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 C3 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)

\[ X = X_0 = \{ \text{AAC, AAT, ACC, ATC, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GCA, GTA, GTC, TTA, TTG} \} \]

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

\[ X_1 = \{ \text{AAG, ACA, AGC, ACT, AGG, ATA, ATG, CCA, CCG, CGA, CGT, GTA, GTC, TAC, TCA, TCC, TCG, TCT, TGC, TTA, TTG} \} \]

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

\[ X_2 = \{ \text{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT, CTA, CIT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TTG} \} \]

The trinucleotide set \( X \) (defined in (1)) coding the reading frame in genes is a maximal (20 trinucleotides) C3 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 translation (Seligmann, 2015, 2017; El Houghami 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).
Table 1
Average codon frequency (%) $F$ (codon usage; Equation (5)) in 1171 bacterial genomes with a total number of 4,148,022 genes of 1,313,192,812 codons.

<table>
<thead>
<tr>
<th>Codon</th>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Codon</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>3.01</td>
<td>CAA</td>
<td>1.58</td>
<td>GAA</td>
</tr>
<tr>
<td>AAG</td>
<td>2.08</td>
<td>CAG</td>
<td>2.08</td>
<td>GAG</td>
</tr>
<tr>
<td>AAT</td>
<td>2.05</td>
<td>CAT</td>
<td>1.04</td>
<td>GAT</td>
</tr>
<tr>
<td>ACA</td>
<td>1.06</td>
<td>CCA</td>
<td>0.77</td>
<td>GCA</td>
</tr>
<tr>
<td>ACC</td>
<td>2.08</td>
<td>CCC</td>
<td>1.13</td>
<td>GCC</td>
</tr>
<tr>
<td>AGG</td>
<td>1.31</td>
<td>CGG</td>
<td>1.70</td>
<td>GGG</td>
</tr>
<tr>
<td>ACT</td>
<td>0.95</td>
<td>CCT</td>
<td>0.85</td>
<td>GCT</td>
</tr>
<tr>
<td>AGA</td>
<td>1.65</td>
<td>CGA</td>
<td>0.42</td>
<td>GGA</td>
</tr>
<tr>
<td>AGC</td>
<td>1.32</td>
<td>CGC</td>
<td>2.08</td>
<td>GGC</td>
</tr>
<tr>
<td>AGT</td>
<td>0.39</td>
<td>CGT</td>
<td>1.07</td>
<td>GGT</td>
</tr>
<tr>
<td>ATA</td>
<td>1.11</td>
<td>CTA</td>
<td>0.58</td>
<td>GTA</td>
</tr>
<tr>
<td>ATT</td>
<td>1.07</td>
<td>CTT</td>
<td>1.42</td>
<td>GTT</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Codon</th>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>26.05</td>
<td>29.35</td>
<td>19.43</td>
</tr>
<tr>
<td>C</td>
<td>21.86</td>
<td>23.18</td>
<td>29.70</td>
</tr>
<tr>
<td>G</td>
<td>36.07</td>
<td>29.35</td>
<td>17.43</td>
</tr>
<tr>
<td>T</td>
<td>16.02</td>
<td>30.03</td>
<td>23.69</td>
</tr>
</tbody>
</table>

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 $10 \pm 5\%$ 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.

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 genomes 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-
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 \( s = (AG)^\ast = AGAGAG... \) leads to a modulo 2 periodicity while on \( B_{RY} \), this sequence \( s = R^\ast = RRR... \) 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 map \( \overline{s}_B \) on these nucleotide alphabets \( B \) are recalled. The DNA complementarity map \( \overline{s}_{DNA} \) is the (canonical) standard one.

**Definition 1.** The DNA complementarity map \( \overline{s}_{DNA} : B_{DNA} \rightarrow B_{DNA} \) is defined by \( \overline{s}_{DNA}(A) = T, \overline{s}_{DNA}(C) = G, \overline{s}_{DNA}(G) = C \) and \( \overline{s}_{DNA}(T) = A \).

There are variants of this canonical complementarity map \( \overline{s}_{DNA} \). They extend the canonical complementarity map \( \overline{s}_{DNA} \) by adding nucleotide pairings.

**Definition 2.** The RNA complementarity map \( \overline{s}_{RNA} : B_{RNA} \rightarrow B_{RNA} \) is defined by \( \overline{s}_{RNA}(A) = U, \overline{s}_{RNA}(C) = G, \overline{s}_{RNA}(G) = \{ C, U \} \) and \( \overline{s}_{RNA}(U) = \{ A, G \} \).

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

**Definition 3.** The RY complementarity map \( \overline{s}_{RY} : B_{RY} \rightarrow B_{RY} \) is defined by \( \overline{s}_{RY}(R) = Y \) and \( \overline{s}_{RY}(Y) = R \).

The map \( \overline{s}_{RY} \) 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 “/=” 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.
A (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 \( p_B : B^n \rightarrow B^n \) is defined by \( p_B(l_1l_2 \ldots l_n) = l_2 \ldots l_n l_1 \) for all \( l_i \in B, i \in \{1, \ldots, n\} \). The \( n-1 \) iterates of \( p_B \) are defined similarly.

**Observation 1.** The permutation map \( p_B \) is (obviously) the identity on the nucleotide alphabets \( B (n = 1) \), i.e. \( p_B(l) \equiv l \) for all \( l \in B \). In other words, \( p_B \) 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 \( \bar{s} \) is denoted by the pair \( p = \{(l, \bar{e}(l)) ; (e(l), \bar{l})\} \) where \( l \in B \) and \( e(l) \in B \) is the nucleotide complementarity map. Under the realistic hypothesis of an equiprobable (balanced) strand pairing, some relations can be obtained in the strand \( s \).

**Property 1.** (DNA pairing rule). On the alphabet \( B_{RNA} \), 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 \( s \):

\[
N(A) + N(T) = N(C) + N(G).
\]

**Remark 1.** When the occurrence number \( N(l) \) of nucleotides is computed both in the 2 strands \( s \) and \( \bar{s} \), 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 \( \bar{s} \) (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 \( s \):

\[
N(A) + N(U) = N(C) + N(G) = N(C) + N(U)
\]

leading to

\[
N(A) = N(G) \quad \text{and} \quad N(C) = N(U).
\]

### 2.2. Codon usage matrix

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

\[
F_j = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{64} F_{ij}.
\]

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 \( F(X_0), F(X_1) \) and \( F(X_2) \) of the circular codes \( X_0 \), \( X_1 \) and \( X_2 \) (defined in (1), (2) and (3), respectively) are computed as follows:

\[
F(X_i) = \sum_{j=1}^{64} F_{ij},
\]

where \( F_j \) is obtained by Equation (5) and \( i \in \{0, 1, 2\} \).

### 2.3. Bacterial gene kingdom

A kingdom of \( |\mathcal{S}| = 1171 \) bacterial genomes \( \mathcal{S} \) is obtained from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, January 2021). In each genome \( \mathcal{S} \), 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 (TAG, TGA); and (iv) their lengths are not modulo 3. In order to obtain a broad but unduplicated sampling of each kingdom, 1 genome \( \mathcal{S} \) 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, MS-90 and NI, etc., but only one sequenced genome was chosen randomly in each organism group. Thus, the 1171 selected bacterial genomes \( \mathcal{S} \) 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 \( B^Z \).

**Definition 5.** The circular code alphabet is defined by \( B^Z = \{X_0, X_1, X_2, Z\} \) 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 $B = \{X_0, X_1, X_2, Z\}$ 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 “/=” 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 $G$. 

C.J. Michel
Table 6
Transfer RNA of arginine with anticodon CCG (tRNA-Arg-CCG) of *Escherichia coli* (*Escherichia coli* coli K-12 str M1655 tRNA-Arg-CCG; 77 bp; chr:3982375-3982451 (+); tRNAviz; Lin et al., 2019). The upper part of the table is the tRNA-Ala-CCG downstream the anticodon CCG at the nucleotide position 35 (in bold). The lower part of the table is the tRNA-Ala-CCG upstream the anticodon CCG. Each part has three rows. The upper row gives the positions of trinucleotides (1st trinucleotide site). The middle row gives the tRNA-Ala-CCG on the standard alphabet (Bahi and Michel, 2008, Section 3.1.2), eukaryotes (Arquès et al., 1997, Table 2). In order to evaluate the relation of circular code pairing (Equation (7)) is verified. Indeed, as \( F(X_1) + F(X_2) = 46.19\% \) then \( F(X_0) = F(X_1) + F(X_2) \).

\[
F(X_0) = F(X_1) + F(X_2)
\]

\( F(X_0) \) is retrieved in genes (reading frame) of prokaryotes (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 \( F(X_1) + F(X_2) = 46.19\% \) then \( F(X_0) = F(X_1) + F(X_2) \).

\[
F(X_0) = F(X_1) + F(X_2)
\]

In order to evaluate the statistical significance of a similar distribution of the frequency of \( X_0 \) and \( X_1 \) in the 1171 bacterial genomes, in each genome \( j \), the frequency \( F_{X_0} \) of \( X_0 \) and the frequency sum \( F_{X_0} + F_{X_1} \) of \( X_0 \) and \( X_1 \) is computed as follows: \( F_{X_0} = \sum_{i \in X_0} F_i \) and \( F_{X_1} = \sum_{i \in X_1} F_i \), where \( F_i \) is the frequency (usage) of codon \( j \) in genome \( i \). 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 3,131,192,812 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 B0RNA, the computed frequency sums \( F(A) + F(T) = 48.19\% \) and \( F(C) + F(G) = 51.81\% \) (Table 2) lead to \( F(A) + F(T) \neq F(C) + F(G) \).

\[
F(A) + F(T) \neq F(C) + F(G)
\]

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 \( F_A + F_T \) of \( A \) and \( T \), and \( F_C + F_G \) of \( C \) and \( G \) are computed in each genome \( i \). Using the 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:

\[
R(Y) = \left| F(Y_0) - \left( F(Y_1) + F(Y_2) \right) \right|
\]
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”, 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 “/=” 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 BRNA. 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 \( Z = \{ AAA, CCC, GGG, TTT \} \).

<table>
<thead>
<tr>
<th>Pos</th>
<th>1</th>
<th>4</th>
<th>7</th>
<th>10</th>
<th>13</th>
<th>16</th>
<th>18</th>
<th>21</th>
<th>24</th>
<th>27</th>
<th>30</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_{RNA} )</td>
<td>TGG</td>
<td>GGT</td>
<td>ATC</td>
<td>GCC</td>
<td>AAG</td>
<td>CG</td>
<td>GTA</td>
<td>AGG</td>
<td>CGG</td>
<td>CGG</td>
<td>ATT</td>
<td>CTG</td>
</tr>
<tr>
<td>( B_X )</td>
<td>( x_0 )</td>
<td>( x_1 )</td>
<td>( x_2 )</td>
<td>( x_0 )</td>
<td>( x_0 )</td>
<td>( x_0 )</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
</tr>
</tbody>
</table>

4.1. Transfer RNA of alanine

The transfer RNA (tRNA) of alanine (tRNA-Ala) with anticodon GCC (tRNA-Ala-GCC) is given in Table 3. As the anticodon GCC is in position 34 (1 modulo 3), the sequence upstream the anticodon GCC starting at position 1 can be divided into 11 trinucleotides. Similarly, the sequence downstream the anticodon GCC ending at position 76 can be divided into 12 trinucleotides, by conventionally excluding the last tetranucleotide ACCA. Interestingly, the tRNA-Ala-GCC 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 \( \frac{\theta}{2} \). Then, the probability to have 0 pairings and 1 mismatch in a stem of 10 trinucleotides, is equal to \( \approx 0.0002 \) with a binomial distribution. The 4 trinucleotides surrounding the anticodon can pair on \( B_{RNA} \): CAT in position 31 is complementary to ATG in position 37, and TGG in position 28 is complementary to CAA in position 40. On the standard alphabet \( B_{RNA} \), 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 \( X_0 \) (GCT) in position 10 pairs with the trinucleotide \( X_1 \) (CCG) in position 61, the trinucleotide \( X_2 \) (CCG) in position 25 pairs with the trinucleotide \( X_2 \) (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-GCC, 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 \( \approx 0.0002 \) with a binomial distribution. In contrast to the tRNA-Ala-GCC, the 2 trinucleotides surrounding the anticodon cannot pair on \( B_{RNA} \): CTG in position 31 is not complementary to AGG in position 37. Interestingly, these 2 trinucleotides pair on \( B_X \) as \( X_1 \) (CTT) is complementary to \( X_1 \) (AGG). In the tRNA-Ala-GGC, \( X_1 \) (TTG) in position 28 pairs with \( X_0 \) (CA), in position 40, and in the tRNA-Ala-TGC, \( X_0 \) (CTG) in position 28 pairs with \( X_1 \) (CA) in position 40. The trinucleotides in position 49 is AGG 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 \( x_1 \) that pair with \( x_1 \) in position 7.

4.2. Transfer RNA of arginine

The transfer RNA of arginine with anticodon ACC (tRNA-Arg-ACC) is given in Table 5. As the anticodon ACC is in position 35 (2 modulo 3), one nucleotide in the sequence upstream the anticodon ACC 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 analyzing the 3 frames downstream the variable loop. Interestingly, Fig. 4 shows that the tRNA-Arg-ACC 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 \( \approx 0.00007 \) 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 \( \approx 0.005 \) 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 \( \approx 0.016 \) 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 \( B_X \). In contrast to the tRNA-Ala-GGC, the 2 trinucleotides surrounding the anticodon do not pair on \( B_{RNA} \): CTG in position 31 is not complementary to AGG in position 37. Interestingly, these 2 trinucleotides pair on \( B_X \) as \( X_2 \) (CTT) is complementary to \( X_2 \) (AGG). In 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 \( \approx 0.011 \) 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).
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 frame in genes and properties of regulation of gene expression. These 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 C0 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 X0, X1 and X2, correctly paired or not, in the 6 studied stems without considering the D-, variable and T-loops are N(X0) = 48, N(X1) = 38 and N(X2) = 34 (120/3 = 40 in the random case), showing an excess of X0 (p-value equal to 0.073 with a one-tailed z-test for a proportion; Woolson, 1987). The numbers N of X0, X1 and X2, correctly paired or not, in the 6 studied stems by now considering the D-, variable and T-loops are N(X0) = 59, N(X1) = 39 and N(X2) = 36 (134/3 = 45 in the random case), showing an excess of X0 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, 2020h). 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 (X0, X1) or (X1, X2) surrounding the anticodon could be more flexible with a partial pairing of the nucleotides constituting the trinucleotides of X0, X1 and X2, 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 YF 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