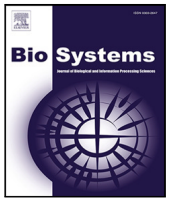
BioSystems 208 (2021) 104485

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems



Robustness against point mutations of genetic code extensions under consideration of wobble-like effects

E. Fimmel *, M. Gumbel, M. Starman, L. Strüngmann

Competence Center in Medicine, Biology, and Biotechnology, Mannheim University of Applied Sciences, 68163 Mannheim, Germany

ARTICLE INFO

Keywords:
Genetic code
Point mutations
Wobble-effect

ABSTRACT

Many theories of the evolution of the genetic code assume that the genetic code has always evolved in the direction of increasing the supply of amino acids to be encoded (Barbieri, 2019; Di Giulio, 2005; Wong, 1975). In order to reduce the risk of the formation of a non-functional protein due to point mutations, nature is said to have built in control mechanisms. Using graph theory the authors have investigated in Blazej et al. (2019) if this robustness is optimal in the sense that a different codon–amino acid assignment would not generate a code that is even more robust. At present, efforts to expand the genetic code are very relevant in biotechnological applications, for example, for the synthesis of new drugs (Anderson et al., 2004; Chin, 2017; Dien et al., 2018; Kimoto et al., 2009; Neumann et al., 2010). In this paper we generalize the approach proposed in Blazej et al. (2019) and will explore hypothetical extensions of the standard genetic code with respect to their optimal robustness in two ways:

(1) We keep the usual genetic alphabet but move from codons to longer words, such as tetranucleotides. This increases the supply of coding words and thus makes it possible to encode non-canonical amino acids.

(2) We expand the genetic alphabet by introducing non-canonical base pairs.

In addition, the approach from Blazej et al. (2019) and Blazej et al. (2018) is extended by incorporating the weights of single point-mutations into the model. The weights can be interpreted as probabilities (appropriately normalized) or degrees of severity of a single point mutation. In particular, this new approach allows us to take a closer look at the wobble effects in the translation of codons into amino acids. According to the results from Blazej et al. (2019) and Blazej et al. (2018), the standard genetic code is not optimal in terms of its robustness to point mutations if the weights of single point mutations are not taken into account. After incorporation into the model weights that mimic the wobble effect, the results of the present work show that it is much more robust, almost optimal in that respect.

We hope, that this theoretical analysis might help to assess extended genetic codes and their abilities to encode new amino acids.

1. Introduction

The origin of the standard genetic code is the subject of many theories and is still not conclusively understood. One point on which almost all theories agree is that its development has followed the direction of including ever increasing numbers of amino acids to be encoded (Barbieri, 2019; Di Giulio, 2005; Wong, 1975). This point is increasingly relevant today because of the development of biotechnology and especially the synthesis of new medicines (Anderson et al., 2004; Chin, 2017; Dien et al., 2018; Kimoto et al., 2009; Neumann et al., 2010). Graph theoretical methods have proved to be extremely helpful and fruitful in the study of the structural properties of the genetic code. Thus, the graph-theoretic approach is ideally suited to the study of the error-detecting properties of the genetic code, which are used, for example, in dealing with the so-called frame-shift problem due to indels (see Fimmel et al., 2016, 2018, 2020a,b). The main goal of this work is to investigate the robustness of the genetic code against point-mutations using graph theory, more specifically as an optimal clustering problem in a suitably modelled graph.

The so-called clustering or sparsest cut problems in graph theory have many applications: To Natural Language Processing, in communications engineering and recently especially due to the growing importance of social media for community detection (Gaertler, 2005; Alev et al., 2017). There are different methods to measure the "quality"

* Corresponding author.
E-mail addresses: e.fimmel@hs-mannheim.de (E. Fimmel), m.gumbel@hs-mannheim.de (M. Gumbel), m.starman@live.com (M. Starman), l.struengmann@hs-mannheim.de (L. Strüngmann).

https://doi.org/10.1016/j.biosystems.2021.104485
Received 16 June 2021; Received in revised form 7 July 2021; Accepted 9 July 2021
Available online 16 July 2021

of a clustering of a graph. One of them is using the conductance measure (see, for instance, Gaertler, 2005; Kannan et al., 2004; Lee et al., 2014). In the paper (Blazej et al., 2018), the conductance approach was applied to the question of testing different variants of the genetic code for their robustness against single point mutations: the conductance measure, a number between 0 and 1, can be seen as a fitness function for modelling an optimal assignment of the amino acids to be encoded to the tuples of bases (usually codons) that encode them.

Recently, the approach has been further developed (Blazej et al., 2019, 2020) and used in particular to assess the robustness of potential extensions of the genetic code against single point-mutations. In all cases, however, the unweighted graph was considered, whereas the term "conductance" in graph theory is defined more generally, for weighted graphs. Of course, with abandonment of edge weights, some degrees of freedom were given up in favour of simpler modelling. This makes the reflection of some biological phenomena, such as the wobble effect, in mathematical modelling at least more difficult, if not impossible. The results of the present work show that the incorporation of weights into the model leads to a better understanding of the origin of the standard genetic code in its modern form, when the wobble effects in the translation of codons into amino acids are taken into account.

An important point in the preceding work is to find the lower bounds for the conductance measure for a given number of classes. These results are important because they make it possible to make statements about the optimality of a partition of the set of codons into subsets of synonymous codons. With the (Higher-Order) Cheeger Inequalities (Alev et al., 2017; Gaertler, 2005; Kwok et al., 2013; Lee et al., 2014), which have received much attention in recent years, such a lower bound has already been known. However, as shown in this paper, it is only sharp in very few cases. In this paper, we develop a more precise method to determine lower bounds.

The paper is structured as follows: In Section 2 we define the basic graph-theoretic approach with an extension to weighted graphs. Moreover, we give the definition of conductance. In the next Section 3.1 we show some results on lower bounds for the conductances using the Higher Order Cheeger Inequalities and improve this by a new combinatorial method. Finally, in Section 3.2 we explicitly determine a model using conductance that incorporates the Wobble-effect showing that the genetic code is almost optimal with respect to its robustness against point mutations. The paper closes with some conclusions.

2. Graph theory and conductance

In this section we generalize an approach developed in Blazej et al. (2018) by extending it to weighted graphs. As explained in the introduction the idea behind this approach and the definition of conductance is to measure the robustness of partitions of sets of words against one-point mutations, especially in the case of genetic information.

Let Σ be a finite alphabet of even cardinality $|\Sigma| = 2n$ for some $n \in \mathbb{N}$. Actually, almost all the results presented here also hold for alphabets of odd cardinality but due to the biological motivation of the present study, we restrict ourselves to alphabets of even cardinality. Moreover, we will use the notation $\mathcal{B} = \{A, C, G, T(U)\}$ for the particularly important special case of $\Sigma = \mathcal{B}$ being the standard genetic alphabet. We first give a generalized version of the approach from Blazej et al. (2018). Recall that for an integer $l \in \mathbb{N}$, Σ^l denotes the set of all words over Σ of length l, i.e. $\Sigma^l = \{v_1 \cdots v_l : v_i \in n\Sigma\}$. For further details on graph theory we refer the reader to Clark and Holton (1991)

Definition 2.1. Let $\ell \in \mathbb{N}$ and $P = \{p_i^{\{N,N'\}} \mid i = 1, \ldots, \ell, N \neq N' \in \Sigma\}$ where $p_i^{\{N,N'\}}$ are non-negative weights. We define a weighted graph $G(V, E) = G_l^P(V, E, w)$ as follows:

(1) $V = \Sigma^\ell$ is the set of vertices (nodes) representing all possible ℓ-letter words over Σ;

(2) E is the set of edges where $(c, c') \in E$ if and only if $c, c' \in V$ and c differs from c' in exactly one position;

(3) The function $w : E \to P$ assigns to every edge $(c, c') \in E$ a weight $p_i^{\{N,N'\}}$ by $w((c, c')) = p_i^{\{N,N'\}}$ if and only if c differs from c' in position $i \in \{1, \ldots, \ell\}$ and $c_i = N, c_i' = N'$ where $c_i = (c_1, \ldots, c_l)$ and $c' = (c_1', \ldots, c_l')$.

If for all $i \in \{1, \ldots, l\}$ the weights $p_i^{\{N,N'\}}$ are independent of the choice of N, N' we will simply denote the weights $p_i^{\{N,N'\}}$ by p_i.

According to Definition 2.1 the graph G is a weighted, undirected and regular graph, i.e. the degree of each node is equal to $\ell \cdot (2n - 1)$. Note that the weight function in (3) of the above Definition 2.1 is well-defined since for any pair $N, N' \in \Sigma$ the sets $\{N, N'\}$ and $\{N', N\}$ and hence the weights $p_i^{\{N,N'\}}$ and $p_i^{\{N',N\}}$ are identical. From a biological point of view, i.e. when $\Sigma = \mathcal{B}$, the graph G has a nice interpretation: 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. Such point-mutations appear quite often and might lead to fatal changes in the encoded proteins The weights p_i (respectively $p_i^{\{N,N'\}}$) can be interpreted, correspondingly standardized, as the probabilities with which a point mutation occurs at position i (respectively occurs at position i and changes N to N'). The above Definition 2.1 also takes into account that these mutation probabilities may depend not only on the position in the tuple, but also on the base pairs. For example, it is conceivable that the mutation $U \to G$ in the third position of a codon is more likely than the mutation $U \to A$ (see also Section 3.2 for further results related to the Wobble effect). Before we give some illustrative examples we state a remark for the convenience of the reader.

Remark 2.2. The mindful reader may have noticed that Definition 2.1 speaks about undirected graphs which implies that the one-point mutations are symmetric in the sense that the probability of e.g. a mutation $U \to G$ has to be the same as that of $G \to U$ but may depend on the position where the mutation happens. In certain settings this might not be the case and thus we remark that Definition 2.1 can easily be turned into a directed version, say $\vec{G}$, by replacing the weights $p_i^{\{N,N'\}}$ by weights $p_i^{(N,N')}$ which then respect the order of the bases N and N'. The definition of conductance and several of our results then also hold in the directed version. However, the authors wanted to avoid even more technicalities and have therefore kept the undirected version.

It is now time to give some examples of graphs that satisfy Definition 2.1.

Example 2.3. Let $\Sigma = \{0, 1\}$ be the binary alphabet and $\ell = 3$. Moreover, choose $p_1^{\{0,1\}} = 1$, $p_2^{\{0,1\}} = 3$ and $p_3^{\{0,1\}} = 5$, then G looks like in Fig. 1:

We have a second example in a biological setting.

Example 2.4. Let $\Sigma = \{A, C, G, U, I, K\}$ be an extended genetic alphabet and $\ell = 2$. Moreover, choose all $p_i^{\{N,N'\}} = 2$, then G looks like in Fig. 2.

And a final example in the genetic code setting.

Example 2.5. Let $\Sigma = \mathcal{B} = \{A, C, G, U\}$ be the genetic alphabet and $\ell = 2$ respectively $l = 3$. Moreover, choose all $p_i^{\{N,N'\}} = 1$, then G looks like in Figs. 3 and 4.

With the help of the graph defined in Definition 2.1 we want to tackle the following biological problem:

Question 2.6. How should the set of all available ℓ-tuples (ℓ-plets) be partitioned into a given number of disjoint subsets corresponding to the amino acids to be encoded such that the influence of single point mutations is minimal?

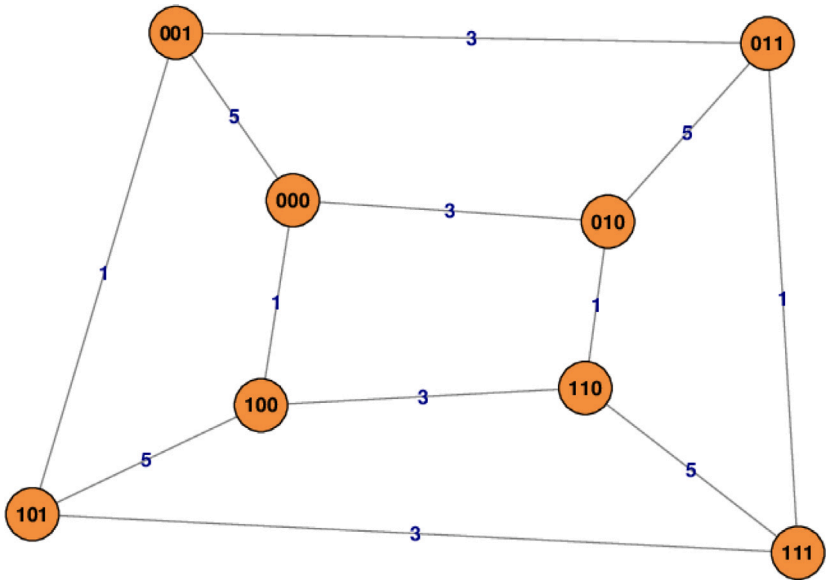


Fig. 1. The graph G for $\Sigma = \{0,1\}$, $l = 3$ and weights $p_1^{\{0,1\}} = 1$, $p_2^{\{0,1\}} = 3$ and $p_3^{\{0,1\}} = 5$.

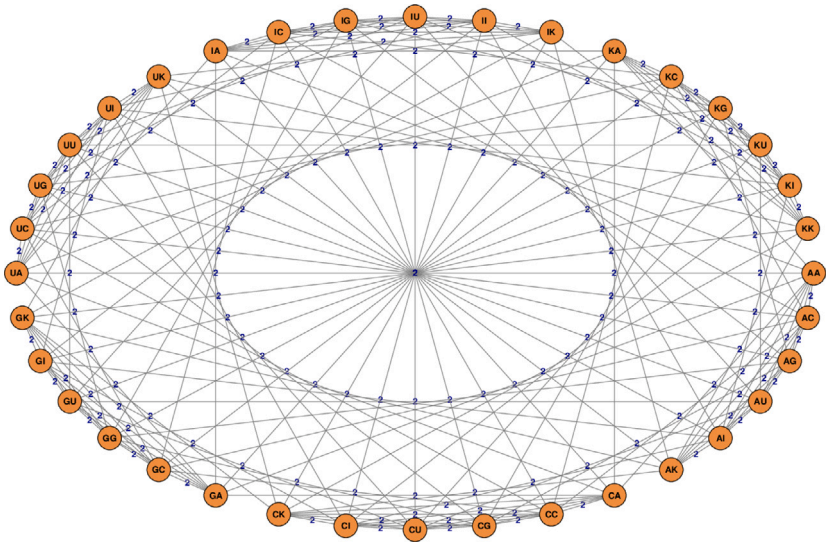


Fig. 2. The graph G for $\Sigma = \{A, C, G, U, I, K\}$, $\ell = 2$ and all weights $p_i^{\{N,N'\}} = 2$. Note that in the centre there is no vertex.

We approach the problem as an optimal clustering problem in a graph. First, we will explain what we mean by partitioning C_k the set V of nodes of the graph G into a fixed number $1 < k \leq (2n)^\ell$ of disjoint non-empty subsets C_k, i.e. k ℓ-letter-word groups:

$$C_k = \{S_1, S_2, \ldots, S_k : \ S_i \cap S_j = \emptyset for all i \neq j \leq k, \ S_1 \cup S_2 \cup \cdots \cup S_k = V\}.$$

The biological interpretation of such a partition is obvious: ℓ-tuples of each subset S_i in the partition synonymously encode one and the same amino acid. The way to measure the "quality" of such a partition with the help of the so-called conductance measure as proposed in Blazej et al. (2018) seems to fit very naturally to the biological task. However, there are easily different ways to define conductance for a given subset of the nodes of a graph.[1] Following Alev et al. (2017), we first define the conductance for a single subset S of V and thus adapt the corresponding definition from Blazej et al. (2019) to the new weighted situation:

Definition 2.7. For a given weighted graph $G = G(V, E, w)$ let S be a subset of $V = \Sigma^\ell$ where $\ell \in \mathbb{N}$. We define the conductance of S as:

$$\phi(S) = \frac{w(E(S, \overline{S}))}{\sum_{c \in S, (c,c') \in E} w((c, c'))}$$

where $w(E(S, \bar{S}))$ is the sum of the weights of edges of G crossing from S to its complement $\bar{S}$:

$$E(S, \bar{S}) := \{(c, c') \in E : |\{c, c'\} \cap S| = 1\} \quad and$$
$$w(E(S, \bar{S})) = \sum_{(c,c') \in E(S,\bar{S})} w((c, c')).$$

Remark 2.8. It is easy to see that multiplying by a positive constant $\lambda > 0$ of all edge weights does not change the conductance of any subset

[1] For example, in a widely used variant of the definition, the conductances of a subset $S \subseteq V$ and of its complement $\bar{S} \subseteq V$ are always equal (Kannan et al., 2004; Gaertler, 2005). Because of our biological motivation, it would not make sense to adopt this definition, since in particular conductances for S with $|S| = 1$ or $|S| = 2n^\ell - 1$ would be the same.

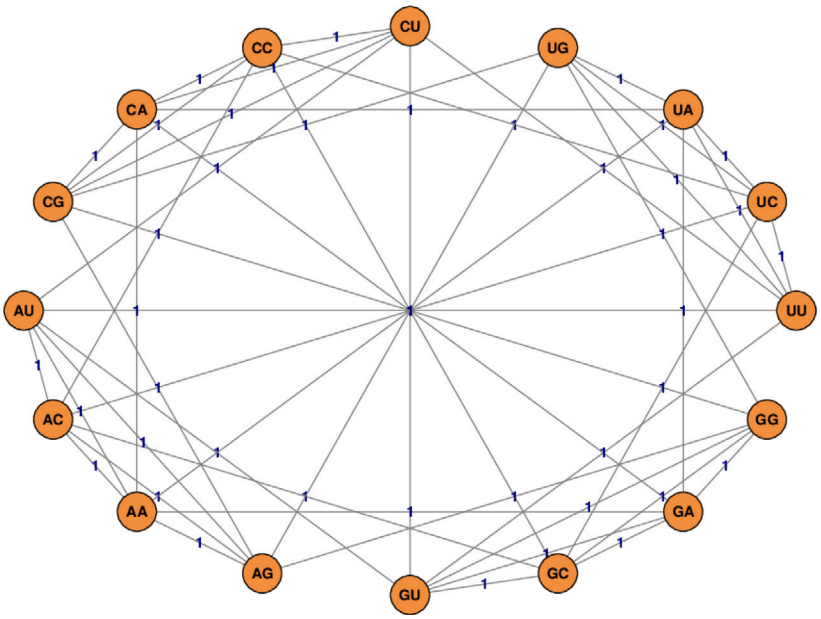


Fig. 3. The graph G for $\Sigma = \mathcal{B}$, $\ell = 2$ and all weights $p_i^{\{N,N'\}} = 1$.

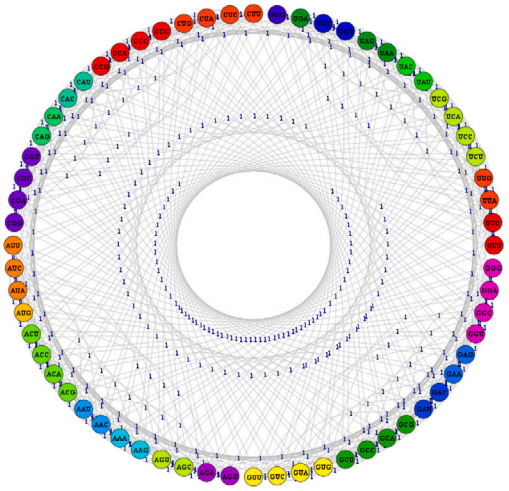


Fig. 4. The graph G for $\Sigma = \mathcal{B}$, $\ell = 3$ and all weights $p_i^{\{N,N'\}} = 1$. Note that in the centre there is no vertex. The colours of the nodes represent the amino acids..

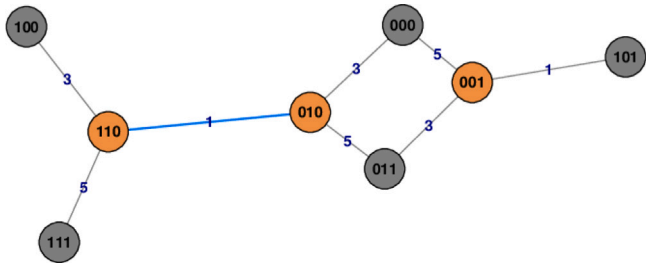


Fig. 5. The graph of S inside the graph of G for $\Sigma = \{0,1\}$, $\ell = 3$, $S = \{010, 110, 001\}$ and weights $p_1^{\{0,1\}} = 1$, $p_2^{\{0,1\}} = 3$ and $p_3^{\{0,1\}} = 5$. $\Phi(S) = 25/27$ where 25 is the sum of the weights of crossing edges and 27 the total sum. Note that the internal edge (blue) is counted twice by definition.

$S \subseteq V$, since the numerator and denominator in the expression from Definition 2.7 are multiplied by the same constant $\lambda > 0$.

Before we give some examples we have a first easy lemma (recall that $|\Sigma| = 2n$).

Lemma 2.9. *Assume that $G = G(V, E, w)$ is a weighted graph as in Definition 2.1 and assume that all weights depend only on the position and not on the letters, i.e. $p_i^{\{N,N'\}} = p_i$ for all $N \neq N' \in \Sigma$. Then we have*

$$\phi(S) = \frac{w(E(S, \overline{S}))}{(2n-1)|S| \sum_{i=1}^{\ell} p_i}.$$

Proof. Easy. □

In the biological setting, $\phi(S)$ has a very interesting interpretation. Assuming that all codons belonging to S encode the same label, i.e. the same amino acid or the stop coding signal, then $\phi(S)$ is the ratio of the total number of non-synonymous single nucleotide substitutions to all possible nucleotide substitutions generated by the codons from S together. We now give some examples.

Example 2.10. Let $\Sigma = \{0,1\}$ be the binary alphabet and $\ell = 3$. Moreover, choose $p_1^{\{0,1\}} = 1$, $p_2^{\{0,1\}} = 3$ and $p_3^{\{0,1\}} = 5$ (see Example 2.3 and Fig. 5) and $S = \{010, 110, 001\}$. Then $\Phi(S) = 25/27 = 0.9259259$.

We continue with a second and third example.

Example 2.11. Let $\Sigma = \{A, C, G, U, I, K\}$ be an extended genetic alphabet, $\ell = 3$. and all $p_i^{\{N,N'\}} = 2$ (see Example 2.4 and Fig. 6). Moreover choose $S = \{ACG, AIG, KIG\}$ then $\Phi(S) = 0.9111111$.

Example 2.12. Let $\Sigma = \mathcal{B} = \{A, C, G, U\}$ be the genetic alphabet, $\ell = 3$ and choose $p_1^{\{N,N'\}} = p_2^{\{N,N'\}} = 1$ for all $N, N' \in \Sigma$ as well as $p_3^{\{U,G\}} = p_3^{\{A,C\}} = 2$ and $p_3^{\{U,C\}} = p_3^{\{A,G\}} = 4$. Moreover, choose $S = \{ACU, ACC, ACA, ACG\}$, then $\Phi(S) = 0.5333333$. See also Fig. 7.

Now that we have clarified what the conductance of a selected subset of the potential codewords means, the next step is to define the conductance for an entire partition of the set of all potential codewords. A reasonable way to do this, which was also adopted in Blazej et al. (2019) and Blazej et al. (2018), was proposed in Lee et al. (2014). The conductance of a partition is defined as the conductance of the "weakest link in the chain":

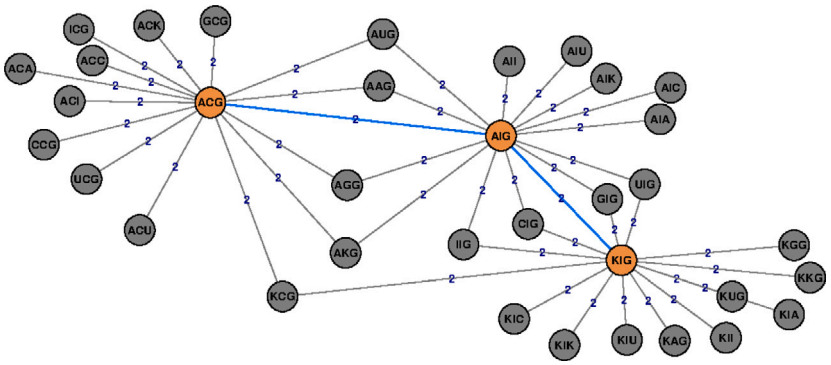


Fig. 6. The graph of S inside the graph of G for $\Sigma = \{A, C, G, U, I, K\}$, $l = 3$, $S = \{ACG, AIG, KIG\}$ and weights $p_1^{\{N,N'\}} = 2$.

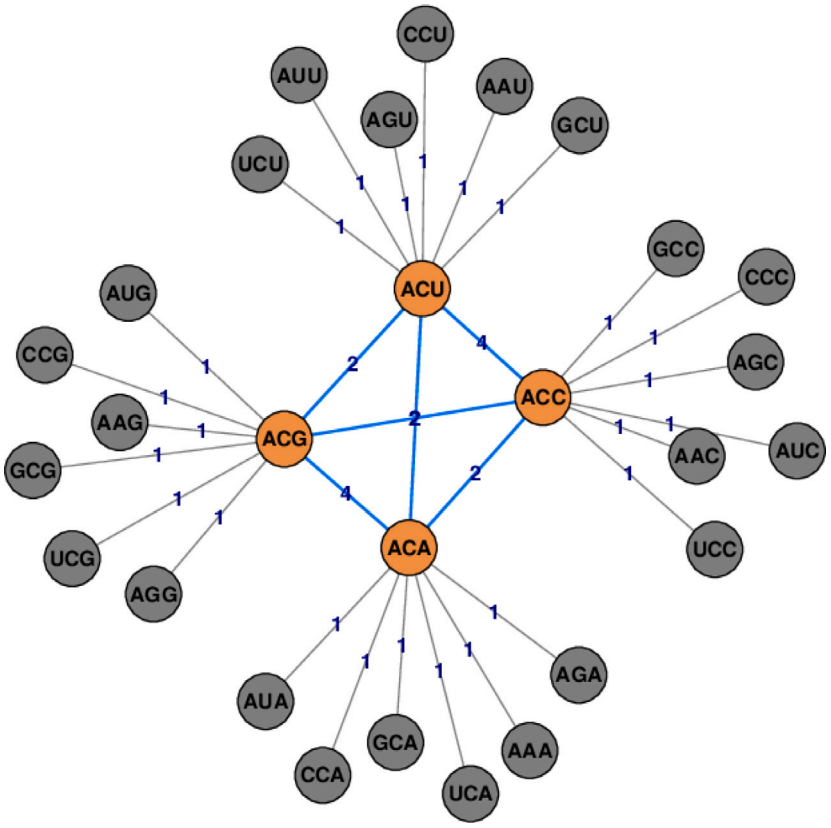


Fig. 7. The graph of S inside the graph G for $\Sigma = \mathcal{B}$, $\ell = 3$, $p_1^{\{N,N'\}} = p_2^{\{N,N'\}} = 1$ for all $N, N' \in \Sigma$, $p_3^{\{U,G\}} = p_3^{\{A,C\}} = 2$ and $p_3^{\{U,C\}} = p_3^{\{A,G\}} = 4$ as well as $S = \{ACU, ACC, ACA, ACG\}$.

Definition 2.13. For a given weighted graph $G = G(V, E, w)$ and an integer $\ell \in \mathbb{N}$ the conductance of a partition C_k of $V = \Sigma^\ell$ is defined as

$$\Phi(C_k) = \max_{S \in C_k} \phi(S).$$

Therefore, the Φ measure gives us a characterization of the quality of a given partition C_k as the set conductance of the worst l-letter group in this partition. What is more, Φ measure involves a question about the structure of the optimal graph partition C_k for a fixed k. In this context, the best partition C_k of the graph G in terms of Φ follows in a natural way and is given by the formula:

$$\Phi_{\min}(k) = \min_{C_k} \Phi(C_k),$$

i.e. we minimize over all possible k-partitions of V with respect to the conductance measure.

It can also be useful in certain cases to take a different conductance definition for a given partition, namely that of the average partition conductance. This can make sense, for example, if the partition is to have some "inconvenient" predefined properties, e.g. that an amino acid (Met) is to be encoded by exactly one tuple, as in the standard genetic code. In this case $\Phi(C_k) = 1$ is always true.

Definition 2.14. For a given weighted graph $G = G(V, E, w)$ and an integer $\ell \in \mathbb{N}$ the average conductance of a partition C_k is defined as

$$\bar{\Phi}(C_k) = \frac{1}{k} \sum_{S \in C_k} \phi(S).$$

Just as in the above case, one can search for an optimal partition of the set Σ^ℓ into k subsets such that the average conductance of the partition is smallest among all partitions of Σ^ℓ in k subsets:

$$\bar{\Phi}_{\min}(k) = \min_{C_k} \bar{\Phi}(C_k).$$

Table 1
Code Variants: the average conductance of the partition ($\bar{\Phi}(C_k)$) as well as the range of conductance values of the single subsets of synonymous codons ($\phi(S)$) are given for different dialects of the genetic code with unique assignment of amino acids to codons. The two variants are calculated on the one hand for the unweighted variant of the graph 2.1 (all weights equal 1), and on the other hand for the weighted variant from Section 3.2.

Name	Unweighted			$p_1 = p_2 = 1, \tilde{p}_3 = 2$		
	Average	Range min–max		Average	Range min–max	
The Standard Code	0.8113	0.6667	1	0.6395	0.4286	1
The Vertebrate Mitochondrial Code	0.8114	0.6667	0.8889	0.6169	0.4286	0.7143
The Yeast Mitochondrial Code	0.8183	0.6667	0.8889	0.6259	0.4286	0.7143
The Mould, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code	0.8078	0.6667	1	0.6236	0.4286	1
The Invertebrate Mitochondrial Code	0.8042	0.6667	0.8889	0.6077	0.4286	0.7143
The Ciliate, Dasycladacean and Hexamita Nuclear Code	0.8131	0.6667	1	0.6474	0.4286	1
The Echinoderm and Flatworm Mitochondrial Code	0.8042	0.6667	1	0.6259	0.4286	1
The Euplotid Nuclear Code	0.8078	0.6667	1	0.6327	0.4286	1
The Bacterial, Archaeal and Plant Plastid Code	0.8113	0.6667	1	0.6395	0.4286	1
The Alternative Yeast Nuclear Code	0.8183	0.6667	1	0.65	0.4286	1
The Ascidian Mitochondrial Code	0.8078	0.6667	0.8889	0.6145	0.4286	0.7143
The Alternative Flatworm Mitochondrial Code	0.8042	0.6667	1	0.6349	0.4286	1
Ter Chlorophycean Mitochondrial Code	0.8138	0.6667	1	0.65	0.4286	1
Trematode Mitochondrial Code	0.8042	0.6667	1	0.6168	0.4286	1
Scenedesmus obliquus Mitochondrial Code	0.826	0.6667	1	0.672	0.4286	1
Thraustochytrium Mitochondrial Code	0.8113	0.6667	1	0.6395	0.4286	1
Pterobranchia Mitochondrial Code	0.8047	0.6667	1	0.6229	0.4286	1
Candidate Division SR1 and Gracilibacteria Code	0.8141	0.6667	1	0.6413	0.4286	1
Pachysolen tannophilus Nuclear Code	0.8198	0.6667	1	0.6522	0.4286	1
Mesodinium Nuclear	0.8078	0.6667	1	0.6372	0.4286	1
Peritrich Nuclear	0.8131	0.6667	1	0.6474	0.4286	1

It is easy to see that for each partition C_k

$$\bar{\Phi}(C_k) \leq \Phi(C_k)$$

applies and, thus, it follows that for every $k \in \{1, \dots, |V| = (2n)^\ell\}$

$$\bar{\Phi}_{\min}(k) \leq \Phi_{\min}(k)$$

takes place.

In Table 1 the average conductance of the partition as well as the range of conductance values of the single subsets of synonymous codons are given for different dialects of the genetic code with unique assignment of amino acids to codons. The two variants are calculated on the one hand for the unweighted variant of the graph 2.1 (all weights equal 1), and on the other hand for the weighted variant taking into account the wobble effect as will be considered in Section 3.2 of this article.

3. Results and discussion

3.1. Calculating lower bounds for conductances

One of the most important goals of the work is to find optimal partitions of the set of all available coding tuples into a given number of subsets for which the maximum conductance of subsets or the average of their conductances is minimal (compare Question 2.6). For this purpose, it is very useful to have lower bounds on the minimum or average conductance possible for a partition. In order to simplify statements we will assume in this subsection that all the weights of the graphs from Definition 2.1 are independent of the letters, i.e. $p_i^{\{N,N'\}} = p_i$ for all i and $N \neq N'$.

One of the most important results presented in Lee et al. (2014) is a generalization of the much noted Cheeger's inequality (see, for example, Alev et al., 2017; Gaertler, 2005; Kwok et al., 2013) to Higher-Order Cheeger Inequalities. Applied to our case using the definitions given above (the authors of Lee et al., 2014 partly use other terms and notations) it reads as follows. Again we refer to Clark and Holton (1991) for details on graph theory:

Higher-Order Cheeger Inequalities 3.1. *Let $G = (V, E)$ be an undirected, d-regular graph, with positive weights $w : E \to (0, \infty)$ on the edges. Consider the normalized Laplacian matrix of G defined by*

$$L = I - \frac{1}{d}A,$$

where A is the adjacency matrix of G[2] and d is the common node degree, and its eigenvalues in ascending order

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{|V|}.$$

Then for all $k \in \{1, \dots, |V|\}$ the following inequality holds:

$$\frac{\lambda_k}{2} \leq \Phi_{min}(k) \leq O(k^2)\sqrt{\lambda_k}. \tag{3.1}$$

The question that naturally arises is whether the lower bound presented is tight. Unfortunately, it turns out that it is not:

Proposition 3.2. *The lower bound in the Higher-Order Cheeger Inequalities 3.1 is not tight.*

Proof. Consider the following example:

Let $n = 2$ and $\Sigma = \mathcal{B} = \{A, C, G, U\}$ be the genetic alphabet, $\ell = 3$ and $p_i = 1$ for all $i \in \{1, \dots, \ell\}$. Then the adjacency matrix of G is displayed in Fig. 8 (see Table 2).

In this case we have:

$$\lambda_1 = 0$$
$$\lambda_k = \frac{4}{9} \qquad k = 2, \dots, 10$$

[2] The adjacency matrix of a weighted graph $G = (V, E)$ is a square $|V| \times |V|$-matrix which is used to represent edges of a graph. We put the corresponding weight as the value at the intersection of the column and the row corresponding to the vertices connected or 0 if the vertices are not connected.

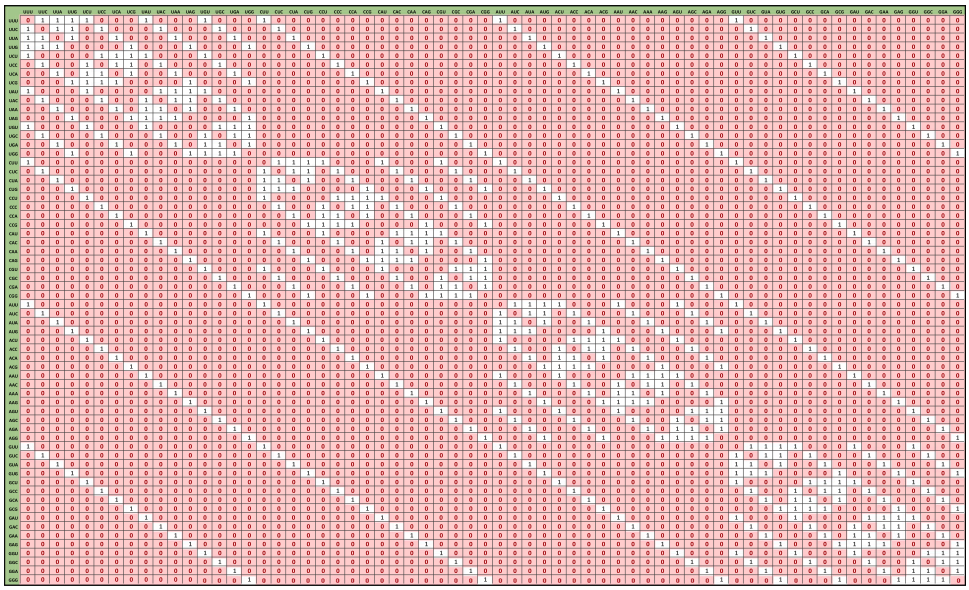


Fig. 8. The adjacency matrix of G.

Table 2
Lower bounds for $\Phi_{min}(k)$ in dependence on the number of classes k ($n = 2, \ell = 3$, see Blazej et al. (2019)) where the numbers marked in red were proved to be tight lower bounds.

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Lower bound	0	$\frac{2}{9}$	$\frac{73}{207}$	$\frac{1}{3}$	$\frac{4}{9}$	$\frac{8}{15}$	$\frac{5}{9}$	$\frac{5}{9}$	$\frac{13}{21}$	$\frac{2}{3}$	$\frac{31}{45}$	$\frac{31}{45}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$
k	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Lower bound	$\frac{7}{9}$	$\frac{7}{9}$	$\frac{7}{9}$	$\frac{7}{9}$	$\frac{7}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{8}{9}$

$$\lambda_k = \frac{8}{9} \qquad k = 11, \dots, 38$$
$$\lambda_k = \frac{4}{3} \qquad k = 39, \dots, 64$$

and, thus, on the one hand, according to the Higher-Order Cheeger Inequalities 3.1

$$\Phi_{min}(1) \geq 0$$
$$\Phi_{min}(k) \geq \frac{2}{9} \qquad k = 2, \dots, 10$$
$$\Phi_{min}(k) \geq \frac{4}{9} \qquad k = 11, \dots, 38$$
$$\Phi_{min}(k) \geq \frac{2}{3} \qquad k = 39, \dots, 64$$

On the other hand, according to the results from Blazej et al. (2019) we have the following lower bounds where the numbers marked in red were proved to be tight lower bounds:

For $k \geq 33$, there must always be at least one one-element subset among the subsets of a partition and thus for this case $\Phi_{min}(k) = 1$ applies. A look at the table above shows that the lower bound in Higher-Order Cheeger's Inequalities is tight only for the two trivial cases $k = 1, 2$. □

The approach from Blazej et al. (2019) for calculating lower bounds for $\Phi_{min}(k)$ obviously yields better results than the Higher-Order Cheeger's Inequalities. Therefore, we will now generalize the corresponding result from Blazej et al. (2019) and adapt it for the case of the weighted graph and arbitrary alphabet and tuple sizes.

In Bezrukov (1999) it was proved that for the unweighted graphs (all weights equal to 1) and a given subset size, the subsets that are lexicographically ordered according to a freely chosen order in the alphabet have the smallest conductance among all subsets of the same cardinality. When weights come into play, this property no longer applies, as the following example shows:

Example 3.3. Let $\ell = 3$, $p_1 = 3, p_2 = 2, p_3 = 1$, and the order of $\mathcal{B}$ be defined as $A < C < G < T$. If we now consider $S = \{AAA, AAC, AAG\}$ *and* $S' = \{AAA, CAA, GAA\}$, it is clear that S is lexicographically ordered with respect to the chosen order, while S' is not. Nevertheless, we have

$$\phi(S') = \frac{3 \cdot (3 \cdot 1 + 3 \cdot 2 + 3)}{3 \cdot (3 \cdot 1 + 3 \cdot 2 + 3 \cdot 3)} = \frac{2}{3} < \phi(S) = \frac{3 \cdot (3 \cdot 2 + 3 \cdot 3 + 1)}{3 \cdot (3 \cdot 1 + 3 \cdot 2 + 3 \cdot 3)} = \frac{8}{9}$$

However, in the case where the weights depend only on the position, one can arrange the positions so that the weights are ascending. Then the above property applies again. For the convenience of the reader we state the result without proof (see Propositions 1 and 2 in Blazej et al., 2018 and Theorem 1 in Bezrukov and Elsässer).

Theorem 3.4. *Let $G = G_l^P(V, E, w)$ be an undirected weighted graph as in Definition 2.1. Let $N_1 < \cdots < N_{2n}$ be an ordering of the alphabet Σ and let $P = \{p_i : i = 1, \dots, \ell\}$ be the set of weights. Then there is a permutation $\pi : \{1, \dots, \ell\} \to \{1, \cdots, \ell\}$ such that the weights $\langle p_{\pi(i)} : i = 1, \dots, \ell \rangle$ are increasing. Let π also denote the induced mapping $\Sigma^\ell \to \Sigma^\ell$ sending a word $a_1 \cdots a_\ell \in \Sigma^\ell$ to $a_{\pi(1)} \cdots a_{\pi(\ell)}$. Then for any m the set S_m has the smallest conductance $\Phi(S)$ among all subsets $S \subseteq \Sigma^\ell$ of size m where $\pi(S_m)$ is the set of the first m words in the lexicographic order induced by $<$ on Σ^ℓ*

Just to illustrate Theorem 3.4 we note that in Example 3.3 one can choose π to be $\pi : \{1, 2, 3\} \to \{1, 2, 3\}$ with $\pi(1) = 3, \pi(2) = 2, \pi(3) = 1$. Then $\pi(S') = S$ which contains the first three codons in lexicographical order.

It is worth mentioning that the above Theorem 3.4 justifies to restrict ourselves in general to the case when the weights are ascending. The following theorem is helpful, since it offers the possibility of recursively determining the minimum possible conductances for subsets of Σ^ℓ of arbitrary size. The following result generalizes Theorem 3.2. from Blazej et al. (2019) to our weighted situation.

Theorem 3.5. *Let $G = G_l^P(V, E, w)$ be an undirected and weighted graph as in Definition 2.1, $p_i \geq 0$, $i = 1, \dots, \ell$ the corresponding edge weights, $N_1 < N_2 < N_3 < \cdots < N_{2n}$ a linear order on the alphabet Σ and $S_m \subseteq V$ the collection of the first $m = 1, 2, \dots, (2n)^\ell$ vertices of the graph G in the lexicographic order induced by $<$. Then the following recursive formula for the sum of the weights of edges of G crossing from S_m to its complement $\bar{S}_m$ holds:*

$$w(E(S_{m+1}, \overline{S_{m+1}})) = w(E(S_m, \overline{S_m})) + \sum_{i=1}^{\ell} p_i(2(n - m_i) - 1),$$

$$w(E(S_1, \overline{S_1})) = (2n-1)\sum_{i=1}^{\ell} p_i$$

where $(m_1, m_2, \dots, m_\ell)_{2n}, m_i \in \{0, 1, 2, \dots, (2n-1)\}$[3] *is the* $2n$*-adic representation of* m *to base* $2n$*, i.e.*

$$m = m_1 \cdot (2n)^{\ell-1} + m_2 \cdot (2n)^{\ell-2} + \cdots + m_\ell \cdot (2n)^0.$$

The conductance of S_m *is accordingly equal to*

$$\phi(S_m) = \frac{w(E(S_m, \overline{S_m}))}{(2n-1) \cdot m \cdot \sum_{i=1}^{\ell} p_i}.$$

Proof. It is clear that $w(E(S_1, \overline{S_1})) = (2n-1)\sum_{i=1}^{\ell} p_i$ since the graph G from Definition 2.1 is $\ell(2n-1)$-regular and for every $i = 1, \dots, \ell$, $(2n-1)$ edges are assigned the weight p_i respectively.

Let us assume now that we already have calculated $w(E(S_m, \overline{S_m}))$ for $m \geq 1$ and we are inserting now the next ℓ-letter word $c \in \Sigma^\ell$ in the lexicographic order. It is easy to see that all ℓ-letter words over Σ ordered in lexicographic order can be rewritten as a sequence of consecutive ℓ-digits numbers to the base $2n$ if we assign, for example, $N_1 \to 0, N_2 \to 1, N_3 \to 2, N_4 \to 3 \ldots, N_{2n} \to 2n-1$. Therefore, newly included ℓ-letter word c has a numeric representation $c = (m_1, m_2, \dots, m_\ell)_{2n}$, where $m_i = 0, 1, 2, \dots, 2n-1$. What is more, m_i, $i = 1, 2, \dots, \ell$ is the number of ℓ-letter words over Σ that differ from c at the position i which are smaller than c in the lexicographic order and the total weight of edges crossing S_m and c is, consequently, equal to $w(E(S_m, \{c\})) = \sum_{i=1}^{\ell} m_i p_i$. As a result, the total sum of the weights of edges between S_{m+1} to its complement fulfils the equation:

$$w(E(S_{m+1}, \overline{S_{m+1}})) = w(E(S_m, \overline{S_m})) - w(E(S_m, \{c\})) + (2n-1)\sum_{i=1}^{\ell} p_i - w(E(S_m, \{c\})) =$$

$$w(E(S_m, \overline{S_m})) + \sum_{i=1}^{\ell} p_i(2(n-m_i)-1).$$

That completes the proof. □

Corollary 3.6. *Let* $G = G_l^P(V, E, w)$ *be an undirected and weighted graph as in Definition 2.1,* $p_i = 1$ *for all* $i = 1, \dots, \ell$, $N_1 < N_2 < N_3 < \cdots < N_{2n}$ *a linear order on the alphabet* Σ *and* $S_m \subseteq V$ *the collection of the first* $m = 1, 2, \dots, (2n)^\ell$ *vertices of the graph* G *in the lexicographic order induced by* $<$*. Then the following recursive formula for the number of edges of* G *crossing from* S_m *to its complement* $\bar{S}_m$ *holds:*

$$w(E(S_{m+1}, \overline{S_{m+1}})) = w(E(S_m, \overline{S_m})) + \ell \cdot (2n-1) - 2 \cdot (m_1 + m_2 + \cdots + m_\ell),$$
$$w(E(S_1, \overline{S_1})) = \ell \cdot (2n-1)$$

where $(m_1, m_2, \dots, m_\ell)_{2n}, m_i \in \{0, 1, 2, \dots, (2n-1)\}$ *is the representation of* m *to base* $2n$*, i.e.*

$$m = m_1 \cdot (2n)^{\ell-1} + m_2 \cdot (2n)^{\ell-2} + \cdots + m_\ell \cdot (2n)^0.$$

The conductance of S_m *is accordingly equal to*[4]

$$\phi(S_m) = \frac{w(E(S_m, \overline{S_m}))}{\ell \cdot (2n-1) \cdot m}.$$

We now derive an explicit presentation of $w(E(S_m, \overline{S_m}))$

Proposition 3.7. *Let* $G = G_l^P(V, E, w)$ *be an undirected and weighted graph as in Definition 2.1,* $p_i \geq 0$, $i = 1, \dots, \ell$ *the corresponding edge weights,* $N_1 < N_2 < N_3 < \cdots < N_{2n}$ *a linear order on the alphabet* Σ *and* $S_m \subseteq V$ *the collection of the first* $m = 1, 2, \dots, (2n)^\ell$ *vertices of the graph* G *in the lexicographic order induced by* $<$*. Then*

$$w(E(S_m, \overline{S_m})) := m \cdot \sum_{i=1}^{\ell} p_i \cdot \lambda_i$$

with

$$\lambda_i := \frac{m_i \cdot (2n)^{\ell-i}}{m}(2n - m_i) + \frac{(m \bmod (2n)^{\ell-i})}{m}(2(n-m_i)-1)$$

where $(m_1, m_2, \dots, m_\ell)_{2n}, m_i \in \{0, 1, 2, \dots, (2n-1)\}$ *is the representation of* m *to base* $2n$*, i.e.*

$$m = m_1 \cdot (2n)^{\ell-1} + m_2 \cdot (2n)^{\ell-2} + \cdots + m_\ell \cdot (2n)^0.$$

and hence

$$\phi(S_m) = \frac{\sum_{i=1}^{\ell} p_i \cdot \lambda_i}{(2n-1) \cdot \sum_{i=1}^{\ell} p_i}.$$

Proof. See in Appendix A.1 □

The above results and formulas can be used to determine conductances algorithmically using computer programmes. We will give some optimization algorithms applied to the genetic code setting in a sequel paper (Fimmel et al., 2021). However, we here deal with a special weight distribution for the genetic code setting that sheds some light on the Wobble-effect in the next Section 3.2.

The conductance values tend to decrease as a function of k, i.e. the larger the subset cardinality, the smaller its conductance value. Therefore in the following proposition $\Phi_{min}(k)$ can be estimated downwards as follows:

Proposition 3.8. *Let* C_k *be a partition of a graph* $G = G_l^P(V, E, w)$ *as in Definition 2.1 where* $k \in \mathbb{N}$*, i.e. with* k *classes and* $0 < p_1 \leq p_2 \leq \cdots \leq p_\ell$*. Moreover, let* $N_1 < N_2 < N_3 < \cdots < N_{2n}$ *be a linear order on the alphabet* Σ *and* $S_m \subseteq V$ *the collection of the first* $m = 1, 2, \dots, (2n)^\ell$ *vertices of the graph* G *in the lexicographic order induced by* $<$*. Then we have the following lower boundary for the conductance of the partition* C_k *:*

$$\Phi(C_k) \geq \phi(S_{m_k}) \text{ with } m_k = (2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor} \cdot \left\lfloor \frac{(2n)^\ell}{k(2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor}} \right\rfloor$$

Proof. See in Appendix A.2 □

The lower bounds for $\Phi(C_k)$ given by the proposal 3.8 are not tight. This is clear from the results from Blazej et al. (2019). However, the results are generally tighter than the bounds from the Higher-Order Cheeger Inequalities 3.1

3.2. Mirrowing the Wobble effect

In this section we are interested in modelling the Wobble-effect by our weighted graph approach from Definition 2.1 and therefore let $\Sigma = \mathcal{B} = \{A, C, G, U\}$ be the genetic alphabet for the rest of this section.

It is well known that the translation of each codon requires a corresponding tRNA molecule with which it can complement via the canonical Watson–Crick pairings A and U or C and G. However, most organisms have only a few more than 40 tRNA types, so some tRNA types must necessarily match several codons that all encode the same amino acid. In 1966, Francis Crick therefore proposed the Wobble hypothesis as a possible solution to this problem. He suggested that the 5' base on the anticodon, which binds to the 3' base on the mRNA, is less spatially restricted than the other two bases are. This, he believed, affected the pairing geometry of the tRNA, allowing for non-standard base pairing. So according to Crick, the first two bases form strong

[3] In the formula, we need for calculations always the 'previous' m. For instance, for calculation of $w(E(S_{(2n)^\ell}, \bar{S}_{(2n)^\ell}))$ we need $m = (2n)^\ell - 1$. This is why we can always represent m as a ℓ-digit number to base $2n$.

[4] It suffices if we calculate minimal conductances for $1 \leq m \leq \frac{(2n)^\ell}{2}$ since in the case of at least one partitioning of Σ^ℓ into at least 2 subsets the size of one of them will be at most $\frac{(2n)^\ell}{2}$.

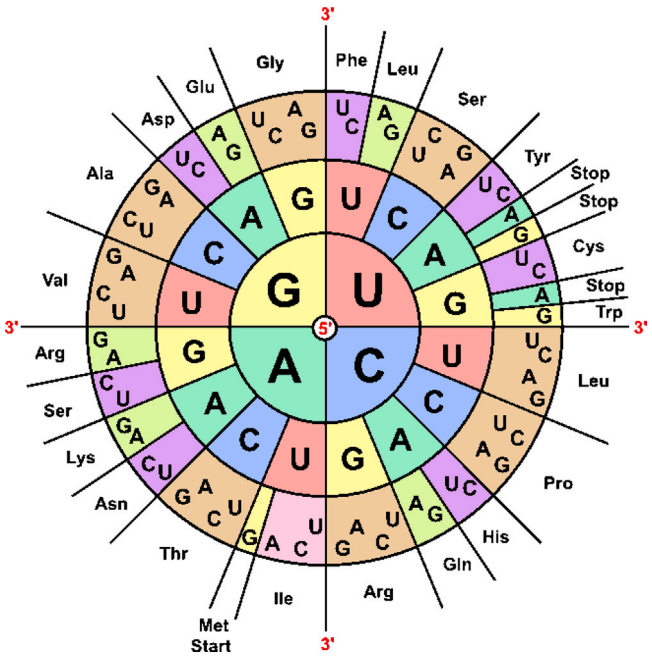


Fig. 9. The standard RNA codon table organized in a wheel. *Source:* https://commons.wikimedia.org/wiki/File:Codons_aminoacids_table.png.

Watson–Crick bonds and therefore bind strongly with the anticodon on the tRNA. However, the third position also allows bonds of the form uracil and guanine. In addition, there is the base hypoxanthine (I, named after the associated nucleoside inosine), which does not occur in mRNA or DNA, but can be incorporated into tRNAs at the wobble position (Alseth et al., 2014). It allows binding to adenine, uracil and cytosine.

According to the above it is visible in the standard genetic code (see the codon wheel in Fig. 9) that mostly amino acids that have a degeneracy of four are encoded by codons of the form XYA, XYC, XYG, XYU and those of degeneracy two are encoded by codons that differ in the third base only and are grouped NOT according to the Wobble base pairs, i.e. XYU and XYC together as well as XYA and XYG together. This aspect has not been considered in the conductance approach so far. In fact, it was shown in Blazej et al. (2019) that the best model of a codon amino-acid assignment with respect to maximal conductance and average conductance at the same time consists of groups of size three mostly. It is displayed in Fig. 10.

The optimality with respect to average conductance is surprising since if one remembers that the first k codons in the lexicographic order provide a set with minimal k-conductance (see Theorem 3.4), it might be expected that a set of two codons and a set of four codons (in the lexicographic order) instead of two sets of size three have a better average conductance. (Recall that degeneracy two and four appear much more often in the standard genetic code than degeneracy three). However, we prove next that indeed the average conductance is the same even in the weighted approach.

Proposition 3.9. *Let $\mathcal{B}$ be the standard genetic alphabet, $G = G_l^P(V, E, w)$ be an undirected weighted graph as in Definition 2.1 and $p_i \geq 0$, $i = 1, \ldots, 3$ the corresponding edge weights in ascending order. Let S_m be the collection of the first m vertices of G in the lexicographic order. Then the average conductance of S_2 and S_4 is the same as the conductance of S_3, i.e.*

$$\frac{1}{2}(\phi(S_2) + \phi(S_4)) = \phi(S_3)$$

	T	C	A	G	
T	TTT	TCT	TAT	TGT	T
T	TTC	TCC	TAC	TGC	C
T	TTA	TCA	TAA	TGA	A
T	TTG	TCG	TAG	TGG	G
C	CTT	CCT	CAT	CGT	T
C	CTC	CCC	CAC	CGC	C
C	CTA	CCA	CAA	CGA	A
C	CTG	CCG	CAG	CGG	G
A	ATT	ACT	AAT	AGT	T
A	ATC	ACC	AAC	AGC	C
A	ATA	ACA	AAA	AGA	A
A	ATG	ACG	AAG	AGG	G
G	GTT	GCT	GAT	GGT	T
G	GTC	GCC	GAC	GGC	C
G	GTA	GCA	GAA	GGA	A
G	GTG	GCG	GAG	GGG	G

Fig. 10. A model of the genetic code that has the minimal possible maximal conductance $\Phi_{min}(21) = 0.6667$ and at the same time the best average conductance $\tilde{\Phi}_{min}(21) = 0.7724868$ - see Blazej et al. (2019). The 21 classes corresponding to the 20 canonical amino acids and a stop signal are distributed into 20 classes of size three and one of size four.

Proof. According to the recursive formula for the conductance (see Theorem 3.5) it is readily seen that we have

$$\phi(S_2) = \frac{2 \cdot (3 \cdot p_1 + 3 \cdot p_2 + 2 \cdot p_3)}{6 \cdot (p_1 + p_2 + p_3)} = \frac{(3 \cdot p_1 + 3 \cdot p_2 + 2 \cdot p_3)}{3 \cdot (p_1 + p_2 + p_3)}$$

$$\phi(S_3) = \frac{3 \cdot (3 \cdot p_1 + 3 \cdot p_2 + p_3)}{9 \cdot (p_1 + p_2 + p_3)} = \frac{(3 \cdot p_1 + 3 \cdot p_2 + p_3)}{3 \cdot (p_1 + p_2 + p_3)}$$

$$\phi(S_4) = \frac{4 \cdot (3 \cdot p_1 + 3 \cdot p_2)}{12 \cdot (p_1 + p_2 + p_3)} = \frac{(3 \cdot p_1 + 3 \cdot p_2)}{3 \cdot (p_1 + p_2 + p_3)}$$

Thus it follows that

$$\frac{1}{2}(\phi(S_2) + \phi(S_4)) = \frac{1}{2}\left(\frac{(3 \cdot p_1 + 3 \cdot p_2 + 2 \cdot p_3)}{3 \cdot (p_1 + p_2 + p_3)} + \frac{(3 \cdot p_1 + 3 \cdot p_2)}{3 \cdot (p_1 + p_2 + p_3)}\right) =$$

$$= \frac{1}{2}\left(\frac{(3 \cdot p_1 + 3 \cdot p_2 + 2 \cdot p_3) + (3 \cdot p_1 + 3 \cdot p_2)}{3 \cdot (p_1 + p_2 + p_3)}\right) =$$

$$= \frac{1}{2}\left(\frac{2 \cdot (3 \cdot p_1 + 3 \cdot p_2 + p_3)}{3 \cdot (p_1 + p_2 + p_3)}\right) =$$

$$= \frac{(3 \cdot p_1 + 3 \cdot p_2 + p_3)}{3 \cdot (p_1 + p_2 + p_3)} = \phi(S_3) \quad \square$$

The above result explains why the standard genetic code *looses* against the code from Fig. 10 in the following sense. One could wonder if the model from Fig. 10 could be improved by replacing two classes of size three by one of size two and one of size four since it might be the case that the average conductance then gets better because although the set of size two has a worse conductance, the set of size four might compensate this. Proposition 3.9 shows that this is not the case. However, as we can see the model from Fig. 10 does not respect the Wobble hypothesis (However, if we allow I to be in the game, then the model is quite good since it groups codons with third bases U, C, A).

Hence we will now consider graphs as in Definition 2.1 fixing two weights p_1 and p_2 and a third constant $\tilde{p}_3$ and defining the weight function w as follows:

Definition 3.10.

(1) $p_1^{\{N,N'\}} = p_1$ for all $N \neq N' \in \mathcal{B}$;
(2) $p_2^{\{N,N'\}} = p_2$ for all $N \neq N' \in \mathcal{B}$;
(3) $p_3^{\{A,C\}} = p_3^{\{A,U\}} = p_3^{\{C,G\}} = p_3^{\{G,U\}} = \tilde{p}_3$;
(4) $p_3^{\{A,G\}} = p_3^{\{C,U\}} = 2\tilde{p}_3$.

The idea here is that now mutations in positions 1 and 2 have a probability that depends on the position only. However, some of the

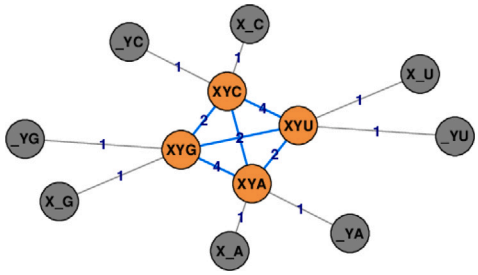


Fig. 11. An example of a subgraph of G from Definition 2.1 with weights $p_1 = p_2 = 1$ and $\tilde{p}_3 = 2$ and $\Sigma = \mathcal{B}$.

edges coming from the third base position that match the Wobble hypothesis get a double weight compared to the other mutations in position 3. An example with $p_1 = p_2 = 1$ and $\tilde{p}_3 = 2$ is illustrated in Fig. 11.

We now calculate the conductance of several sets of codons of sizes at most 6 that correspond to groups of codons that encode the same amino acid in the standard genetic code (see the Codon Wheel in Fig. 9). As remarked above for weights not being in ascending order also here it will turn out that now the first k codons in the lexicographic order do not any longer provide sets with the smallest conductance.

Proposition 3.11. *$G = G_l^P(V, E, w)$ be an undirected weighted graph as in Definition 2.1 and the corresponding weights $p_1 = p_2 = 1$ and $\tilde{p}_3 = 2$. Moreover, let $X, Y \in \mathcal{B}$. Then*

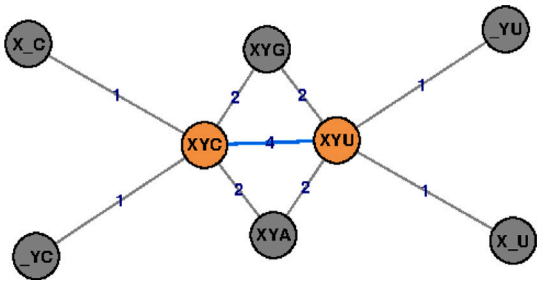
(1) If $S = \{XYZ\}$ for some $Z \in \mathcal{B}$, then $\phi(S) = 1$
(2) If $S = \{XYU, XYC\}$ or $S = \{XYA, XYG\}$, then $\phi(S) = \frac{15}{21}$
(3) If $S = \{XYA, XYC\}$ or $S = \{XYU, XYG\}$ or $\{XYA, XYU\}$ or $\{XYC, XYG\}$, then $\phi(S) = \frac{18}{21}$
(4) If $S = \{XYA, XYC, XYU\}$, then $\phi(S) = \frac{13}{21}$
(5) If $S = \{XYA, XYC, XYG, XYU\}$, then $\phi(S) = \frac{9}{21}$

Moreover,

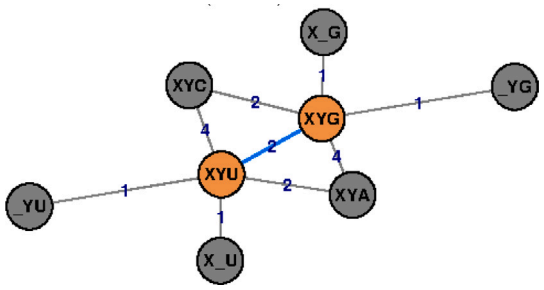
(1) If $S = \{UAA, UAG, UGA\}$, then $\phi(S) = \frac{16}{21}$ (this set corresponds to the Stop signal)
(2) If $S = \{CUA, CUC, CUG, CUU, UUA, UUG\}$, then $\phi(S) = \frac{10}{21}$ (this set corresponds to Leucine)
(3) If $S = \{UCU, UCG, UCA, UCC, AGU, AGC\}$, then $\phi(S) = \frac{11}{21}$ (this corresponds to Serine)
(4) If $S = \{CGA, CGC, CGG, CGU, AGA, AGG\}$, then $\phi(S) = \frac{10}{21}$ (this set corresponds to Arginine)

Proof. For the proof we give the detailed calculation of edges in the nominator and denominators and illustrate it. The value of the denominator is always the size of the set times $(3 + 3 + (2 + 2 + 4))$.

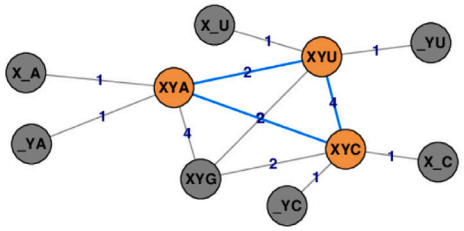
(1) If $S = \{XYZ\}$ for some $Z \in \mathcal{B}$, then $\phi(S) = \frac{3+3+(2+2+4)}{3+3+(2+2+4)} = 1$
(2) If $S = \{XYU, XYC\}$ or $S = \{XYA, XYG\}$, then $\phi(S) = \frac{2\cdot3+2\cdot3+(2+2)+(2+2)}{2\cdot3+2\cdot3+2\cdot(2+2+4)} = \frac{15}{21}$



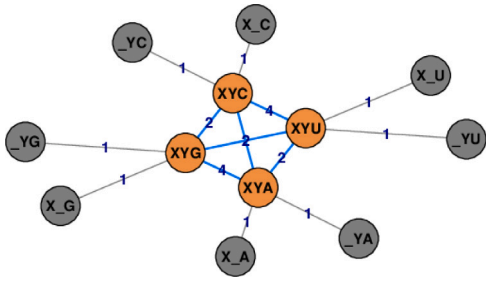
(3) If $S = \{XYA, XYC\}$ or $S = \{XYU, XYG\}$ or $\{XYA, XYU\}$ or $\{XYC, XYG\}$, then $\phi(S) = \frac{2\cdot3+2\cdot3+2\cdot(2+4)}{2\cdot3+2\cdot3+2\cdot(2+2+4)} = \frac{18}{21}$



(4) If $S = \{XYA, XYC, XYU\}$, then $\phi(S) = \frac{3\cdot3+3\cdot3+(2+2+4)}{3\cdot3+3\cdot3+3\cdot(2+2+4)} = \frac{13}{21}$

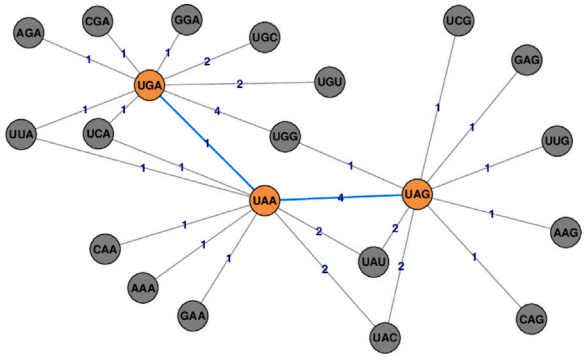


(5) If $S = \{XYA, XYC, XYG, XYU\}$, then $\phi(S) = \frac{4\cdot3+4\cdot3}{4\cdot3+4\cdot3+4\cdot(2+2+4)} = \frac{9}{21}$

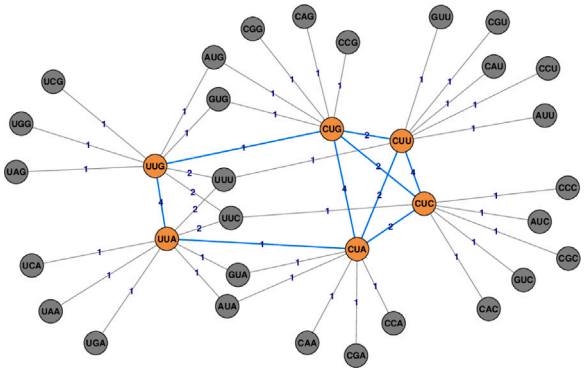


Moreover,

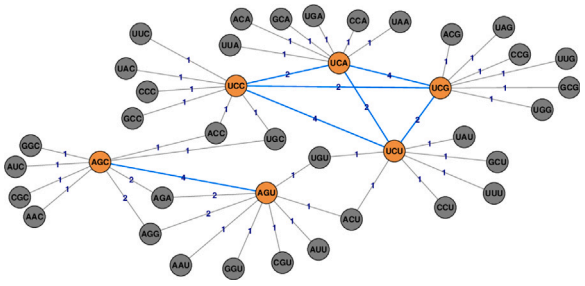
(a) If $S = \{UAA, UAG, UGA\}$, then $\phi(S) = \frac{3\cdot3+2\cdot2+3+(2+2)+(2+2)+(2+2+4)}{3\cdot3+3\cdot3+3\cdot(2+2+4)} = \frac{16}{21}$



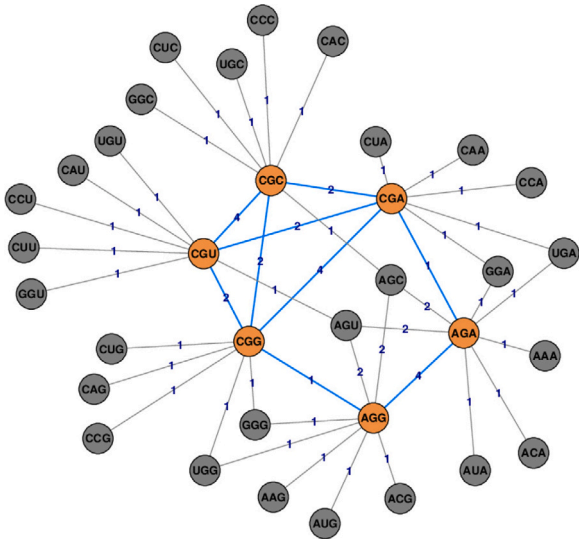
(b) If $S = \{CUA, CUC, CUG, CUU, UUA, UUG\}$, then $\phi(S) = \frac{(2+2+3+3+2+2)+6\cdot3+(2\cdot2+2\cdot2)}{6\cdot(3+3+8)} = \frac{10}{21}$ (this corresponds to Leucine)



(c) If $S = \{UCU, UCG, UCA, UCU, AGU, AGC\}$, then $\phi(S) = \frac{(6\cdot3+6\cdot3)+(2\cdot2+2\cdot2)}{6\cdot(3+3+8)} = \frac{11}{21}$ (this corresponds to Serine)



(d) If $S = \{CGA, CGC, CGG, CGU, AGA, AGG\}$, then $\phi(S) = \frac{(2+2+3+3+2+2)+6\cdot3+(2\cdot2+2\cdot2)}{6(3+3+8)} = \frac{10}{21}$ (this corresponds to Arginine) □



As an immediate corollary we have that now the minimal average conductance of a set of size two and a set of size four is now smaller than that of a set of size three.

Corollary 3.12. *$G = G_l^P(V, E, w)$ be an undirected weighted graph as in Definition 2.1 and the corresponding weights $p_1 = p_2 = 1$ and $\tilde{p}_3 = 2$. Then the minimal average conductance of a set of size two and a set of size four is $\frac{1}{2}(\frac{15}{21} + \frac{9}{21}) = \frac{12}{21}$ while the minimal conductance of a set of size three is $\frac{13}{21}$.*

A more interesting corollary is the total average conductance of the genetic code using this weighted approach.

Corollary 3.13. *$G = G_l^P(V, E, w)$ be an undirected weighted graph as in Definition 2.1 and the corresponding weights $p_1 = p_2 = 1$ and $\tilde{p}_3 = 2$. Then the total average conductance of the standard genetic code is*

$$\frac{1}{24}(\frac{13}{21} + \frac{16}{21} + 12\cdot\frac{15}{21} + 8\cdot\frac{9}{21} + 1 + 1) = \frac{323}{504} = 0,6408$$

if we separate amino acids of degeneracy 6 into two of degeneracy 2 and 4.

Moreover, the total average conductance of the standard genetic code is

$$\frac{1}{21}(\frac{13}{21} + \frac{16}{21} + 9\cdot\frac{15}{21} + 5\cdot\frac{9}{21} + 1 + 1 + 2\cdot\frac{10}{21} + \frac{11}{21}) = \frac{237}{441} = 0.5374$$

if we allow degeneracy 6.

To conclude this section we now turn back to the model from Fig. 10 and calculate again the conductance of the sets of size three and four that appear there.

Proposition 3.14. *$G = G_l^P(V, E, w)$ be an undirected weighted graph as in Definition 2.1 and the corresponding weights $p_1 = p_2 = 1$ and $\tilde{p}_3 = 2$. Moreover, let $X, Y \in \mathcal{B}$. then*

(1) If $S = \{XYA, XYC, XYG\}$ or $S = \{XYA, XYC, XYU\}$, then $\phi(S) = \frac{3\cdot3+3\cdot3+(2+2+4)}{3\cdot3+3\cdot3+3\cdot(2+2+4)} = \frac{13}{21}$

(2) If $S = \{AXY, CXY, UXY\}$, then $\phi(S) = \frac{3+3\cdot3+3\cdot(2+2+4)}{3\cdot3+3\cdot3+3\cdot(2+2+4)} = \frac{18}{21}$

(3) If $S = \{GAG, GCG, GGG, GUG\}$, then $\phi(S) = \frac{4\cdot3+4\cdot8}{4\cdot3+4\cdot3+4\cdot(2+2+4)} = \frac{11}{14}$

Again we have the average conductance of the model from Fig. 10.

Corollary 3.15. *$G = G_l^P(V, E, w)$ be an undirected weighted graph as in Definition 2.1 and the corresponding weights $p_1 = p_2 = 1$ and $\tilde{p}_3 = 2$. Then the total average conductance of the model from Fig. 10 is*

$$\frac{1}{21}(16\cdot\frac{13}{21} + \frac{11}{14} + 4\cdot\frac{8}{21}) = \frac{513}{882} = 0,5816$$

Finally we see that the genetic code now performs better than the model from Fig. 10.

Corollary 3.16. *$G = G_l^P(V, E, w)$ be an undirected weighted graph as in Definition 2.1 and the corresponding weights $p_1 = p_2 = 1$ and $\tilde{p}_3 = 2$. Then the total average conductance of the standard genetic code is smaller than the average conductance of the model from Fig. 10 since $\frac{237}{441} < \frac{513}{882}$.*

As we can see once we incorporate the Wobble hypothesis in our graph model, the standard genetic code performs better and it is almost optimal as we will be shown in the sequel paper (Fimmel et al., 2021). In fact, our results in Fimmel et al. (2021) using optimization algorithms will show that the particular choice of weights ($p_1 = p_2 = 1$ and $\tilde{p}_3 = 2$) is not important for the genetic code to be almost optimal.

4. Conclusions

In the present work, the graph-theoretical approach of Blazej et al. (2019, 2018, 2020) was adopted and further developed to investigate the robustness of different variants of the genetic code to single point mutations. The generalization of the graph-theoretic model took place in several different directions.

First, the model now allows (genetic) alphabets of arbitrary (even) cardinality and, correspondingly, words of arbitrary length to be considered over it. Thus, the graph theoretic approach now enables to consider arbitrary extensions of the genetic code, be it by extending the genetic alphabet or the tuple length of the codons.

Second, weights have now been assigned to the edges of the graph under consideration, allowing the probabilities or severities of individual point mutations to be modelled. In particular, this change in the model now allows the inclusion of the wobble effect in the translation of codons into amino acids.

This second change proved particularly fruitful in the analysis of the structure of the standard genetic code. This analysis showed that the standard genetic code, which according to the results of Blazej et al. (2018) (without the weights of single point-mutations) is not

optimal in terms of robustness to point-mutations, performs better than the optimal variant of it found in Blazej et al. (2018) when the wobble effects are taken into account. Biologists may be interested in Table 1 in which, for all known dialects of the genetic code with a unique assignment of amino acids, their conductances without and with consideration of the wobble effect are brought together.

A large part of the work is devoted to finding the lower bounds for the conductance of a partition for a given number of amino acids to be encoded. This part of the work is important for assessing the optimality of the found partitions of the genetic code into synonymous codons and may be useful in further investigations. Among other things, it was shown that the lower bounds obtained with Higher Order Cheeger's Inequalities (see Lee et al., 2014) are not tight, and another method, which is, however, tailored to the graph under consideration, was demonstrated for this purpose.

In summary, the authors have further developed and extended the graph-theoretical approach initiated in Blazej et al. (2019) in several ways in order to be applicable to a wider (biological) setting that sheds light on the robustness of the genetic code. The authors hope that their work will also open up new possibilities for research into the structure and origin of the genetic code.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

A.1. Proof of Proposition 3.7

Proof. Let $\tilde{w}(E(S_m, \overline{S_m})) = \sum_{i=1}^{\ell} p_i \cdot \lambda_i$ with λ_i and p_i as stated in the proposition, i.e.

$$\lambda_i := \frac{m_i \cdot (2n)^{\ell-i}}{m}(2n - m_i) + \frac{(m \bmod (2n)^{\ell-i})}{m}(2(n - m_i) - 1)$$

where $(m_1, m_2, \dots, m_\ell)_{2n}, m_i \in \{0, 1, 2, \dots, (2n-1)\}$ is the representation of m to base $2n$, i.e.

$$m = m_1 \cdot (2n)^{\ell-1} + m_2 \cdot (2n)^{\ell-2} + \cdots + m_\ell \cdot (2n)^0.$$

We will show by induction that the sequence

$$m \cdot \tilde{w}(E(S_m, \overline{S_m})) \quad 1 \le m \le (2n)^l$$

satisfies the recursion from Theorem 3.5, i.e. we will show that

$$m \cdot \tilde{w}(E(S_m, \overline{S_m})) + \sum_{i=1}^{\ell} p_i(2(n - m_i) - 1) = (m+1) \cdot \tilde{w}(E(S_{m+1}, \overline{S_{m+1}})) \quad \text{(A.1)}$$

and

$$\tilde{w}(E(S_1, \overline{S_1})) = (2n-1)\sum_{i=1}^{\ell} p_i.$$

It then immediately follows that $w(E(S_m, \overline{S_m})) = m \cdot \sum_{i=1}^{\ell} p_i \cdot \lambda_i$ as claimed.

Assume that $m = 1$, then $m_1 = \cdots = m_{l-1} = 0$ and $m_l = 1$ in the $(2n)$-adic presentation of m. Moreover, we easily see that $\lambda_1 = \cdots = \lambda_{l-1} = \lambda_l = (2n-1)$. Hence

$$\tilde{w}(E(S_1, \overline{S_1})) = (2n-1)\sum_{i=1}^{\ell} p_i$$

as claimed.

Now assume that $m > 1$. We need to show Eq. (A.1)

This will be done in two steps by calculating the left and the right sides of the above equation separately. We start with the left side and calculate the term

$$m \cdot \tilde{w}(E(S_m, \overline{S_m})) + \sum_{i=1}^{\ell} p_i(2(n - m_i) - 1)$$

$$\begin{aligned}
&m \cdot \tilde{w}(E(S_m, \overline{S_m})) + \sum_{i=1}^{\ell} p_i(2(n - m_i) - 1) \\
&= m \cdot \sum_{i=1}^{\ell} p_i \cdot \lambda_i + \sum_{i=1}^{\ell} p_i(2(n - m_i) - 1) = \sum_{i=1}^{\ell} p_i \cdot (m \cdot \lambda_i + 2(n - m_i) - 1) \\
&= \sum_{i=1}^{\ell} p_i \cdot (m \cdot (\frac{m_i \cdot (2n)^{\ell-i}}{m}(2n - m_i) + \frac{m \ (mod\ (2n)^{\ell-i})}{m} \\
&\quad \times (2(n - m_i) - 1)) + 2(n - m_i) - 1) \\
&= \sum_{i=1}^{\ell} p_i \cdot (m_i \cdot (2n)^{\ell-i}(2n - m_i) + m \ (mod\ (2n)^{\ell-i})(2(n - m_i) - 1)) \\
&\quad + (2(n - m_i) - 1) \\
&= \sum_{i=1}^{\ell} p_i \cdot (m_i \cdot (2n)^{\ell-i}(2n - m_i) + (m \ (mod\ (2n)^{\ell-i}) + 1)(2(n - m_i) - 1)
\end{aligned} \quad \text{(A.2)}$$

Next we calculate the term $(m+1)\cdot\tilde{w}(E(S_{m+1}), \overline{S_{m+1}})$ and let $((m+1)_1, (m+1)_2, \dots, (m+1)_\ell)_{2n}, (m+1)_i \in \{0, 1, 2, \dots, (2n-1)\}$ be the representation of $(m+1)$ to base $2n$, i.e.

$$m + 1 = (m+1)_1 \cdot (2n)^{\ell-1} + (m+1)_2 \cdot (2n)^{\ell-2} + \cdots + (m+1)_\ell \cdot (2n)^0.$$

It the follows that

$$\begin{aligned}
&(m+1) \cdot \tilde{w}(E(S_{m+1}), \overline{S_{m+1}}) \\
&= \sum_{i=1}^{\ell} p_i \cdot (m+1) \cdot (\frac{(m+1)_i \cdot (2n)^{\ell-i}}{(m+1)}(2n - (m+1)_i) \\
&\quad + \frac{((m+1) \ mod\ (2n)^{\ell-i})}{(m+1)}(2(n - (m+1)_i) - 1)) \\
&= \sum_{i=1}^{\ell} p_i \cdot ((m+1)_i \cdot (2n)^{\ell-i}(2n - (m+1)_i) \\
&\quad + ((m+1) \ mod\ (2n)^{\ell-i})(2(n - (m+1)_i) - 1))
\end{aligned} \quad \text{(A.3)}$$

To prove Eq. (A.1) we will show that the differences of the coefficients in Eq. (A.2) and Eq. (A.3) of the p_is is always zero, i.e. for each i we show that

$$\begin{aligned}
&(m_i \cdot (2n)^{\ell-i}(2n - m_i) + (m \ (mod\ (2n)^{\ell-i}) + 1)(2(n - m_i) - 1)) = \\
&((m+1)_i \cdot (2n)^{\ell-i}(2n - (m+1)_i) + (m+1) \ (mod\ (2n)^{\ell-i})(2(n - (m+1)_i) - 1))
\end{aligned} \quad \text{(A.4)}$$

We have to distinguish three cases that might occur if one passes from m to $m+1$ in the $(2n)$-adic presentation: The case (i) where $m_i - (m+1)_i = 0$, (ii) where $m_i - (m+1)_i = -1$ and (iii) $m_i - (m+1)_i = 2n-1$. We start with case (i).

Case (i) Assume that $m_i - (m+1)_i = 0$. It then follows that

$$(m \ (mod\ (2n)^{\ell-i}) + 1) = (m+1) \ (mod\ (2n)^{\ell-i})$$

and hence we see immediately that in Eq. (A.4) equality holds.

Case (ii) Assume that $m_i - (m+1)_i = -1$. It then follows that

$$(m+1) \ (mod\ (2n)^{\ell-i}) = 0$$

and

$$(m \ (mod\ (2n)^{\ell-i}) + 1) = (2n)^{\ell-i}$$

which shows that the term in Eq. (A.2) can be rewritten as:

$$(2n)^{l-i} \cdot ((2(n - m_i) - 1) + m_i \cdot (2n - m_i))$$

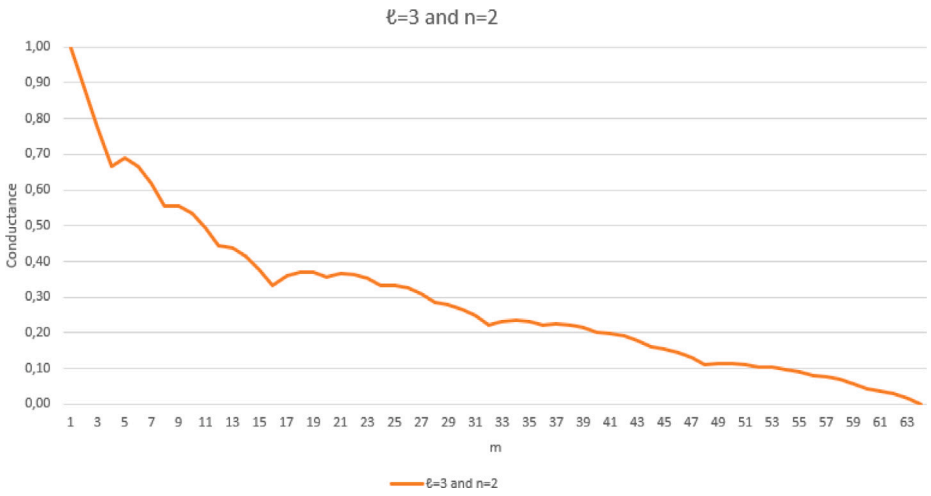


Fig. 12. The graph shows the conductance for $\phi(S_m)$ for all $m \in 1, \dots, 64$ where $n = 2$ and $\ell = 3$. The local minimums are at $m \in \{1, 2, 3, 4, 8, 12, 16, 32, 48, 64\}$.

and the term in Eq. (A.3) reduces to:

$$(2n)^{l-i} \cdot (m_i + 1) \cdot (2n - m_i - 1)$$

It is now an easy exercise to see that this implies that the equality in Eq. (A.2) holds.

Case (iii) Assume that $m_i - (m+1)_i = 2n - 1$. Consequently we have $m_i = 2n - 1$ and $(m+1)_i = 0$. As in Case (ii) we conclude

$$(m+1)\ (mod\ (2n)^{\ell-i}) = 0$$

and

$$(m\ (mod\ (2n)^{\ell-i}) + 1) = (2n)^{\ell-i}$$

which shows that the terms in Eq. (A.2) and Eq. (A.3) are equal to zero. This can be seen from the reduced forms from above:

$$(2n)^{l-i} \cdot ((2(n - m_i) - 1) + m_i \cdot (2n - m_i)) = 0$$

$$(2n)^{l-i} \cdot (m+1)_i \cdot (2n - m_i - 1) = 0.$$

Again it follows that the equality in Eq. (A.2) holds.

This finishes the proof. □

A.2. Proof of Proposition 3.8

Proof. For any $k \in \mathbb{N}$ we define

$$m_k = (2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor} \cdot \left\lceil \frac{(2n)^\ell}{k(2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor}} \right\rceil.$$

First note that the quotient $\frac{(2n)^\ell}{k}$ is the average size of a set in any partition C_k of the graph G into k classes. Thus the integer m_k represents up to a coefficient the highest $(2n)$-power that is below $\frac{(2n)^\ell}{k}$. In the special case that k is a $(2n)$-power itself, say $k = (2n)^s$, then $\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor = l - s$ and $m_k = \frac{(2n)^\ell}{k}$. However, in general $m_k \geq \frac{(2n)^\ell}{k}$.

Recall from Proposition 3.7 that

$$w(E(S_{m_k}, \overline{S_{m_k}})) = m_k \cdot \tilde{w}(E(S_{m_k}, \overline{S_{m_k}})) := m_k \cdot \sum_{i=1}^{\ell} p_i \cdot \lambda_{k,i}$$

with

$$\lambda_{k,i} := \frac{m_{k,i} \cdot (2n)^{\ell-i}}{m_k}(2n - m_{k,i}) + \frac{m_k\ (mod\ (2n)^{\ell-i})}{m_k}(2(n - m_{k,i}) - 1)$$

where $(m_{k,1}, m_{k,2}, \dots, m_{k,\ell})_{2n}, m_{k,i} \in \{0, 1, 2, \dots, (2n-1)\}$ is the representation of m_k to base $2n$, i.e.

$$m_k = m_{k,1} \cdot (2n)^{\ell-1} + m_{k,2} \cdot (2n)^{\ell-2} + \cdots + m_{k,\ell} \cdot (2n)^0.$$

It is easy to see that the value of $\left\lceil \frac{(2n)^\ell}{k(2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor}} \right\rceil$ is a natural number less than or equal to $(2n)$. Without loss of generality we may assume that $\left\lceil \frac{(2n)^\ell}{k(2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor}} \right\rceil < (2n)$ - the case of equality uses a similar arguments. Let $j_k = \ell - \lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor$. In the case where $\left\lceil \frac{(2n)^\ell}{k(2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor}} \right\rceil = (2n)$ we have to increase j_k by one $j_k = 1 + \ell - \lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor$. Then it immediately follows that in the above presentation of m_k we have

$$m_{k,1} = \cdots = m_{k,j_k - 1} = m_{k,j_k+1} = \cdots = m_{k,l} = 0.$$

Note that $m_{k,s}$ is the coefficient of $(2n)^{l-s}$ which is the reason why we had to define j_k as $\ell - \lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor$. Moreover, we also easily see that

$$m_k\ (mod\ (2n)^{\ell-i}) = m_k$$

for $i \in \{1, \dots, j_k - 1\}$ and similarly

$$m_k\ (mod\ (2n)^{\ell-i}) = 0$$

for $i \in \{j_k + 1, \dots, \ell\}$. Consequently, the coefficients $\lambda_{k,s}$ satisfy

$$\lambda_{k,1} = \cdots = \lambda_{k,j_k - 1} = 2n - 1$$

and

$$\lambda_{k,j_k+1} = \cdots = \lambda_{k,\ell} = 0.$$

Thus

$$\tilde{w}(E(S_{m_k}, \overline{S_{m_k}})) = \sum_{i=1}^{j_k - 1} p_i(2n-1) + \lambda_{k,j_k} p_{j_k} = (2n-1) \cdot \sum_{i=1}^{j_k - 1} p_i + \lambda_{k,j_k} p_{j_k}.$$

It is now immediate that the values of $\tilde{w}(E(S_{m_k}, \overline{S_{m_k}}))$ for $k = 1, \dots, (2n)^l$ form a decreasing sequence. Note that for increasing k, either the value of j_k decreases or the value of λ_{k,j_k} decreases (in the cases that the value of j_k remains the same).

We finally claim that *(I)* the $\tilde{w}(E(S_{m_k}, \overline{S_{m_k}}))$ for $k = 1, \dots, (2n)^l$ also form local minima of the function $\tilde{w}(\cdot)$, i.e. that for any $k = 1, \dots, (2n)^l$ we have $\tilde{w}(E(S_{m_k}, \overline{S_{m_k}})) < \tilde{w}(E(S_{\bar{m}}, \overline{S_{\bar{m}}}))$ for any $0 < \bar{m} < m_k$ and *(II)* for any $k = 1, \dots, (2n)^l$ there is at least one $S \in C_k$ so that $|S| \leq m$.

claim (I) For $k \in \mathbb{N}$ choose any $0 < \bar{m} < m_k$ and let

$$\bar{m} = \bar{m}_1 \cdot (2n)^{\ell-1} + \bar{m}_2 \cdot (2n)^{\ell-2} + \cdots + \bar{m}_{j_k} \cdot (2n)^{\ell-j_k} + \cdots + \bar{m}_\ell \cdot (2n)^0$$

be the $2n$-adic presentation of $\bar{m}$. Since $\bar{m} < m_k$ it immediately follows as above that $\bar{m}_1 = \cdots = \bar{m}_{j_k - 1} = 0$ and that the term $\bar{m}\ (mod\ (2n)^{\ell-i}) = \bar{m}$ for $i \in \{1, \dots, j_k - 1\}$. Consequently, the coefficients

$$\bar{\lambda}_1 = \cdots = \bar{\lambda}_{j_k - 1} = 2n - 1.$$

as above. It follows that

$$\tilde{w}(E(S_{\bar{m}}, \overline{S_{\bar{m}}})) = (2n-1) \cdot \sum_{i=1}^{j_k-1} p_i + \sum_{i=j_k}^{l} \bar{\lambda}_i p_i.$$

Since the coefficient $\bar{\lambda}_s$ cannot be negative (see Theorem 3.5), we can surmise that these points must be local minimums, i.e. $\tilde{w}(E(S_{m_k}, \overline{S_{m_k}})) < \tilde{w}(E(S_{\bar{m}}, \overline{S_{\bar{m}}}))$. Fig. 12 illustrates the graph of $\phi(S)$ for $\ell = 3$ and $n = 2$.

claim (II) To complete the proof we must show that there is at least one $S \in C_k$ so that $|S| \leq m_k$. To confirm this claim, we use the expression

$$\frac{(2n)^\ell}{k} = (2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor} \cdot \frac{(2n)^\ell}{k(2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor}}$$

$$\leq (2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor} \cdot \left\lceil \frac{(2n)^\ell}{k(2n)^{\lfloor \log_{2n}(\frac{(2n)^\ell}{k}) \rfloor}} \right\rceil = m_k.$$

Thus, $|S| > m_k$ for all $S \in C_k$ would imply that the union of all $S \in C_k$ has size greater than $k \cdot m_k \geq (2n)^l$ - a contradiction. Hence there must be one $S \in C_k$ so that $|S| \leq m_k$.

To complete the proof we note that for S from claim (II) we have

$$\Phi(C_k) \geq \Phi(S) \geq \Phi(S_{m_k}) = \frac{w(E(S_{m_k}, \overline{S_{m_k}}))}{\ell \cdot (2n-1) \cdot m_k} = \frac{\tilde{w}(E(S_{m_k}, \overline{S_{m_k}}))}{\ell \cdot (2n-1)}$$

since $\tilde{w}(E(S_{m_k}, \overline{S_{m_k}})) \leq \tilde{w}(E(S, \overline{S}))$ by claim (I). □

References

Alev, Vedat Levi, Anari, Nima, Lau, Lap Chi, Gharan, Shayan Oveis, 2017. Graph clustering using effective resistance. ArXiv e-prints, arXiv:1711.06530.

Alseth, I., Dalhus, B., Bjøra, M., 2014. Inosine in DNA and RNA. Curr. Opin. Genetics Dev. 26, 116–123. http://dx.doi.org/10.1016/j.gde.2014.07.008.

Anderson, J.C., Wu, N., Santoro, S.W., Lakshman, V., King, D.S., Schultz, P.G., 2004. An expanded genetic code with a functional quadruplet codon. Proc. Natl. Acad. Sci. USA 101, 7566–7571. http://dx.doi.org/10.1073/pnas.0401517101.

Barbieri, M., 2019. Evolution of the genetic code: the ambiguity-reduction theory. Biosystems 185, 104024.

Bezrukov, S.L., 1999. Edge isoperimetic problems on graphs. Graph Theory Comb. Biol. 7, 157–197, Akademia Kiado, Budapest.

Bezrukov, S.L., Elsässer, R., Edge-isoperimetric problems for cartesian powers of regular graphs. Theor. Comput. Sci. 307 (3) 473–492.

Blazej, P., Fimmel, E., Gumbel, M., 2019. The quality of genetic code models in terms of their robustness against point mutations. Bull. Math. Biol. 81 (7), 2239–2257.

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

Blazej, P., Wnetrzak, M., Mackiewicz, D., Mackiewicz, P., 2020. Basic principles of the genetic code extension R. Soc. Open Sci. 7191384. http://dx.doi.org/10.1098/rsos.191384.

Chin, J.W., 2017. Expanding and reprogramming the genetic code. Nature 550, 53–60. http://dx.doi.org/10.1038/nature24031.

Clark, J., Holton, D.A., 1991. First Look at Graph Theory. World Scientific Publishing Co Pte Ltd, ISBN: 978-9810204891.

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

Dien, V.T., Morris, S.E., Karadeema, R.J., Romesberg, F.E., 2018. Expansion of the genetic code via expansion of the genetic alphabet. Curr. Opin. Chem. Biol. 46, 196–202. http://dx.doi.org/10.1016/j.cbpa.2018.08.009.

Fimmel, E., Gumbel, M., Starman, M., Strüngmann, L., 2021. Computational analysis of genetic code variations optimized for the robustness against point mutations with wobble-like effects (in preparation).

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Starman, M., Strüngmann, L., 2020a. The relation between k-circularity and circularity of codes. Bull. Math. Biol. 82, 105. http://dx.doi.org/10.1007/s11538-020-00770-7.

Fimmel, E., Michel, Ch.J., Starman, M., Strüngmann, L., 2018. Self-complementary circular codes in coding theory. Theory Biosci. 37 (1), 51–65.

Fimmel, E., Michel, C.J., Strüngmann, L., 2016. *n*-nucleotide circular codes in graph theory. Phil. Trans. R. Soc. A 374, 20150058.

Fimmel, E., Starman, M., Strüngmann, L., 2020b. Circular tessera codes in the evolution of the genetic code. Bull. Math. Biol. 82, 48. http://dx.doi.org/10.1007/s11538-020-00724-z.

Gaertler, M., 2005. Clustering. In: Brandes, U., Erlebach, T. (Eds.), Network Analysis. In: Lecture Notes in Computer Science, vol. 3418, Springer, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-540-31955-9_8.

Kannan, R., Vempala, S., Vetta, A., 2004. On clusterings: Good, bad and spectral. J. ACM 51 (3), 497–515. http://dx.doi.org/10.1145/990308.990313.

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. http://dx.doi.org/10.1093/nar/gkn956.

Kwok, Tsz Chiu, Lau, Lap Chi, Lee, Yin Tat, Gharan, Shayan Oveis, 2013. Improved Cheeger's inequality: Analysis of spectral partitioning algorithms through higher order spectral gap. arXiv:1301.5584 [cs.DS].

Lee, J.R., Gharan, S.O., Trevisan, L., 2014. Multiway spectral partitioning and higher-order cheeger inequalities. J. ACM 61 (6), 37. http://dx.doi.org/10.1145/2665063.

Neumann, H., Wang, K., Davis, L., Garcia-Alai, M., Chin, J.W., 2010. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. Nature 464, 441–444. http://dx.doi.org/10.1038/nature08817.

Wong, J.T., 1975. A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA 72 (5), 1909–1912.