Journal of Theoretical Biology 365 (2015) 164–174

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

An extended genetic scale of reading frame coding

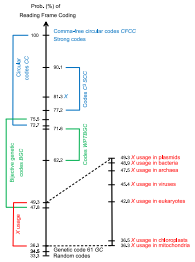
Christian J. Michel

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

HIGHLIGHTS

- Genetic scale of reading frame coding of usage of trinucleotide codes.
- Reading frame coding of the C^3 self-complementary circular code *X*.
- Genes with genetic information for reading frame coding.
- Genes of bacteria and plasmids with the highest efficiencies for reading frame coding.
- Gene evolution by coding.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 22 May 2014
Received in revised form
29 August 2014
Accepted 30 September 2014
Available online 13 October 2014

Keywords:
Reading frame coding
Trinucleotide circular code
Circular code usage
Genetic scale
Coding evolution

ABSTRACT

The reading frame coding (RFC) of codes (sets) of trinucleotides is a genetic concept which has been largely ignored during the last 50 years. An extended definition of the statistical parameter *PrRFC* (Michel, 2014) is proposed here for analysing the probability (efficiency) of reading frame coding of usage of any trinucleotide code. It is applied to the analysis of the RFC efficiency of usage of the C^3 self-complementary trinucleotide circular code *X* identified in prokaryotic and eukaryotic genes (Arquès and Michel, 1996). The usage of *X* is called usage *XU*. The highest RFC probabilities of usage *XU* are identified in bacterial plasmids and bacteria (about 49.0%). Then, by decreasing values, the RFC probabilities of usage *XU* are observed in archaea (47.5%), viruses (45.4%) and nuclear eukaryotes (42.8%). The lowest RFC probabilities of usage *XU* are found in mitochondria and chloroplasts (about 36.5%). Thus, genes contain information for reading frame coding. Such a genetic property which to our knowledge has never been identified, may bring new insights in the origin and evolution of the genetic code.

1. Introduction

Codon usage within and among species has been attributed to various factors such as expression level, GC content, recombination rates, RNA stability, codon position in the gene, gene length and others including environmental stress and population size (reviewed in Behura and Severson, 2013). To our knowledge, the reading frame coding (RFC) of usage of codes (set) of trinucleotides is a concept which has never been studied.

The reading frame coding of a trinucleotide code, e.g. the genetic code, is a fascinating and open problem. It is also an old problem. Almost sixty years ago (in 1957), before the discovery of the genetic code, a class of trinucleotide codes, called comma-free codes was proposed by Crick et al. (1957) for explaining how the reading of a series of trinucleotides could code amino acids. By excluding the four periodic permuted trinucleotides {AAA, CCC, GGG, TTT} and by gathering the 60 remaining trinucleotides in 20 classes of three trinucleotides such that, in each class, three trinucleotides are deduced from each other by the circular permutation map, e.g. ACG, CGA and GAC, we see that a comma-free code has only one trinucleotide per class and therefore contains at most 20 trinucleotides. This trinucleotide number is

E-mail address: c.michel@unistra.fr
URL: http://dpt-info.u-strasbg.fr/~c.michel/

http://dx.doi.org/10.1016/j.jtbi.2014.09.040
0022-5193/

identical to the amino acid number, thus leading to a code assigning one trinucleotide per amino acid without ambiguity. However, no trinucleotide comma-free code was identified in genes statistically. Furthermore, the discovery that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes phenylalanine (Nirenberg and Matthaei, 1961), led to the abandonment of the concept of comma-free code.

In 1996, a statistical analysis of occurrence frequencies of the 64 trinucleotides {AAA,…,TTT} in the three frames of genes of both prokaryotes and eukaryotes showed that the trinucleotides are not uniformly distributed in these three frames (Arquès and Michel, 1996). By convention here, the frame 0 is the reading frame in a gene, and the frames 1 and 2 are the reading frame 0 shifted by one and two nucleotides in the 5′–3′ direction, respectively. By excluding the four periodic permuted trinucleotides and by assigning each trinucleotide to a preferential frame (frame of its highest occurrence frequency), three subsets $X = X_0$, X_1 and X_2 of 20 trinucleotides are found in the frames 0, 1 and 2, respectively, simultaneously of two large gene populations (protein coding regions): prokaryotes (13,686 sequences, 4708,758 trinucleotides) and eukaryotes (26,757 sequences, 11,397,678 trinucleotides) (Arquès and Michel, 1996). This set X contains the 20 following

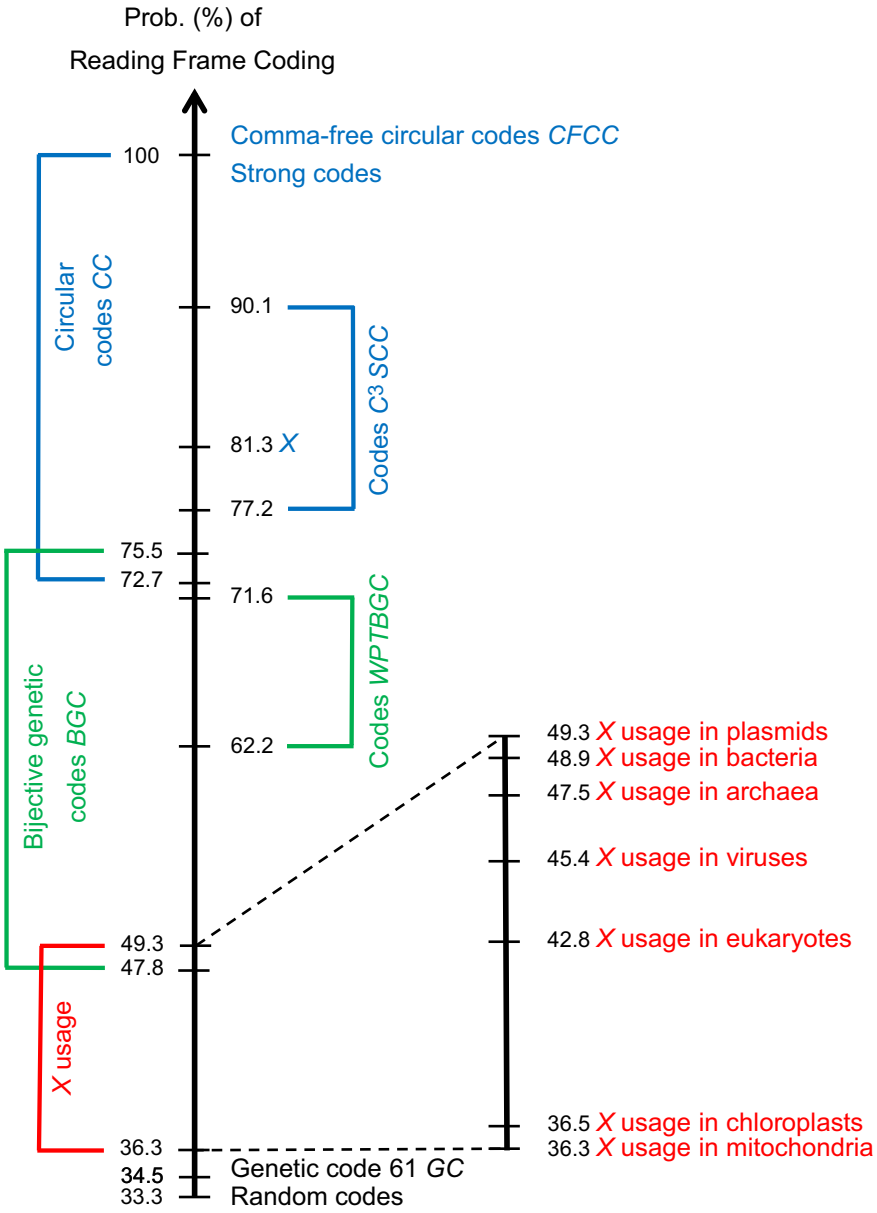


Fig. 1. An extended genetic scale of reading frame coding of usage of trinucleotide codes. Reading frame coding probability $PrRFC(X)$ of the C^3 self-complementary circular code X equal to 81.3% (Table 2). Reading frame coding probability $PrRFC(X,K)$ (Eq. (4)) of usage of the C^3 self-complementary circular code X in gene kingdoms K: $PrRFC(X,A) = 47.5\%$ in archaea A (357,142 genes, 101,350,970 trinucleotides), $PrRFC(X,B) = 48.9\%$ in bacteria B (7862,438 genes, 2484,909,928 trinucleotides), $PrRFC(X,E) = 42.8\%$ in nuclear eukaryotes E (1891,168 genes, 940,289,792 trinucleotides), $PrRFC(X,V) = 45.4\%$ in viruses V (184,995 genes, 45,871,186 trinucleotides), $PrRFC(X,M) = 36.3\%$ in mitochondrion M (1164 genes, 217,899 trinucleotides), $PrRFC(X,C) = 36.5\%$ in chloroplasts C (1495 genes, 395,768 trinucleotides) and $PrRFC(X,P) = 49.3\%$ in bacterial plasmids P (238,368 genes, 68,492,239 trinucleotides). The reading frame coding probability of trinucleotide codes from Fig. 1 in Michel (2014) is recalled: (i) the 12,964,440 circular codes CC including the 408 comma-free circular codes $CFCC$ and the 216 C^3 self-complementary circular codes C^3SCC; (ii) the 339,738,624 bijective genetic codes BGC of 20 trinucleotides coding the 20 amino acids including the 52 bijective genetic codes $WPTBGC$ without permuted trinucleotides (WPT). The genetic scale of reading frame coding of usage of trinucleotide codes ranges according to Proposition 1 from a probability equal to 0 to a probability equal to 1 with the comma-free codes and the strong codes (the reading frame is always retrieved). The random codes (one chance out of three to retrieve the reading frame among the three possible frames in genes) have a reading frame coding probability equal to 1/3.

trinucleotides

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

The two sets X_1 and X_2, of 20 trinucleotides each, in the shifted frames 1 and 2, respectively, of genes can be deduced from X by the circular permutation map (see below). These three trinucleotide sets present several strong mathematical properties, particularly the fact that X is a C^3 self-complementary trinucleotide circular code (Arquès and Michel, 1996). A trinucleotide circular code has the fundamental property to always retrieve the reading frame in any position of any sequence generated with the circular code. In particular, initiation and stop trinucleotides as well as any frame signals are not necessary to define the reading frame. Indeed, a window of a few nucleotides, whose nucleotide length depends on the circular code, positioned anywhere in a sequence generated with the circular code always retrieves the reading frame (Lassez, 1976; Berstel and Perrin, 1985). For crossing the largest ambiguous words generated with the circular code X (words, not necessarily unique, in two or three frames), this window needs a length of 13 nucleotides with X (Michel, 2012, Fig. 3). A window of 13 nucleotide length is the largest window of X to retrieve the reading frame for all the ambiguous words generated with X.

Gonzalez et al. (2011), by defining a statistical function analysing the covering capability of a circular code, showed on a gene data set from 13 classes of proteins that the circular code X has, on average, the best covering capability among the whole class of the 216 C^3 self-complementary trinucleotide circular codes (Arquès and Michel, 1996; list given in Tables 4a, 5a and 6a in Michel et al., 2008). A review of this circular code X gives some additional properties (Michel, 2008). Recently, X motifs, i.e. motifs generated with the circular code X, are identified in the 5′ and/or 3′ regions of 16 isoaccepting tRNAs of prokaryotes and eukaryotes (Michel, 2013). Seven X motifs of length greater or equal to 15 nucleotides are also found in 16S rRNAs, in particular in the ribosome decoding center which recognizes the codon-anticodon helix in A-tRNA (Michel, 2012; El Soufi and Michel, 2014). A 3D visualization of X motifs in the ribosome shows several spatial configurations involving mRNA X motifs, A-tRNA and E-tRNA X motifs, and 16S rRNA X motifs (Michel, 2012; El Soufi and Michel, 2014). These results led to the concept of a possible translation (framing) code based on circular code (Michel, 2012).

There are two mathematical approaches for proving that a trinucleotide code is circular or not: a classical proof based on the flower automaton (Lassez, 1976; Berstel and Perrin, 1985) and a modern proof, more refined, using the necklaces *5LDCN* (Pirillo, 2003) and *nLDCCN* (Michel and Pirillo, 2010). Indeed, the necklace proof allows not only to decide if a trinucleotide code is circular or not, but also to classify the circular codes. The most general hierarchy of circular codes is given in Proposition 4.1 in Michel and Pirillo (2011). However, these proofs do not allow a quantitative measure of the property of reading frame coding for trinucleotide codes, inside as well as between the classes of circular and non-circular codes. Important trinucleotide non-circular codes studied in Michel (2014) are the bijective genetic codes of 20 trinucleotides coding the 20 amino acids.

In Michel (2014), we have defined a statistical parameter, called *PrRFC*, for analysing the probability (efficiency) of reading frame coding (RFC) of any trinucleotide code C, circular or not. The RFC probability $PrRFC(C)$ of a trinucleotide code C is the ratio of the occurrence probability of C in frame 0 to the occurrence probabilities of C in the three frames 0, 1 and 2. A genetic scale of reading frame coding (Michel, 2014, Fig. 1) is proposed for two main classes of trinucleotide codes C:

(i) The 12,964,440 circular codes *CC* and their two subclasses. The necklace subclass contains the comma-free codes and six other necklace circular codes based on different necklace lengths. The map subclass is based on to the complementarity and circular permutation maps. It includes the 216 C^3 self-complementary circular codes.
(ii) The 339,738,624 bijective genetic codes *BGC* of 20 trinucleotides coding the 20 amino acids and some subclasses. These codes *BGC* are not circular.

However, the *PrRFC* definition for analysing the efficiency of reading frame coding of trinucleotide codes is based on trinucleotide probabilities chosen to be equiprobable and of sum equal to 1. It does not allow measuring the reading frame coding of a trinucleotide code which is used in genes, such as the observed trinucleotide frequencies of the circular code X in genes of prokaryotes and eukaryotes.

We extend here the previous *PrRFC* definition by relaxing the condition that the sum of trinucleotide probabilities of a trinucleotide code C must be equal to 1. The newly *PrRFC* definition is applied to measure the reading frame coding of the C^3 self-complementary circular code X in genes of large kingdoms of archaea, bacteria, nuclear eukaryotes, viruses, mitochondrion, chloroplasts and bacterial plasmids. An extended genetic scale of reading frame coding is proposed here.

2. Method

2.1. Definitions

We briefly recall a few classical definitions in order to understand the property of reading frame coding (RFC) of trinucleotide circular codes.

Notation 1. The letters (or nucleotides or bases) define the genetic alphabet $A_4 = \{\text{A, C, G, T}\}$. The set of non-empty words (words, respectively) over A_4 is denoted by A_4^+ (A_4^*, respectively). The set of the 64 words of length 3 (trinucleotides or triletters) on A_4 is denoted by $A_4^3 = \{\text{AAA}, \ldots, \text{TTT}\}$. Let $x_1 \cdots x_n$ be the concatenation of the words x_i for $i = 1, \ldots, n$, the symbol "$\cdot$" being the concatenation operator.

Notation 2. In genes, there are three frames f. By convention here, the reading frame $f = 0$ is established by a start codon $\{\text{ATG, GTG, TTG}\}$, and the frames $f = 1$ and $f = 2$ are the reading frame $f = 0$ shifted by one and two nucleotides in the $5' - 3'$ direction, respectively.

There are two important biological maps involved in codes in genes on A_4.

Definition 1. The *nucleotide complementarity map* $\mathcal{C} : A_4 \to A_4$ is defined by $\mathcal{C}(\text{A}) = \text{T}$, $\mathcal{C}(\text{C}) = \text{G}$, $\mathcal{C}(\text{G}) = \text{C}$, $\mathcal{C}(\text{T}) = \text{A}$. According to the property of the complementary and antiparallel double helix, the *trinucleotide complementarity map* $\mathcal{C} : A_4^3 \to A_4^3$ is defined by $\mathcal{C}(l_0 \cdot l_1 \cdot l_2) = \mathcal{C}(l_2) \cdot \mathcal{C}(l_1) \cdot \mathcal{C}(l_0)$ for all $l_0, l_1, l_2 \in A_4$, e.g. $\mathcal{C}(\text{ACG}) = \text{CGT}$. By extension to a trinucleotide set S, the *set complementarity map* $\mathcal{C} : S \to S$ is defined by $\mathcal{C}(S) = \{v \mid u, v \in A_4^3, u \in S, v = \mathcal{C}(u)\}$, i.e. a complementary trinucleotide set $\mathcal{C}(S)$ is obtained by applying the complementarity map $\mathcal{C}$ to all its trinucleotides, e.g. $\mathcal{C}(\{\text{ACG, AGT}\}) = \{\text{ACT, CGT}\}$.

Definition 2. The *trinucleotide circular permutation map* $\mathcal{P} : A_4^3 \to A_4^3$ is defined by $\mathcal{P}(l_0 \cdot l_1 \cdot l_2) = l_1 \cdot l_2 \cdot l_0$ for all $l_0, l_1, l_2 \in A_4$, e.g. $\mathcal{P}(\text{ACG}) = \text{CGA}$. The 2nd iterate of $\mathcal{P}$ is denoted $\mathcal{P}^2$, e.g. $\mathcal{P}^2(\text{ACG}) = \text{GAC}$. By extension to a trinucleotide set S, the *set circular permutation map* $\mathcal{P} : S \to S$ is defined by $\mathcal{P}(S) = \{v \mid u, v \in A_4^3, u \in S, v = \mathcal{P}(u)\}$, i.e. a permuted trinucleotide set $\mathcal{P}(S)$ is obtained by applying

the circular permutation map $\mathcal{P}$ to all its trinucleotides, e.g. $\mathcal{P}(\{ACG,\ AGT\}) = \{CGA, GTA\}$ and $\mathcal{P}^2(\{ACG,\ AGT\}) = \{GAC, TAG\}$.

Definition 3. A set $S \subset A_4^+$ of words is a *code* if, for each $x_1, \ldots, x_n, y_1, \ldots, y_m \in S$, $n, m \geq 1$, the condition $x_1 \cdots x_n = y_1 \cdots y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \ldots, n$.

Definition 4. As the set $A_4^3 = \{AAA, \ldots, TTT\}$ is a code, its non-empty subsets are codes and called *trinucleotide codes C*.

Definition 5. A trinucleotide code $C \subset A_4^3$ is *circular* and called *CC* if, for each $x_1, \ldots, x_n, y_1, \ldots, y_m \in C$, $n, m \geq 1$, $r \in A_4^*$, $s \in A_4^+$, the conditions $sx_2 \cdots x_n r = y_1 \cdots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word) and $x_i = y_i$ for $i = 1, \ldots, n$.

Remark 1. A trinucleotide code C containing either one periodic permuted trinucleotide $PPT = \{AAA, CCC, GGG, TTT\}$ or two non-periodic permuted trinucleotides $NPPT = \{t, \mathcal{P}(t)\}$ for a trinucleotide $t \in A_4^3 \backslash PPT$ cannot be circular. Thus, the two trinucleotide codes A_4^3 and $A_4^3 \backslash PPT$ are not circular.

Remark 2. The fundamental property of a circular code is the ability to retrieve the reading (original or construction) frame of any sequence generated with this circular code. A circular code is a set of words over an alphabet such that any sequence written on a circle (the next letter after the last letter of the sequence being the first letter) has a unique decomposition (factorization) into words of the circular code (Michel, 2012, both Fig. 1 for a graphical representation of the circular code definition and Fig. 2 for an example). The reading frame in a sequence (gene) is retrieved after the reading of a certain number of letters (nucleotides), called the window of the circular code. The length of this window for retrieving the reading frame is the letter length of the longest ambiguous word, not necessarily unique, which can be read in at least two frames, plus one letter (Michel, 2012, Fig. 3 for an example).

Definition 6. A trinucleotide circular code $CC \subset A_4^3$ is *self-complementary* and called *SCC* if, for each $y \in CC$, $\mathcal{C}(y) \in CC$.

Definition 7. A trinucleotide circular code $CC \subset A_4^3$ is C^3 and called C^3CC if the two permuted trinucleotide sets $CC_1 = \mathcal{P}(CC)$ and $CC_2 = \mathcal{P}^2(CC)$ are trinucleotide circular codes.

Definition 8. A trinucleotide circular code $CC \subset A_4^3$ is C^3 *self-complementary* and called C^3SCC if CC, $CC_1 = \mathcal{P}(CC)$ and $CC_2 = \mathcal{P}^2(CC)$ are trinucleotide circular codes satisfying the following properties $CC = \mathcal{C}(CC)$ (self-complementary), $\mathcal{C}(CC_1) = CC_2$ and $\mathcal{C}(CC_2) = CC_1$ (CC_1 and CC_2 are complementary).

The trinucleotide set $X = X_0$ (Eq. (1)) coding the reading frame (frame 0) in prokaryotic and eukaryotic genes is a maximal (20 trinucleotides) C^3 self-complementary circular code C^3SCC with a window length equal to 13 nucleotides for biinfinite words (Arquès and Michel, 1996) and 12 nucleotides for right infinite words (Michel, 2012). The circular code $X_1 = \mathcal{P}(X)$ contains the 20 following trinucleotides

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

and the circular code $X_2 = \mathcal{P}^2(X)$, the 20 following trinucleotides

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

Thus, X, $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ are trinucleotide circular codes verifying $X = \mathcal{C}(X)$, $\mathcal{C}(X_1) = X_2$ and $\mathcal{C}(X_2) = X_1$.

2.2. Probability of reading frame coding of usage of a trinucleotide code

A simple quantitative definition is proposed for measuring the probability $PrRFC(C)$ of reading frame coding (RFC) for any code C of trinucleotides t with any assigned probabilities. The RFC probability $PrRFC(C)$ of a trinucleotide code C is the ratio of the

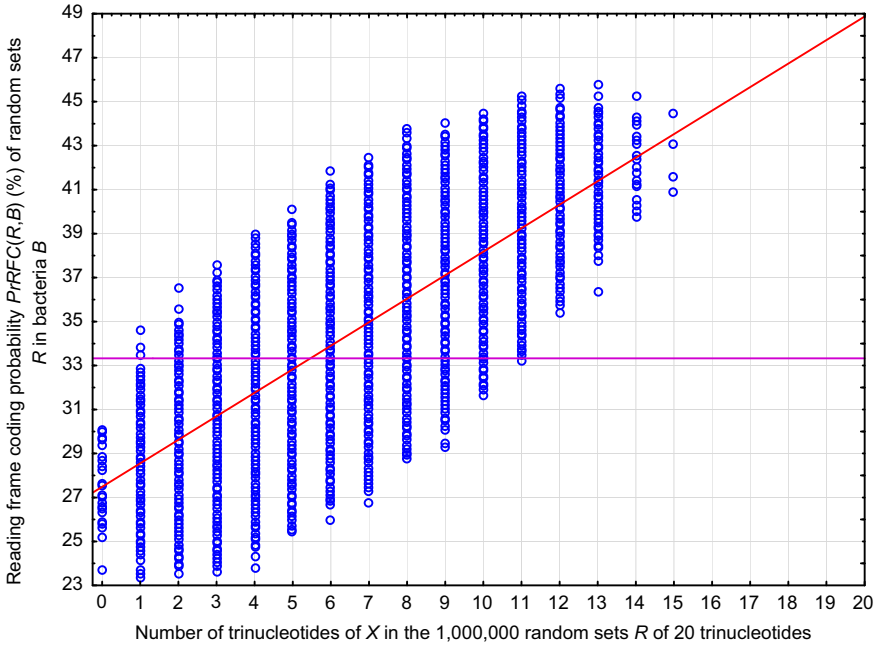


Fig. 2. Reading frame coding probability $PrRFC(R, B)$ (Eq. (4)) of one million random sets R of 20 trinucleotides from 61 trinucleotides, i.e. without the three stop codons TAA, TAG and TGA, in bacteria B. The horizontal line (violet) is the reading frame coding probability $PrRFC(Rand) = 1/3$ of random codes *Rand* (Fig. 1). The linear adjustment equation $y = 1.074x + 27.4652$ (red) shows that the estimated set $\hat{R}$ containing 20 trinucleotides of X has a probability $PrRFC\left(\hat{R}, B\right) = 1.074 \times 20 + 27.4652 \approx 48.9\%$ which is the value (to the first decimal place) of the probability $PrRFC(X, B) = 48.9\%$ of X in bacteria B (Fig. 1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

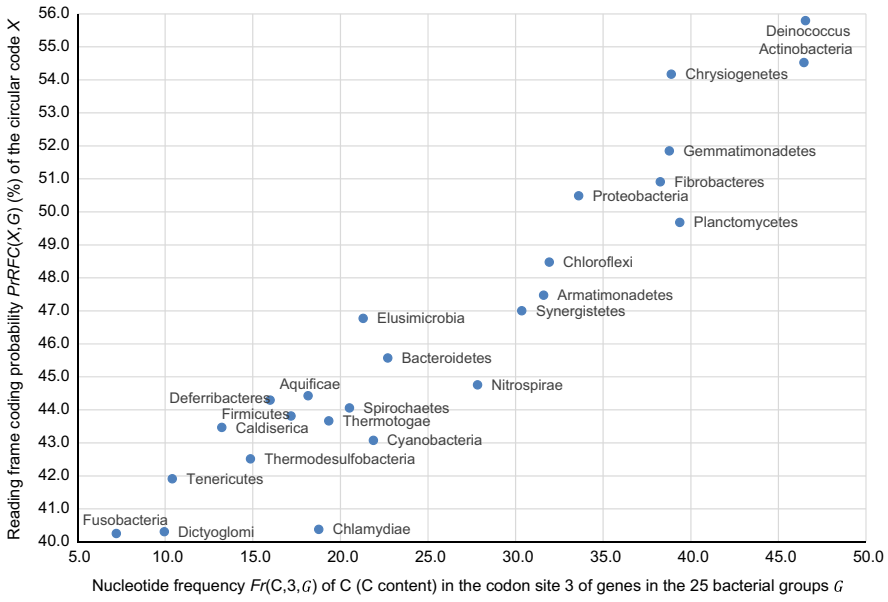


Fig. 3. Reading frame coding probability $PrRFC(X,G)$ (Eq. (4)) of the C^3 self-complementary circular code X as a function of the frequency $Fr(C,3,G)$ of the nucleotide C in the codon site 3 of genes in the 25 bacterial groups G (correlation $r=0.946$, Table 5).

occurrence probability of C in frame 0 to the occurrence probabilities of C in the three frames 0, 1 and 2.

Let $Pr(t_f,K)$ be the frequency of a trinucleotide t_f of a code C occurring in a frame $f\in\{0,1,2\}$ of a (protein coding) gene kingdom K. For a code C of 20 trinucleotides, e.g. the circular code X, there are $3\times 20=60$ trinucleotide frequencies $Pr(t_f,K)$ in K. The probability $Pr(C_f,K)$ of a trinucleotide code C_f in a frame $f\in\{0,1,2\}$ of a gene kingdom K is equal to

$$Pr(C_f,K)=\sum_{t_f\in C} Pr(t_f,K) \tag{2}$$

with $\sum_{t_f\in C} Pr(t_f,K)\le 1$. Then, the reading frame coding probability $PrRFC(C)$ (efficiency) of a trinucleotide code C_f in a frame $f\in\{0,1,2\}$ of a gene kingdom K is equal to

$$PrRFC(C_f,K)=\frac{Pr(C_f,K)}{\sum_{f=0}^{2} Pr(C_f,K)}. \tag{3}$$

This simple formula allows to measure the reading frame coding efficiency of any trinucleotide code C_f in a given frame f.

Proposition 1. $0\le PrRFC(C_f,K)\le 1$ according to Eq. (3).

The more the RFC probability $PrRFC(C_f,K)$ value is raised, the more the trinucleotide code C_f in frame f has efficiency for coding its frame f.

(i) $PrRFC(C_f,K)=1$ if $0<Pr(C_f,K)\le 1, Pr\left(C_{f'},K\right)=Pr(C_{f''},K)=0$ (same conditions with Eq. (6) in Michel, 2014, associated to the comma-free codes and the strong codes). The code C_f in frame f only codes its frame f and its frame f is always retrieved.

(ii) $PrRFC(C_f,K)=0$ if $Pr(C_f,K)=0$ with: (iia) $Pr\left(C_{f'},K\right)=0$ and $0<Pr(C_{f''},K)\le 1$; or (iib) $0<Pr\left(C_{f'},K\right)\le 1$ and $0\le Pr(C_{f''},K)\le 1$. The code C_f has no occurrence in frame f and the frame f is not coded.

(iii) $PrRFC(C_f,K)=1/3$ if $Pr(C_{f''},K)=2Pr(C_f,K)-Pr\left(C_{f'},K\right)$ with: (iiia) $Pr(C_f,K)=Pr\left(C_{f'},K\right)=1$ (same conditions, i.e. $Pr(C_f,K)=Pr\left(C_{f'},K\right)=Pr(C_{f''},K)=1$, with Eq. (6) in Michel, 2014, associated to the random codes *Rand*); or (iiib) $0<Pr(C_f,K)\le 1/2$ and $0\le Pr\left(C_{f'},K\right)\le 2Pr(C_f,K)$; or (iiic) $1/2<Pr(C_f,K)<1$ and $2Pr(C_f,K)-1\le Pr\left(C_{f'},K\right)\le 1$.

Proposition 2. Let us denote $PrRFC(C_f,K;Pr(t_f,K))$ the probability $PrRFC(C_f,K)$ of a code C_f in a frame f of K as a function of its trinucleotides probabilities $Pr(t_f,K)$. Let λ be a scalar. Then, according to Eq. (3)

$$PrRFC(C_f,K;\lambda\times Pr(t_f,K))=PrRFC(C_f,K;Pr(t_f,K)).$$

Proposition 2 allows to measure the RFC probability $PrRFC(C_f,K)$ of a code C_f in a frame f of a gene kingdom K regardless its absolute trinucleotide probabilities $Pr(t_f,K)$. Only its relative trinucleotide probabilities $Pr(t_f,K)$ determine the RFC probability $PrRFC(C_f,K)$. Thus, the normalization of trinucleotide probabilities $Pr(t_f,K)$ is not necessary for comparing different RFC efficiencies, either for a given trinucleotide code with different usage or for usage of different trinucleotide codes.

Remark 3. For a trinucleotide code C (from a point of view of coding theory), e.g. the circular code X, its trinucleotide probabilities in frame 0 are $Pr(t_0)=Pr(t)=1/20$ and its trinucleotide probabilities $Pr(t_f)$ in the two shifted frames $f\in\{1,2\}$ are estimated from the product of trinucleotide probabilities $Pr(t')$ and $Pr(t'')$ in frame 0 for all the di-trinucleotides $t't''\in C^2$ and with the simplest hypothesis of independent events. In contrast, for a trinucleotide code usage CU in genes, its trinucleotide probabilities $Pr(t_f)$ in the three frames $f\in\{0,1,2\}$ are the observed trinucleotide frequencies in the three frames of a gene kingdom K. Its trinucleotide probabilities $Pr(t_f)$ in the two shifted frames $f\in\{1,2\}$ are not estimated from a probability product as this information is available.

Remark 4. The RFC probability $PrRFC(C_f,K)$ of a trinucleotide code C_f in a frame f of a gene kingdom K can be measured as long as the sum of its trinucleotide probabilities $Pr(t_f,K)$ is less or equal to 1 (Eq. (2)). In Michel (2014), the sum of trinucleotide probabilities $Pr(t)$ of a trinucleotide code C must be equal to 1 strictly. Indeed, the RFC measure of codes C of 20 trinucleotides, e.g. the circular code X, is based on the simplest hypothesis that each trinucleotide t of C occurs with the same probability, i.e. $Pr(t)=1/20$ for all $t\in C$ leading to $Pr(C)=1$. Thus, the definition

Table 1

Observed trinucleotide frequencies $Pr(t_0,K)$ ($Pr(t_1,K)$ and $Pr(t_2,K)$, respectively) of the C^3 self-complementary circular code X in reading frame $f=0$ (usage XU) (in the shifted frames $f=1$ and $f=2$, respectively) of genes in kingdoms K of archaea A (357,142 genes, 101,350,970 trinucleotides), bacteria B (7862,438 genes, 2484,909,928 trinucleotides), nuclear eukaryotes E (1891,168 genes, 940,289,792 trinucleotides), viruses V (184,995 genes, 45,871,186 trinucleotides), mitochondrion M (1164 genes, 217,899 trinucleotides), chloroplasts C (1495 genes, 395,768 trinucleotides) and bacterial plasmids P (238,368 genes, 68,492,239 trinucleotides).

K	Archaea A			Bacteria B			Eukaryotes E			Viruses V			Mitochondrion M			Chloroplasts C			Plasmids P		
$t\in X$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$
AAC	1.90	1.21	1.34	1.79	1.46	1.14	2.00	1.42	1.18	2.45	1.61	1.51	1.26	1.41	2.24	1.19	1.75	2.10	1.77	1.35	0.99
AAT	1.86	1.21	2.58	1.93	1.26	1.60	2.19	1.23	1.48	2.72	1.71	1.97	3.30	2.04	4.15	3.61	2.44	3.65	1.70	1.09	1.24
ACC	1.53	1.59	0.85	2.12	1.65	0.74	1.54	1.72	1.13	1.62	1.61	1.00	1.12	1.00	1.53	0.78	0.73	1.28	2.21	1.74	0.71
ATC	2.30	1.21	0.93	2.71	1.62	0.73	1.87	1.43	1.16	2.17	1.52	1.11	1.87	1.58	1.84	1.49	2.32	1.77	3.02	1.72	0.73
ATT	2.22	1.25	1.73	2.44	1.40	1.49	1.92	1.36	1.36	2.50	1.83	1.84	4.35	2.60	5.02	4.58	2.85	3.61	1.72	1.19	1.16
CAG	1.45	1.81	1.26	2.18	1.45	1.11	2.73	2.39	1.95	1.72	1.78	1.04	0.55	2.25	0.53	0.75	1.92	0.47	2.35	1.33	1.28
CTC	2.47	0.78	1.20	1.73	0.74	0.99	1.71	1.63	1.65	1.28	1.11	0.99	1.14	1.36	1.17	0.62	1.68	0.59	2.19	0.73	1.09
CTG	1.85	1.80	0.77	3.66	1.53	0.93	2.77	2.79	1.16	1.91	2.19	0.78	0.68	1.67	0.53	0.53	2.08	0.35	3.91	1.39	1.10
GAA	3.52	1.42	2.33	3.47	0.72	1.87	3.24	1.18	2.92	3.58	0.97	2.29	2.65	1.78	1.17	4.12	2.07	1.88	2.99	0.69	1.96
GAC	3.17	0.54	1.50	2.63	0.44	1.44	2.38	0.68	1.32	2.92	0.57	1.37	0.80	0.84	0.64	0.79	0.86	0.68	2.90	0.51	1.75
GAG	4.18	1.25	1.44	2.62	0.65	0.65	3.64	1.34	1.51	2.67	0.94	0.81	0.86	1.47	0.59	1.24	1.41	0.68	2.88	0.64	0.77
GAT	2.76	0.66	2.47	2.80	0.40	2.36	2.77	0.60	1.67	3.22	0.64	1.81	2.08	1.53	1.27	3.20	1.27	1.60	2.57	0.42	2.53
GCC	2.42	0.87	1.11	3.54	1.61	1.63	2.17	1.35	1.49	2.03	0.96	1.12	1.18	0.52	0.54	0.75	0.41	0.50	4.16	1.88	1.86
GGC	2.29	0.97	2.22	3.34	1.27	3.00	1.80	1.08	2.03	1.94	0.80	1.94	0.67	0.38	0.83	0.61	0.56	0.95	3.68	1.60	3.46
GGT	1.59	0.73	2.04	1.76	0.54	2.33	1.43	0.67	1.60	2.11	0.71	1.57	2.19	0.60	0.80	2.52	0.78	0.90	1.38	0.61	2.36
GTA	1.56	0.90	0.87	1.08	0.95	0.59	0.85	0.85	0.67	1.41	1.41	0.91	2.59	1.58	0.44	2.25	1.85	0.45	0.80	0.70	0.52
GTC	2.50	0.52	0.75	2.04	0.79	0.86	1.41	0.87	1.18	1.54	0.89	1.05	0.77	0.85	0.49	0.58	1.21	0.57	2.50	0.77	0.99
GTT	2.16	0.58	1.58	1.52	0.86	1.67	1.59	0.80	1.41	1.96	0.97	1.73	1.96	1.53	1.27	2.26	1.81	1.23	1.25	0.67	1.53
TAC	1.98	0.91	1.03	1.32	0.75	1.07	1.44	0.77	0.99	1.75	0.91	1.47	0.96	2.11	1.92	0.77	1.68	1.95	1.26	0.60	0.86
TTC	2.14	1.29	1.08	1.94	1.33	1.15	1.90	1.62	1.57	1.87	1.35	1.30	2.35	2.36	2.03	1.89	2.89	2.40	2.22	1.34	1.02

in Michel (2014) cannot measure the reading frame coding of trinucleotide codes with different usage in contrast to the new RFC definition presented here.

Remark 5. It is important to stress that Eq. (3), as Eq. (6) in Michel (2014), satisfies the combinatorial properties of trinucleotides codes. In particular, the RFC probability is equal to 1 with the comma-free codes and the strong codes where the trinucleotides are only in the frame 0 and the RFC probability is equal to 1/3 with the random codes where the trinucleotides are in the three frames 0, 1 and 2 equiprobably.

Eq. (3) is applied here to the trinucleotide code X which is a C^3 self-complementary circular code and to the frame $f=0$ as X is identified in the reading frame of genes. Thus, $C_f=X_0=X$. Then, the reading frame coding probability $PrRFC(X,K)$ (efficiency) of the C^3 self-complementary circular code X in a gene kingdom K is equal to

$$PrRFC(X,K)=\frac{Pr(X,K)}{Pr(X,K)+\sum_{f=1}^{2}Pr(X_f,K)}. \quad (4)$$

Remark 6. Eq. (6) in Michel (2014) is a particular case of Eq. (4) with $Pr(X,K)=1$.

Remark 7. In a concept similar to codon usage in genes, $Pr(X,K)$ is the usage of the circular code X in genes and is called usage XU.

The RFC probability $PrRFC(X,K)$ (Eq. (4)) can very easily be programmed in a computer language or in a spreadsheet. It can also easily be extended to genetic motifs of any finite length (dinucleotide codes, tetranucleotide codes, etc.).

2.3. Usage of the C^3 self-complementary circular code X

Genes of kingdoms K of archaea, bacteria, nuclear eukaryotes, viruses, mitochondrion, chloroplasts and bacterial plasmids are extracted from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, May 2014). Usual preliminary tests exclude genes with nucleotides different from A_4, without start codons {ATG, GTG, TTG}, without stop codons {TAA, TAG, TGA} and with nucleotide lengths non-modulo 3.

Table 1 gives the 20 observed trinucleotide frequencies $Pr(t_0,K)$ of the C^3 self-complementary circular code X in reading frame $f=0$ (usage XU) of gene kingdoms K of archaea A (357,142 genes, 101,350,970 trinucleotides), bacteria B (7862,438 genes, 2484,909,928 trinucleotides), nuclear eukaryotes E (1891,168 genes, 940,289,792 trinucleotides), viruses V (184,995 genes, 45,871,186 trinucleotides), mitochondrion M (1164 genes, 217,899 trinucleotides), chloroplasts C (1495 genes, 395,768 trinucleotides) and bacterial plasmids P (238,368 genes, 68,492,239 trinucleotides). Table 1 also gives the $2\times 20=40$ observed trinucleotide frequencies $Pr(t_1,K)$ and $Pr(t_2,K)$ of X in the shifted frames $f=1$ and $f=2$, respectively, in the different kingdoms K. The $3(64-20)=132$ trinucleotide frequencies not related to the trinucleotides of X, i.e. $A_4^3\backslash X$, are not given here as they are not necessary for computing the RFC probability $PrRFC(X,K)$ (Eq. (4)). We have chosen here to compute the usage of the circular code X on large gene populations in order to have reference values of usage of X with stable and average trinucleotide frequencies $Pr(t_f,K)$.

2.4. Explained example

The reading frame coding probability of usage XU of X in bacteria B is equal to $PrRFC(X,B)=48.9\%$. Indeed, from Table 1 with the bacterial kingdom $K=B$, $Pr(X,B)=Pr(X_0,B)=46.6\%$ (sum of the 20 trinucleotide frequencies $Pr(t_0,B)$ of X in frame $f=0$ of B), $Pr(X_1,B)=21.4\%$ (sum of the 20 trinucleotide frequencies $Pr(t_1,B)$ of X in frame $f=1$ of B) and $Pr(X_2,B)=27.3\%$ (sum of the 20 trinucleotide frequencies $Pr(t_2,B)$ of X in frame $f=2$ of B). Thus, $PrRFC(X,B)=46.6/(46.6+21.4+27.3)=48.9\%$.

3. Results

3.1. Reading frame coding of the C^3 self-complementary circular code X

The reading frame coding probability of the C^3 self-complementary circular code X is equal to $PrRFC(X)=81.3\%$ (Table 2). This reference value is obtained from Eq. (4) with:

(i) equiprobable trinucleotide probabilities $Pr(t_0) = Pr(t)$ of X in frame 0, i.e. $Pr(t) = 1/20$ (20 trinucleotides in X) leading to a probability of X in frame 0 equal to $Pr(X) = \sum_{t \in X} Pr(t) = 1$;

(ii) trinucleotide probabilities $Pr(t_1)$ and $Pr(t_2)$ of X in frames $f = 1$ and $f = 2$, respectively, given in Table 2 (not given in Michel, 2014) which are estimated (see Remark 3) from the product of two trinucleotide probabilities in frame 0 according to Eq. (5) in Michel (2014) leading to a probability of X in frames $f = 1$ and $f = 2$ equal to $Pr(X_1) = Pr(X_2) = 11.5\%$ (see also Michel, 2014, (via) and (vib) in Section 2.2.2).

The value 81.3% of reading frame coding probability of X can also be obtained from Eq. (6) in Michel (2014).

3.2. Reading frame coding of usage of the C^3 self-complementary circular code X

Fig. 1 gives the reading frame coding probability $PrRFC(X,K)$ of usage XU of the C^3 self-complementary circular code X in gene kingdoms K of archaea A, bacteria B, nuclear eukaryotes E, viruses V, mitochondrion M, chloroplasts C and bacterial plasmids P. Recall that the trinucleotide probabilities $Pr(t_0) = Pr(t)$, $Pr(t_1)$ and $Pr(t_2)$ are the observed trinucleotide frequencies in the three frames of a gene kingdom K (see Remark 3).

The RFC probabilities $PrRFC(X,K)$ of usage XU in these seven kingdoms are significantly lower than the reference value $PrRFC(X) = 81.3\%$ of X. This observation may be related to the fact that today genes were subjected to a large number of mutations. Thus, and in particular, the four trinucleotides of the subset $\tilde{X} = \{\mathrm{CAG}, \mathrm{CTC}, \mathrm{CTG}, \mathrm{GAG}\}$ of X which is a C^3 self-complementary trinucleotide comma-free code (Michel, 2012) do not occur in the shifted frames (Table 2). But, the four trinucleotides of this subset $\tilde{X}$ occur in the shifted frames of today genes with frequencies different from 0 (Table 1).

The highest RFC probabilities of usage XU are observed in bacterial plasmids P and bacteria B with $PrRFC(X,P) \approx PrRFC(X,B) \approx 49.0\%$. Then, by decreasing values, the RFC probabilities of usage XU are found in archaea A with $PrRFC(X,A) = 47.5\%$, then viruses V with $PrRFC(X,V) = 45.4\%$ and nuclear eukaryotes E with $PrRFC(X,E) = 42.8\%$. The lowest RFC probabilities of usage XU are obtained in mitochondria M and chloroplasts C with $PrRFC(X,M) \approx PrRFC(X,C) \approx 36.5\%$, i.e. in the organelles of eukaryotes. These two lowest probabilities of mitochondria M and chloroplasts C are close but still higher than $PrRFC(Rand) = 1/3$ with the random codes $Rand$. Note that the RFC efficiency of usage XU in nuclear eukaryotes E is in the middle interval [36.3, 49.3] of usage XU in the seven kingdoms.

Fig. 1 extends the genetic scale of reading frame coding of trinucleotide codes (Michel, 2013, Fig. 1) by including the reading frame coding of usage XU of the C^3 self-complementary circular code X in the different gene kingdoms K. The genetic scale of reading frame coding of usage of trinucleotide codes ranges according to Proposition 1 from a probability equal to 0 to a probability equal to 1 with the comma-free codes and the strong codes (the reading frame is always retrieved). The random codes (one chance out of three to retrieve the reading frame among the three possible frames in genes) have a probability equal to 1/3 in the genetic scale.

3.3. Adequacy of the parameter PrRFC(X,K) for analysing the coding property of genes

From a theoretical point of view, the probability $PrRFC(X,K)$ is the "best" parameter for analysing the coding property of genes. Indeed, it is based on the code X which is the set of 20 trinucleotides having in average the highest occurrence in genes (reading frame) compared to the two shifted frames of both prokaryotes and eukaryotes (Arquès and Michel, 1996). Furthermore, this parameter $PrRFC(X,K)$ is, in addition, associated to a mathematical property as the code X is a C^3 self-complementary circular code (Arquès and Michel, 1996). Thus, these statistical and mathematical properties of $PrRFC(X,K)$ makes this parameter "unique" for analysing the coding property of genes.

Table 2
Reading frame coding probability of the C^3 self-complementary circular code X equal to $PrRFC(X) = 81.3\%$.

$t \in X$	$f = 0$ $Pr(t_0)$	$f = 1$ $Pr(t_1)$	$f = 2$ $Pr(t_2)$
AAC	5	0.75	0.5
AAT	5	0.5	1
ACC	5	2.25	0
ATC	5	1.5	0
ATT	5	1	0.5
CAG	5	0	0
CTC	5	0	0
CTG	5	0	0
GAA	5	0	1.5
GAC	5	0	0.75
GAG	5	0	0
GAT	5	0	1.5
GCC	5	0.75	0
GGC	5	0	0.75
GGT	5	0	2.25
GTA	5	1.25	0.75
GTC	5	0.75	0
GTT	5	0.5	0.75
TAC	5	0.75	1.25
TTC	5	1.5	0
$Pr(X_f)$	100	11.5	11.5
$PrRFC(X)$	81.3		

In order to have a statistical evaluation of the adequacy of the parameter $PrRFC(X,K)$ for its coding property, random sets R of 20 trinucleotides from 61 trinucleotides are generated by computer and their reading frame coding probabilities $PrRFC(R,B)$ are determined in the chosen kingdom of bacteria $K = B$. The 61 trinucleotides are the 64 trinucleotides without the three stop codons TAA, TAG and TGA which do not occur in reading frame of genes, except in a few rare cases explaining that their frequencies are not always equal to 0 in reading frame. Thus, this computer analysis needs the 20 trinucleotide probabilities $Pr(t_0,B)$, $Pr(t_1,B)$ and $Pr(t_2,B)$ of X in the three frames of genes in bacteria B given in Table 1 and the remaining $64 - 20 - 3 = 41$ trinucleotide probabilities $Pr(t_0,B)$, $Pr(t_1,B)$ and $Pr(t_2,B)$ of $A_4^3 \backslash \{X, \mathrm{TAA}, \mathrm{TAG}, \mathrm{TGA}\}$ which are not given here. There are $\binom{61}{20} \approx 6 \times 10^{15}$ possible sets of 20 trinucleotides among 61 trinucleotides. As they cannot all be enumerated (with a personal computer), a large sample of one million random sets R is generated.

The 1000,000 reading frame coding probabilities $PrRFC(R,B)$ of random sets R in bacteria B follow a Gaussian distribution (not shown) with a mean of 34.5%, a standard deviation of 2.6%, a 25th percentile of 32.7%, a 75th percentile of 36.3%, a minimum of 23.4% and a maximum of 45.8%. The mean RFC value 34.5% of random sets R retrieves the value $PrRFC(61GC) = 3721/10779 \approx 34.5\%$ (Michel, 2014, (v) in Section 2.2.2) of the genetic code $61GC$ with 61 equiprobable codons without the three stop codons {TAA, TAG, TGA} (Fig. 1 here and also Michel 2014, Fig. 1). The maximal RFC value 45.8% with this random sample is less than the RFC probability $PrRFC(X,B) = 48.9\%$ of X in bacteria B (Fig. 1). Its associated random set R (not shown) has 13 trinucleotides in common with X. The minimal value 23.4% with this random sample is associated to a random set R (not shown) with one

Table 3
Twenty five groups G of genes of bacteria extracted from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, May 2014) with their number of genes and trinucleotides.

Group G	Nb of genes	Nb of trinucleotides
Actinobacteria	1089,730	357,508,023
Aquificae	20,280	6165,531
Armatimonadetes	2809	1017,309
Bacteroidetes	318,160	110,404,488
Caldiserica	1581	481,735
Chlamydiae	127,066	43,201,999
Chloroflexi	51,703	17,418,781
Chrysiogenetes	2571	858,437
Cyanobacteria	283,213	87,868,703
Deferribacteres	9387	3041,832
Deinococcus	50,870	15,707,697
Dictyoglomi	3654	1177,023
Elusimicrobia	1529	494,013
Fibrobacteres	41,927	15,301,826
Firmicutes	1692,429	502,060,799
Fusobacteria	15,867	4998,875
Gemmatimonadetes	3935	1420,091
Nitrospirae	11,182	3380,667
Planctomycetes	27,692	10,270,528
Proteobacteria	3873,667	1225,804,951
Spirochaetes	114,930	38,196,569
Synergistetes	8903	2837,236
Tenericutes	63,690	20,717,741
Thermodesulfobacteria	3791	1199,562
Thermotogae	31,196	10,032,466

trinucleotide in common with X. These minimal and maximal values suggest that the RFC probabilities $PrRFC(R,B)$ of random sets R increase as a function of their number of trinucleotides of X they contain. In order to test this hypothesis, Fig. 2 shows the distribution of the RFC probabilities $PrRFC(R,B)$ of random sets R as a function of their number of trinucleotides of X. Very interestingly, the linear adjustment equation $y=1.074x+27.4652$ of 1000,000 random sets R shows that the estimated set $\hat{R}$ containing 20 trinucleotides of X has a probability $PrRFC\left(\hat{R},B\right)=1.074\times 20+27.4652\approx 48.9\%$ (Fig. 2) which is the value (to the first decimal place) of the probability $PrRFC(X,B)=48.9\%$ of X in bacteria B (Fig. 1). This statistical approach also demonstrates that a random set R of 20 trinucleotides is not appropriate for the analysis of the coding property of genes.

3.4. Correlation of the parameter PrRFC(X, B) with the nucleotide frequencies in the codon sites of bacteria B

Twenty five groups G of genes in the chosen kingdom of bacteria are extracted from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, May 2014) (Table 3). As in Section 2.3, usual preliminary tests exclude genes with nucleotides different from A_4, without start codons {ATG, GTG, TTG}, without stop codons {TAA, TAG, TGA} and with nucleotide lengths non-modulo 3. Bacterial groups having less than 1500 genes are excluded from this data sample.

Table 4a gives the number of nucleotides of the C^3 self-complementary circular code X per trinucleotide site. It is a simple counting of the letters in the three factors of length 1 of X. As the circular code X is self-complementary, the number of A (C, G and T, respectively) in the site 1 of X is equal to the number of T (G, C and A, respectively) in the site 3 of X and is equal to 5 (3, 10 and 2, respectively). Also as a consequence of the self-complementarity of X, the number of A (C, respectively) in the site 2 of X is equal to the number of T (G, respectively) in the site 2 of X and is equal to 8 (2, respectively). Table 4b represents this information in terms of

Table 4a
Number of nucleotides of the C^3 self-complementary circular code X per trinucleotide site.

Nucleotide	Site 1	Site 2	Site 3	Sum
A	5	8	2	15
C	3	2	10	15
G	10	2	3	15
T	2	8	5	15
Sum	20	20	20	

Table 4b
Frequency (%) of nucleotides of the C^3 self-complementary circular code X per trinucleotide site.

Nucleotide	Site 1	Site 2	Site 3
A	25	40	10
C	15	10	50
G	50	10	15
T	10	40	25

nucleotide frequency per trinucleotide site. Some nucleotides in the trinucleotide sites of X occur with significant higher frequencies. The site 1 of X is mainly related to G (50%), the site 2 of X, to A and T (40% for each nucleotide) and the site 3 of X, to C (50%) (Table 4b). Thus, some correlations between the reading frame coding probability $PrRFC(X,G)$ of the circular code X in the 25 bacterial groups G and some nucleotides in the trinucleotide sites of X can be expected.

It should be reminded, without being detailed here, that the C^3 self-complementary circular code X cannot be generated from the nucleotide frequencies in the trinucleotide sites (Lacan and Michel, 2001; Koch and Lehmann, 1997; Fimmel et al., 2014).

Table 5 gives the reading frame coding probability $PrRFC(X,G)$ of the circular code X in the 25 bacterial groups G. The lowest RFC value is observed with $PrRFC(X,Fusobacteria)=40.3\%$, the highest RFC value, with $PrRFC(X,Deinococcus)=55.8\%$ and the average value without group ponderation is 46.4%.

Table 5 also gives the frequency of nucleotides A, C, G, T and GC (frequency sum of C and G, also called GC content) in the codon sites 1, 2, 3 and {1,2,3} (per codon), noted $Fr(n,s,G)$ where $n\in\{A, C, G, T, GC\}$ and $s\in\{1, 2, 3, \{1,2,3\}\}$, of genes in the 25 bacterial groups G. It also provides the correlation noted $r(PrRFC(X), Fr(n,s))$ between the RFC probability $PrRFC(X,G)$ and each nucleotide frequency $Fr(n,s,G)$ in bacterial genes (last line in Table 5). Note obviously that $r(PrRFC(X), Fr(AT,s))=-r(PrRFC(X), Fr(GC,s))$ whatever the codon site $s\in\{1, 2, 3, \{1,2,3\}\}$, meaning that the correlation results with the GC content allow the correlation results for the AT content to be deduced.

The highest correlation between the RFC probability of the circular code X and the nucleotide frequency in the trinucleotide sites of X is observed with $r(PrRFC(X), Fr(C,3))=0.946$, i.e. with the nucleotide C in the codon site 3 of bacterial genes. This result is expected from a theoretical property of the C^3 self-complementary circular code X as C occurs with the highest frequency (50%) in the trinucleotide site 3 of X (see above). Fig. 3 represents the reading frame coding probability $PrRFC(X,G)$ of the circular code X as a function of the frequency $Fr(C,3,G)$ of the nucleotide C in the codon site 3 of genes in the 25 bacterial groups G. Then, by decreasing absolute values, $|r(PrRFC(X), Fr(T,\{1,2,3\}))|=0.933$ with the nucleotide T per codon, $|r(PrRFC(X), Fr(T,1))|=0.930$ with the nucleotide T in the codon site 1, $r(PrRFC(X), Fr(GC,3))=0.928$ with the GC content in the codon site 3, until the lowest

Table 5

Reading frame coding probability $PrRFC(X,G)$ (%) (Eq. (4)) of the C^3 self-complementary circular code X and frequency $Fr(n,s,G)$ (%) of nucleotides n (A, C, G, T and GC) in the codon sites s (1, 2, 3 and {1,2,3} (per codon)) of genes in the 25 groups G of bacteria. The last line gives the correlation r between the reading frame coding probability $PrRFC(X,G)$ of X and the frequency $Fr(n,s,G)$ of each nucleotide in bacterial genes.

Bacterial groups G	$PrRFC$	A in site				C in site				G in site				T in site				GC in site			
		1	2	3	{1,2,3}	1	2	3	{1,2,3}	1	2	3	{1,2,3}	1	2	3	1 2 3	1	2	3	{1,2,3}
Actinobacteria	54.5	18.6	23.5	7.0	16.4	26.8	28.6	46.5	33.9	42.3	20.4	37.2	33.3	12.3	27.6	9.3	16.4	69.1	49.0	83.7	67.2
Aquificae	44.4	33.4	34.3	33.0	33.6	15.5	18.3	18.2	17.3	33.0	14.6	19.2	22.3	18.1	32.8	29.6	26.9	48.4	32.9	37.3	39.6
Armatimonadetes	47.5	22.1	27.3	16.6	22.0	27.3	24.7	31.6	27.9	35.3	18.8	28.1	27.4	15.2	29.2	23.8	22.7	62.6	43.5	59.7	55.3
Bacteroidetes	45.6	30.7	33.3	25.7	29.9	18.0	20.8	22.7	20.5	32.3	15.7	21.4	23.1	18.9	30.2	30.1	26.4	50.4	36.5	44.1	43.7
Caldiserica	43.5	34.8	33.4	35.4	34.5	14.5	19.1	13.2	15.6	32.1	13.6	14.3	20.0	18.6	34.0	37.2	29.9	46.6	32.6	27.5	35.6
Chlamydiae	40.4	26.6	30.2	27.6	28.1	20.6	23.1	18.8	20.8	31.6	15.9	18.9	22.1	21.2	30.9	34.7	28.9	52.2	38.9	37.7	42.9
Chloroflexi	48.5	22.6	25.7	15.4	21.2	26.2	24.7	31.9	27.6	36.8	19.5	30.5	28.9	14.4	30.1	22.2	22.2	63.0	44.2	62.4	56.5
Chrysiogenetes	54.2	25.0	28.9	11.9	21.9	26.7	21.8	38.9	29.2	34.8	18.7	29.7	27.7	13.5	30.6	19.6	21.2	61.5	40.6	68.6	56.9
Cyanobacteria	43.1	26.6	30.2	26.3	27.7	22.0	22.4	21.9	22.1	32.9	17.3	20.5	23.6	18.5	30.1	31.3	26.7	54.9	39.7	42.4	45.7
Deferribacteres	44.3	34.8	34.8	30.3	33.3	14.2	18.4	16.0	16.2	32.4	14.3	18.4	21.7	18.5	32.4	35.4	28.8	46.6	32.7	34.3	37.9
Deinococcus	55.8	18.1	24.5	6.2	16.3	30.2	25.1	46.5	33.9	40.3	21.1	39.3	33.5	11.4	29.3	8.0	16.3	70.5	46.1	85.8	67.5
Dictyoglomi	40.3	35.1	34.3	36.1	35.2	13.3	17.2	9.9	13.5	31.5	14.8	15.2	20.5	20.1	33.7	38.7	30.9	44.7	32.0	25.2	34.0
Elusimicrobia	46.8	33.5	33.1	29.6	32.0	14.0	21.4	21.3	18.9	33.3	14.8	17.6	21.9	19.3	30.7	31.5	27.2	47.2	36.3	38.9	40.8
Fibrobacteres	50.9	23.8	27.2	11.7	20.9	24.8	25.8	38.3	29.6	36.6	18.6	32.7	29.3	14.8	28.5	17.4	20.2	61.4	44.4	70.9	58.9
Firmicutes	43.8	31.1	33.4	30.2	31.6	16.9	20.3	17.2	18.1	33.5	14.9	19.4	22.6	18.5	31.4	33.2	27.7	50.4	35.2	36.6	40.7
Fusobacteria	40.3	38.3	37.4	42.8	39.5	10.2	17.0	7.2	11.5	31.7	13.5	12.1	19.1	19.8	32.0	37.9	29.9	41.9	30.5	19.3	30.6
Gemmatimonadetes	51.8	19.6	22.9	8.8	17.1	27.3	28.4	38.8	31.5	40.2	20.7	38.4	33.1	12.8	28.0	14.0	18.3	67.5	49.1	77.2	64.6
Nitrospirae	44.8	25.6	28.1	19.5	24.4	23.6	23.3	27.8	24.9	34.5	17.9	28.6	27.0	16.3	30.7	24.1	23.7	58.1	41.2	56.4	51.9
Planctomycetes	49.7	20.8	26.1	11.8	19.6	27.3	26.3	39.4	31.0	37.8	19.9	32.7	30.1	14.1	27.7	16.1	19.3	65.1	46.2	72.1	61.1
Proteobacteria	50.5	24.1	27.9	15.0	22.3	24.4	24.0	33.6	27.3	36.5	18.4	30.9	28.6	15.1	29.7	20.5	21.7	60.9	42.4	64.5	55.9
Spirochaetes	44.1	30.9	32.3	26.9	30.0	17.6	20.7	20.5	19.6	32.1	15.6	20.5	22.7	19.4	31.5	32.1	27.7	49.7	36.2	41.0	42.3
Synergistetes	47.0	26.5	27.8	18.1	24.1	20.2	21.7	30.4	24.1	37.2	18.9	31.6	29.2	16.1	31.6	20.0	22.6	57.4	40.6	61.9	53.3
Tenericutes	41.9	37.7	38.5	38.9	38.4	12.5	18.1	10.4	13.7	27.4	12.0	9.2	16.2	22.3	31.4	41.5	31.8	39.9	30.0	19.6	29.8
Thermodesulfobacteria	42.5	31.0	33.2	33.4	32.5	17.3	19.0	14.9	17.1	32.3	14.9	14.6	20.6	19.4	32.8	37.2	29.8	49.6	33.9	29.4	37.7
Thermotogae	43.7	33.3	33.3	31.6	32.8	14.8	18.4	19.3	17.5	33.5	14.8	20.1	22.8	18.4	33.5	28.9	26.9	48.3	33.2	39.5	40.3
Correlation r		−0.805	−0.801	−0.911	−0.869	0.816	0.771	0.946	0.904	0.856	0.835	0.887	0.884	−0.930	−0.732	−0.926	−0.933	0.870	0.819	0.928	0.904

value $|r(PrRFC(X), Fr(\mathrm{T},2))| = 0.732$ with the nucleotide T in the codon site 2. However, two correlation asymmetries observed in bacterial genes could be significant and in contradiction with the self-complementarity of X, in particular, $r(PrRFC(X), Fr(\mathrm{C},3)) = 0.946 > r(PrRFC(X), Fr(\mathrm{G},1)) = 0.856$ and $|r(PrRFC(X), Fr(\mathrm{A},2))| = 0.801 > |r(PrRFC(X), Fr(\mathrm{T},2))| = 0.732$. This correlation asymmetry could be related to the well-known asymmetry between the circular codes $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ observed in prokaryotic genes (Bahi and Michel, 2008, Section 3.1.2) and eukaryotic genes (Arquès et al., 1997 both Fig. 2 and Section 2.2; Bahi and Michel, 2004, Section 1.2.2) or to other combinatorial or biological properties. It should be investigated in future.

4. Conclusion

The results in Fig. 1 demonstrate that today genes of archaea, bacteria, nuclear eukaryotes, viruses, mitochondrion, chloroplasts and bacterial plasmids contain genetic information for reading frame coding (compared to the random codes). To our knowledge, such a genetic property has never been identified in genes (see e.g. the review of Behura and Severson, 2013).

Genes of bacterial plasmids and bacteria have the highest efficiencies (about 49.0%) for reading frame coding, then, by decreasing values, genes of archaea (47.5%), viruses (45.4%) and nuclear eukaryotes (42.8%). The lowest reading frame coding efficiencies are observed in mitochondria and chloroplast genes (about 36.5%).

The lower reading frame coding efficiency of mitochondrial genes may be related to the fact that the circular code X has not been found in mitochondrial genes (Arquès and Michel, 1997). Two reasons may explain this mitochondrial case: (i) the very small sample of mitochondrial genes (1303 genes with 350,963 trinucleotides in 1997; 1164 genes with 217,899 trinucleotides only extracted from complete mitochondrial genomes in 2014); (ii) the coding process in mitochondrial genes known to have a great diversity in symmetric and asymmetric nucleotide exchanges during RNA transcription (Seligmann, 2013a,b; Michel and Seligmann, 2014).

The circular code X is found in various genes from two large kingdoms, the prokaryotes and the eukaryotes, i.e. the circular code X is a "universal" trinucleotide set occurring with a frequency higher than the random one in today genes. With the realistic hypothesis that today genes have been subjected to (mainly random) mutations then the law of large numbers asserts that the circular code X had a frequency in "primitive" genes, i.e. in the past, greater than in today genes. In other words, the 20 trinucleotides of the circular code X can be assumed to be the basic words of primitive genes (genes before mutations). Continuing the reasoning and according to my point of view, primitive life conditions could have selected initially the 20 trinucleotides of X among 64. Chemical selection of the 20 trinucleotides of X could have occurred at several levels: (i) during the concatenation of the three nucleotides l_0, l_1 and l_2 for forming a trinucleotide $l_0 \cdot l_1 \cdot l_2 \in X$; (ii) during the simplest (primitive) concatenation of a trinucleotide $l_0l_1l_2 \in X$ with itself, i.e. $l_0l_1l_2 \cdot l_0l_1l_2 \cdots = (l_0l_1l_2)^+$; (iii) up to a complex (independent or markov) mixing (a primitive soup) of trinucleotides of X leading, after mutations, to the observed circular code X in today genes; (iv) during the concatenation of a trinucleotide $l_0l_1l_2 \in X$ with its complementary trinucleotide $\mathcal{C}(l_0l_1l_2) \in X$ for forming the complementary and antiparallel double helix or for pairing with the tRNA and the 16S rRNA X motifs (Michel, 2012; El Soufi and Michel, 2014). It would be interesting to analyse the 20 trinucleotides of X or the 10 complementary pairs of trinucleotides of X with reagents, light, temperature, pH, etc., and to compare their chemical properties with the 44 remaining trinucleotides. Primitive life conditions could also have selected the 20 trinucleotides of X together with their amino acids they code. The circular code X codes 12 amino acids $AA = \{\mathrm{Ala}, \mathrm{Asn}, \mathrm{Asp}, \mathrm{Gln}, \mathrm{Glu}, \mathrm{Gly}, \mathrm{Ile}, \mathrm{Leu}, \mathrm{Phe}, \mathrm{Thr}, \mathrm{Tyr}, \mathrm{Val}\}$ (Arquès and Michel, 1996, Table 4a). As a consequence, the 12 amino acids AA could be more primitive compared to the eight remaining amino acids. Particular chemical properties of the 12 amino acids AA related to primitive life as well as their presence in other planets and moons in our solar system, exoplanets, meteorites, etc. could confirm or reject such an assumption. Furthermore, the genetic code which assigns trinucleotides with amino acids could also have been subjected to evolution from an ancestral coding related to the circular code X and the 12 amino acids AA to a modern coding associating the 61 trinucleotides to the 20 amino acids. The variability of the genetic code is a realistic hypothesis as mitochondrial genes have different codes (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes). In the yeast mitochondrial code (NCBI, code 3), two trinucleotides CTC and CTG of X have a different amino acid assignment. Indeed, CTC and CTG code Thr in the code 3 instead of Leu in the standard code. Furthermore, in the code 3, the circular code X does not code Leu and only codes for 11 amino acids. Thus, the maintenance of the circular code X in the yeast mitochondrial genes may cause an evolutionary disadvantage. The variability of mitochondrial codes could be a cause of the lack of the circular code X in mitochondrial genes. However, as mentioned above, the statistical data with a very small sample of mitochondrial genes may also be a reason. This biological question remains to be investigated.

The statistical and mathematical properties of the reading frame coding probability $PrRFC(X, K)$ of the C^3 self-complementary circular code X makes this parameter "unique" for analysing the coding property of genes. Indeed, the code X is the set of 20 trinucleotides having in average the highest occurrence in genes (reading frame) compared to the two shifted frames of both prokaryotes and eukaryotes and, in addition, the code X is a C^3 self-complementary circular code (Arquès and Michel, 1996). In order to have a statistical evaluation of the adequacy of the parameter $PrRFC(X, K)$ of the circular code X for its coding property, a computer program is developed for studying the RFC probability of random sets R of 20 trinucleotides in the chosen kingdom of bacteria B. The linear adjustment equation of 1000,000 random sets R shows that the estimated set $\hat{R}$ containing 20 trinucleotides of X has a RFC probability which is the value (to the first decimal place) of the RFC probability of X in bacteria B (Figs. 1 and 2).

The correlation between the reading frame coding probability $PrRFC(X, G)$ of the C^3 self-complementary circular code X and the frequency $Fr(n, s, G)$ of nucleotides n (A, C, G, T and GC) in the codon sites s (1, 2, 3 and {1,2,3}) of genes in the 25 groups G of the chosen kingdom of bacteria, has been also investigated. The number and frequency of nucleotides of X per trinucleotide site is given in Tables 4a and 4b. This property of the C^3 self-complementary circular code X is used for the first time here. The highest correlation 0.946 between the RFC probability of the circular code X and the nucleotide frequency is obtained with the nucleotide C in the codon site 3 of bacterial genes. It is a result expected from a theoretical property of X (Section 3.4). However, two correlation asymmetries in bacterial genes may be significant with the nucleotide C in the codon site 3 and the nucleotide G in the codon site 1 as well as with the nucleotide A in the codon site 2 and the nucleotide T in the codon site 2. This correlation asymmetry which could be related to the well-known asymmetry between the circular codes $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ observed in prokaryotic genes (Bahi and Michel, 2008, Section 3.1.2) and eukaryotic genes (Arquès et al., 1997 both Fig. 2 and Section 2.2; Bahi and Michel, 2004, Section 1.2.2) should be investigated in future.

The measure of reading frame coding of usage of the C^3 self-complementary circular code X is computed here on large gene populations in order to have reference values of usage of X. Obviously, this RFC parameter can be applied to individual genes or particular sets of genes having the same biological function within and among species. It can also be determined for particular regions of genes such as first exons, last exons, beginning of genes, end of genes, etc. Finally, it can be easily associated to any other classical genetic parameter, such as the nucleotide frequencies, the GC content, the length of genes, etc., and to any physical properties of genes, genomes and cells, such as temperature, pH, etc.

Finally, the reading frame coding measure can be a new approach to study gene evolution by coding.

Acknowledgment

I thank the two reviewers for their advice, and Denise Besch, Svetlana Gorchkova, Elisabeth Michel and Jean-Marc Vassards for their 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 circular code in the protein coding genes of mitochondria. J. Theor. Biol. 189, 273–290.
Arquès, D.G., Fallot, J.-P., Michel, C.J., 1997. An evolutionary model of a complementary circular code. J. Theor. Biol. 185, 241–253.
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.
Berstel, J., Perrin, D., 1985. Theory of Codes. Pure and Applied Mathematics. vol. 117. Academic Press, London, UK.
Behura, S.K., Severson, D.W., 2013. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol. Rev. 88, 49–61.
Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Nat. Acad. Sci. U.S.A. 43, 416–421.
El Soufi, K., Michel, C.J., 2014. Circular code motifs in the ribosome decoding center. Comput. Biol. Chem. 52, 9–17.
Fimmel, E., Giannerini, S., Gonzalez, D.L., Strüngmann, L., 2014. Circular codes, symmetries and transformations. J. Math. Biol. 10.1007/s00285-014-0806-7.
Gonzalez, D.L., Giannerini, S., Rosa, R., 2011. Circular codes revisited: a statistical approach. J. Theor. Biol. 275, 21–28.
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.
Lassez, J.-L., 1976. Circular codes and synchronization. Int. J. Comput. Inf. Sci. 5, 201–208.
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., Pirillo, G., 2010. Identification of all trinucleotide circular codes. Comput. Biol. Chem. 34, 122–125.
Michel, C.J., Pirillo, G., 2011. Strong trinucleotide circular codes. Int. J. Comb., 1–14 (Article ID 659567).
Michel, C.J., Pirillo, G, Pirillo, M.A., 2008. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theor. Comput. Sci. 401, 17–26.
Michel, C.J., Seligmann, H., 2014. Bijective transformation circular codes and nucleotide exchanging RNA transcription. Biosystems 118, 39–50.
Nirenberg, M.W., Matthaei, J.H., 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc. Nat. Acad. Sci. U.S.A. 47, 1588–1602.
Pirillo, G., 2003. A characterization for a set of trinucleotides to be a circular code. In: Pellegrini, C., Cerrai, P., Freguglia, P., Benci, V., Israel, G. (Eds.), Determinism, Holism, and Complexity. Kluwer, Boston, MA.
Seligmann, H., 2013a. Systematic asymmetric nucleotide exchanges produce human mitochondrial RNAs cryptically encoding for overlapping protein coding genes. J. Theor. Biol. 324, 1–20.
Seligmann, H., 2013b. Polymerization of non-complementary RNA: systematic symmetric nucleotide exchanges mainly involving uracil produce mitochondrial RNA transcripts coding for cryptic overlapping genes. BioSystems 111, 156–174.