Optimality of circular codes versus the genetic code after frameshift errors

Julie D. Thompson, Gopal Dila, Christian Michel

Version of Record: https://www.sciencedirect.com/science/article/pii/S0303264720300381
Manuscript_30843a35915e06c934728246b0c4db74

Optimality of circular codes versus the genetic code after frameshift errors

Gopal Dila, Christian J. Michel *, Julie D. Thompson *

Department of Computer Science, ICube, CNRS, University of Strasbourg, Strasbourg, France

* To whom correspondence should be addressed; Email: c.michel@unistra.fr

* Corresponding authors

Email: d.gopal@outlook.com, c.michel@unistra.fr, thompson@unistra.fr

Keywords: circular code, standard genetic code, frameshift, optimization, evolution.

Abstract

The standard genetic code (SGC) describes how 64 trinucleotides (codons) encode 20 amino acids and the stop translation signal. Biochemical and statistical studies have shown that the standard genetic code is optimized to reduce the impact of errors caused by incorporation of wrong amino acids during translation. This is achieved by mapping codons that differ by only one nucleotide to the same amino acid or one with similar biochemical properties, so that if misincorporation occurs, the structure and function of the translated protein remain relatively unaltered. Some previous studies have extended the analysis of SGC optimality to the effect of frameshift errors on the conservation of amino acids. Here, we compare the optimality of the SGC with a set of circular codes, and in particular the X circular code identified in genes, on the basis of various biochemical properties over all possible frameshift errors. We show that the X circular code is more optimized to minimize the impact of frameshift errors than the SGC for the chosen amino acid properties. Furthermore, in the context of a problem that has been unresolved since 1996, we also demonstrate that the X circular code has a frameshift optimality in its combinatorial class of 216 maximal self-complementary C^3 circular codes. To our knowledge, this is the first demonstration of the role of the X circular code in mitigation of translation errors. These results lead us to discuss the potential role of the X circular code in the evolution of the standard genetic code.

1. Introduction

One of the most intriguing questions in molecular biology is how the basic structures of life as we know them evolved over 4 billion years and what were the evolutionary pressures acting on them? The genetic code is one such structure that defines the set of rules needed to translate the information in DNA into proteins. Virtually all living organisms use the same standard genetic code (SGC) to determine how the 64 DNA trinucleotides (codons) are translated into 20 amino acids and the stop signal. Many hypotheses have been put forward to explain the origin of the genetic code (e.g. reviewed in Koonin and Novozhilov, 2009), including the frozen accident theory that proposes that the genetic code was created randomly and stayed frozen ever since, the stereochemical theory that suggests some kind of stereochemical relationship existed between amino acids and specific codons (Pelc and Welton, 1966; Yarus, 2017), the adaptive theory that suggests the genetic code was shaped to be maximally robust (Freeland and Hurst, 1998), and the coevolution theory of the genetic code with amino acid biosynthetic pathways (Wong, 1975). However, it is likely that all these aspects combined to play a part in the evolution of the SGC.

In this article, we will focus on the adaptive theory which suggests that the SGC was optimized to minimize the effects of errors during transcription and translation, originally proposed by Woese (1965). The most common source of translation errors, known as missense errors, is the incorrect reading of a codon and the resulting incorporation of the wrong amino acid. The per-codon missense error rate has been estimated to be between 10^{-4} and 10^{-3} (Garofalo et al., 2019). It is generally accepted that the SGC is optimized to reduce the effects of these errors. First, base changes at the third position of the codon, known as the wobble position, are generally synonymous, i.e. they code for the

same amino acid. Second, amino acids with similar physicochemical properties are located in close proximity in the genetic code table and differ usually by only one substitution. For example, hydrophobic amino acids are usually coded by codons with thymine (*T*) in the second position and hydrophilic amino acids by those with adenine (*A*) in this position. It has been shown previously that the SGC outperforms most theoretical alternative codes in terms of reducing the effects of missense errors, when amino acid similarity is measured by polarity (Haig and Hurst, 1991; Freeland and Hurst, 1998; Kumar and Saini, 2016), by polarity and volume (Wnętrzak et al., 2019), or by using empirical data of substitution frequencies (Freeland et al., 2000). Another important source of translation errors is ribosomal frameshifting, which occurs with an error rate of around 10^{-5} (Drummond and Wilke, 2009). Since the genetic code has a non-overlapping structure, the codons in a DNA sequence must be decoded in the correct reading frame in order to produce the correct amino acid sequence. A shift of one or two bases into the $+1$ or $+2$ (-1)[1] frames respectively, can have severe effects, including termination of translation if a stop codon is encountered out-of-frame, or production of a non-functional protein sequence otherwise (Figure 1).

Figure 1. Original reading frame in comparison to the two shifted frames $+1$ and $+2$ (-1) results in different read out of amino acids.

The "ambush hypothesis" proposes that out-of-frame stop codons (also known as hidden stops) allow rapid termination of frameshifted translations and are selected for (Seligmann and Pollock, 2004; Itzkovitz and Alon, 2007; Abrahams and Hurst, 2018; Seligmann, 2019). Furthermore, it has been suggested recently that the SGC is also optimized to reduce the effects of frameshift errors when no out-of-frame stop codon is encountered (Geyer and Madany Mamlouk, 2018). Thus, to minimize the costs of errors, organisms evolve either by implementing “increased accuracy” or “increased robustness”. The question remains of how these optimizations evolved and which mechanisms are responsible for them. The robustness of the SGC to frameshift errors represents an attractive problem from a coding theory point of view. One of the first solutions was suggested by Crick (Crick et al., 1957), who proposed that the genetic code was a comma-free code in order to explain how 64 codons could code for 20 amino acids and how the correct reading frame could be retrieved and maintained at the same time. Using a comma-free code, codons in the reading frame make sense, while codons in the shifted frames 1 and 2 make nonsense. However, it was later proved that the standard genetic code could not be a comma-free code (Nirenberg and Matthaei, 1961), when it was discovered that *TTT*, which codes for phenylalanine cannot belong to a comma-free code.

[1] The shifted frame $+2$ classically used in circular code theory is called -1 in biology.

Another possible solution to the frameshift problem is the X circular code (Arquès and Michel, 1996). Circular codes are a weaker version of comma-free codes, where any word written on a circle (the last letter becoming the first in the circle) has a unique decomposition into trinucleotides of the circular code (reviewed in Michel, 2008; Fimmel and Strüngmann, 2018). A circular code naturally excludes the periodic trinucleotides $\{AAA, CCC, GGG, TTT\}$. It also excludes trinucleotides related by circular permutation, e.g. AAC and ACA, since the concatenation of AAC with itself …$AACAAC$…, for example, can be decomposed in two ways: …AAC,AAC… or …A,ACA,AC… By excluding the periodic trinucleotides and dividing the 60 remaining trinucleotides into three disjoint classes, a circular code of trinucleotides has at most 20 trinucleotides (called a maximal circular code). There exist 12,964,440 maximal circular codes, although it has been shown that there is no maximal circular code that can code for 20 or 19 amino acids and only 10 can code for 18 amino acids (Michel and Pirillo, 2013). Remarkably, one of the maximal circular codes, called the X circular code, was found to be overrepresented in the reading frame of protein coding genes from bacteria, archaea, eukaryotes, plasmids and viruses (Arquès and Michel, 1996; Michel, 2015, 2017). The X circular code consists of 20 trinucleotides

$$X = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\} \quad (1)$$

and codes the 12 following amino acids (three and one letter notation)

$$\mathcal{X} = \{Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr, Val\} = \{A, N, D, Q, E, G, I, L, F, T, Y, V\}. \quad (2)$$

This X circular code has in addition several strong mathematical properties. It is self-complementary: if a trinucleotide belongs to X then its complementary trinucleotide also belongs to X. Moreover, the +1/-2 and +2/-1 circular permutations of X, denoted X_1 and X_2 respectively, are also maximal circular codes (C^3) and are complementary to each other (see Section 2.1). There exist 216 maximal C^3 self-complementary trinucleotide circular codes named $\mathbb{X}$ (Arquès and Michel, 1996), and X belongs to $\mathbb{X}$. Any class of circular codes, like comma-free codes, also have the property of synchronization, i.e. they are hypothesized to retrieve and maintain the reading frame by using an appropriate window of nucleotides. In any sequence generated by a trinucleotide comma-free code, the reading frame can be determined in a window length of at most 3 consecutive nucleotides, while for the X circular code, at most 13 consecutive nucleotides are enough to always retrieve the reading frame. In other words, a sequence 'motif' containing several consecutive X trinucleotides is sufficient to determine the correct reading frame. It has been shown recently that X motifs are enriched in the reading frame of modern genes (Michel et al., 2017; Dila et al., 2019a), as well as in tRNA sequences (Michel, 2012, 2013) and in functional regions of rRNA involved in mRNA translation (Michel, 2012; Dila et al., 2019b). Furthermore, a circular code periodicity 0 modulo 3 was identified in the 16S rRNA, covering the region that corresponds to the primordial proto-ribosome decoding center and containing numerous sites that interact with the tRNA and mRNA during translation (Michel and Thompson, 2020). Based on the mathematical properties of the X circular code and the enrichment of X motifs in the main actors

involved in translation, it has been suggested that the X circular code was an ancestor code of the SGC that was used to code amino acids and simultaneously to identify and maintain the reading frame (Dila et al., 2019b). Furthermore, it has been suggested that the X circular code arose from selection for non-redundant overlap coding in short nucleotide sequences (Michel, 2019; Demongeot and Seligmann, 2020). This is in line with the hypothesis that the primordial genes maximized the number of coded amino acids over the shortest length, since these primordial genes, called RNA rings, are biased towards codons belonging to X (Demongeot and Seligmann, 2019).

In this study, we test for the first time the hypothesis that the X circular code has the additional property of minimizing the effects of frameshift errors. To achieve this, we compare the optimality of the X circular code with the SGC, as well as its combinatorial class of 216 maximal self-complementary C^3 circular codes. The effects of frameshift errors are estimated in terms of the resulting differences in various physicochemical properties of the translated amino acids. We defined two different measures of code optimality: (i) a code score, e.g. a code Y, where the frameshift is analysed according to code permutations Y_1 and Y_2; and (ii) a code motif score, precisely a code dicodon score, where the frameshift is analysed according to 1 or 2 base shifts in a dicodon (in reading frame) generated from a code.

2. Method

2.1. Definition of codes

We recall a few definitions without detailed explanation (i.e. without examples and figures) that are necessary for understanding the main properties of 216 maximal C^3 self-complementary trinucleotide circular codes $\mathbb{X}$.

Notation 1. Let us denote the nucleotide 4-letter alphabet $B = \{A, C, G, T\}$ where A stands for adenine, C stands for cytosine, G stands for guanine and T stands for thymine. The trinucleotide set over B is denoted by $B^3 = \{AAA, \dots, TTT\}$. The set of non-empty words (words, respectively) over B is denoted by B^+ (B^*, respectively).

Notation 2. Genes or motifs in reading frame have three frames f. By convention here, the reading frame $f = 0$ is set up by a start trinucleotide, classically ATG, and the frames $f = 1$ and $f = 2$ are the reading frame $f = 0$ shifted by one and two nucleotides in the $5' - 3'$ direction (to the right), respectively.

Two biological maps are involved in gene coding.

Definition 1. According to the complementary property of the DNA double helix, the *nucleotide complementarity map* $\mathcal{C}: B \to B$ is defined by $\mathcal{C}(A) = T, \mathcal{C}(C) = G, \mathcal{C}(G) = C, \mathcal{C}(T) = A$. According to the complementary and antiparallel properties of the DNA double helix, the *trinucleotide complementarity map* $\mathcal{C}: B^3 \to B^3$ is defined by $\mathcal{C}(l_0 l_1 l_2) = \mathcal{C}(l_2)\mathcal{C}(l_1)\mathcal{C}(l_0)$ for all $l_0, l_1, l_2 \in B$. By extension to a trinucleotide set S, the *set complementarity map* $\mathcal{C}: \mathbb{P}(B^3) \to \mathbb{P}(B^3)$, $\mathbb{P}$ being the set of all subsets of B^3, is defined by $\mathcal{C}(S) = \{v : u, v \in B^3, u \in S, v = \mathcal{C}(u)\}$, e.g. $\mathcal{C}(\{CGA, GAT\}) = \{ATC, TCG\}$.

Definition 2. The *trinucleotide circular permutation map* $\mathcal{P}$: $B^3 \to B^3$ is defined by $\mathcal{P}(l_0 l_1 l_2) = l_1 l_2 l_0$ for all $l_0, l_1, l_2 \in B$. The 2nd iterate of $\mathcal{P}$ is $\mathcal{P}^2(l_0 l_1 l_2) = l_2 l_0 l_1$. By extension to a trinucleotide set S, the *set circular permutation map* $\mathcal{P}$: $\mathbb{P}(B^3) \to \mathbb{P}(B^3)$ is defined by $\mathcal{P}(S) = \{v : u, v \in B^3, u \in S, v = \mathcal{P}(u)\}$, e.g. $\mathcal{P}(\{CGA, GAT\}) = \{ATG, GAC\}$ and $\mathcal{P}^2(\{CGA, GAT\}) = \{ACG, TGA\}$.

Definition 3. A set $S \subseteq B^+$ is a *code* if, for each $x_1, \dots, x_n, y_1, \dots, y_m \in S, n, m \geq 1$, the condition $x_1 \cdots x_n = y_1 \cdots y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \dots, n$.

Definition 4. Any non-empty subset of the code B^3 is a code and called *trinucleotide code*.

Definition 5. A trinucleotide code $X \subseteq B^3$ is *self-complementary* if, for each $t \in X$, $\mathcal{C}(t) \in X$, i.e. $X = \mathcal{C}(X)$.

Definition 6. A trinucleotide code $X \subseteq B^3$ is *circular* if, for each $x_1, \dots, x_n, y_1, \dots, y_m \in X$, $n, m \geq 1$, $r \in B^*$, $s \in B^+$, the conditions $s x_2 \cdots x_n r = y_1 \cdots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word) and $x_i = y_i$ for $i = 1, \dots, n$.

The proofs to decide that a code is circular or not are not recalled here, the reader is referred to the proofs based on the flower automaton (Arquès and Michel, 1996), the necklace 5*LDCN* (Letter Diletter Continued Necklace) (Pirillo, 2003), the necklace nLDCCN (Letter Diletter Continued Closed Necklace) with $n \in \{2,3,4,5\}$ (Michel and Pirillo, 2010), and the graph theory (Fimmel et al., 2016).

Definition 7. A trinucleotide circular code $X \subseteq B^3$ is maximal if for all trinucleotide circular codes $Y \subseteq B^3$, we have $|Y| \leq |X|$.

Thus, a trinucleotide circular code $X \subseteq B^3$ has obviously at most 20 trinucleotides and the maximality is 20 trinucleotides on B^3.

Definition 8. A trinucleotide circular code $X \subseteq B^3$ is C^3 *self-complementary* if X, $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ are trinucleotide circular codes such that $X = \mathcal{C}(X)$ (self-complementary), $\mathcal{C}(X_1) = X_2$ and $\mathcal{C}(X_2) = X_1$ (X_1 and X_2 are complementary).

The trinucleotide set X (defined in (1)) coding the reading frame ($f = 0$) in genes is a maximal (20 trinucleotides) C^3 self-complementary trinucleotide circular code (Arquès and Michel, 1996) where the maximal circular code $X_1 = \mathcal{P}(X)$ coding the frame $f = 1$ contains the 20 following trinucleotides

$$X_1 = \{AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG, GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG\} \quad (3)$$

and the maximal circular code $X_2 = \mathcal{P}^2(X)$ coding the frame $f = 2$ contains the 20 following trinucleotides

$$X_2 = \{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT, CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT\}. \quad (4)$$

The trinucleotide circular codes X_1 and X_2 are related by the permutation map, i.e. $X_2 = \mathcal{P}(X_1)$ and $X_1 = \mathcal{P}^2(X_2)$, and by the complementary map, i.e. $X_1 = \mathcal{C}(X_2)$ and $X_2 = \mathcal{C}(X_1)$ (Bussoli et al., 2012).

Several classes of methods were developed for identifying the circular code X in genes over the last 20 years: frame frequency methods (Arquès and Michel, 1996; Frey and Michel, 2003, 2006), correlation function per frame (Arquès and Michel, 1997) and occurrence probability of a

complementary/permutation (CP) trinucleotide set at the gene population level (Michel, 2015) and at the gene level (Michel, 2017).

There exists 216 maximal C^3 self-complementary trinucleotide circular codes $\mathbb{X}$ (Definition 8; Arquès and Michel, 1996; list given in Tables 4a, 5a and 6a in Michel et al., 2008), including the X circular code observed in genes.

2.2. Reading frame and frameshift errors

The translation of a nucleotide sequence into a protein sequence begins at the start codon (generally ATG). The ribosome then reads the following codons in the correct (reading) frame and translates them into amino acids, according to the standard genetic code SGC. Translation is terminated when a stop codon (generally TAA, TAG and TGA) is encountered. If the ribosome shifts on the nucleotide sequence by only one or two bases in either direction, the protein sequence can change dramatically (as illustrated in Table 1). Ribosomal frameshift errors can lead to abnormally short proteins if an out of frame stop codon is read or to non-functional proteins if the out of frame codons are translated into amino acids.

Table 1. Four classes of ribosomal frameshift errors, N being any nucleotide on $B = \{A, C, G, T\}$.

	Frameshift	Trinucleotide sequence
Reading frame	0	ATT CAG GTC GCC
Forward 1 base shift	+1	TTC AGG TCG CCN
Forward 2 base shift	+2	TCA GGT CGC CNN
Backward 1 base shift	-1	NAT TCA GGT CGC
Backward 2 base shift	-2	NNA TTC AGG TCG

We defined two different scores to measure the optimality of a given code to minimise the effects of frameshift errors. First, a code score takes into account all codons (trinucleotides) of a code Y and its two permutated codes Y_1 and Y_2. For example, in the case of a maximal C^3 self-complementary trinucleotide circular code, the $60 = 3 \times 20$ codons of Y, Y_1 and Y_2 are considered. This approach can also be viewed as a codon score. Second, a dicodon score, where the frameshift is analysed according to 1 or 2 base shifts in a dicodon (in reading frame) generated from a code. The code score is defined in Section 2.4 and the dicodon score is defined in Section 2.5. Both measures are based on the average differences in various physicochemical properties between the amino acids (AA) in the original reading frame and the frameshifted amino acids. The matrices used to define the amino acid properties are described in Section 2.3. Section 2.6 defines a multi-objective score based on either the code score or the dicodon score taking into account several amino acid properties simultaneously.

2.3. Amino acid substitution matrices

The effect of a frameshift error is estimated by calculating the absolute difference between the physicochemical properties of the amino acid encoded by the codon in reading frame and the amino acid

encoded by the frameshifted codons in frames $+1$ and -1. We used 11 amino acid properties $\mathbb{P}$: charge $\mathbb{P}_C$, hydrophobicity $\mathbb{P}_H$, isoelectric point $\mathbb{P}_{IP}$, melting point $\mathbb{P}_{MP}$, molecular weight $\mathbb{P}_{MW}$, optical rotation $\mathbb{P}_{OR}$, polarity $\mathbb{P}_{Pr}$, polarizability $\mathbb{P}_{Pz}$, size $\mathbb{P}_{Si}$, steric $\mathbb{P}_{St}$ and volume $\mathbb{P}_V$, extracted from the AAindex database (Kawashima and Kanehisa, 2000) (Table 2 in Appendix). In the AAIndex, a physicochemical property $\mathbb{P}$ is defined by a set of 20 numerical values, representing the absolute or relative value of the property for each amino acid (Table 3 in Appendix). Let us denote an AAindex vector as $\mathbf{V}_{1\times 20}(\mathbb{P})$ for a physicochemical property $\mathbb{P}$ where each element $v_i(\mathbb{P})$ is associated with an amino acid $i \in AA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$.

Example 1. $v_G(\mathbb{P}_V) = 36.3$ is the score of the amino acid glycine G for the volume property $\mathbb{P}_V$ (Table 3 in Appendix).

For a physicochemical property $\mathbb{P}$, we construct an amino acid substitution matrix $\mathbf{M}_{20\times 20}(\mathbb{P})$ of absolute differences $m_{ij}(\mathbb{P})$ between the physicochemical values $v_i(\mathbb{P})$ of the amino acid i and $v_j(\mathbb{P})$ of the amino acid j:

$$m_{ij}(\mathbb{P}) = \left|v_i(\mathbb{P}) - v_j(\mathbb{P})\right| \tag{5}$$

where $v_i(\mathbb{P})$ and $v_j(\mathbb{P})$ are the physicochemical values of the amino acids i and j, $i, j \in AA$.

The matrices $\mathbf{M}(\mathbb{P})$ are symmetric with diagonal elements equal to zero.

Example 2. An example of an amino acid substitution matrix $\mathbf{M}(\mathbb{P}_V)$ for the volume property $\mathbb{P}_V$ is provided in Table 4 in Appendix.

Example 3. For the volume property $\mathbb{P}_V$, the substitution value for the amino acids glycine G and proline P is equal to $m_{GP}(\mathbb{P}_V) = m_{PG}(\mathbb{P}_V) = |v_G(\mathbb{P}_V) - v_P(\mathbb{P}_V)| = |36.3 - 73.6| = 37.3$ (Table 4 in Appendix).

The values of the different amino acid properties have different scales (Table 3 in Appendix). For example, the 20 amino acids have a mean value of 262.7 and a standard deviation of 43.6 for the melting point property $\mathbb{P}_{Mp}$ while they have a mean value of -10.6 and a standard deviation of 24.3 for the optical rotation property $\mathbb{P}_{Or}$. To allow direct comparisons between the various amino acid properties, each amino acid substitution matrix $\mathbf{M}_{20\times 20}(\mathbb{P})$ is normalized by dividing each element of the given matrix by the sum of the whole matrix, leading to the normalized amino acid substitution matrix $\widehat{\mathbf{M}}_{20\times 20}(\mathbb{P})$:

$$\widehat{m}_{ij}(\mathbb{P}) = \frac{1000}{\sum_{i=1}^{20}\sum_{j=1}^{20} m_{ij}(\mathbb{P})} m_{ij}(\mathbb{P}) \tag{6}$$

where $m_{ij}(\mathbb{P})$ is defined in Equation (9) for the amino acids i and j, $i, j \in AA$.

The matrices $\widehat{\mathbf{M}}(\mathbb{P})$ are also symmetric with diagonal elements equal to zero.

Example 4. An example of a normalized amino acid substitution matrix $\mathbf{M}(\mathbb{P}_V)$ for the volume property $\mathbb{P}_V$ is provided in Table 5 in Appendix.

Example 5. With Example 3, the normalized substitution value for the amino acids glycine G and proline P for the volume property $\mathbb{P}_V$ is equal to $\widehat{m}_{GP}(\mathbb{P}_V) = \widehat{m}_{PG}(\mathbb{P}_V) = \frac{1000}{\sum_{i=1}^{20}\sum_{j=1}^{20} m_{ij}(\mathbb{P})} m_{PG}(\mathbb{P}_V) = \frac{1000}{10790.8} 37.3 = 3.5$.

2.4. Code score for measuring frameshift optimality

The code score considers the frameshift errors from a code Y point of view. The codes Y analysed are: (i) the maximal C^3 self-complementary trinucleotide circular code X identified in genes (defined in (1)); (ii) the 215 circular codes $\mathbb{X}\backslash X$; and (iii) the standard genetic code SGC. A codon $c = l_0 l_1 l_2$ of a code $Y \subseteq B^3$ is associated with the reading frame $f = 0$, the shifted codon $\mathcal{P}(c) = l_1 l_2 l_0$ of the code $Y_1 = \mathcal{P}(Y) \subseteq B^3$ is obviously associated with the shifted frame $f = 1$ (+1) and the shifted codon $\mathcal{P}^2(c) = l_2 l_0 l_1$ of the code $Y_2 = \mathcal{P}^2(Y) \subseteq B^3$ is obviously associated with the shifted frame $f = -1$ (+2). In short, the code Y is associated with the reading frame $f = 0$, the shifted code Y_1 is associated with the shifted frame $f = 1$ and the shifted code Y_2 is associated with the shifted frame $f = -1$.

The code score is defined by the average difference for a given amino acid property $\mathbb{P}$ when all codons of a given code Y are substituted into all shifted codons of a shifted code Y_1 or Y_2. Thus, two code scores will be defined: one for the shifted frame $f = 1$ and one for the shifted frame $f = -1$. These two scores will be measured for the three classes of codes Y defined above.

As the definition is based on an amino acid property, only the sense codons (i.e. codons coding for an amino acid) are considered in a code Y, thus the three stop codons $S = \{TAA, TAG, TGA\}$ are excluded. The two permutation sets of S are $S_1 = \mathcal{P}(S) = \{AAT, AGT, GAT\}$ and $S_2 = \mathcal{P}^2(S) = \{ATA, ATG, GTA\}$.

The code score $CS_{+1}(Y)$ in a $+1$ frameshift of a code Y is defined by

$$CS_{+1}(Y, \mathbb{P}) = \frac{1}{|Y\backslash(S \cup S_2)|} \sum_{c \in Y\backslash(S \cup S_2)} \widehat{m}_{ij}(\mathbb{P}) \tag{7}$$

where the codon $c \in Y\backslash(S \cup S_2)$ belongs to the code Y excluding the stop codons S and the codons S_2 (as S_2 in frame 0 leads to $\mathcal{P}(S_2) = S$ in $+1$ frameshift), $\widehat{m}_{ij}(\mathbb{P})$ is the value of the normalized substitution matrix (Equation (6)) of an AA property $\mathbb{P}$ where i and j are the amino acids coded by the codons $c \in Y$ and $\mathcal{P}(c) \in Y_1 = \mathcal{P}(Y)$ (remember that $\widehat{\mathbf{M}}$ is symmetric).

Similarly, the code score $CS_{-1}(Y)$ in a -1 frameshift of a code Y is defined by

$$CS_{-1}(Y, \mathbb{P}) = \frac{1}{|Y\backslash(S \cup S_1)|} \sum_{c \in Y\backslash(S \cup S_1)} \widehat{m}_{ij}(\mathbb{P}) \tag{8}$$

where the codon $c \in Y\backslash(S \cup S_1)$ belongs to the code Y excluding the stop codons S and the codons S_1 (as S_1 in frame 0 leads to $\mathcal{P}^2(S_1) = S$ in -1 frameshift), $\widehat{m}_{ij}(\mathbb{P})$ is the value of the normalized substitution matrix (Equation (6)) of an AA property $\mathbb{P}$ where i and j are the amino acids coded by the codons $c \in Y$ and $\mathcal{P}^2(c) \in Y_2 = \mathcal{P}^2(Y)$.

Remark 1. For the circular code $Y = X$, $X \cap S = \emptyset$ (X has 20 sense codons, defined in (1)), $X \cap S_2 = \{GTA\}$ (X_1 has 19 sense codons and one stop codon $\mathcal{P}(\{GTA\}) = \{TAG\}$, defined in (3)) and $X \cap S_1 = \{AAT, GAT\}$ (X_2 has 18 sense codons and two stop codons $\mathcal{P}^2(\{AAT, GAT\}) = \{TAA, TGA\}$, defined in (4)). Thus, for Equation (7), $X\backslash(S \cup S_2) = X\backslash\{GTA\}$ and $|X\backslash\{GTA\}| = 20 - 1 = 19$ and for Equation (8), $X\backslash(S \cup S_1) = X\backslash\{AAT, GAT\}$ and $|X\backslash\{AAT, GAT\}| = 20 - 2 = 18$.

Remark 2. For the standard genetic code $Y = \text{SGC} = B^3$, $Y \cap S = S$ (Y has 61 sense codons and three stop codons S), $Y \cap S_2 = S_2$ (Y_1 has 61 sense codons and three stop codons $\mathcal{P}(S_2) = S$) and $X \cap S_1 = S_1$

(Y_2 has 61 sense codons and three stop codons $\mathcal{P}^2(S_1) = S$). Thus, for Equation (7), $Y\backslash(S \cup S_2) = B^3\backslash\{ATA, ATG, GTA, TAA, TAG, TGA\}$ and $|Y\backslash(S \cup S_2)| = 64 - 6 = 58$ and for Equation (8), $Y\backslash(S \cup S_1) = B^3\backslash\{AAT, AGT, GAT, TAA, TAG, TGA\}$ and $|Y\backslash(S \cup S_1)| = 64 - 6 = 58$.

Remark 3. For the 215 circular codes $\mathbb{X}\backslash X$, the codes having none, one or several stop codons are analysed similarly.

2.5. Dicodon score for measuring frameshift optimality

The dicodon score considers the frameshift errors from a code motif point of view, precisely a motif with two consecutive trinucleotides, called a dicodon, from a code Y. As with the code score, the codes Y analysed are: (i) the maximal C^3 self-complementary trinucleotide circular code X identified in genes (defined in (1)); (ii) the 215 circular codes $\mathbb{X}\backslash X$; and (iii) the standard genetic code SGC. A codon $c = l_0l_1l_2$ of a code $Y \subseteq B^3$ is associated with the reading frame $f = 0$. The shifted frames $f = 1$ (+1) and $f = -1$ (+2) are obtained from the dicodons. Let a dicodon $c \cdot c' = l_0l_1l_2 \cdot l'_0l'_1l'_2$ such that the codon $c' = l'_0l'_1l'_2$ also belongs to the code $Y \subseteq B^3$. Let the map $Q: B^3 \times B^3 \to B^3$. Then, the shifted codon $Q(c \cdot c') = l_1l_2l'_0$ is obviously associated with the shifted frame $f = 1$ and the shifted codon $Q^2(c \cdot c') = l_2l'_0l'_1$ is obviously associated with the shifted frame $f = -1$. In contrast to the code score, the shifted codon $Q(c \cdot c')$ does not necessarily belong to the code $Y_1 = \mathcal{P}(Y) \subseteq B^3$ and the shifted codon $Q^2(c \cdot c')$ does not necessarily belong to the code $Y_2 = \mathcal{P}^2(Y) \subseteq B^3$.

The dicodon score is defined by the average difference for a given amino acid property $\mathbb{P}$ when all codons $c = l_0l_1l_2$ of all dicodons $c \cdot c' = l_0l_1l_2 \cdot l'_0l'_1l'_2$ of a given code Y are "substituted" into the shifted codons $Q(c \cdot c') = l_1l_2l'_0$ or $Q^2(c \cdot c') = l_2l'_0l'_1$. As with the code score, only the sense codons are considered in the dicodons of a code Y. Let us denote the set of dicodons containing a stop codon as $DS = \{c \cdot c'\}$, where $c \in S$ or $c' \in S$. The two sets of dicodons that result in a stop codon are $DS_1 = \{NTA.ANN, NTA.GNN, NTG.ANN\}$ for the +1 frameshift and $DS_2 = \{NNT.AAN, NNT.AGN, NNT.GAN\}$ for the -1 frameshift, N being any letter on B^3.

The dicodon score $CS_{+1}(Y)$ in a +1 frameshift of a code Y is defined by

$$DS_{+1}(Y, \mathbb{P}) = \frac{1}{|Y^2\backslash(DS \cup DS_1)|} \sum_{c \cdot c' \in Y^2\backslash(DS \cup DS_1)} \widehat{m}_{ij}(\mathbb{P}) \tag{9}$$

where the dicodon $c \cdot c'$ belong to the code Y^2 excluding the stop codons DS and DS_1, $\widehat{m}_{ij}(\mathbb{P})$ is the value of the normalized substitution matrix (Equation (6)) of an AA property $\mathbb{P}$ where i and j are the amino acids coded by the codons $c \in Y$ and $Q(c \cdot c')$.

Similarly, the dicodon score $DS_{-1}(Y)$ in a -1 frameshift of a code Y is defined by

$$DS_{-1}(Y, \mathbb{P}) = \frac{1}{|Y^2\backslash(DS \cup DS_2)|} \sum_{c \cdot c' \in Y^2\backslash(DS \cup DS_2)} \widehat{m}_{ij}(\mathbb{P}) \tag{10}$$

where the dicodon $c \cdot c'$ belong to the code Y^2 excluding the stop codons DS and DS_2, $\widehat{m}_{ij}(\mathbb{P})$ is the value of the normalized substitution matrix (Equation (6)) of an AA property $\mathbb{P}$ where i and j are the amino acids coded by the codons $c \in Y$ and $Q^2(c \cdot c')$.

2.6. Multi-objective optimality score

The multi-objective score is based on either the code score or the dicodon score and takes into account several amino acid properties simultaneously. To compare the optimality of the X circular code with the $|\mathbb{X}| = 216$ maximal C^3 self-complementary circular codes $\mathbb{X}$ when a combination of the $|\mathbb{P}| = 11$ AA properties $\mathbb{P}$ is taken into account, we calculated the number N_i, for $i = 0, \ldots, |\mathbb{P}|$, of AA properties that were optimized better with the codes $x, x \in \mathbb{X}\backslash\mathrm{X}$, than with the circular code X. Hence, for $i = 0, \ldots, |\mathbb{P}|$,

$$N_i(\mathcal{S}) = \sum_{x \in \mathbb{X}} \Delta_i \left(\sum_{j=1}^{|\mathbb{P}|} \delta\left(x, \mathbb{P}_j\right) \right) \tag{11}$$

where

$$\delta\left(x, \mathbb{P}_j\right) = \begin{cases} 1 & \text{if } \mathcal{S}\left(x, \mathbb{P}_j\right) \leq \mathcal{S}\left(X, \mathbb{P}_j\right) \\ 0 & \text{otherwise} \end{cases},$$

$$\Delta_i(k) = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{otherwise} \end{cases},$$

the code score $\mathcal{S} \in \{CS_{+1}, CS_{-1}, DS_{+1}, DS_{-1}\}$ and

$j \in \mathbb{P} = \{\mathbb{P}_C, \mathbb{P}_H, \mathbb{P}_{IP}, \mathbb{P}_{MP}, \mathbb{P}_{MW}, \mathbb{P}_{OR}, \mathbb{P}_{Pr}, \mathbb{P}_{Pz}, \mathbb{P}_{Si}, \mathbb{P}_{St}, \mathbb{P}_V\}$.

Remark 4. If $x = X$ then $\delta\left(x, \mathbb{P}_j\right) = 1$ for any $\mathbb{P}_j$, thus $\sum_{j=1}^{|\mathbb{P}|} \delta\left(x, \mathbb{P}_j\right) = |\mathbb{P}|$ and $N_{|\mathbb{P}|}(\mathcal{S}) \geq 1$.

Remark 5. If $N_{|\mathbb{P}|}(\mathcal{S}) = 1$ then the X circular code is optimal among its combinatorial class of the 216 maximal C^3 self-complementary circular codes $\mathbb{X}$.

Remark 6. $\sum_{i=0}^{|\mathbb{P}|} N_i(\mathcal{S}) = |\mathbb{X}|$.

3. Results

The section is divided into two main parts. In the first part, we estimate the capacity of the X circular code to reduce the effects of a frameshift error, and compare it to the capacity of the standard genetic code (SGC). In the second part, we investigate the frameshift optimality of the 216 maximal self-complementary C^3 circular codes $\mathbb{X}$. Indeed, since the discovery of the X circular code in genes in 1996, the question remains of why this particular code was chosen among its combinatorial class $\mathbb{X}$ of 216 maximal self-complementary C^3 circular codes. Despite numerous combinatorial studies, this approach has not provided any answers. In particular, transformations of the X circular code by letter invariance with respect to complementarity lead to circular codes in $\mathbb{X}$ with combinatorial properties identical to that of X. Unexpectedly, we will show that a solution to this problem is of biological and biochemical origin.

From a biological point of view, forward $(+1)$ and backward (-1) frameshifts are fundamentally different events (Abrahams and Hurst, 2018). Forward frameshifts are assumed to be the more frequent form of accidental ribosomal slippage. As translation occurs in the 5' to 3' direction, the molecular mechanics required to halt and reverse the direction of translation during a backward frameshift are likely to be more complex and require greater energy than for a ribosome to skip to the $+1$ frame in the

same direction. We therefore considered $+1$ and -1 frameshifts independently in the following analyses.

3.1. Frameshift code score of the *X* circular code and the standard genetic code SGC

To estimate the effects of either a $+1$ or -1 frameshift error on the encoded amino acids (AA), we first computed the code scores $CS_{+1}(Y)$ (Equation (7)) in a $+1$ frameshift and $CS_{-1}(Y)$ (Equation (8)) in a -1 frameshift of a code Y, where $Y = X$ for the X circular code and $Y = \text{SGC}$ for the standard genetic code, for a set of 11 fundamental AA amino acid properties (Table 2 in Appendix). These scores measure the difference between the physicochemical properties for the AA coded by the non-shifted codons of Y and the shifted codons of Y_1 for the $+1$ frameshift and of Y_2 for the -1 frameshift. Thus, a smaller score indicates a smaller effect of the frameshift error, and hence a better optimality of the code. The results for the X circular code and the standard genetic code SGC are shown in Figure 2.

The code scores obtained for the SGC are the same for the $+1$ and -1 frameshifts with the 11 AA properties $\mathbb{P}$, as expected due to the symmetry of the 64 codons (Figure 2A,B). Thus, for all $\mathbb{P}$, $CS_{+1}(\text{SGC}, \mathbb{P}) = CS_{-1}(\text{SGC}, \mathbb{P})$. However, the code scores of X are clearly different for $+1$ and -1 frameshifts (Figure 2A,B), i.e. for all $\mathbb{P}$, $CS_{+1}(X, \mathbb{P}) \neq CS_{-1}(X, \mathbb{P})$.

In the case of a $+1$ frameshift, the code scores obtained for polarity $\mathbb{P}_{Pr}$, molecular weight $\mathbb{P}_{MW}$, isoelectric point $\mathbb{P}_{IP}$, polarizability $\mathbb{P}_{Pz}$, volume $\mathbb{P}_{V}$, size $\mathbb{P}_{Si}$ and charge $\mathbb{P}_{C}$ are smaller for X than for SGC (Figure 2A), i.e. for $\mathbb{P} \in \{\mathbb{P}_{Pr}, \mathbb{P}_{MW}, \mathbb{P}_{IP}, \mathbb{P}_{Pz}, \mathbb{P}_{V}, \mathbb{P}_{Si}, \mathbb{P}_{C}\}$,

$$CS_{+1}(X, \mathbb{P}) < CS_{+1}(\text{SGC}, \mathbb{P}). \quad (12)$$

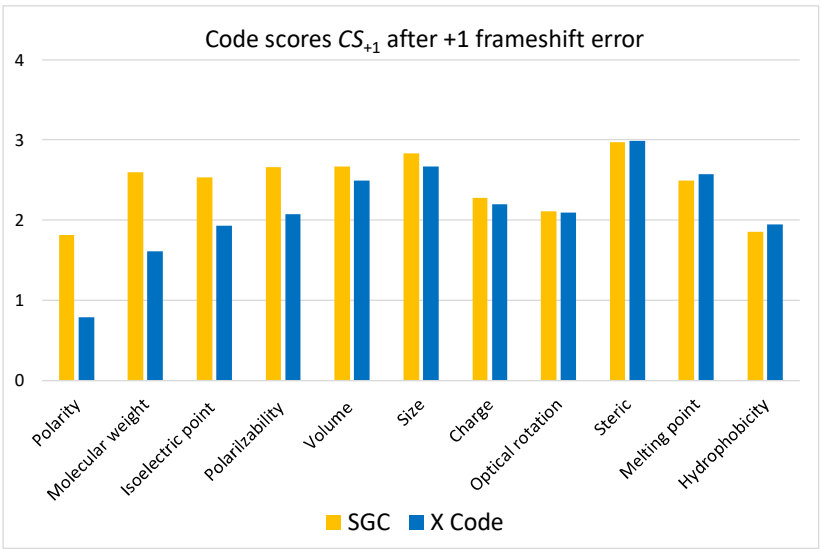
The remaining properties, namely optical rotation $\mathbb{P}_{OR}$, steric $\mathbb{P}_{St}$, melting point $\mathbb{P}_{MP}$ and hydrophobicity $\mathbb{P}_{H}$, are similar for both codes X and SGC (Figure 2A), i.e. for $\mathbb{P} \in \{\mathbb{P}_{OR}, \mathbb{P}_{St}, \mathbb{P}_{MP}, \mathbb{P}_{H}\}$,

$$CS_{+1}(X, \mathbb{P}) \approx CS_{+1}(\text{SGC}, \mathbb{P}). \quad (13)$$

In contrast, after a -1 frameshift, the code scores for most of the properties $\mathbb{P}$ are larger for X than for SGC (Figure 2B), with the exception of the optical rotation $\mathbb{P}_{OR}$, i.e. for $\mathbb{P} \neq \mathbb{P}_{OR}$,

$$CS_{-1}(X, \mathbb{P}) > CS_{-1}(\text{SGC}, \mathbb{P}). \quad (14)$$

A



B

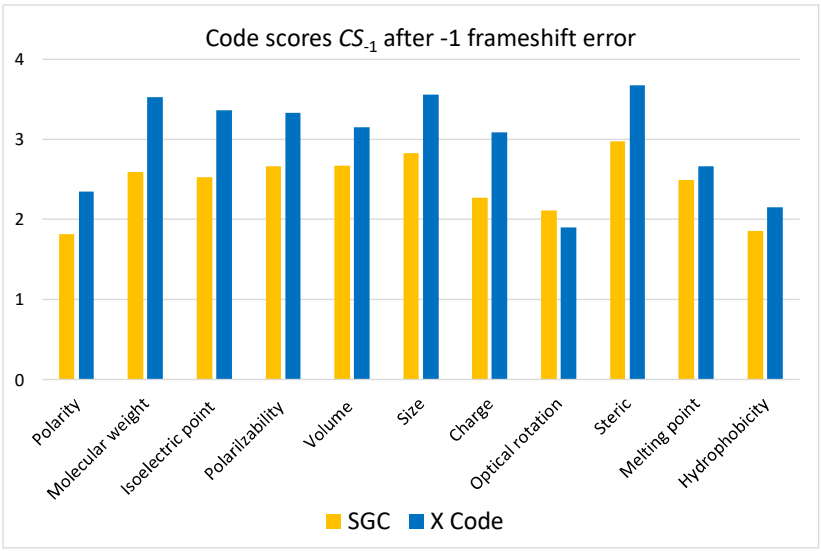


Figure 2. Frameshift code score of the X circular code and the standard genetic code SGC. **A**. Code score CS_{+1} (Equation (7)) after a $+1$ frameshift error. The 11 AA properties $\mathbb{P}$ are ordered according to the difference between the code scores for the SGC and X. **B**. Code score CS_{-1} (Equation (8)) after a -1 frameshift error. For comparison purposes, the AA properties $\mathbb{P}$ are shown in the same order as **A**.

In summary, in a $+1$ frameshift, the X circular code is better optimized than the standard genetic code SGC for 7 AA properties: polarity, molecular weight, isoelectric point, polarizability, volume, size and charge.

3.2. Frameshift dicodon score of the *X* circular code and the standard genetic code SGC

As mentioned in the Introduction, circular codes have the ability to retrieve and maintain the reading frame using an appropriate window of nucleotides, for example with the X circular code, a window of at most 13 consecutive nucleotides is sufficient. This led us to consider the code optimality for the same AA properties at the motif level, and more specifically with the dicodon scores $DS_{+1}(Y)$ (Equation (9)) in a $+1$ frameshift and $DS_{-1}(Y)$ (Equation (10)) in a -1 frameshift of a code Y, where $Y = X$ for the X circular code and $Y = \text{SGC}$ for the standard genetic code, for a set of 11 fundamental amino acid properties (Table 2 in Appendix). Again, we observe the same optimality scores in case of the SGC for the $+1$ and -1 frameshifts with the 11 AA properties $\mathbb{P}$ (Figure 3A,B), as expected, $DS_{+1}(\text{SGC}, \mathbb{P}) = DS_{-1}(\text{SGC}, \mathbb{P})$. Again, the dicodon scores of X are clearly different for $+1$ and -1 frameshifts (Figure 3A,B), i.e. for all $\mathbb{P}$, $DS_{+1}(X, \mathbb{P}) \neq DS_{-1}(X, \mathbb{P})$.

After a $+1$ frameshift, the X circular code has smaller scores than the SGC for all AA properties except hydrophobicity $\mathbb{P}_H$ (Figure 3A), i.e. for $\mathbb{P} \neq \mathbb{P}_H$,

$$DS_{+1}(X, \mathbb{P}) < DS_{+1}(\text{SGC}, \mathbb{P}). \quad (15)$$

In contrast, after a -1 frameshift, the SGC achieves smaller scores than the X circular code for all AA properties (Figure 3B), except for the optical rotation $\mathbb{P}_{OR}$ and the melting point $\mathbb{P}_{MP}$, i.e. for $\mathbb{P} \neq \{\mathbb{P}_{OR}, \mathbb{P}_{MP}\}$,

$$DS_{-1}(X, \mathbb{P}) > DS_{-1}(\text{SGC}, \mathbb{P}). \quad (16)$$

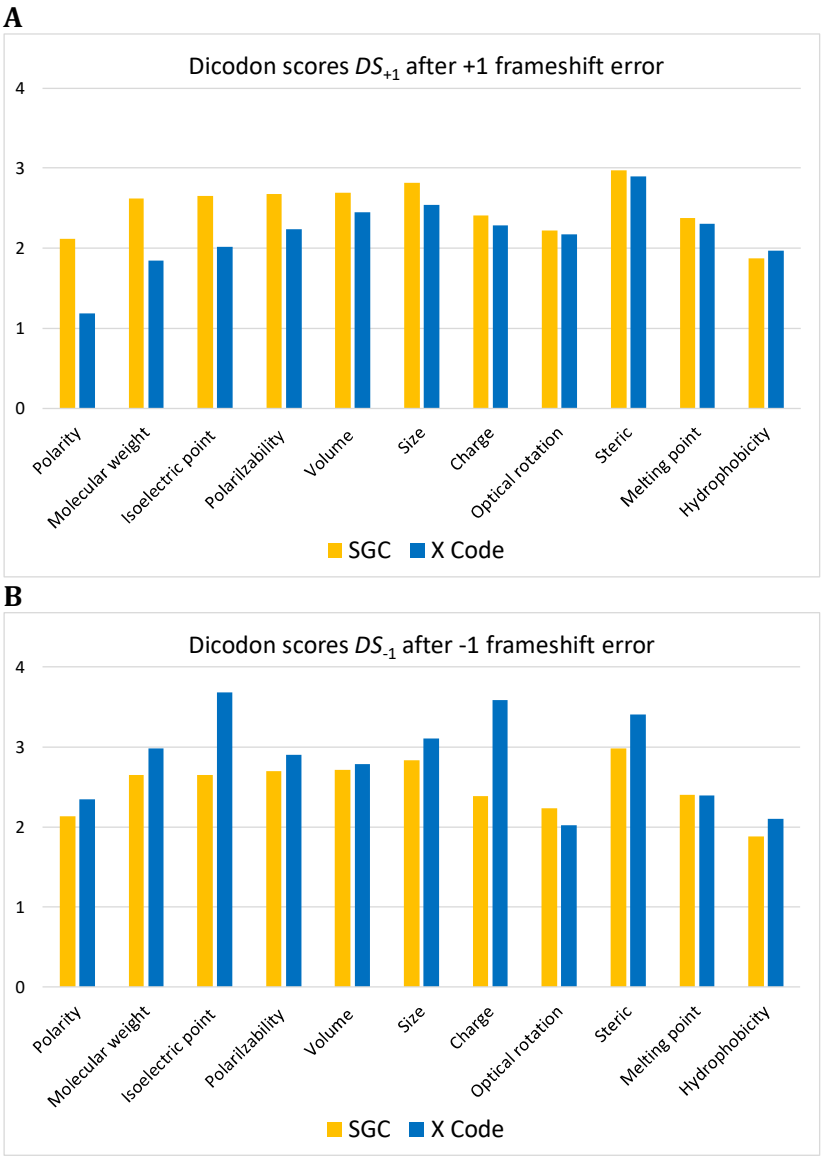


Figure 3. Frameshift dicodon score of the X circular code and the standard genetic code SGC. **A**. Dicodon score DS_{+1} (Equation (9)) after a $+1$ frameshift error. **B**. Dicodon score DS_{-1} (Equation (10)) after a -1 frameshift error. For comparison purposes, the AA properties $\mathbb{P}$ are shown in the same order as in Figure 2.

In summary, in a $+1$ frameshift, the X circular code is better optimized than the standard genetic code SGC for 10 AA properties (except hydrophobicity $\mathbb{P}_H$). Thus, two different classes of results obtained from the X circular code (Figure 2) and the motifs (dicodons) of the X circular code (Figure 3) lead us to conclude that the X circular code is better optimized to minimize the effects of $+1$ frameshift errors than the standard genetic code SGC.

3.3. Frameshift code score of the 216 maximal complementary C^3 circular codes $\mathbb{X}$

In the next two sections, we explore the capacity of the 216 maximal self-complementary C^3 circular codes $\mathbb{X}$ (including the X circular code) to minimize the effects of frameshift errors. Using the same method as above, we calculated the frameshift code scores and the frameshift dicodon scores, using the same set of amino acid (AA) properties $\mathbb{P}$. As mentioned in the previous section, these scores measure

differences between the physicochemical properties for the AA and therefore a smaller score indicates a smaller effect of the frameshift error, and hence a better optimality of the code. For each individual AA property measured either with the code score or with the dicodon score in the $+1$ or -1 frameshifts, there exists a different circular code x among the 215 with better optimality than the X circular code (data not shown).

Since specific circular codes exist that are more optimized for individual AA properties, we wanted to test the hypothesis that the X circular code is optimized to minimize a combination of the AA properties $\mathbb{P}$, rather than a single one. To achieve this, for each of the 216 maximal self-complementary C^3 circular codes x, we calculated a multi-objective score N_i (Equation (11)) corresponding to the number i of AA properties that were optimized better with this code than with the X circular code.

We first considered the multi-objective code score (Figure 4). After a $+1$ frameshift, a significant number of 216 maximal C^3 self-complementary circular codes $\mathbb{X}$ optimize a combination of up to 5 AA properties $\mathbb{P}$ better than the X circular code (i.e. circular codes $\mathbb{X}$ with $N_i(CS_{+1}) \leq 5$; Figure 4A). However, when more than 6 AA properties $\mathbb{P}$ are taken into account, the X code is one of the best 18 codes, i.e. the X code is in the top 8% of the 216 codes $\mathbb{X}$. Furthermore, no other circular codes $\mathbb{X}$ achieve the same optimality as the X code for 10 or 11 AA properties $\mathbb{P}$ ($N_{11}(CS_{+1}) = 1$ and $N_{10}(CS_{+1}) = 0$; Figure 4A). In the case of a -1 frameshift, 39 of the 216 codes $\mathbb{X}$ (18%) are more optimal than the X code when up to 10 AA properties are combined, and only one other code x achieves the best optimality for all 11 AA properties ($N_{11}(CS_{-1}) = 2$ and $N_{10}(CS_{-1}) = 39$; Figure 4B). The code x consists of the following 20 trinucleotides:

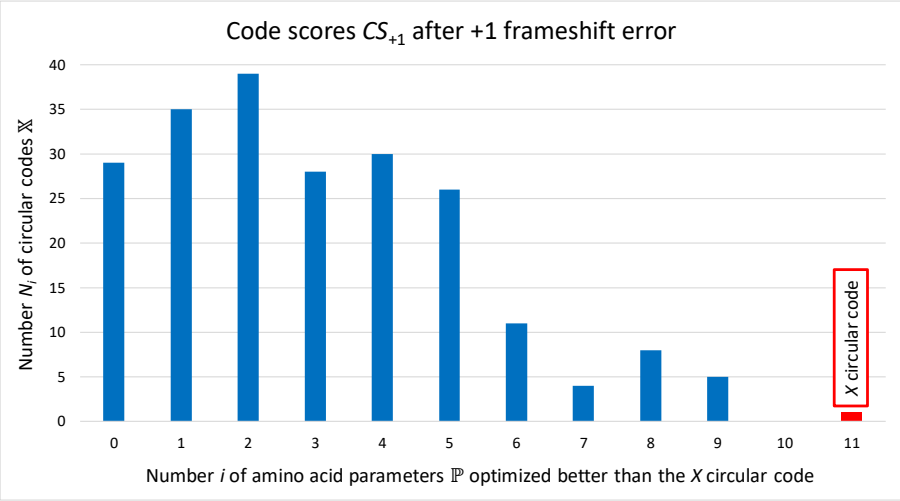
$$x = \{ATC, CAA, CAC, CAG, CTG, GAA, GAC, GAT, GCC, GGA, GGC, GTA, GTC, GTG, TAA, TAC, TCC, TTA, TTC, TTG\} \quad (17)$$

and codes the stop codon TAA and the 12 following amino acids:

$$\{Ala, Asp, Gln, Glu, Gly, His, Ile, Leu, Phe, Ser, Tyr, Val\}.$$

However, this maximal circular code x with a stop codon cannot exist in the reading frame of genes. Thus, the maximal circular code X could be considered optimal.

A.



B.

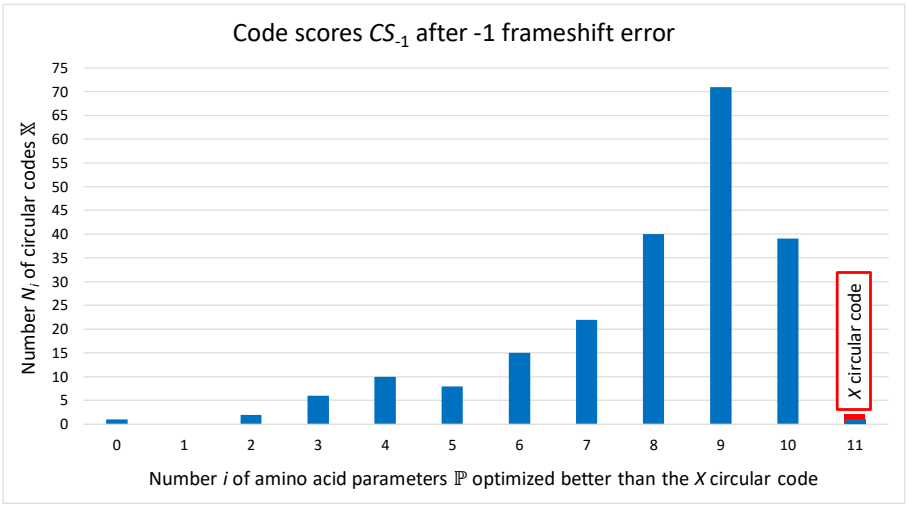


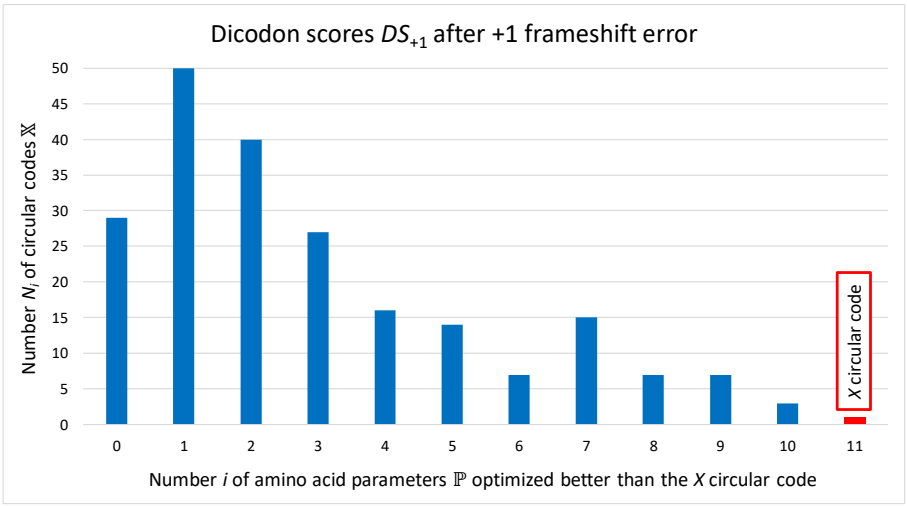
Figure 4. Number $N_i(\mathcal{S})$ (Equation (11)) of 216 maximal C^3 self-complementary circular codes $\mathbb{X}$ that optimize a combination of AA properties $\mathbb{P}$ better than or equal to the X circular code, for a number i of amino acid properties varying from 0 to 11. **A**. Multi-objective code score $N_i(CS_{+1})$ (Equations (7) and (11)) after a $+1$ frameshift error. **B**. Multi-objective code score $N_i(CS_{-1})$ (Equations (8) and (11)) after a -1 frameshift error.

We conclude that the X circular code is the best maximal C^3 self-complementary circular codes $\mathbb{X}$, in terms of minimizing the overall effects of $+1$ frameshift events on the translated AA sequence.

3.4. Frameshift dicodon score of the 216 maximal complementary C^3 circular codes $\mathbb{X}$

We then considered the multi-objective dicodon score (Figure 5). We observe very similar distributions of optimal codes after a $+1$ or -1 frameshift. After a $+1$ frameshift, only 3 of the 216 codes $\mathbb{X}$ (1%) optimize 10 AA properties $\mathbb{P}$ better than the X circular code, and again the code X achieves the best optimality for all 11 AA properties ($N_{11}(DS_{+1}) = 1$ and $N_{10}(DS_{+1}) = 3$; Figure 5A). After a -1 frameshift, 12 of the 216 codes $\mathbb{X}$ (6%) optimize 10 AA properties better than the X circular code, and only one other code x, the same code described by Equation (17), achieves the best optimality for all 11 AA properties ($N_{11}(DS_{-1}) = 2$ and $N_{10}(DS_{-1}) = 12$; Figure 5B).

A.



B.

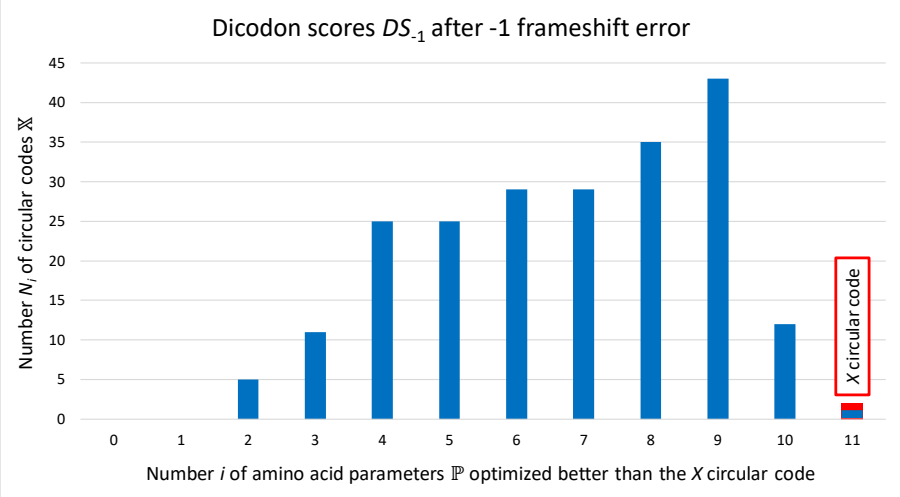


Figure 5. Number $N_i(\mathcal{S})$ (Equation (11)) of 216 maximal C^3 self-complementary circular codes $\mathbb{X}$ that optimize a combination of AA properties $\mathbb{P}$ better than or equal to the X circular code, for a number i of amino acid properties varying from 0 to 11. **A**. Multi-objective dicodon score $N_i(DS_{+1})$ (Equations (9) and (11)) after a $+1$ frameshift error. **B**. Multi-objective dicodon score $N_i(DS_{-1})$ (Equations (10) and (11)) after a -1 frameshift error.

4. Conclusion

Translation of mRNA sequences to proteins is one of the most error-prone processes affecting all domains of life and evidence shows that translation errors reduce the fitness of an organism (Wilke, 2015). Therefore, to minimize the costs of errors, organisms have evolved complex mechanisms for either error prevention by reducing the frequency of errors leading to increased translational accuracy, or error mitigation by minimizing the consequences of errors leading to increased robustness (Drummond and Wilke, 2009). For example, it is widely accepted that the standard genetic code (SGC) is optimized to reduce the impact of errors caused by incorporation of wrong amino acids or by ribosomal frameshifting.

The work described in this paper addresses the question of whether the X circular code is also optimized in some way to minimize frameshift errors. We recall that the main property of a circular code is to retrieve the reading frame. We performed a comprehensive evaluation of the optimality of different codes, and measured the differences in the amino acid (AA) sequences produced after a frameshift. While most previous studies of code optimization have estimated AA differences in terms of changes in polarity or volume, here we considered a wider range of properties, including charge, hydrophobicity, isoelectric point, melting point, molecular weight, optical rotation, polarity, polarizability, size, steric effect and volume. This set of 11 properties provide a better picture of the potential changes to the physico-chemical properties of the translated protein sequence. Furthermore, the chosen properties are associated with the fundamental chemistry of the amino acid regarded as an elementary unit, i.e. chemical properties which would have acted in a primitive environment (Earth, solar and extrasolar planets, etc.). However, numerous other amino acid properties in extant proteins could be considered in the future, in particular those associated with the 3-dimensional structure such as preferences for

alpha-helix or beta-sheet conformations (Chou and Fasman, 1978), surface accessibility (Chothia, 1976), etc. For example, the X circular code is not optimal compared to the SGC and the other circular codes $\mathbb{X}$ for the alpha-helix and beta-sheet preference properties (data not shown).

We introduced two scores that estimate the optimality of the codes. First, a code score is calculated over all codons of a code Y, where the frameshift is represented by a circular permutation of the code. Second, a dicodon score is calculated over all possible dicodons generated from a code Y, where a frameshift results in a 1 or 2 base shift of the reading frame. The dicodon score was designed to investigate the effects of frameshifts in a DNA sequence motif. In this work, we restricted the sequence motif to a length of two codons, but in the future this could be extended to longer motifs.

We also considered the events of forward $(+1)$ and backward (-1) frameshifts separately, since it is known that the biological mechanisms involved in the two types of frameshift are very different. Indeed, $+1$ frameshifts are more energy efficient and are generally much more frequent than -1 frameshifts. Using both code-level and dicodon-level scores, we have shown that the X circular code is more optimized than the SGC to reduce the effects of $+1$ frameshifts, in particular with respect to the AA volume, size and molecular weight, as well as the polarity, isoelectric point, polarizability, and charge properties. In contrast, in case of a -1 frameshift, the SGC was generally more optimized than the X circular code. Furthermore, we have shown that the X code is the most optimized of the 216 C^3 self-complementary circular codes (1st with $+1$ frameshifts, 2nd with -1 frameshifts), when all the AA properties are taken into account, thus providing a solution to a question that has been open since 1996. Based on these results, it is tempting to suggest that, in addition to its frameshift synchronization property, the X circular code may also play a role in error mitigation of the more frequent $+1$ frameshift events. In contrast, the rarity of -1 frameshift events means that reduction of their effects would be less useful.

The presence of out-of-frame stop codons in the coding sequences has also been proposed to be a frameshift catch and destroy mechanism, limiting the effects of frameshift errors by terminating the translation as soon as possible after the frameshift event. However, this mechanism requires a sophisticated molecular apparatus for stop codon recognition, including a set of protein release factors (Adio et al., 2018). We have hypothesized that circular codes represented an important step in the emergence of the modern genetic code, allowing simultaneous coding of amino acids as well as synchronization of the reading frame in primitive translation systems, prior to the advent of more sophisticated mechanisms (Dila et al., 2019b). The X circular code does not contain stop codons and would have allowed the detection and mitigation of frameshift errors in primitive systems before the evolution of the stop codon recognition machinery.

In addition to further exploring the possibility that the X circular code is the possible ancestor of the modern genetic code, our ongoing studies are now focused on the hypothesis that circular code motifs continue to act as functional elements within the coding regions of extant genomes.

REFERENCES

Abrahams, L., Hurst, L.D., 2018. Refining the ambush hypothesis: Evidence that GC- and AT-rich bacteria employ different frameshift defence strategies. Genome Biology and Evolution 10, 1153-1173.

Adio, S., Sharma, H., Senyushkina, T., Karki, P., Maracci, C., Wohlgemuth, I., Holtkamp, W., Peske, F., Rodnina, M.V., 2018. Dynamics of ribosomes and release factors during translation termination in *E. coli*. eLife 7, e34252.

Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. Journal of Theoretical Biology 182, 45-58.

Arquès, D.G., Michel, C.J., 1997. A code in the protein coding genes. Biosystems 44, 107-134.

Bussoli, L., Michel, C.J., Pirillo, G., 2012. On conjugation partitions of sets of trinucleotides. Applied Mathematics 3, 107-112.

Chothia, C., 1976. The nature of the accessible and buried surfaces in proteins. Journal Molecular Biology 105, 1-14.

Chou, P.Y., Fasman, G.D., 1978. Prediction of the secondary structure of proteins from their amino acid sequence. Advances in Enzymology 47, 45-148.

Crick, F.H., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proceedings of the National Academy of Sciences of the United States of America 43, 416-421.

Demongeot, J., Seligmann, H., 2019. Spontaneous evolution of circular codes in theoretical minimal RNA rings. Gene 705, 95-102.

Demongeot, J., Seligmann, H., 2020. Pentamers with Non-redundant Frames: Bias for Natural Circular Code Codons. Journal Molecular Evolution 88, 194-201.

Dila, G., Michel, C.J., Poch, O., Ripp, R., Thompson, J.D., 2019a. Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes. Biosystems 175, 57-74.

Dila, G., Mayer, C., Ripp, R., Poch, O., Michel, C.J., Thompson, J.D., 2019b. Circular code motifs in the ribosome: a missing link in the evolution of translation? RNA 25, 1714-1730.

Drummond, D.A., Wilke, C.O., 2009. The evolutionary consequences of erroneous protein synthesis. Nature Reviews Genetics 10, 715-724.

Fimmel, E., Michel, C.J., Strüngmann, L., 2016. *n*-Nucleotide circular codes in graph theory. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374, 20150058.

Fimmel, E., Strüngmann, L., 2018. Mathematical fundamentals for the noise immunity of the genetic code. Biosystems 164, 186-198.

Freeland, S.J., Hurst, L.D., 1998. The genetic code is one in a million. Journal of Molecular Evolution 47, 238-248.

Freeland, S.J., Knight, R.D., Landweber, L.F., Hurst, L.D., 2000. Early fixation of an optimal genetic code. Molecular Biology Evolution 17, 511–518.

Frey, G., Michel, C.J., 2003. Circular codes in archaeal genomes. Journal of Theoretical Biology 223, 413-431.

Frey, G., Michel, C.J., 2006. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Computational Biology and Chemistry 30, 87-101.

Garofalo, R., Wohlgemuth, I., Pearson, M., Lenz, C., Urlaub, H., Rodnina, M.V., 2019. Broad range of missense error frequencies in cellular proteins. Nucleic Acids Research 47, 2932-2945.

Geyer, R., Madany Mamlouk, A., 2018. On the efficiency of the genetic code after frameshift mutations. PeerJ 6, e4825.

Haig, D., Hurst, L.D., 1991. A quantitative measure of error minimization in the genetic code. Journal of Molecular Evolution 33, 412-417.

Itzkovitz, S., Alon, U., 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Research 17, 405-412.

Kawashima, S., Kanehisa, M., 2000. AAindex: amino acid index database. Nucleic Acids Research 28, 374.

Koonin, E.V., Novozhilov, A.S., 2009. Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61, 99-111.

Kumar, B., Saini, S., 2016. Analysis of the optimality of the standard genetic code. Molecular Biosystems 12, 2642-2651.

Michel, C.J., 2008. A 2006 review of circular codes in genes. Computer and Mathematics with Applications 55, 984-988.

Michel, C.J., 2012. Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Computational Biology and Chemistry 37, 24-37.

Michel, C.J., 2013. Circular code motifs in transfer RNAs. Computational Biology and Chemistry 45, 17-29.

Michel, C.J., 2015. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. Journal of Theoretical Biology 380, 156-177.

Michel, C.J., 2017. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. Life 7, 20, 1-16.

Michel, C.J., 2019. Single-frame, multiple-frame and framing motifs in genes. Life. 9, 18.

Michel, C.J., Nguefack Ngoune, V., Poch, O., Ripp, R., Thompson, J.D., 2017. Enrichment of circular code motifs in the genes of the yeast Saccharomyces cerevisiae. Life 7, 52, 1-20.

Michel, C.J., Pirillo, G., 2010. Identification of all trinucleotide circular codes. Computational Biology and Chemistry 34, 122-125.

Michel, C.J., Pirillo, G., 2013. A permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes. Journal of Theoretical Biology 319, 116-121.

Michel, C.J., Pirillo, G., Pirillo, M.A., 2008. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theoretical Computer Science 401, 17-26.

Michel, C.J., Thompson, J.D., 2020. Identification of a circular code periodicity in the bacterial ribosome: origin of codon periodicity in genes? RNA Biology in press.

Nirenberg, M.W., Matthaei, J.H., 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proceedings of the National Academy of Sciences of the United States of America 47, 1588-1602.

Pelc, S.R., Welton, M.G., 1966. Stereochemical relationship between coding triplets and amino-acids. Nature 209, 868-870.

Pirillo, G., 2003. A characterization for a set of trinucleotides to be a circular code. In: Benci, V., Cerrai, P., Freguglia, P., Israel, G., Pellegrini, C., (eds) Determinism, Holism, and Complexity. Springer, Boston, MA.

Seligmann, H., Pollock, D.D., 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA and Cell Biology 23, 701-705.

Seligmann, H., 2019. Localized context-dependent effects of the "ambush" hypothesis: More off-frame stop codons downstream of shifty codons. DNA and Cell Biology 38, 786-795.

Yarus, M., 2017. The genetic code and RNA-amino acid affinities. Life 7.

Wilke, C.O., 2015. Evolutionary paths of least resistance. Proceedings of the National Academy of Sciences of the United States of America 112, 12553-12554.

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

Woese, C.R., 1965. Order in the genetic code. Proceedings of the National Academy of Sciences of the United States of America 54, 71-75.

Wong, J.T., 1975. A co-evolution theory of the genetic code. Proceedings of the National Academy of Sciences of the United States of America 72, 1909-1912.

APPENDIX

Table 2. Eleven classical amino acid indices (AAindex database at http://www.genome.ad.jp/aaindex/).

Property $\mathbb{P}$	AAindex name	Reference
Charge $\mathbb{P}_C$	KLEP840101	Klein et al., 1984
Hydrophobicity $\mathbb{P}_H$	FASG890101	Fasman, 1989
Isoelectric point $\mathbb{P}_{IP}$	ZIMJ680104	Zimmerman et al., 1968
Melting point $\mathbb{P}_{MP}$	FASG760102	Fasman, 1976
Molecular weight $\mathbb{P}_{MW}$	FASG760101	Fasman, 1976
Optical rotation $\mathbb{P}_{OR}$	FASG760103	Fasman, 1976
Polarity $\mathbb{P}_{Pr}$	ZIMJ680103	Zimmerman et al., 1968
Polarizability $\mathbb{P}_{Pz}$	CHAM820101	Charton-Charton, 1982
Size $\mathbb{P}_{Si}$	DAWD720101	Dawson, 1972
Steric $\mathbb{P}_{St}$	CHAM810101	Charton, 1981
Volume $\mathbb{P}_V$	BIGC670101	Bigelow, 1967

Table 3. Amino acid property vectors for indices mentioned in Table 2 where $AA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ is the 20 amino acid alphabet.

Property $\mathbb{P}$	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Charge $\mathbb{P}_C$	0	0	-1	-1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
Hydrophobicity $\mathbb{P}_H$	-0.2	-6.0	1.4	2.3	-4.7	0.0	-1.2	-4.8	3.9	-4.7	-3.7	1.0	0.8	1.5	2.1	1.7	0.8	-3.5	-3.3	-1.0
Isoelectric point $\mathbb{P}_{IP}$	6.0	5.1	2.8	3.2	5.5	6.0	7.6	6.0	9.7	6.0	5.7	5.4	6.3	5.7	10.8	5.7	5.7	6.0	5.9	5.7
Melting point $\mathbb{P}_{MP}$	297	178	270	249	284	290	277	284	224	337	283	236	222	185	238	228	253	293	282	344
Molecular weight $\mathbb{P}_{MW}$	89.1	121.2	133.1	147.1	165.2	75.1	155.2	131.2	146.2	131.2	149.2	132.1	115.1	146.2	174.2	105.1	119.1	117.2	204.2	181.2
Optical rotation $\mathbb{P}_{OR}$	1.8	-16.5	5.1	12.0	-34.5	0.0	-38.5	12.4	14.6	-11.0	-10.0	-5.6	-86.2	6.3	12.5	-7.5	-28.0	5.6	-33.7	-10.0
Polarity $\mathbb{P}_{Pr}$	0.0	1.5	49.7	49.9	0.4	0.0	51.6	0.1	49.5	0.1	1.4	3.4	1.6	3.5	52.0	1.7	1.7	0.1	2.1	1.6
Polarizability $\mathbb{P}_{Pz}$	0.05	0.13	0.11	0.15	0.29	0.00	0.23	0.19	0.22	0.19	0.22	0.13	0.13	0.18	0.29	0.06	0.11	0.14	0.41	0.30
Size $\mathbb{P}_{Si}$	2.5	3.0	2.5	5.0	6.5	0.5	6.0	5.5	7.0	5.5	6.0	5.0	5.5	6.0	7.5	3.0	5.0	5.0	7.0	7.0
Steric $\mathbb{P}_{St}$	0.52	0.62	0.76	0.68	0.70	0.00	0.70	1.02	0.68	0.98	0.78	0.76	0.36	0.68	0.68	0.53	0.50	0.76	0.70	0.70
Volume $\mathbb{P}_V$	52.6	68.3	68.4	84.7	113.9	36.3	91.9	102.0	105.1	102.0	97.7	75.7	73.6	89.7	109.1	54.9	71.2	85.1	135.4	116.2

Table 4. Amino acid substitution matrix $\mathbf{M}(\mathbb{P}_V)$ for the volume property $\mathbb{P}_V$ where $AA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ is the 20 amino acid alphabet.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	15.7	15.8	32.1	61.3	16.3	39.3	49.4	52.5	49.4	45.1	23.1	21.0	37.1	56.5	2.3	18.6	32.5	82.8	63.6
C	15.7	0	0.1	16.4	45.6	32.0	23.6	33.7	36.8	33.7	29.4	7.4	5.3	21.4	40.8	13.4	2.9	16.8	67.1	47.9
D	15.8	0.1	0	16.3	45.5	32.1	23.5	33.6	36.7	33.6	29.3	7.3	5.2	21.3	40.7	13.5	2.8	16.7	67.0	47.8
E	32.1	16.4	16.3	0	29.2	48.4	7.2	17.3	20.4	17.3	13.0	9.0	11.1	5.0	24.4	29.8	13.5	0.4	50.7	31.5
F	61.3	45.6	45.5	29.2	0	77.6	22.0	11.9	8.8	11.9	16.2	38.2	40.3	24.2	4.8	59.0	42.7	28.8	21.5	2.3
G	16.3	32.0	32.1	48.4	77.6	0	55.6	65.7	68.8	65.7	61.4	39.4	37.3	53.4	72.8	18.6	34.9	48.8	99.1	79.9
H	39.3	23.6	23.5	7.2	22.0	55.6	0	10.1	13.2	10.1	5.8	16.2	18.3	2.2	17.2	37.0	20.7	6.8	43.5	24.3
I	49.4	33.7	33.6	17.3	11.9	65.7	10.1	0	3.1	0.0	4.3	26.3	28.4	12.3	7.1	47.1	30.8	16.9	33.4	14.2
K	52.5	36.8	36.7	20.4	8.8	68.8	13.2	3.1	0	3.1	7.4	29.4	31.5	15.4	4.0	50.2	33.9	20.0	30.3	11.1
L	49.4	33.7	33.6	17.3	11.9	65.7	10.1	0.0	3.1	0	4.3	26.3	28.4	12.3	7.1	47.1	30.8	16.9	33.4	14.2
M	45.1	29.4	29.3	13.0	16.2	61.4	5.8	4.3	7.4	4.3	0	22.0	24.1	8.0	11.4	42.8	26.5	12.6	37.7	18.5
N	23.1	7.4	7.3	9.0	38.2	39.4	16.2	26.3	29.4	26.3	22.0	0	2.1	14.0	33.4	20.8	4.5	9.4	59.7	40.5
P	21.0	5.3	5.2	11.1	40.3	37.3	18.3	28.4	31.5	28.4	24.1	2.1	0	16.1	35.5	18.7	2.4	11.5	61.8	42.6
Q	37.1	21.4	21.3	5.0	24.2	53.4	2.2	12.3	15.4	12.3	8.0	14.0	16.1	0	19.4	34.8	18.5	4.6	45.7	26.5
R	56.5	40.8	40.7	24.4	4.8	72.8	17.2	7.1	4.0	7.1	11.4	33.4	35.5	19.4	0	54.2	37.9	24.0	26.3	7.1
S	2.3	13.4	13.5	29.8	59.0	18.6	37.0	47.1	50.2	47.1	42.8	20.8	18.7	34.8	54.2	0	16.3	30.2	80.5	61.3
T	18.6	2.9	2.8	13.5	42.7	34.9	20.7	30.8	33.9	30.8	26.5	4.5	2.4	18.5	37.9	16.3	0	13.9	64.2	45.0
V	32.5	16.8	16.7	0.4	28.8	48.8	6.8	16.9	20.0	16.9	12.6	9.4	11.5	4.6	24.0	30.2	13.9	0	50.3	31.1
W	82.8	67.1	67.0	50.7	21.5	99.1	43.5	33.4	30.3	33.4	37.7	59.7	61.8	45.7	26.3	80.5	64.2	50.3	0	19.2
Y	63.6	47.9	47.8	31.5	2.3	79.9	24.3	14.2	11.1	14.2	18.5	40.5	42.6	26.5	7.1	61.3	45.0	31.1	19.2	0

Table 5. Normalized amino acid substitution matrix $\widehat{\mathbf{M}}(\mathbb{P}_V)$ for the volume property $\mathbb{P}_V$ where $AA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ is the 20 amino acid alphabet.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	1.5	1.5	3.0	5.7	1.5	3.6	4.6	4.9	4.6	4.2	2.1	1.9	3.4	5.2	0.2	1.7	3.0	7.7	5.9
C	1.5	0	0.0	1.5	4.2	3.0	2.2	3.1	3.4	3.1	2.7	0.7	0.5	2.0	3.8	1.2	0.3	1.6	6.2	4.4
D	1.5	0.0	0	1.5	4.2	3.0	2.2	3.1	3.4	3.1	2.7	0.7	0.5	2.0	3.8	1.3	0.3	1.5	6.2	4.4
E	3.0	1.5	1.5	0	2.7	4.5	0.7	1.6	1.9	1.6	1.2	0.8	1.0	0.5	2.3	2.8	1.3	0.0	4.7	2.9
F	5.7	4.2	4.2	2.7	0	7.2	2.0	1.1	0.8	1.1	1.5	3.5	3.7	2.2	0.4	5.5	4.0	2.7	2.0	0.2
G	1.5	3.0	3.0	4.5	7.2	0	5.2	6.1	6.4	6.1	5.7	3.7	3.5	4.9	6.7	1.7	3.2	4.5	9.2	7.4
H	3.6	2.2	2.2	0.7	2.0	5.2	0	0.9	1.2	0.9	0.5	1.5	1.7	0.2	1.6	3.4	1.9	0.6	4.0	2.3
I	4.6	3.1	3.1	1.6	1.1	6.1	0.9	0	0.3	0.0	0.4	2.4	2.6	1.1	0.7	4.4	2.9	1.6	3.1	1.3
K	4.9	3.4	3.4	1.9	0.8	6.4	1.2	0.3	0	0.3	0.7	2.7	2.9	1.4	0.4	4.7	3.1	1.9	2.8	1.0
L	4.6	3.1	3.1	1.6	1.1	6.1	0.9	0.0	0.3	0	0.4	2.4	2.6	1.1	0.7	4.4	2.9	1.6	3.1	1.3
M	4.2	2.7	2.7	1.2	1.5	5.7	0.5	0.4	0.7	0.4	0	2.0	2.2	0.7	1.1	4.0	2.5	1.2	3.5	1.7
N	2.1	0.7	0.7	0.8	3.5	3.7	1.5	2.4	2.7	2.4	2.0	0	0.2	1.3	3.1	1.9	0.4	0.9	5.5	3.8
P	1.9	0.5	0.5	1.0	3.7	3.5	1.7	2.6	2.9	2.6	2.2	0.2	0	1.5	3.3	1.7	0.2	1.1	5.7	3.9
Q	3.4	2.0	2.0	0.5	2.2	4.9	0.2	1.1	1.4	1.1	0.7	1.3	1.5	0	1.8	3.2	1.7	0.4	4.2	2.5
R	5.2	3.8	3.8	2.3	0.4	6.7	1.6	0.7	0.4	0.7	1.1	3.1	3.3	1.8	0	5.0	3.5	2.2	2.4	0.7
S	0.2	1.2	1.3	2.8	5.5	1.7	3.4	4.4	4.7	4.4	4.0	1.9	1.7	3.2	5.0	0	1.5	2.8	7.5	5.7
T	1.7	0.3	0.3	1.3	4.0	3.2	1.9	2.9	3.1	2.9	2.5	0.4	0.2	1.7	3.5	1.5	0	1.3	5.9	4.2
V	3.0	1.6	1.5	0.0	2.7	4.5	0.6	1.6	1.9	1.6	1.2	0.9	1.1	0.4	2.2	2.8	1.3	0	4.7	2.9
W	7.7	6.2	6.2	4.7	2.0	9.2	4.0	3.1	2.8	3.1	3.5	5.5	5.7	4.2	2.4	7.5	5.9	4.7	0	1.8
Y	5.9	4.4	4.4	2.9	0.2	7.4	2.3	1.3	1.0	1.3	1.7	3.8	3.9	2.5	0.7	5.7	4.2	2.9	1.8	0