BioSystems 197 (2020) 104201

Contents lists available at ScienceDirect

BioSystems

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

The maximality of circular codes in genes statistically verified

Christian J. Michel

Theoretical Bioinformatics, ICube, CNRS, University of Strasbourg, 300 Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Keywords:
Circular codes
Maximality
Trinucleotide per frame
Codon usage
Historical context

ABSTRACT

The maximality of circular codes in genes has 20 preferential trinucleotides in each frame. This combinatorial property is statistically verified in the genes of both bacteria and eukaryotes, and by two approaches computing the trinucleotide occurrence frequencies in the 3 frames at the gene population level (classical method) and at the gene level (recent method). Several remarks explain why the codon usage parameter is unable to identify the circular codes. Some historical and theoretical considerations on comma-free and circular codes are presented. An evolutionary process by trinucleotide permutation is proposed to describe the transformation of a circular code (and its motifs) into another circular code.

1. Introduction

A circular code X is a set of words such that any motif from X, called X motif, allows to retrieve, maintain and synchronize the original (construction) frame. The circular code X identified in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Michel, 2017, 2015; Arquès and Michel, 1996) contains the 20 following trinucleotides in reading frame (frame 0)

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

the 20 following trinucleotides in frame 1 (reading frame shifted by 1 nucleotide in the 5’ – 3’ direction, i.e. to the right)

$$X_1 = \{AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG, GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG\} \tag{2}$$

and the 20 following trinucleotides in frame 2 (reading frame shifted by 2 nucleotides in the 5’ – 3’ direction)

$$X_2 = \{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT, CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT\}. \tag{3}$$

The trinucleotide set X (defined in (1)) coding the reading frame in genes is a maximal (20 trinucleotides) C^3 self-complementary trinucleotide circular code (Arquès and Michel, 1996; reviewed in Michel, 2008; Fimmel and Strüngmann, 2018).

From a mathematical point of view, the identification of circular codes in genes led to about 200 theorems obtained in the different research fields of circular codes: the flower automaton (Arquès and Michel, 1996; and subsequent works), the probability approach (Koch and Lehmann, 1997; Lacan and Michel, 2001), the necklace 5*LDCN* (Letter Diletter Continued Necklace) (Pirillo, 2003), the necklace *n*LDCCN (Letter Diletter Continued Closed Necklace) with $n \in \{2, 3, 4, 5\}$ (Michel and Pirillo, 2010; and subsequent works), the group theory (Fimmel et al., 2014; and subsequent works) and the graph theory (Fimmel et al., 2016; and subsequent works).

From a biological point of view, the X circular code motifs (or briefly X motifs), i.e. motifs of the X circular code, allow to retrieve, maintain and synchronize the reading frame in genes. The concept, the statistical analyses and the biological studies of X circular code motifs have been introduced in Michel (2012).[1] It has been shown recently that the X motifs are enriched in the reading frame of extant genes (El Soufi and Michel, 2016; Michel et al., 2017; Dila et al., 2019a), as well as in tRNA sequences (Michel, 2012, 2013; El Soufi and Michel, 2015) and in functional regions of rRNA involved in mRNA translation (Michel, 2012; El Soufi and Michel, 2014, 2015; 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

E-mail address: c.michel@unistra.fr.

[1] This research field appeared very late after the identification of circular codes in 1996.

https://doi.org/10.1016/j.biosystems.2020.104201
Received 10 May 2020; Received in revised form 22 June 2020; Accepted 22 June 2020
Available online 2 July 2020

been suggested that the X circular code was an ancestor code of the standard genetic code that was used to code amino acids and simultaneously to identify and maintain the reading frame (Dila et al., 2019b).

In order to verify the maximality of circular codes in genes, we will reformulate in this work the presentation of the classical method (Arquès and Michel, 1996; Michel, 2015) and the recent method (Michel, 2017). Precisely, we will demonstrate here that the maximality of the 3 circular codes X (with 20 trinucleotides in reading frame, defined in (1)), X_1 (with 20 trinucleotides in frame 1, defined in (2)) and X_2 (with 20 trinucleotides in frame 2, defined in (3)), that have been assigned by inspection in Arquès and Michel (1996), is statistically verified.

2. Method

2.1. Gene kingdoms

Gene kingdoms $\mathbb{K}$ of bacteria $\mathbb{B}$ and eukaryotes $\mathbb{E}$ are obtained from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, January 2020). Computer tests exclude genes when: (i) their nucleotides do not belong to the alphabet $B = \{A, C, G, T\}$ where A stands for adenine, C stands for cytosine, G stands for guanine and T stands for thymine; (ii) they do not begin with a start trinucleotide ATG; (iii) they do not end with a stop trinucleotide $\{TAA, TAG, TGA\}$; and (iv) their lengths are not modulo 3. In order to obtain a broad but unduplicated sampling of each kingdom, we randomly selected 1 genome from each organism group. In bacteria, for example, there are several sequenced genomes of *Bacillus*: *amyloliquefaciens*, *anthracis*, *atrophaeus*, *cellulosilyticus*, *cereus*, etc., but only one was chosen randomly. Similarly in eukaryotes, for example, there are several sequenced genomes of *Drosophila*: *busckii*, *melanogaster*, *miranda*, *pseudoobscura*, *sechellia*, *simulans*, *yakuba*, etc., but only one was chosen randomly. Table 1 gives some basic information about these two studied gene kingdoms.

2.2. Classical method (reformulation of Arquès and Michel's method in 1996, and Michel's method in 2015)

2.2.1. Principle

The classical method is a computation at the gene population level. For all available genes in a given genome of a kingdom, the 192 trinucleotides frequencies are computed in the 3 frames and the preferential frame for the 64 trinucleotides are determined by assigning each trinucleotide in its preferential frame (frame associated with its highest trinucleotide frequency).

2.2.2. Formalism

In order to analyse the maximality of circular codes, the 4 periodic trinucleotides $P = \{AAA, CCC, GGG, TTT\}$ are not considered. Let $P_f(t, \mathscr{G})$ be the occurrence frequency of a trinucleotide $t \in \tilde{B}^3 = \{AAA, \ldots, TTT\} \backslash P$ in a frame $f \in \{0, 1, 2\}$ of all available genes (see Section 2.1) in a genome $\mathscr{G}$ of a kingdom $\mathbb{K}$ (Table 1). Thus, there are $3 \times 64 = 192$ trinucleotide occurrence frequencies $P_f(t, \mathscr{G})$ in the 3 frames f of genes in the genome $\mathscr{G}$ (see for example Arquès and Michel, 1996, Tables 1(a) and 1(b)). The trinucleotide assignment per frame, done by inspection in 1996, is now formulated in mathematical expressions. Then, the preferential frame $F(t, \mathscr{G}) \in \{0, 1, 2\}$ of a trinucleotide t in the genes of the genome $\mathscr{G}$ is the frame of maximal occurrence frequency $P_f(t, \mathscr{G})$ among the 3 frames f

$$F(t, \mathscr{G}) := \arg \max_{f \in \{0,1,2\}} \mathscr{F}(f, t, \mathscr{G}) \tag{4}$$

where the generic (see also method in Section 2.3.2) function $\mathscr{F}(f, t, \mathscr{G}) = P_f(t, \mathscr{G})$.

Table 1
Kingdoms $\mathbb{K}$ of genes extracted from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, January 2020) with their symbol and their numbers of genomes, genes and trinucleotides.

Kingdom	$\mathbb{K}$ (symbol)	Nb of genomes	Nb of genes	Nb of trinucleotides
Bacteria	$\mathbb{B}$	613	2,067,464	678,532,578
Eukaryotes	$\mathbb{E}$	146	3,556,369	2,049,298,998

Remark 1. Equation (4) allows the classification of each trinucleotide t in its preferential frame $F(t, \mathscr{G})$. This classification in 1996 led to the circular code X with 20 trinucleotides in reading frame (defined in (1)), the circular code X_1 with 20 trinucleotides in frame 1 (defined in (2)) and the circular code X_2 with 20 trinucleotides in frame 2 (defined in (3)).

The indicator function $\delta_f(F(t, \mathscr{G})) \in \{0, 1\}$ is equal to 1 if the preferential frame $F(t, \mathscr{G})$ of a trinucleotide t is equal to the frame f of genes in the genome $\mathscr{G}$, 0 otherwise

$$\delta_f(F(t, \mathscr{G})) = \begin{cases} 1 & \text{if } F(t, \mathscr{G}) = f \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $F(t, \mathscr{G})$ is defined in Equation (4).

Remark 2. At the genome level, a trinucleotide t has different occurrence frequencies $Pr_f(t, \mathscr{G})$ in the 3 frames f.

The number $N_f(\mathscr{G}) \in \mathbb{N}$ of preferential frames for the 60 trinucleotides $t \in \tilde{B}^3$ for each frame $f \in \{0, 1, 2\}$ in the genes of the genome $\mathscr{G}$ is simply obtained by summing the indicator function for the 60 trinucleotides t

$$N_f(\mathscr{G}) = \sum_{t \in \tilde{B}^3} \delta_f(F(t, \mathscr{G})) \tag{6}$$

where $\delta_f(F(t, \mathscr{G}))$ is defined in Equation (5).

Remark 3. $\sum_{f \in \{0,1,2\}} N_f(\mathscr{G}) = \left|\tilde{B}^3\right| = 60.$

2.3. Recent method (reformulation of Michel's method in 2017)

2.3.1. Principle

The recent method is a computation at the gene level. For a given available gene in a given genome of a kingdom, the 192 trinucleotides frequencies are computed in the 3 frames and the preferential frame of the 64 trinucleotides are determined by assigning the number 1 for the trinucleotide in its preferential frame (frame associated with its highest trinucleotide frequency) and the number 0 for the 2 other frames. These numbers 0 and 1 are summed for all available genes in the genome. Then, the preferential frame for the 64 trinucleotides are determined by assigning each trinucleotide in its preferential frame (frame associated with its highest number).

It should be stressed that in the recent and improved method, all the genes, i.e. of large and small lengths, in a genome are assigned with the same statistical weight.

2.3.2. Formalism

Let $Pr_f(t, g)$ be the occurrence frequency of a trinucleotide $t \in \tilde{B}^3 = \{AAA, \ldots, TTT\} \backslash P$ in a frame $f \in \{0, 1, 2\}$ of an available gene g (see Section 2.1) in a genome $\mathscr{G}$ of a kingdom $\mathbb{K}$ (Table 1). Thus, there are $3 \times 64 = 192$ trinucleotide occurrence frequencies $Pr_f(t, g)$ in the 3 frames f of the gene g. Then, the preferential frame $F(t, g) \in \{0, 1, 2\}$ of the trinucleotide t in the gene g is the frame of maximal occurrence frequency $Pr_f(t, g)$ among the 3 frames f

$$F(t, g) = \arg \max_{f \in \{0,1,2\}} Pr_f(t, g). \tag{7}$$

The indicator function $\delta_f(F(t, g)) \in \{0, 1\}$ is equal to 1 if the preferential frame $F(t, g)$ of the trinucleotide t is equal to the frame f of the gene g, 0 otherwise

$$\delta_f(F(t,g)) = \begin{cases} 1 & \text{if } F(t,g) = f \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $F(t,g)$ is defined in Equation (7).

Remark 4. As mentioned in Michel (2017), at the gene level and particularly for the genes g of short lengths, a trinucleotide t may have an identical occurrence frequency $Pr_f(t,g)$ in 2 or 3 frames f. In this case, a set of 2 or 3 preferential frames $F(t,g)$ are assigned to the trinucleotide t (2 or 3 numbers 1). Besides, if a trinucleotide t is absent in a gene g, mainly for the genes g of very short lengths, then no preferential frame is attributed to t (3 numbers 0).

The number $M_f(t, \mathscr{G}) \in \mathbb{N}$ of preferential frames of the trinucleotide t for each frame f in the genome $\mathscr{G}$ is simply obtained by summing the indicator function for all the genes g belonging to the genome $\mathscr{G}$

$$M_f(t, \mathscr{G}) = \sum_{g \in \mathscr{G}} \delta_f(F(t,g)) \quad (9)$$

where $\delta_f(F(t,g))$ is defined in Equation (8).

Then, we apply Equation (4) with the generic function $\mathscr{F}(f,t,\mathscr{G}) = M_f(t, \mathscr{G})$ (Equation (9)) and then Equations (5) and (6).

In Section 3, the distribution of the numbers $N_f(\mathscr{G})$ (Equation (6)) for the 3 frames f are studied for the 613 bacterial genomes (Section 3.1) and the 146 eukaryotic genomes (Section 3.2) with the classical method (Equation (6) with the trinucleotide occurrence frequencies $P_f(t, \mathscr{G})$ at the gene population level) and the recent method (Equation (6) with the trinucleotide occurrence frequencies $Pr_f(t,g)$ at the gene level).

3. Results

3.1. Bacterial genes

With the classical method (trinucleotide occurrence frequencies at the gene population level), the mean numbers of trinucleotides in the frames 0 (reading frame), 1 and 2 of genes in the 613 bacterial genomes are 20.28, 19.83 and 19.90, respectively (Table 2). These important statistical results are the first demonstration of the maximality of circular codes with 20 preferential trinucleotides in average for each frame. The standard deviation is about 2 trinucleotides for each frame with a range between 16 and 27 trinucleotides for the frame 0, a range between 12 and 25 trinucleotides for the frame 1, and a range between 15 and 27 trinucleotides for the frame 2 (see Table 2 and the frequency histogram in Fig. 1). Table 3 reports the 12 particular bacterial genomes among 643 with minimal or maximal trinucleotide numbers in the three frames.

Interestingly, these results are confirmed with the recent method (trinucleotide occurrence frequencies at the gene level) which is a sharper approach by considering the genes of large and small lengths with the same statistical weight. Indeed, the mean numbers of trinucleotides in the frames 0, 1 and 2 of genes in the 613 bacterial genomes are 20.07, 19.85 and 20.08, respectively (Table 4). The standard deviation is also about 2 trinucleotides for each frame with a range between 15 and 27 trinucleotides for the frame 0, a range between 12 and 25 trinucleotides for the frame 1, and a range between 15 and 26 trinucleotides for the frame 2 (see Table 4 and the frequency histogram in Fig. 2). Table 5 reports the 9 particular bacterial genomes among 643 with minimal or maximal trinucleotide numbers in the three frames, some of them being already cited in Table 3.

Table 2
Basic statistics of the numbers of trinucleotides (minimum, maximum, mean, standard deviation) in the frames 0 (reading frame), 1 and 2 of the genes in the 613 bacterial genomes (Table 1) with the classical method (Section 2.2; computation of the trinucleotide occurrence frequencies at the gene population level).

	Numbers of trinucleotides		
	Frame 0	Frame 1	Frame 2
Minimum	16	12	15
Maximum	27	25	27
Mean	20.28	19.83	19.90
Standard deviation	1.82	2.19	1.88

3.2. Eukaryotic genes

We adopt a similar presentation. With the classical method, the mean numbers of trinucleotides in the frames 0 (reading frame), 1 and 2 of genes in the 146 eukaryotic genomes are 19.83, 20.46 and 19.71, respectively (Table 6). These statistical results with the eukaryotic genes confirm the maximality of circular codes that is observed with the bacterial genes. The standard deviation is about 2 trinucleotides for each frame with a range between 16 and 24 trinucleotides for the frame 0, a range between 16 and 26 trinucleotides for the frame 1, and a range between 15 and 24 trinucleotides for the frame 2 (see Table 6 and the frequency histogram in Fig. 3). Table 7 reports the 9 particular eukaryotic genomes among 146 with minimal or maximal trinucleotide numbers in the three frames.

These results are also retrieved with the recent method. Indeed, the mean numbers of trinucleotides in the frames 0, 1 and 2 of genes in the 146 eukaryotic genomes are 19.65, 20.54 and 19.81, respectively (Table 8). The standard deviation is also about 2 trinucleotides for each frame with a range between 16 and 25 trinucleotides for the frame 0, a range between 17 and 24 trinucleotides for the frame 1, and a range between 15 and 24 trinucleotides for the frame 2 (see Table 8 and the frequency histogram in Fig. 4). Table 9 reports the 26 particular eukaryotic genomes among 146 with minimal or maximal trinucleotide numbers in the three frames, some of them being already cited in Table 7.

4. The codon usage parameter unable to identify the circular codes

The concept and the method by which the X circular code in genes was identified, and in particular the determination of its maximality property (20 trinucleotides), is a question very often asked by the reader. In particular, the reader is astonished not to find the circular codes with the classical parameter analysing the codon usage (CU). This CU parameter is based on the frequencies of trinucleotides in reading frame. I would like to make a few responses, which will also help to explain the historical context that identified the X circular code in the genes.

From a mathematical point of view, the codon usage parameter was never involved, directly or indirectly, in any of the 200 theorems obtained in the different research fields of circular codes (described in Introduction).

From a statistical point of view, the codon usage parameter does not have the property to identify maximal circular codes for the following reasons, which are briefly recalled here.

(i) The CU parameter is a (classical) parameter of a code, e.g. the genetic code. The main objective of this CU parameter, used massively by biologists, is to relate the codon frequencies to the amino acid frequencies and to identify their over- and under-representations according to the genome, the function of genes,

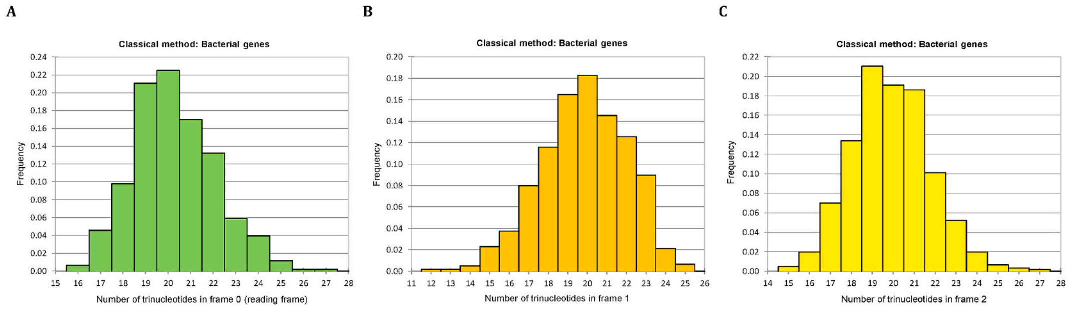

Fig. 1. Frequency histogram of trinucleotides in the three frames of the genes in the 613 bacterial genomes (Table 1) with the classical method (Section 2.2; computation of the trinucleotide occurrence frequencies at the gene population level): **A**: Frame 0 (reading frame). **B**: Frame 1. **C**: Frame 2.

Table 3
List of the 12 bacterial genomes having a minimum or a maximum number (in bold; from Table 2) of trinucleotides in the frames 0, 1 and 2 of the genes among the 613 bacterial genomes (Table 1) with the classical method.

Genome	Numbers of trinucleotides		
	Frame 0	Frame 1	Frame 2
Desulfurobacterium thermolithotrophum	**16**	21	23
Eggerthella lenta	**16**	**25**	19
Gordonibacter pamelaeae	**16**	**25**	19
Ilyobacter polytropus	**16**	20	24
Macrococcus caseolyticus	**27**	15	18
Mesotoga prima	25	**12**	23
Fibrella aestuarina	19	**25**	16
Rhodothermus marinus	17	**25**	18
Chelativorans	23	22	**15**
Desulfovibrio africanus	23	22	**15**
Sulfurospirillum barnesii	23	22	**15**
Exiguobacterium antarcticum	18	15	**27**

Table 4
Basic statistics of the numbers of trinucleotides (minimum, maximum, mean, standard deviation) in the frames 0 (reading frame), 1 and 2 of the genes in the 613 bacterial genomes (Table 1) with the recent method (Section 2.3; computation of the trinucleotide occurrence frequencies at the gene level).

	Numbers of trinucleotides		
	Frame 0	Frame 1	Frame 2
Minimum	15	12	15
Maximum	27	25	26
Mean	20.07	19.85	20.08
Standard deviation	1.86	2.11	1.86

Table 5
List of the 9 bacterial genomes having a minimum or a maximum number (in bold; from Table 4) of trinucleotides in the frames 0, 1 and 2 of the genes among the 613 bacterial genomes (Table 1) with the recent method.

Genome	Numbers of trinucleotides		
	Frame 0	Frame 1	Frame 2
Ilyobacter polytropus	**15**	21	24
Macrococcus caseolyticus	**27**	15	18
Mesotoga prima	25	**12**	23
Rhodothermus marinus	17	**25**	18
Desulfovibrio africanus	23	22	**15**
Sulfurospirillum barnesii	23	22	**15**
Exiguobacterium antarcticum	18	16	**26**
Thermoclostridium stercorarium	17	17	**26**
Thermotoga maritima	21	13	**26**

Table 6
Basic statistics of the number of trinucleotides (minimum, maximum, mean, standard deviation) in the frames 0 (reading frame), 1 and 2 of the genes in the 146 eukaryotic genomes (Table 1) with the classical method (Section 2.2; computation of the trinucleotide occurrence frequencies at the gene population level).

	Numbers of trinucleotides		
	Frame 0	Frame 1	Frame 2
Minimum	16	16	15
Maximum	24	26	24
Mean	19.83	20.46	19.71
Standard deviation	1.74	1.87	1.56

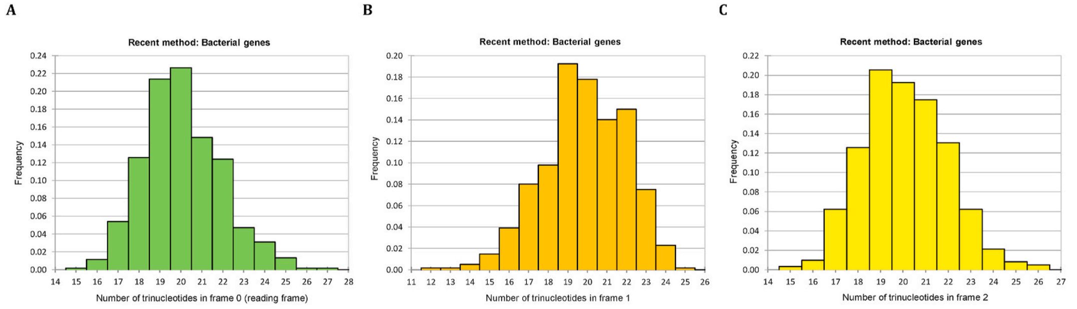

Fig. 2. Frequency histogram of trinucleotides in the three frames of the genes in the 613 bacterial genomes (Table 1) with the recent method (Section 2.3; computation of the trinucleotide occurrence frequencies at the gene level): **A**: Frame 0 (reading frame). **B**: Frame 1. **C**: Frame 2.

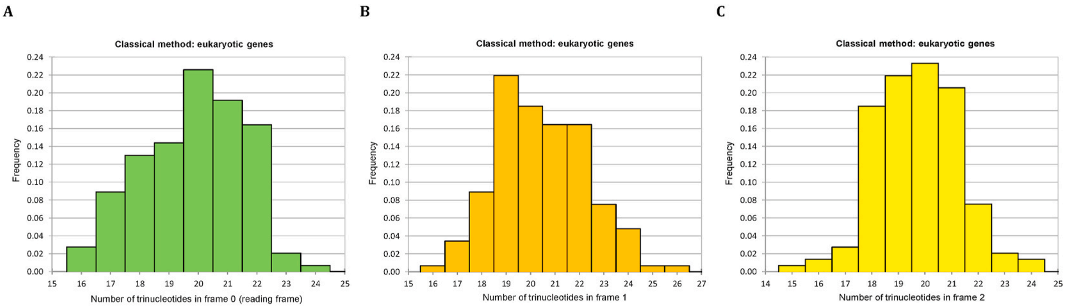

Fig. 3. Frequency histogram of trinucleotides in the three frames of the genes in the 146 eukaryotic genomes (Table 1) with the classical method (Section 2.2; computation of the trinucleotide occurrence frequencies at the gene population level): **A**: Frame 0 (reading frame). **B**: Frame 1. **C**: Frame 2.

Table 7
List of the 9 eukaryotic genomes having a minimum or a maximum number (in bold; from Table 6) of trinucleotides in the frames 0, 1 and 2 of the genes among the 146 eukaryotic genomes (Table 1) with the classical method.

Genome	Numbers of trinucleotides		
	Frame 0	Frame 1	Frame 2
Larimichthys crocea	**16**	23	21
Nothobranchius furzeri	**16**	23	21
Salmo salar	**16**	24	20
Xiphophorus couchianus	**16**	22	22
Callithrix jacchus	**24**	17	19
Candida glabrata	20	**16**	**24**
Carassius auratus	18	**26**	16
Bathycoccus prasinos	23	22	**15**
Yarrowia lipolytica	19	17	**24**

Table 8
Basic statistics of the number of trinucleotides (minimum, maximum, mean, standard deviation) in the frames 0 (reading frame), 1 and 2 of the genes in the 146 eukaryotic genomes (Table 1) with the recent method (Section 2.3; computation of the trinucleotide occurrence frequencies at the gene level).

	Numbers of trinucleotides		
	Frame 0	Frame 1	Frame 2
Minimum	16	17	15
Maximum	25	24	24
Mean	19.65	20.54	19.81
Standard deviation	1.93	1.77	1.69

the chemical property of the amino acid, etc. The CU parameter is not a parameter specific of a code which is circular. In particular, a set of trinucleotides with a high (or the highest) CU frequency does not imply circularity. In other words, the property of a high trinucleotide frequency in reading frame does not imply reading frame retrieval. Thus, this property does not lead to a code that is circular.

(ii) A circular code has the property to retrieve the reading frame among the three frames in genes. However, the CU parameter does not consider the two shifted frames.

(iii) A C^3 circular code, such as X defined in (1), has two permuted codes which are circular. The CU parameter does not consider additional codes related to the frames 1 and 2.

(iv) The maximality of a circular code with 3-letter words (e.g. the trinucleotides) on a 4-letter alphabet (e.g. the genetic alphabet) is 20 trinucleotides. The CU parameter does not reveal such a number. And even if one had thought of such a number, the 20 most frequent trinucleotides in the reading frame contain permuted trinucleotides (trinucleotides such as AAC and ACA are permuted) and periodic trinucleotides ($\{AAA, CCC, GGG, TTT\}$) which are excluded in a circular code.[2] Furthermore, another problem appears immediately. Assuming that a circular code, maximal or not, can be determined with the most frequent trinucleotides using the CU parameter. Then, such a circular code will obviously be assigned to the reading frame. But what about the next trinucleotides? How can they be assigned to the frame 1 or 2? And what criteria to use for the number of trinucleotides to be assigned to the frame 1 or 2? The CU parameter is once again unable to determine such properties.

(v) The X circular code identified in genes (Arquès and Michel, 1996) as well as the variant circular codes (Frey and Michel, 2006a,b) were never identified by the CU parameter. The four main methods developed since 1996: trinucleotide occurrence per frame (Arquès and Michel, 1996), correlation function per frame (Arquès and Michel, 1997) and improved statistical functions at the gene population level (Michel, 2015) and at the gene level (Michel, 2017); always analyse the three frames simultaneously. The reader can verify that Equation (4) in the classical method (Section 2.2) and Equations (7) and (4) (Section 2.3) consider the three frames simultaneously.

(vi) In the research work in 1996, the circularity property of the code X was not immediately discovered but some time after the identification of the three sets of 20 trinucleotides in the 3 frames of genes and their basic properties: complementarity, permutation, same number (15) of nucleotides A, C, G and T, prefixes and suffixes of trinucleotides, etc.

As a side note, if this very simple codon usage parameter had been specific to the circular codes, the circular codes would have been identified well before 1996, as the concept of comma-free codes has been introduced in 1957 from a biological point of view (Crick et al., 1957) and in 1958 from a mathematical point of view (Golomb et al., 1958). Astonishingly for the authors in 1996, no statistical studies prior to 1996, even basic frequency analyses, were developed in order to identify properties in the two shifted frames of genes. The codon usage parameter analysing the reading frame was the central dogma at that time and still today.

In conclusion, for the many reasons given above, the codon usage parameter is not able to identify the circular codes, which does not imply

[2] The four periodic trinucleotides also have a property identified by the method analysing the trinucleotide occurrence frequencies in the 3 frames, but which cannot be revealed by the CU parameter. Indeed, in Tables 1(a) and 1(b) in Arquès and Michel (1996), the trinucleotide AAA and its complementary trinucleotide TTT occur preferentially in frame 0 (similarly to the X circular code which is self-complementary), the trinucleotide CCC occurs preferentially in frame 1 and its complementary trinucleotide GGG occurs preferentially in frame 2 (similarly to the circular codes X_1 and X_2 which are complementary) (Table 2(a) in Arquès and Michel, 1996).

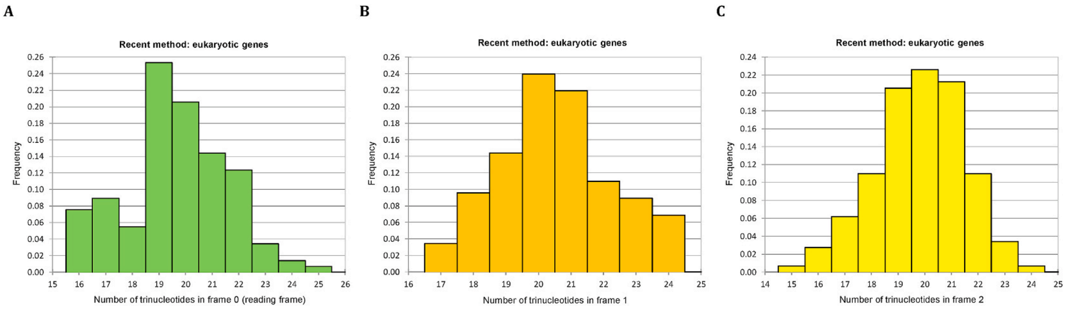

Fig. 4. Frequency histogram of trinucleotides in the three frames of the genes in the 146 eukaryotic genomes (Table 1) with the recent method (Section 2.3; computation of the trinucleotide occurrence frequencies at the gene level): **A**: Frame 0 (reading frame). **B**: Frame 1. **C**: Frame 2.

that this CU parameter cannot bring some information to the codes that are circular (remember that the circular codes are particular classes of codes). The codon usage parameter can be considered as a restricted case to the reading frame of the method analysing the trinucleotide occurrence frequencies in the 3 frames.

5. Some historical and theoretical considerations on comma-free and circular codes

In 1957, the comma-free codes have been proposed to the amino acid coding, i.e. a model of the genetic code before its experimental discovery (Crick et al., 1957). The idea was to find a code of 20 trinucleotides (codons) in the reading frame of genes coding for 20 amino acids such that no trinucleotides of the code exist in one of the two shifted frames, i.e. such that the trinucleotides of the code appear only in the reading frame – the comma-free property. The four nucleotides $\{A, C, G, T\}$ as well as the 16 dinucleotides $\{AA, \ldots, TT\}$ are simple codes which are not appropriate for coding 20 amino acids. The 64 trinucleotides $\{AAA, \ldots, TTT\}$ induce a redundancy in their coding. The determination of a code of 20 trinucleotides forming a comma-free code has several necessary conditions:

(i) A periodic trinucleotide $\{AAA, CCC, GGG, TTT\}$ must be excluded from such a code. Indeed, the concatenation of *AAA* with itself, for instance, does not allow the reading frame to be retrieved as there are three possible decompositions: … *AAA*,*AAA*,*AAA*, … (original reading frame), … *A*,*AAA*,*AAA*,*AA* … and … *AA*,*AAA*, *AAA*,*A* …, the commas showing the adopted decomposition.

(ii) Two non-periodic permuted trinucleotides, e.g. *ACG* and *CGA*, must also be excluded from such a code. Indeed, the concatenation of *ACG* with itself, for instance, does not allow the reading frame to be retrieved as there are two possible decompositions: … *ACG*,*ACG*,*ACG*, … (original reading frame) and … *A*,*CGA*,*CGA*, *CG* … Therefore, by excluding the four periodic trinucleotides and by gathering the 60 remaining trinucleotides in 20 classes of 3 trinucleotides such that, in each class, the 3 trinucleotides are deduced from each other by a circular permutation, e.g. *ACG*, *CGA* and *GAC*, we see that a comma-free code can contain only one trinucleotide from each class and thus has at most 20 trinucleotides. This trinucleotide number is identical to the amino acid number, thus leading to a code assigning one trinucleotide per amino acid without ambiguity. However, a comma-free code coding for 20 amino acids does not exist (see below).

Table 9
List of the 26 eukaryotic genomes having a minimum or a maximum number (in bold; from Table 8) of trinucleotides in the frames 0, 1 and 2 of the genes among the 146 eukaryotic genomes (Table 1) with the recent method.

Genome	Numbers of trinucleotides		
	Frame 0	Frame 1	Frame 2
Brachypodium distachyon	**16**	22	22
Carassius auratus	**16**	23	21
Cynoglossus semilaevis	**16**	21	23
Larimichthys crocea	**16**	22	22
Ogataea parapolymorpha	**16**	21	23
Panicum hallii	**16**	22	22
Salmo salar	**16**	22	22
Schizosaccharomyces pombe	**16**	**24**	20
Setaria italica	**16**	22	22
Sorghum bicolor	**16**	23	21
Xiphophorus couchianus	**16**	22	22
Monodelphis domestica	**25**	**17**	18
Candida glabrata	21	**17**	22
Cryptosporidium parvum	22	**17**	21
Kluyveromyces lactis	22	**17**	21
Yarrowia lipolytica	19	**17**	**24**
Babesia_bigemina	19	**24**	17
Caenorhabditis elegans	20	**24**	16
Eremothecium cymbalariae	19	**24**	17
Musa acuminata	20	**24**	16
Ornithorhynchus anatinus	19	**24**	17
Papaver somniferum	18	**24**	18
Takifugu rubripes	17	**24**	19
Thermothelomyces thermophilus	18	**24**	18
Thielavia terrestris	18	**24**	18
Elaeis guineensis	24	21	**15**

In 1996, a statistical study of trinucleotides in the 3 frames of genes has been developed to identify universal trinucleotide properties associated with the 3 frames. The approach carried out was initially neither an analysis of the amino acid coding nor a study of the synchronization properties of the reading frame. However, an important condition in this statistical approach was the search for trinucleotide properties "independent" of the taxonomy and function of the genes. Such a principle has already been investigated a few years ago with the identification of universal trinucleotide properties in the reading frames of genes on the purine-pyrimidine alphabet, leading to a model of codon evolution at three successive steps (Michel, 1989). By inspection, three codes X, X_1 and X_2 of 20 trinucleotides are identified for each frame. They have the main properties of complementarity, permutation, maximality and circularity: X coding the reading frame in genes is a maximal C^3 self-complementary trinucleotide circular code.

The relation between the comma-free codes and the circular codes was an open problem because each class of codes had a different definition. In 2008, we identified a mathematical relation between these two classes of codes by constructing a hierarchy of comma-free and circular codes (Michel et al., 2008b). Thereafter, this relation can be defined very simply in graph theory (Fimmel et al., 2016; and subsequent works). Thus, the comma-free codes belong to the class of circular codes and represent a more constrained class of circular codes.

The maximality of comma-free codes is 20 trinucleotides and there are 408 maximal comma-free codes (Golomb et al., 1958) coding for a maximum of 13 amino acids (Table 4 in Michel, 2014). Furthermore, the maximality of comma-free codes which are self-complementary, or C^3, or C^3 self-complementary, is only 16 trinucleotides (Table 3a,b, 4a,b and 5 in Michel et al., 2008a). There are 4 maximal self-complementary comma-free codes (Table 3a,b in Michel et al., 2008a) coding for a maximum of 11 amino acids. Thus, the model of a genetic code based on comma-free codes could not satisfy the coding condition of 20 amino acids.

The maximality of circular codes, that include the comma-free codes, is 20 trinucleotides and there are 12,964,440 maximal circular codes (Arquès and Michel, 1996) coding for a maximum of 18 amino acids (see Introduction in Michel and Pirillo, 2013, Table 4 in Michel, 2014). The maximality of C^3 self-complementary circular codes that include the X circular code observed in genes, is 20 trinucleotides and there are 216 (Arquès and Michel, 1996) of them coding for a maximum of 14 amino acids (Table 4 in Michel, 2014). The X circular code codes for 12 amino acids.

Self-complementary is an important property for the codes as it is associated with the DNA double helix (see Fig. 4 in Arquès and Michel, 1996). As the maximality of self-complementary comma-free codes is only 16 trinucleotides, a model of a genetic code based on self-complementary comma-free codes must not consider 4 trinucleotides per frame, thus a total of 12 trinucleotides.

6. Conclusion

We presented several scientific arguments that the codon usage parameter has limitations for the identification of circular codes. We also demonstrated the maximality of circular codes with 20 preferential trinucleotides in average for each frame, both in the bacterial and eukaryotic genes, and both with the classical method based on the trinucleotide occurrence frequencies at the gene population level and the recent method founded on the trinucleotide occurrence frequencies at the gene level.

As expected, the statistical results obtained here show that some genomes differ from the mean number 20 of trinucleotides per frame. The choice, voluntary or involuntary, of such particular genomes would prevent the identification of a maximal circular code. It could be that the average circular code X is hidden by several specific genes in these genomes.

The transformation of the average circular code X (and its X motifs) into another circular code Y (and its Y motifs) in a particular genome is an interesting open problem. From my point of view, an evolutionary process by trinucleotide permutation seems simpler than a process by mutation. Without loss of generality, if the trinucleotide *ATC*, for example, of the average maximal circular code X is replaced by the trinucleotide *TCA* of a variant maximal circular code, then the transformation of *ATC* into *TCA* by permutation can be achieved in “one event”, while the transformation by mutation requires 3 non-random modifications in the 3 trinucleotides sites: *A* into *T*, *T* into *C*, and *C* into *A*. However, evolution by permutation has never been mentioned to my knowledge, even though the ribosome could have such an evolutionary function with the ribosomal frameshifting.

In 1996, the data available in the gene databases came from various and incomplete genomes representing an average sample that led to the identification of an average maximal circular code X in genes. To date, there is no experimental evidence of a biological function of the X circular code and its X motifs in the modern genes. The molecular mechanism is also unknown.[3]

Acknowledgments

This work is dedicated to Marie Denise Besch for her constant support.

References

Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182, 45–58.

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

Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Natl. Acad. Sci. U. S. A. 43, 416–421.

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.

El Soufi, K., Michel, C.J., 2014. Circular code motifs in the ribosome decoding center. Comput. Biol. Chem. 52, 9–17.

El Soufi, K., Michel, C.J., 2015. Circular code motifs near the ribosome decoding center. Comput. Biol. Chem. 59, 158–176.

El Soufi, K., Michel, C.J., 2016. Circular code motifs in genomes of eukaryotes. J. Theor. Biol. 408, 198–212.

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

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

Fimmel, E., Michel, C.J., Strüngmann, L., 2016. *n*-Nucleotide circular codes in graph theory. Phil. Trans. Math. Phys. Eng. Sci. 374, 20150058.

Frey, G., Michel, C.J., 2006a. An analytical model of gene evolution with 6 mutation parameters: an application to archaeal circular codes. Comput. Biol. Chem. 30, 1–11.

Frey, G., Michel, C.J., 2006b. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Comput. Biol. Chem. 30, 87–101.

Golomb, S.W., Gordon, B., Welch, L.R., 1958. Comma-free codes. Can. J. Math. 10, 202–209.

Koch, A.J., Lehmann, J., 1997. About a symmetry of the genetic code. J. Theor. Biol. 189, 171–174.

Lacan, J., Michel, C.J., 2001. Analysis of a circular code model. J. Theor. Biol. 213, 159–170.

Michel, C.J., 1989. A study of the purine/pyrimidine codon occurrence with a reduced centered variable and an evaluation compared to the frequency statistic. Math. Biosci. 97, 161–177.

Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.

Michel, C.J., 2012. Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Comput. Biol. Chem. 37, 24–37.

Michel, C.J., 2013. Circular code motifs in transfer RNAs. Comput. Biol. Chem. 45, 17–29.

Michel, C.J., 2014. A genetic scale of reading frame coding. J. Theor. Biol. 355, 83–94.

Michel, C.J., 2015. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. J. Theor. Biol. 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., Pirillo, G., 2010. Identification of all trinucleotide circular codes. Comput. Biol. Chem. 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. J. Theor. Biol. 319, 116–121.

Michel, C.J., Thompson, J.D., 2020. Identification of a circular code periodicity in the bacterial ribosome: origin of codon periodicity in genes? RNA Biol. 17, 571–583.

Michel, C.J., Pirillo, G., Pirillo, M.A., 2008a. Varieties of comma free codes. Comput. Math. Appl. 55, 989–996.

Michel, C.J., Pirillo, G., Pirillo, M.A., 2008b. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theor. Comput. Sci. 401, 17–26.

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.

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.

[3] The lack of design of any biological experiments since 1996 explains why this question remains open. A futuristic experiment that could confirm or disprove the theory of circular code in genes, would consist of attaching a nanocamera with a nanophone and nanoseismograph to the ribosome to analyse its dynamics when it encounters the X motifs. Does the ribosome speed up (more noise and vibrations) or slow down when it encounters the X motifs which retrieve and synchronize the reading frame in genes?.