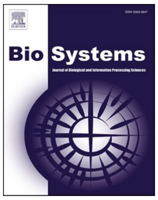
BioSystems 206 (2021) 104431

Contents lists available at ScienceDirect

BioSystems

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



Genes on the circular code alphabet

Christian J. Michel

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

ARTICLE INFO

Keywords:
Circular code
Alphabet
Trinucleotide
Pairing rule
Protein coding gene
Transfer RNA gene

ABSTRACT

The X motifs, motifs from the circular code X, are enriched in the (protein coding) genes of bacteria, archaea, eukaryotes, plasmids and viruses, moreover, in the minimal gene set belonging to the three domains of life, as well as in tRNA and rRNA sequences. They allow to retrieve, maintain and synchronize the reading frame in genes, and contribute to the regulation of gene expression. These results lead here to a theoretical study of genes based on the circular code alphabet. A new occurrence relation of the circular code X under the hypothesis of an equiprobable (balanced) strand pairing is given. Surprisingly, a statistical analysis of a large set of bacterial genes retrieves this relation on the circular code alphabet, but not on the DNA alphabet. Furthermore, the circular code X has the strongest balanced circular code pairing among 216 maximal C^3 self-complementary trinucleotide circular codes, a new property of this circular code X. As an application of this theory, different tRNAs studied on the circular code alphabet reveal an unexpected stem structure. Thus, the circular code X would have constructed a coding stem in tRNAs as an outline of the future gene structure and the future DNA double helix.

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, 2015, 2017; Arquès and Michel, 1996) contains the 20 following trinucleotides in reading frame (frame 0)

$$X = X_0 = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}, \quad (1)$$

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

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

and the 20 following trinucleotides in frame 2 (reading frame shifted by 2 nucleotides in the $5' - 3'$ direction)

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

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). More formal definitions of the mathematical properties (theorems, etc.) of the X circular code are available in a number of reviews (Michel, 2008; Fimmel and Strüngmann, 2018) and recent works (Fimmel et al., 2019, 2020). They are not necessary to understand the methods and results obtained in this work.

The concept, the statistical analyses and the biological studies of X motifs (motifs constructed with the circular code X defined in (1)) have been introduced in Michel (2012). It has been shown recently that the X motifs are enriched in the 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 has been 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). The X motifs are significantly enriched in the minimal gene set belonging to the three domains of life, and in codon-optimized genes (Thompson et al., 2021). The X codons also regulate systematic deletions of nucleotides during mitochondrial transcription (Seligmann, 2015, 2017; El Houmami and Seligmann 2017; Warthi and Seligmann, 2019; Seligmann and Warthi, 2020). Theoretical minimal RNA rings, candidate 22-nucleotide-long ancestral protogenes rationally designed for non-redundant coding, are also X enriched, as are even shorter nucleotide pentamers avoiding redundant coding (Michel, 2019; Demongeot and Seligmann, 2019a, 2020a).

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

https://doi.org/10.1016/j.biosystems.2021.104431
Received 9 April 2021; Received in revised form 15 April 2021; Accepted 15 April 2021
Available online 22 April 2021

Table 1
Average codon frequency (%) $\overline{F}$ (codon usage; Equation (5)) in 1171 bacterial genomes with a total number of 4,148,022 genes of 1,313,192,812 codons.

AAA	3.01	CAA	1.58	GAA	3.56	TAA	0.00
AAC	1.74	CAC	1.01	GAC	2.58	TAC	1.32
AAG	2.08	CAG	2.08	GAG	2.71	TAG	0.00
AAT	2.05	CAT	1.04	GAT	2.90	TAT	1.70
ACA	1.06	CCA	0.77	GCA	1.74	TCA	0.81
ACC	2.08	CCC	1.13	GCC	3.49	TCC	1.01
ACG	1.31	CCG	1.70	GCG	2.72	TCG	1.04
ACT	0.95	CCT	0.85	GCT	1.66	TCT	0.87
AGA	0.65	CGA	0.42	GGA	1.35	TGA	0.01
AGC	1.32	CGC	2.08	GGC	3.14	TGC	0.54
AGG	0.39	CGG	1.07	GGG	1.23	TGG	1.21
AGT	0.85	CGT	1.01	GGT	1.78	TGT	0.39
ATA	1.11	CTA	0.58	GTA	1.14	TTA	1.65
ATC	2.66	CTC	1.74	GTC	1.99	TTC	1.88
ATG	2.32	CTG	3.39	GTG	2.47	TTG	1.46
ATT	2.48	CTT	1.42	GTT	1.61	TTT	2.11

Table 2
Average nucleotide frequency (%) in the 1,313,192,812 codons and their 3 sites (deduced from Table 1) of 4,148,022 genes in 1171 bacterial genomes.

	Site 1	Site 2	Site 3	Codon
A	26.05	29.35	19.43	24.94
C	21.86	23.18	29.70	24.91
G	36.07	17.44	27.18	26.90
T	16.02	30.03	23.69	23.25

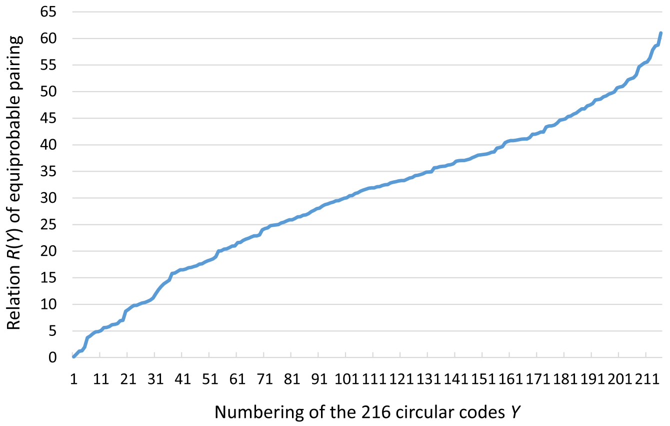


Fig. 1. Parameter $R(Y)$ (Equation (10) in %) of circular code pairing with the 216 maximal C^3 self-complementary trinucleotide circular codes Y. The circular code X observed in genes (defined in (1)) with $R(X) = 0.12\%$ (numbering 1 in the figure) has the strongest balanced circular code pairing among the 216 circular codes.

Furthermore, the density of X motifs generally correlates with experimental measures of translation efficiency and mRNA stability (Thompson et al., 2021). Thus, the X motifs may represent a genetic signal contributing to the maintenance of the correct reading frame and the optimization and regulation of gene expression. Furthermore, motifs of unitary trinucleotide circular codes (sets of one trinucleotide), i.e. repeated trinucleotides, are also observed in the non-coding genomes of eukaryotes (El Soufi and Michel, 2017).

All the results mentioned above suggest a genetic information unit of genomes that is based on trinucleotides (i.e. words of 3 letters on the 4-letter genetic alphabet) of the circular code X. Such a property is fully verified for bacterial and viral genomes. Indeed, genomes of bacteria and viruses, as well as organelles (mitochondria, chloroplasts, plasmids) all possess a compact architecture where genes coding proteins and RNAs (tRNAs, rRNAs, etc.) represent about 95% of a genome (Bobay and Ochman, 2017). Even in the case of the eukaryotic genomes where genes only constitute 10 ± 5% of a genome, repeated trinucleotides (unitary trinucleotide circular codes) are also observed in the non-coding regions (El Soufi and Michel, 2017). In this paper, I propose a theoretical study of genes based on the circular code alphabet.

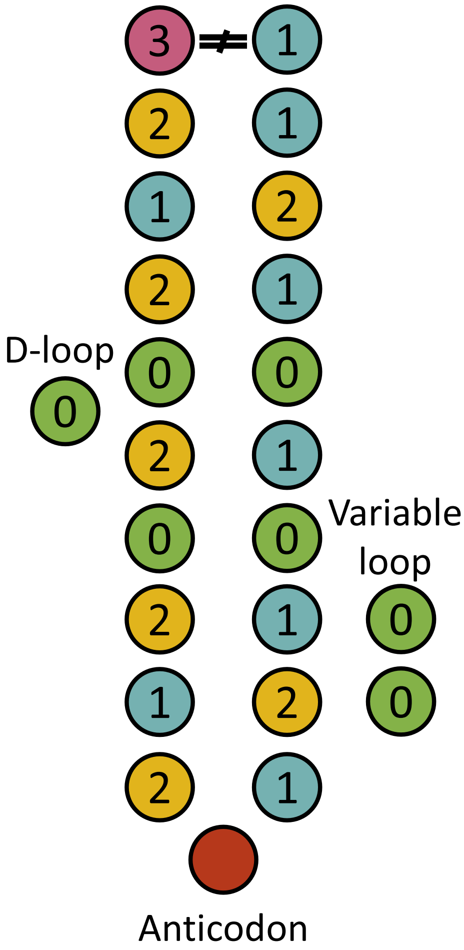


Fig. 2. Representation of the transfer RNA of alanine with anticodon *GGC* (tRNA-Ala-GGC) of *Escherichia coli* (from Table 3) on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$ is defined in (1) and noted "0" with a green color, X_1 is defined in (2) and noted "1" with a blue color, X_2 is defined in (3) with an orange color and noted "2", $Z = \{AAA, CCC, GGG, TTT\}$ and noted "3" with a purple color, and the anticodon *GGC* is in red. By definition, X_0 pairs with itself, and X_1 pairs with X_2, and reciprocally. The symbol "≠" means a mismatch. The tRNA-Ala-GGC on the circular code alphabet is represented by a stem with an anticodon loop made of a single trinucleotide which is the anticodon, a D-loop with one trinucleotide and a variable loop with two trinucleotides.

Section 2 recalls the DNA, RNA and RY genetic alphabets, the nucleotide complementary map on these three alphabets, the n-nucleotide circular permutation, and gives some relations of nucleotide occurrence under the realistic hypothesis of an equiprobable (balanced) strand pairing. The average codon usage of a circular code is formalized. After recalling the properties of the complementarity and circular permutation maps, Section 3 gives a new relation of the occurrence of the circular code X under the hypothesis of an equiprobable strand pairing. A statistical analysis of a large set of bacterial genes will show that the equiprobable strand pairing is observed on the circular code alphabet, but not on the DNA alphabet. Furthermore, the circular code X has the strongest balanced circular code pairing among 216 maximal C^3 self-

Table 3
Transfer RNA of alanine with anticodon *GGC* (tRNA-Ala-GGC) of *Escherichia coli* (Escherichia coli str K-12 substr MG1655 tRNA-Ala-GGC; 76 bp; chr:2518041-2518116 (−); tRNAviz; Lin et al., 2019). The upper part of the table is the tRNA-Ala-GGC upstream the anticodon *GGC* at the nucleotide position 34 (in bold). The lower part of the table is the tRNA-Ala-GGC downstream the anticodon *GGC*. Each part has three rows. The upper row gives the position of trinucleotides (1st trinucleotide site). The middle row gives the tRNA-Ala-GGC on the standard alphabet B_{DNA}. The lower row represents the tRNA-Ala-GGC on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$, X_1 and X_2 are defined in (1), (2) and (3), respectively, and $Z = \{AAA, CCC, GGG, TTT\}$.

Pos	1	4	7	10	13	16	19	22	25	28	31	**34**
B_{DNA}	*GGG*	*GCT*	*ATA*	*GCT*	*CAG*	*CTG*	*GGA*	*GAG*	*CGC*	*TTG*	*CAT*	*GGC*
B_X	Z	X_2	X_1	X_2	X_0	X_0	X_2	X_0	X_2	X_1	X_2	X_0
Pos	37	40	43	46	49	52	55	58	61	64	67	70	73
B_{DNA}	*ATG*	*CAA*	*GAG*	*GTC*	*AGC*	*GGT*	*TCG*	*ATC*	*CCG*	*CTT*	*AGC*	*TCC*	*ACC*
B_X	X_1	X_2	X_0	X_0	X_1	X_0	X_1	X_0	X_1	X_2	X_1	X_1	X_0

Table 4
Transfer RNA of alanine with anticodon *TGC* (tRNA-Ala-TGC) of *Escherichia coli* (Escherichia coli str K-12 substr MG1655 tRNA-Ala-TGC; 76 bp; chr:225500-225575 (+); tRNAviz; Lin et al., 2019). The upper part of the table is the tRNA-Ala-TGC upstream the anticodon *TGC* at the nucleotide position 34 (in bold). The lower part of the table is the tRNA-Ala-TGC downstream the anticodon *TGC*. Each part has three rows. The upper row gives the position of trinucleotides (1st trinucleotide site). The middle row gives the tRNA-Ala-TGC on the standard alphabet B_{DNA}. The lower row represents the tRNA-Ala-TGC on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$, X_1 and X_2 are defined in (1), (2) and (3), respectively, and $Z = \{AAA, CCC, GGG, TTT\}$. The trinucleotide positions of tRNA-Ala-TGC in bold are trinucleotides which differ from tRNA-Ala-GGC (see Table 3).

Pos	1	4	7	10	13	16	19	22	25	**28**	**31**	**34**
B_{DNA}	*GGG*	*GCT*	*ATA*	*GCT*	*CAG*	*CTG*	*GGA*	*GAG*	*CGC*	*CTG*	*CTT*	*TGC*
B_X	Z	X_2	X_1	X_2	X_0	X_0	X_2	X_0	X_2	X_0	X_2	X_1
Pos	**37**	**40**	43	46	**49**	52	55	58	61	**64**	67	70	73
B_{DNA}	*ACG*	*CAG*	*GAG*	*GTC*	*TGC*	*GGT*	*TCG*	*ATC*	*CCG*	*CAT*	*AGC*	*TCC*	*ACC*
B_X	X_1	X_0	X_0	X_0	X_1	X_0	X_1	X_0	X_1	X_2	X_1	X_1	X_0

complementary trinucleotide circular codes. Finally, as an application of this theory, the circular code alphabet is applied to different transfer RNAs (tRNAs) in which a stem structure is unexpectedly identified. Comments on the biological results obtained are given in Discussion.

2. Method

2.1. Classical definitions

2.1.1. Genetic alphabets

A few classical notations of the DNA, RNA and RY nucleotide alphabets B are recalled.

Notation 1. The nucleotide 4-letter alphabet is denoted by $B_{DNA} = \{A, C, G, T\}$ where A stands for adenine, C stands for cytosine, G stands for guanine and T stands for thymine.

Notation 2. The ribonucleotide 4-letter alphabet is denoted by $B_{RNA} = \{A, C, G, U\}$ where A stands for adenine, C stands for cytosine, G stands for guanine and U stands for uracil.

Notation 3. The purine-pyrimidine 2-letter alphabet is denoted by $B_{RY} = \{R, Y\}$ where the purine nucleotide $R = \{A, G\}$ and the pyrimidine nucleotide $Y = \{C, T, U\}$.

It is important to remember that the genetic information on a 2-letter alphabet may be very different from that on a 4-letter alphabet. For example on B_{DNA}, the sequence $s = (AG)^+ = AGAGAG\ldots$ leads to a modulo 2 periodicity while on B_{RY}, this sequence $s = R^+ = RRR\ldots$ leads to an uniform signal. Genetic information, i.e. genes, introns, etc., on these different alphabets was intensively studied in the 1980s (not detailed).

2.1.2. Nucleotide complementarity

The nucleotide complementarity maps $\mathscr{C}_B$ on these nucleotide alphabets B are recalled. The DNA complementarity map $\mathscr{C}_{DNA}$ is the canonical (standard) one.

Definition 1. The DNA complementarity map $\mathscr{C}_{DNA} : B_{DNA} \to B_{DNA}$ is defined by $\mathscr{C}_{DNA}(A) = T$, $\mathscr{C}_{DNA}(C) = G$, $\mathscr{C}_{DNA}(G) = C$ and $\mathscr{C}_{DNA}(T) = A$.

There are variants of this canonical complementarity map $\mathscr{C}_{DNA}$. They extend the canonical complementarity map $\mathscr{C}_{DNA}$ by adding nucleotide pairings.

Definition 2. The RNA complementarity map $\mathscr{C}_{RNA} : B_{RNA} \to B_{RNA}$ is defined by $\mathscr{C}_{RNA}(A) = U$, $\mathscr{C}_{RNA}(C) = G$, $\mathscr{C}_{RNA}(G) = \{C, U\}$ and $\mathscr{C}_{RNA}(U) = \{A, G\}$.

The map $\mathscr{C}_{RNA}$ differs, in particular, from $\mathscr{C}_{DNA}$ with the additional wobble pairing (G, U) observed in the 2D and 3D structures of extant RNAs.

Definition 3. The RY complementarity map $\mathscr{C}_{RY} : B_{RY} \to B_{RY}$ is defined by $\mathscr{C}_{RY}(R) = Y$ and $\mathscr{C}_{RY}(Y) = R$.

The map $\mathscr{C}_{RY}$ on B_{RY} involves a pairing between a purine (chemical compound with a double ring) and a pyrimidine (chemical compound with a single ring). In addition to the pairing (G, T) ((G, U)), the pairing

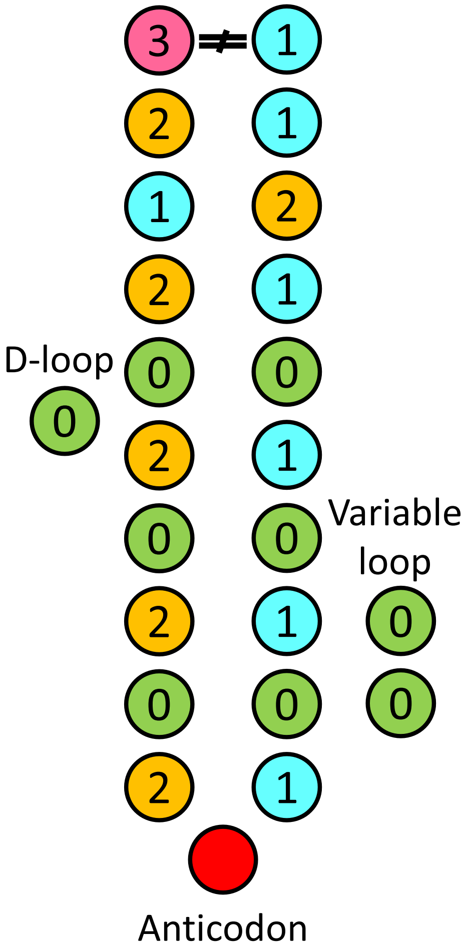


Fig. 3. Representation of the transfer RNA of alanine with anticodon *TGC* (tRNA-Ala-TGC) of *Escherichia coli* (from Table 4) on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$ is defined in (1) and noted "0" with a green color, X_1 is defined in (2) and noted "1" with a blue color, X_2 is defined in (3) with an orange color and noted "2", $Z = \{AAA, CCC, GGG, TTT\}$ and noted "3" with a purple color, and the anticodon *TGC* is in red. By definition, X_0 pairs with itself, and X_1 pairs with X_2, and reciprocally. The symbol "$\neq$" means a mismatch. The tRNA-Ala-TGC on the circular code alphabet is represented by a stem with an anticodon loop made of a single trinucleotide which is the anticodon, a D-loop with one trinucleotide and a variable loop with two trinucleotides.

Table 5
Transfer RNA of arginine with anticodon *ACG* (tRNA-Arg-ACG) of *Escherichia coli* (Escherichia coli str K-12 substr MG1655 tRNA-Arg-ACG; 77 bp; chr:2817784-2817860 (−); tRNAviz; Lin et al., 2019). The upper part of the table is the tRNA-Arg-ACG upstream the anticodon *ACG* at the nucleotide position 35 (in bold). The lower part of the table is the tRNA-Arg-ACG downstream the anticodon *ACG*. Each part has three rows. The upper row gives the position of trinucleotides (1st trinucleotide site). The middle row gives the tRNA-Ala-ACG on the standard alphabet B_{DNA}. The lower row represents the tRNA-Ala-ACG on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$, X_1 and X_2 are defined in (1), (2) and (3), respectively, and $Z = \{AAA, CCC, GGG, TTT\}$.

Pos	1	4	7	10	13	16	19	20	23	26	29	32	**35**
B_{DNA}	*GCA*	*TCC*	*GTA*	*GCT*	*CAG*	*CTG*	*G*	*ATA*	*GAG*	*TAC*	*TCG*	*GCT*	*ACG*
B_X	X_2	X_1	X_0	X_2	X_0	X_0		X_1	X_0	X_0	X_1	X_2	X_1

Pos	38	41	44	47	50	51	54	57	60	63	66	69	72	75
B_{DNA}	*AAC*	*CGA*	*GCG*	*GTC*	*G*	*GAG*	*GTT*	*CGA*	*ATC*	*CTC*	*CCG*	*GAT*	*GCA*	*CCA*
B_X	X_0	X_2	X_1	X_0		X_0	X_0	X_2	X_0	X_0	X_1	X_0	X_2	X_1

(A, C) is admitted as possible, for example in primitive conditions of life.

2.1.3. *n*-nucleotide circular permutation

Definition 4. The n-nucleotide circular permutation map $\mathscr{P}_B : B^n \to B^n$ is defined by $\mathscr{P}_B(l_1l_2...l_n) = l_2...l_nl_1$ for all $l_i \in B$, $i \in \{1,...,n\}$. The $n-1$ iterates of $\mathscr{P}_B$ are defined similarly.

Observation 1. The permutation map $\mathscr{P}_B$ is (obviously) the identity on the nucleotide alphabets B ($n = 1$), i.e. $\mathscr{P}_B(l) = l$ for all $l \in B$. In other words, $\mathscr{P}_B$ does not operate on the nucleotide alphabets.

2.1.4. Relation of nucleotide pairing

Let $N(l)$ be the occurrence number of a nucleotide l in a strand s on an alphabet B. A strand pairing p between a strand s and its complementary strand $\mathscr{C}(s)$ is denoted by the two pairs $p = \{(l, \mathscr{C}(l)), (\mathscr{C}(l), l)\}$ where l, $\mathscr{C}(l) \in B$ and $\mathscr{C}$ is the nucleotide complementarity map. Under the realistic hypothesis of an equiprobable (balanced) strand pairing, some relations can be easily obtained in the strand s.

Property 1. (*DNA pairing rule*). *On the alphabet B_{DNA}, the 2 equiprobable nucleotide pairings $p_1 = \{(A,T),(T,A)\}$ and $p_2 = \{(C,G),(G,C)\}$ leads to the following relation in the strand s*:

$$N(A) + N(T) = N(C) + N(G). \tag{4}$$

Remark 1. *When the occurrence number $N(l)$ of nucleotides is computed both in the 2 strands s and $\mathscr{C}(s)$, the 2 sums of* Equation (4) *are (obviously) multiplied by 2.*

Remark 2. *If in addition the pairings (A,T), (T,A), (C,G) and (G,C) are equiprobable then $N(A) = N(T) = N(C) = N(G)$ in the strand s.*

Remark 3. Equation (4) *does not imply the Chargaff's rule stating that $N(A) = N(T)$ and $N(C) = N(G)$ in the strand s. This Chargaff's rule is obviously verified by considering the 2 strands s and $\mathscr{C}(s)$ (called first parity rule). In the following, only* Equation (4) *of DNA pairing will be considered.*

Property 2. (*RNA pairing rule*). *On the alphabet B_{RNA}, the 3 equiprobable nucleotide pairings $p_1 = \{(A,U),(U,A)\}$, $p_2 = \{(C,G),(G,C)\}$ and $p_3 = \{(G,U),(U,G)\}$ leads to the following relation in the strand s*:

$$N(A) + N(U) = N(C) + N(G) = N(G) + N(U)$$

leading to

$$N(A) = N(G) \text{ and } N(C) = N(U).$$

2.2. Codon usage matrix

The codon set over B_{DNA} is denoted by $B^3 = \{AAA, ..., TTT\}$ of cardinality $|B^3| = 64$. Let $\mathscr{G}$ be a genome in a kingdom of $|\mathscr{G}| = 1171$ bacterial genomes (see Section 2.3). A codon usage matrix $\mathbf{M} = [m_{ij}]_{1 \le i \le |\mathscr{G}|, 1 \le j \le 64}$ of size $|\mathscr{G}| \times |B^3|$ where the 1171 rows are associated with the bacterial genomes $\mathscr{G}$ and the 64 columns are associated with the codons B^3, is defined such that $\mathbf{M}$ has element m_{ij} in row i and column j referring to the frequency F_{ij} (usage) of codon j in genome $\mathscr{G}_i$ from all the available genes in $\mathscr{G}_i$ (see Section 2.3). Then, for each codon j, the average codon usage $\overline{F}$ in a genome kingdom is computed simply by:

$$\overline{F}_j = \frac{1}{|\mathscr{G}|} \sum_{i=1}^{|\mathscr{G}|} F_{ij}. \tag{5}$$

Equation (5) computes the average codon usage with the same weight for each bacterial genome, i.e. whatever the number of genes and the number of codons in each bacterial genome.

Then, the average codon usage $\overline{F}(X_0)$, $\overline{F}(X_1)$ and $\overline{F}(X_2)$ of the circular codes X_0, X_1 and X_2 (defined in (1), (2) and (3), respectively) are computed as follows:

$$\overline{F}(X_i) = \sum_{j \in X_i} \overline{F}_j \tag{6}$$

where $\overline{F}_j$ is obtained by Equation (5) and $i \in \{0, 1, 2\}$.

2.3. Bacterial gene kingdom

A kingdom of $|\mathscr{G}| = 1171$ bacterial genomes $\mathscr{G}$ is obtained from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, January 2021). In each genome $\mathscr{G}$, the available genes are extracted. Computer tests exclude genes when: (i) their nucleotides do not belong to the alphabet $B = \{A, C, G, T\}$; (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, 1 genome $\mathscr{G}$ is randomly selected from each organism group. In bacteria, there are several sequenced genomes in an organism group, for example *Bacteroides fragilis*: 638R, 9343 and YCH46, *Brucella_melitensis*: ATCC 23457, bv. 1 str. 16 M, M28, M5-90 and NI, etc., but only one sequenced genome was chosen randomly in each organism group. Thus, the 1171 selected bacterial genomes $\mathscr{G}$ have a total number of 4,148,022 genes of 1,313,192,812 codons.

3. Results

3.1. Circular code alphabet

I define the circular code alphabet B_X.

Definition 5. The circular code alphabet is defined by $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$, X_1 and X_2 are defined in (1), (2) and (3),

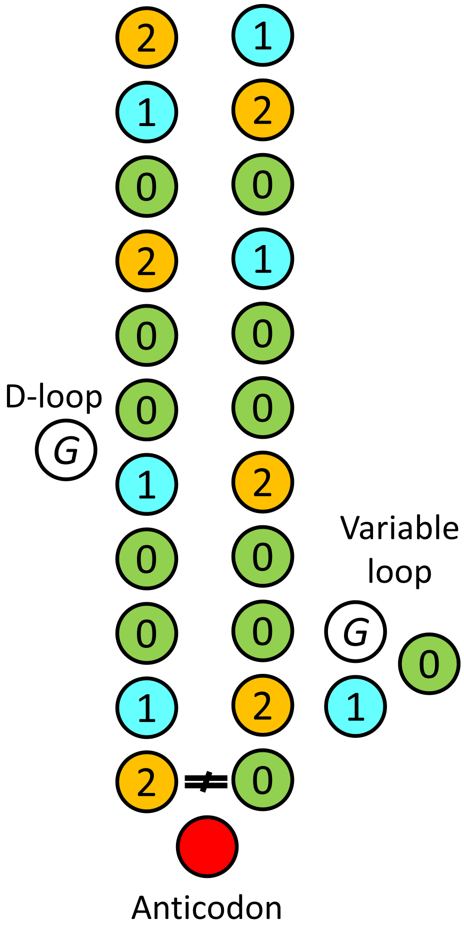


Fig. 4. Representation of the transfer RNA of arginine with anticodon *ACG* (tRNA-Arg-ACG) of *Escherichia coli* (from Table 5) on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$ is defined in (1) and noted "0" with a green color, X_1 is defined in (2) and noted "1" with a blue color, X_2 is defined in (3) with an orange color and noted "2", $Z = \{AAA, CCC, GGG, TTT\}$ and noted "3" with a purple color, and the anticodon *ACG* is in red. By definition, X_0 pairs with itself, and X_1 pairs with X_2, and reciprocally. The symbol "$\neq$" means a mismatch. The tRNA-Arg-ACG on the circular code alphabet is represented by a stem with an anticodon loop made of three trinucleotides, a D-loop with a single nucleotide *G* and a variable loop with two trinucleotides and one nucleotide *G*.

Table 6

Transfer RNA of arginine with anticodon *CCG* (tRNA-Arg-CCG) of *Escherichia coli* (Escherichia coli str K-12 substr MG1655 tRNA-Arg-CCG; 77 bp; chr:3982375-3982451 (+); tRNAviz; Lin et al., 2019). The upper part of the table is the tRNA-Ala-CCG upstream the anticodon *CCG* at the nucleotide position 35 (in bold). The lower part of the table is the tRNA-Ala-CCG downstream the anticodon *CCG*. Each part has three rows. The upper row gives the position of trinucleotides (1st trinucleotide site). The middle row gives the tRNA-Ala-CCG on the standard alphabet B_{DNA}. The lower row represents the tRNA-Ala-CCG on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$, X_1 and X_2 are defined in (1), (2) and (3), respectively, and $Z = \{AAA, CCC, GGG, TTT\}$. The nucleotide positions of tRNA-Arg-CCG in bold are trinucleotides which differ from tRNA-Arg-ACG.

Pos	**1**	**4**	7	10	13	16	19	20	23	**26**	**29**	**32**	**35**
B_{DNA}	*GCG*	*CCC*	*GTA*	*GCT*	*CAG*	*CTG*	*G*	*ATA*	*GAG*	*CGC*	*TGC*	*CCT*	*CCG*
B_X	X_1	Z	X_0	X_2	X_0	X_0		X_1	X_0	X_2	X_1	X_2	X_1

Pos	**38**	**41**	**44**	47	**50**	**53**	54	55	58	**61**	**64**	**67**	**70**	73
B_{DNA}	*GAG*	*GCA*	*GAG*	*GTC*	*TCA*	*GGT*	*T*	*CGA*	*ATC*	*CTG*	*TCG*	*GGC*	*GCG*	*CCA*
B_X	X_0	X_2	X_0	X_0	X_1	X_0		X_2	X_0	X_0	X_1	X_0	X_1	X_1

respectively, and $Z = B^3 \backslash \{X_0, X_1, X_2\} = \{AAA, CCC, GGG, TTT\}$.

Interesting properties are observed in B_X, Property 5 being new.

Property 3. *The circular code complementarity map* $\mathscr{C}_X : B_X \rightarrow B_X$ *has the following relations* (Arquès and Michel, 1996):

$$\mathscr{C}_X(X_0) = X_0,\ \mathscr{C}_X(X_1) = X_2,\ \mathscr{C}_X(X_2) = X_1,\ \mathscr{C}_X(Z) = Z.$$

Property 4. *The circular code permutation map* $\mathscr{P}_X : B_X \rightarrow B_X$ *has the following relations* (Arquès and Michel, 1996):

$$\mathscr{P}_X(X_0) = X_1, \mathscr{P}_X(X_1) = X_2, \mathscr{P}_X(X_2) = X_0, P_X(Z) = Z.$$

Remark 4. *The complementarity and permutation maps* $\mathscr{C}_X$ *and* $\mathscr{P}_X$ *on* B_X *both have the double functions, identity and not-identity.*

Remark 5. *In biology, the property of complementarity is classic. It is important to re-emphasize here that on* B_X, *the permutation property is just as important as the complementarity property since the two maps* $\mathscr{C}_X$ *and* $\mathscr{P}_X$ *are obviously related by the relations* $\mathscr{P}_X(X) = \mathscr{C}_X(\mathscr{P}_X^2(X))$ *and* $\mathscr{P}_X^2(X) = \mathscr{C}_X(\mathscr{P}_X(X))$ (Michel, 2015).

Property 5. (*circular code pairing rule*). *On the circular code alphabet* B_X, *the* 2 *equiprobable strand pairings* $p_1 = \{(X_0, X_0)\}$ *and* $p_2 = \{(X_1, X_2), (X_2, X_1)\}$ *leads to the following relation in the strand* s:

$$N(X_0) = N(X_1) + N(X_2). \tag{7}$$

This interesting and new relation, not mentioned until now in the absence of any hypotheses to research it, is statistically investigated in Section 3.2 *in order to have a foundation of a circular code alphabet in genes.*

Remark 6. Definition 5 *and* Property 5 *will be applied to the class of the* 216 *maximal* C^3 *self-complementary trinucleotide circular codes.*

3.2. The relation of circular code pairing with the X circular code

From Table 1 and Equation (6), the average codon frequencies $\overline{F}(X_0)$, $\overline{F}(X_1)$ and $\overline{F}(X_2)$ of the circular codes X_0, X_1 and X_2 computed in the bacterial genomes are:

$$\overline{F}(X_0) = 46.31\%, \overline{F}(X_1) = 25.59\%, \overline{F}(X_2) = 20.60\%$$

The well-known asymmetry $\overline{F}(X_1) > \overline{F}(X_2)$ between the circular codes X_1 and X_2 is retrieved in genes (reading frame) of prokaryotes (Bahi and Michel, 2008, Section 3.1.2), eukaryotes (Arquès et al., 1997 both Fig. 2 and Section 2.2; Bahi and Michel, 2004, Section 1.2.2; Michel et al., 2017), plasmids (Michel, 2015) and viruses (Michel, 2015). This asymmetry has no biological explanation so far.

Very interestingly, the relation of circular code pairing (Equation (7)) is verified. Indeed, as $\overline{F}(X_1) + \overline{F}(X_2) = 46.19\%$ then

$$\overline{F}(X_0) \approx \overline{F}(X_1) + \overline{F}(X_2). \tag{8}$$

In order to evaluate the statistical significance of a similar distribution of the frequency of X_0 and the frequency sum of X_1 and X_2 in the 1171 bacterial genomes, in each genome $\mathscr{G}_i$ the frequency F_{iX_0} of X_0 and the frequency sum $F_{iX_1} + F_{iX_2}$ of X_1 and X_2 is computed as follows: $F_{iX_0} = \sum_{j \in X_0} F_{ij}$ and $F_{iX_1} + F_{iX_2} = \sum_{j \in X_1} F_{ij} + \sum_{j \in X_2} F_{ij}$, where F_{ij} is the frequency (usage) of codon j in genome $\mathscr{G}_i$ (see Section 2.2). A two-tailed Wilcoxon signed-rank test in this paired sample (Wilcoxon, 1945; Woolson, 1987, page 172) has a p-value equal to 0.81. Thus, the null hypothesis H_0: "both samples follow the same distribution law" cannot be rejected.

Deduced from Table 1, the average nucleotide frequencies in the 1,313,192,812 codons and their 3 sites (4,148,022 genes in 1171 bacterial genomes) are given in Table 2. In order to evaluate the relation of nucleotide pairing (Equation (4)) on B_{DNA}, the computed frequency sums $\overline{F}(A) + \overline{F}(T) = 48.19\%$ and $\overline{F}(C) + \overline{F}(G) = 51.81\%$ (Table 2) lead to

$$\overline{F}(A) + \overline{F}(T) \neq \overline{F}(C) + \overline{F}(G). \tag{9}$$

In order to evaluate the statistical significance of a different distribution of the frequency sum of A and T and the frequency sum of C and G in the 1171 bacterial genomes, the frequency sums $F_{iA} + F_{iT}$ of A and T, and $F_{iC} + F_{iG}$ of C and G are computed in each genome $\mathscr{G}_i$. Using a same statistical approach as above, a two-tailed Wilcoxon signed-rank test in this paired sample leads to a p-value around 10^{-6}. Thus, the 2 distributions "$N(A) + N(T)$" and "$N(C) + N(G)$" are very different.

In conclusion, this statistical analysis showed that the equiprobable (balanced) strand pairing is observed on the circular code alphabet, but not on the DNA alphabet.

3.3. The relation of circular code pairing with the 216 maximal C^3 self-complementary trinucleotide circular codes

For each maximal C^3 self-complementary trinucleotide circular code Y among 216, the parameter $R(Y)$ (%) of circular code pairing is computed with the following parameter:

$$R(Y) = \left| \overline{F}(Y_0) - \left(\overline{F}(Y_1) + \overline{F}(Y_2) \right) \right|. \tag{10}$$

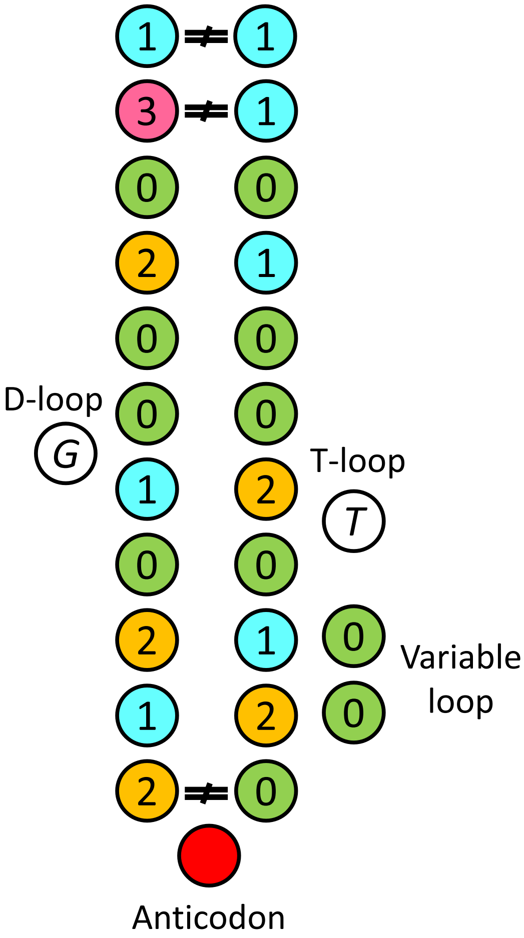


Fig. 5. Representation of the transfer RNA of arginine with anticodon *CCG* (tRNA-Arg-CCG) of *Escherichia coli* (from Table 6) on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$ is defined in (1) and noted "0" with a green color, X_1 is defined in (2) and noted "1" with a blue color, X_2 is defined in (3) with an orange color and noted "2", $Z = \{AAA, CCC, GGG, TTT\}$ and noted "3" with a purple color, and the anticodon *CCG* is in red. By definition, X_0 pairs with itself, and X_1 pairs with X_2, and reciprocally. The symbol "≠" means a mismatch. The tRNA-Arg-CCG on the circular code alphabet is represented by a stem with an anticodon loop made of three trinucleotides, a D-loop with a single nucleotide *G*, a variable loop with two trinucleotides and a T-loop with a single nucleotide *T*.

Table 7
Transfer RNA of asparagine with anticodon *GTT* (tRNA-Asn-GTT) of *Escherichia coli* (Escherichia coli str K-12 substr MG1655 tRNA-Asn-GTT; 76 bp; chr:2044549-2044624 (+); tRNAviz; Lin et al., 2019). The upper part of the table is the tRNA-Asn-GTT upstream the anticodon *GTT* at the nucleotide position 34 (in bold). The lower part of the table is the tRNA-Asn-GTT downstream the anticodon *GTT*. Each part has three rows. The upper row gives the position of trinucleotides (1st trinucleotide site). The middle row gives the tRNA-Asn-GTT on the standard alphabet B_{DNA}. The lower row represents the tRNA-Asn-GTT on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$, X_1 and X_2 are defined in (1), (2) and (3), respectively, and $Z = \{AAA, CCC, GGG, TTT\}$.

Pos	1	4	7	10	13	16	19	22	25	28	31	**34**
B_{DNA}	*TCC*	*TCT*	*GTA*	*GTT*	*CAG*	*TCG*	*GTA*	*GAA*	*CGG*	*CGG*	*ACT*	*GTT*
B_X	X_1	X_1	X_0	X_0	X_0	X_1	X_0	X_0	X_2	X_2	X_1	X_0
Pos	37	40	43	46	49	52	55	58	61	64	67	70
B_{DNA}	*AAT*	*CCG*	*TAT*	*GTC*	*ACT*	*GGT*	*TCG*	*AGT*	*CCA*	*GTC*	*AGA*	*GGA*
B_X	X_0	X_1	X_2	X_0	X_1	X_0	X_1	X_2	X_1	X_0	X_2	X_2

The parameter $R(Y)$ of balanced circular code pairing is all the stronger as its values are close to 0. For the circular code X observed in genes (defined in (1)), $R(X) = |46.31 - 46.19| = 0.12\%$. Fig. 1 shows that the circular code X has the strongest balanced circular code pairing (numbering 1) among the 216 circular codes. The parameter $R(Y)$ ranges between 0.12% and 61.05%. This result is a new property for the circular code X in its class of 216 maximal C^3 self-complementary trinucleotide circular codes.

4. An application of the circular code alphabet to the transfer RNAs

A few tRNAs of *Escherichia coli* (str K-12 substr MG1655) obtained by tRNAviz (Lin et al., 2019) are studied on the circular code alphabet. A

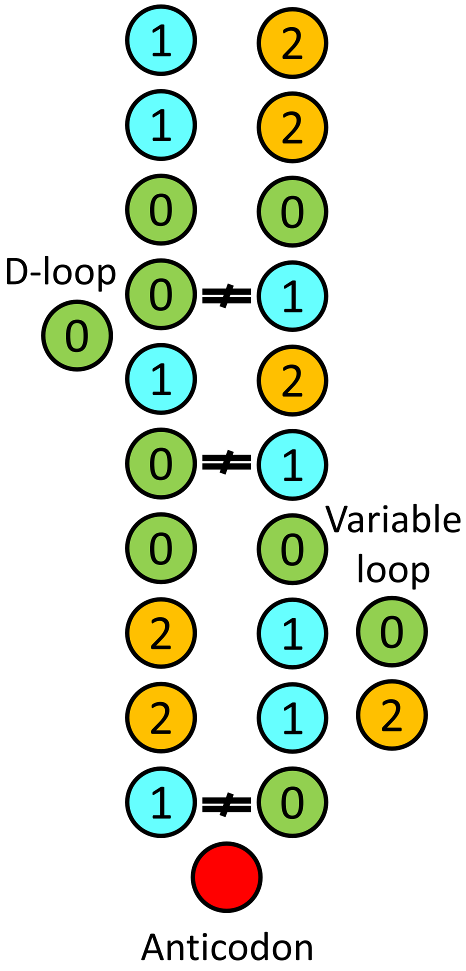


Fig. 6. Representation of the transfer RNA of asparagine with anticodon *GTT* (tRNA-Asn-GTT) of *Escherichia coli* (from Table 7) on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$ is defined in (1) and noted "0" with a green color, X_1 is defined in (2) and noted "1" with a blue color, X_2 is defined in (3) with an orange color and noted "2", $Z = \{AAA, CCC, GGG, TTT\}$ and noted "3" with a purple color, and the anticodon *GTT* is in red. By definition, X_0 pairs with itself, and X_1 pairs with X_2, and reciprocally. The symbol "$\neq$" means a mismatch. The tRNA-Asn-GTT on the circular code alphabet is represented by a stem with an anticodon loop made of three trinucleotides, a D-loop with one trinucleotide and a variable loop with two trinucleotides.

Table 8

Transfer RNA of glutamine with anticodon *CTG* (tRNA-Gln-CTG) of *Escherichia coli* (Escherichia coli str K-12 substr MG1655 tRNA-Gln-CTG; 75 bp; chr:696430-696504 (−); tRNAviz; Lin et al., 2019). The upper part of the table is the tRNA-Gln-CTG upstream the anticodon *CTG* at the nucleotide position 33 (in bold). The lower part of the table is the tRNA-Gln-CTG downstream the anticodon *CTG*. Each part has three rows. The upper row gives the position of trinucleotides (1st trinucleotide site). The middle row gives the tRNA-Gln-CTG on the standard alphabet B_{DNA}. The lower row represents the tRNA-Gln-CTG on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$, X_1 and X_2 are defined in (1), (2) and (3), respectively, and $Z = \{AAA, CCC, GGG, TTT\}$.

Pos	1	4	7	10	13	16	18	21	24	27	30	**33**		
B_{DNA}	*TGG*	*GGT*	*ATC*	*GCC*	*AAG*	*CG*	*GTA*	*AGG*	*CAC*	*CGG*	*ATT*	***CTG***		
B_X	X_2	X_0	X_0	X_0	X_1		X_0	X_1	X_2	X_2	X_0	X_0		
Pos	36	39	42	45	48	51	54	55	58	61	64	67	70	73
B_{DNA}	*ATT*	*CCG*	*GCA*	*TTC*	*CGA*	*GGT*	*T*	*CGA*	*ATC*	*CTC*	*GTA*	*CCC*	*CAG*	*CCA*
B_X	X_0	X_1	X_2	X_0	X_2	X_0		X_2	X_0	X_0	X_0	Z	X_0	X_1

complete analysis of tRNAs would require the development of specific algorithms, which is not the purpose here.

4.1. Transfer RNA of alanine

The transfer RNA (tRNA) of alanine (tRNA-Ala) with anticodon *GGC* (tRNA-Ala-GGC) is given in Table 3. As the anticodon *GGC* is in position 34 (1 modulo 3), the sequence upstream the anticodon *GGC* starting at position 1 can be divided into 11 trinucleotides. Similarly, the sequence downstream the anticodon *GGC* ending at position 76 can be divided into 12 trinucleotides, by conventionally excluding the last tetranucleotide ACCA. Interestingly, the tRNA-Ala-GGC on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$, where $X_0 = X$, X_1 and X_2 are defined in (1), (2) and (3), respectively, and $Z = \{AAA, CCC, GGG, TTT\}$ (Table 3), can be represented by a stem with an anticodon loop made of a single trinucleotide which is the anticodon, a D-loop made of one trinucleotide and a variable loop made of two trinucleotides (Fig. 2). By excluding the D-loop and the variable loop, the circular code pairing from the anticodon leads to a stem as the main structure for this tRNA. To get an order of magnitude for the statistical significance, the probability that a given X_i is correctly paired, is equal to $\frac{20}{64}$. Then, the probability to have 9 pairings and 1 mismatch in a stem of 10 trinucleotides, is equal to ≈ 0.0002 with a binomial distribution. The 4 trinucleotides surrounding the anticodon can pair on B_{DNA}: *CAT* in position 31 is complementary to *ATG* in position 37, and *TTG* in position 28 is complementary to *CAA* in position 40. On the standard alphabet B_{DNA}, the trinucleotide *GCT* in position 10 belonging to the D-stem is complementary to the trinucleotide *AGC* in position 23. In contrast, on the circular code alphabet B_X, the trinucleotide X_2 (*GCT*) in position 10 pairs with the trinucleotide X_1 (*CCG*) in position 61, the trinucleotide X_2 (*CGC*) in position 25 pairs with the trinucleotide X_1 (*AGC*) in position 49, etc.

The transfer RNA of alanine with anticodon *TGC* (tRNA-Ala-TGC) is given in Table 4. As the anticodon is also in position 34 (1 modulo 3), the same reasoning as previously is applied. The tRNA-Ala-TGC on the circular code alphabet B_X (Table 4) can also be represented, as the tRNA-Ala-GGC, by a stem with an anticodon loop made of a single trinucleotide which is the anticodon, a D-loop made of one trinucleotide and a variable loop made of two trinucleotides (Fig. 3). As previously, the statistical significance is equal to ≈ 0.0002 with a binomial distribution. In contrast to the tRNA-Ala-GGC, the 2 trinucleotides surrounding the anticodon cannot pair on B_{DNA}: *CTT* in position 31 is not complementary to *ACG* in position 37. Interestingly, these 2 trinucleotides pair on B_X as X_2 (*CTT*) is complementary to X_1 (*ACG*). In the tRNA-Ala-GGC, X_1 (*TTG*) in position 28 pairs with X_2 (*CAA*) in position 40, and in the tRNA-Ala-TGC, X_0 (*CTG*) in position 28 pairs with X_0 (*CAG*) in position 40. The trinucleotides in position 49 is *AGC* in tRNA-Ala-GGC and *TGC* in tRNA-Ala-TGC, but both trinucleotides belong to X_1 that pair with X_2 in position 25. Similarly, the trinucleotides in position 64 is *CTT* in tRNA-Ala-GGC and *CAT* in tRNA-Ala-TGC, but both trinucleotides belong to X_2 that pair with X_1 in position 7.

4.2. Transfer RNA of arginine

The transfer RNA of arginine with anticodon *ACG* (tRNA-Arg-ACG) is given in Table 5. As the anticodon *ACG* is in position 35 (2 modulo 3), one nucleotide in the sequence upstream the anticodon *ACG* must not be included in the series of trinucleotides on the circular code alphabet B_X. The nucleotide *G* in the D-loop will generate the frameshift. Without preconceived ideas, I have searched for a circular code pairing between the upstream and downstream sequences from the anticodon by analysing the 3 frames downstream the variable loop. Interestingly, Fig. 4 shows that the tRNA-Arg-ACG on the circular code alphabet can also be represented by a stem with an anticodon loop made of three trinucleotides, a D-loop with a single nucleotide *G* and a variable loop with two trinucleotides and one nucleotide *G*. The statistical significance is equal to ≈ 0.00007 with a binomial distribution.

The transfer RNA of arginine with anticodon *CCG* (tRNA-Arg-CCG) is given in Table 6. As the anticodon *CCG* is also in position 35 (2 modulo 3), a similar reasoning is applied as above. Fig. 5 shows that the tRNA-Arg-CCG on the circular code alphabet can be represented by a stem with an anticodon loop made of three trinucleotides, a D-loop with a single nucleotide *G*, a variable loop with two trinucleotides and a T-loop with a single nucleotide *T* in position 54 (see below the tRNA of glutamine in Section 4.4). The statistical significance is equal to ≈ 0.005 with a binomial distribution.

4.3. Transfer RNA of asparagine

The transfer RNA of asparagine with anticodon *GTT* (tRNA-Asn-GTT) is given in Table 7. As the anticodon *GTT* is in position 34 (1 modulo 3), a reasoning similar to the transfer RNA of alanine is applied. Fig. 6 shows that the tRNA-Asn-GTT on the circular code alphabet can be represented by a stem with an anticodon loop made of three trinucleotides, a D-loop with one trinucleotide and a variable loop with two trinucleotides. Compared to the tRNA of alanine, the stem has two additional mismatches and a D-loop occurring one trinucleotide earlier. The statistical significance is equal to ≈ 0.016 with a binomial distribution.

4.4. Transfer RNA of glutamine

The transfer RNA of glutamine with anticodon *CTG* (tRNA-Gln-CTG) is given in Table 8. It is also an interesting case as the anticodon *CTG* is in position 33 (0 modulo 3). Thus, one dinucleotide in the sequence upstream the anticodon *CTG* must not be included in the series of trinucleotides on the circular code alphabet B_X. Fig. 7 shows that the tRNA-Gln-CTG on the circular code alphabet can be represented by a stem with an anticodon loop made of a single trinucleotide which is the anticodon, a D-loop with one dinucleotide *CG*, a variable loop with one trinucleotide and a T-loop with a single nucleotide *T*. The statistical significance is equal to ≈ 0.011 with a binomial distribution. Interestingly, there is the same T-loop, with the same nucleotide *T* at the same position 54, in the tRNA-Gln-CTG and tRNA-Arg-CCG (see Fig. 5).

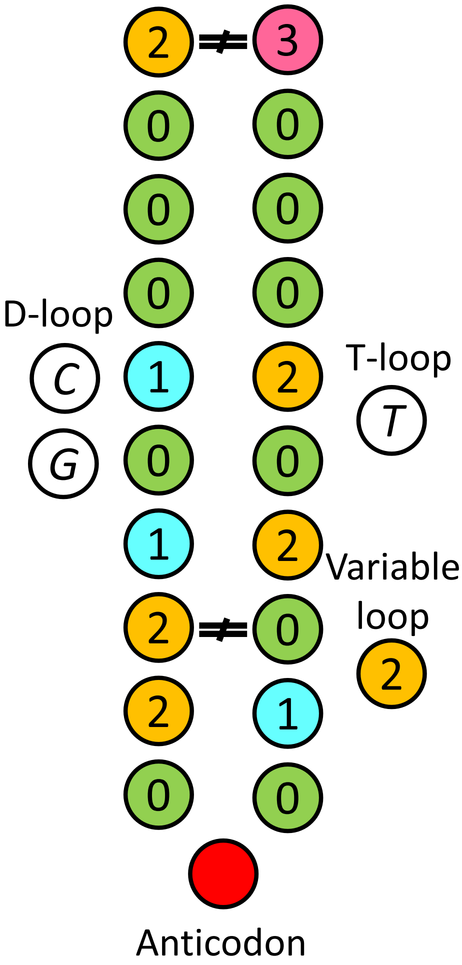


Fig. 7. Representation of the transfer RNA of glutamine with anticodon *CTG* (tRNA-Gln-CTG) of *Escherichia coli* (from Table 8) on the circular code alphabet $B_X = \{X_0, X_1, X_2, Z\}$ where $X_0 = X$ is defined in (1) and noted "0" with a green color, X_1 is defined in (2) and noted "1" with a blue color, X_2 is defined in (3) with an orange color and noted "2", $Z = \{AAA, CCC, GGG, TTT\}$ and noted "3" with a purple color, and the anticodon *CTG* is in red. By definition, X_0 pairs with itself, and X_1 pairs with X_2, and reciprocally. The symbol "$\neq$" means a mismatch. The tRNA-Gln-CTG on the circular code alphabet is represented by a stem with an anticodon loop made of a single trinucleotide which is the anticodon, a D-loop with a single dinucleotide *CG*, a variable loop with one trinucleotide and a T-loop with one nucleotide *T*.

5. Conclusion

The *X* motifs, motifs from the circular code *X*, are enriched in the protein coding genes of bacteria, archaea, eukaryotes, plasmids and viruses, moreover, in the minimal gene set belonging to the three domains of life, and are found in tRNA and rRNA sequences. As described in Introduction, they have properties of synchronisation of the reading frame in genes and properties of regulation of gene expression. These different results suggested to me to analyse the genes on the circular code alphabet. In this approach based on the circular code information, the structure of trinucleotides in 1st, 2nd and 3rd sites is no longer differentiated. Such differentiation could have occurred later in the evolutionary process of the genetic code, in particular the degeneracy of 3rd codon site, by unbalancing the balanced circular code pairing (Equation (8)) leading to an asymmetrical nucleotide pairing (Equation (9)) observed in genes on the DNA alphabet.

A new relation of occurrence of the circular code *X* under the hypothesis of an equiprobable strand pairing is given from a theoretical point of view. A statistical analysis of a large set of 4,148,022 genes in 1171 bacterial genomes (1,313,192,812 codons) retrieves this relation on the circular code alphabet but, surprisingly, not on the DNA alphabet. These positive and negative results are confirmed by two two-tailed Wilcoxon signed-rank tests in a paired sample. Furthermore, the circular code *X* has the strongest balanced circular code pairing among 216 maximal C^3 self-complementary trinucleotide circular codes, a new property of this circular code *X*. As an application of this theoretical work, different tRNAs are studied on the circular code alphabet. The few tRNAs studied are very different, not only in their sequence but also in the position of their anticodon which can occur in the 3 "frames" with respect to their sequence start, e.g. the anticodon of tRNA-Ala is in "reading frame" (position in 1 modulo 3), the anticodon of tRNA-Arg is in "frameshift 1" (position in 2 modulo 3) and the anticodon of tRNA-Gln is in "frameshift 2" (position in 0 modulo 3). Despite these many differences, a stem structure is unexpectedly identified in these tRNAs with a strong statistical significance according to a binomial distribution. It differs from the stem structure proposed by Hopfield (Hopfield, 1978, Fig. 1). Furthermore, the numbers *N* of X_0, X_1 and X_2, correctly paired or not, in the 6 studied stems without considering the D-, variable and T-loops are $N(X_0) = 48$, $N(X_1) = 38$ and $N(X_2) = 34$ ($120/3 = 40$ in the random case), showing an excess of X_0 (*p*-value equal to 0.073 with a one-tailed *z*-test for a proportion; Woolson, 1987). The numbers *N* of X_0, X_1 and X_2, correctly paired or not, in the 6 studied stems by now considering the D-, variable and T-loops are $N(X_0) = 59$, $N(X_1) = 39$ and $N(X_2) = 36$ ($134/3 \approx 45$ in the random case), showing an excess of X_0 with statistical significance (*p*-value equal to 0.006 with a one-tailed *z*-test for a proportion). It is important to stress that if a stem structure is indeed identified in the studied tRNAs, I do not claim that the structure found is the optimal solution for a stem and also that all tRNAs verify this property, since no specific algorithm has been developed for this purpose and only a very few tRNAs have been analysed. Interestingly, this stem structure in tRNAs opens stimulating theoretical questions with the theoretical minimal RNA rings of 22 nucleotide-long proposed for the tRNA loops (Demongeot and Seligmann, 2021; and previous works). Indeed, the RNA ring theory assumes nucleotide pairing within RNA rings and recovers tRNA-like properties for sequences designed according to genetic code coding properties, which suggests a dual function, as tRNA-like, and as gene-like (Demongeot and Moreira, 2007). This dual function was confirmed by further analyses of the RNA ring sequences (Demongeot and Seligmann, 2019b, 2019c, 2020b). However, from my point of view, this stem structure in tRNAs could be the ancestral traces of the DNA double helix structure (its construction) which is also antiparallel and complementary. In the tRNAs analysed, the anticodon loop has either one trinucleotide, that is the anticodon, or three trinucleotides forming a loop of 9 nucleotides. While an anticodon loop of one trinucleotide seems to be impossible according to the chemical bonds on a standard genetic alphabet, a circular code pairing (X_0, X_0) or (X_1, X_2) surrounding the anticodon could be more flexible with a partial pairing of the nucleotides constituting the trinucleotides of X_0, X_1 and X_2, for example only one or two nucleotides of trinucleotides or by extending the canonical DNA pairing to (G, T) ((G, U)) and (A, C) as observed in the RNA and RY alphabets (see Section 2.1.1). Otherwise, a ring loop of 9 nucleotides (anticodon loop with three trinucleotides) could also be associated with the pitch of the DNA double helix, such a motif size has already been involved in the DNA topology (Arquès and Michel, 1987). In the context of the circular code alphabet, the D-loop, the variable loop and the T-loop by inserting one nucleotide or one dinucleotide would be involved to position the anticodon in "reading frame" and to construct a coding stem in tRNAs as an outline of the future gene structure and the future DNA double helix.[1]

[1] A concept that the reader may not agree with.

Declaration of competing interest

The author reports no conflict of interest.

Acknowledgments

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

References

Arquès, D.G., Fallot, J.-P., Michel, C.J., 1997. An evolutionary model of a complementary circular code. J. Theor. Biol. 185, 241–253.

Arquès, D.G., Michel, C.J., 1987. A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups. J. Theor. Biol. 128, 457–461.

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

Bahi, J.M., Michel, C.J., 2004. A stochastic gene evolution model with time dependent mutations. Bull. Math. Biol. 66, 763–778.

Bahi, J.M., Michel, C.J., 2008. A stochastic model of gene evolution with chaotic mutations. J. Theor. Biol. 255, 53–63.

Bobay, L.-M., Ochman, H., 2017. The evolution of bacterial genome architecture. Front. Genet. 8, 1–6.

Demongeot, J., Moreira, A., 2007. A possible circular RNA at the origin of life. J. Theor. Biol. 249, 314–324.

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

Demongeot, J., Seligmann, H., 2019b. The uroboros theory of Life's origin: 22-nucleotide theoretical minimal RNA rings reflect evolution of genetic code and tRNA-rRNA translation machineries. Acta Biotheor. 67, 273–297.

Demongeot, J., Seligmann, H., 2019c. RNA rings strengthen hairpin accretion hypotheses for tRNA evolution: a reply to commentaries by Z.F. Burton and M. Di Giulio. J. Mol. Evol. 88, 243–252.

Demongeot, J., Seligmann, H., 2020a. Pentamers with non-redundant frames: bias for natural circular code codons. J. Mol. Evol. 88, 194–201.

Demongeot, J., Seligmann, H., 2020b. The primordial tRNA acceptor stem code from theoretical minimal RNA ring clusters. BMC Genet. 21 (7), 1–13.

Demongeot, J., Seligmann, H., 2021. Codon assignment evolvability in theoretical minimal RNA rings. Gene 769 (1452), 1–10.

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 Houmami, N., Seligmann, H., 2017. Evolution of nucleotide punctuation marks: from structural to linear signals. Front. Genet. 8 (36), 1–11.

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.

El Soufi, K., Michel, C.J., 2017. Unitary circular code motifs in genomes of eukaryotes. Biosystems 153, 45–62.

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

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Strüngmann, L., 2019. Mixed circular codes. Math. Biosci. 317, 1–14, 108231.

Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Starman, M., Strüngmann, L., 2020. The relation between *k*-circularity and circularity of codes. Bull. Math. Biol. 82 (105), 1–34.

Hopfield, J.J., 1978. Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proofreading. Proceedings of the National Academy of Sciences of the USA 75, 4334–4338.

Lin, B.Y., Chan, P.P., Lowe, T.M., 2019. tRNAviz: explore and visualize tRNA sequence features. Nucleic Acids Res. 47 (Issue W1), W542–W547.

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., 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., 2019. Single-frame, multiple-frame and framing motifs in genes. Life 9 (18), 1–22.

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., 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.

Seligmann, H., 2015. Codon expansion and systematic transcriptional deletions produce tetra-, pentacoded mitochondrial peptides. J. Theor. Biol. 387, 154–165.

Seligmann, H., 2017. Reviewing evidence for systematic transcriptional deletions, nucleotide exchanges, and expanded codons, and peptide clusters in human mitochondria. Biosystems 160, 10–24.

Seligmann, H., Warthi, G., 2020. Natural pyrrolysine-biased translation of stop codons in mitochondrial peptides entirely coded by expanded codons. Biosystems 196, 1–11, 104180.

Thompson, J.D., Ripp, R., Mayer, C., Poch, O., Michel, C.J., 2021. Potential role of the X circular code in the regulation of gene expression. Biosystems 203, 1–15, 104368.

Warthi, G., Seligmann, H., 2019. Transcripts with systematic nucleotide deletion of 1-12 nucleotide in human mitochondrion suggest potential non-canonical transcription. PloS One 14, 1–23 e0217356.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics 1, 80–83.

Woolson, R.F., 1987. Statistical Methods for the Analysis of Biomedical Data. In: Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc.