basis for division between protein and non-protein amino-acids. Nature 210(5042):1267–1268

Epstein CJ (1966) Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. Nature 210(5031):25–28

Fimmel E, Strüngmann L (2016) Yury Borisovich Rumer and his biological papers on the genetic code. Philos Trans R Soc A374:20150228

Fimmel E, Danielli A, Strüngmann L (2013) dichotomic classes and bijections of the genetic code. J Theor Biol 336:221–230

Fimmel E, Giannerini S, Gonzalez D, Strüngmann L (2014) Circular codes, symmetries and transformations. J Math Biol 70(7):1623–1644

Fimmel E, Michel CJ, Strüngmann L (2016) *n*-nucleotide circular codes in graph theory. Philos Trans A 374:20150058

Fimmel E, Michel CJ, Strüngmann L (2017) Strong comma-free codes in genetic information. Bull Math Biol 79(8):1796–1819. https://doi.org/10.1007/s11538-017-0307-0

Fimmel E, Michel CJ, Starman M, Strüngmann L (2018) Self-complementary circular codes in coding theory. Theory Biosci 137(1):51–65. https://doi.org/10.1007/s12064-018-0259-4

Freeland SJ, Hurst LD (1998a) The genetic code is one in a million. J Mol Evol 47(3):238–248

Freeland SJ, Hurst LD (1998b) Load minimization of the genetic code: history does not explain the pattern. Proc R Soc B Biol Sci 265(1410):2111–2119

Giannerini S, Gonzalez DL, Rosa R (2012) DNA, dichotomic classes and frame synchronization: a quasi-crystal framework. Philos Trans R Soc 370:2987–3006

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

José M, Zamudio GS, Morgado ER (2017) A unified model of the standard genetic code. R Soc Open Access. https://doi.org/10.1098/rsos.160908

Khorana HG, Buchi H, Ghosh H, Gupta N, Jacob TM, Kossel H, Morgan R, Narang SA, Ohtsuka E, Wells RD (1966) Polynucleotide synthesis and the genetic code. Cold Spring Harb Symp Quant Biol 31:39–49

Lee JR, Gharan SO, Trevisan L (2014) Multiway spectral partitioning and higher-order cheeger inequalities. J ACM 61(6):37. https://doi.org/10.1145/2665063

Levin DA, Peres Y, Wilmer EL (2009) Markov chains and mixing times. American Mathematical Society, Providence

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

Pelc SR, Welton MGE (1966) Stereochemical relationship between coding triplets and amino-acids. Nature 209(5026):868–870

Rumer YB (2016a) Translation of systematization of codons in the genetic code [I] by Yu. B. Rumer (1966). Philos Trans R Soc A374:20150446
Rumer YB (2016b) Translation of systematization of codons in the genetic code [II] by Yu. B. Rumer (1968). Philos Trans R Soc A374:20150447
Rumer YB (2016c) Translation of systematization of codons in the genetic code [III] by Yu. B. Rumer (1969). Philos Trans R Soc A374:20150448
Santos J, Monteagudo A (2010) Study of the genetic code adaptability by means of a genetic algorithm. J Theor Biol 264(3):854–865
Schönauer S, Clote P (1997) How optimal is the genetic code? In: Frishman D, Mewes HW (eds) Computer science and biology proceedings of the german conference on bioinformatics (GCB'97), pp 65–67
Tlusty T (2010) A colorful origin for the genetic code: information theory, statistical mechanics and the emergence of molecular codes. Phys Life Rev 7(3):362–376. https://doi.org/10.1016/j.plrev.2010.06.002
Wong JT (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72(5):1909–1912
Yarus M, Caporaso JG, Knight R (2005) Origins of the genetic code: the escaped triplet theory. Annu Rev Biochem 74:179–198