Journal of Theoretical Biology 355 (2014) 83–94

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

# A genetic scale of reading frame coding

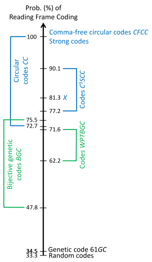
Christian J. Michel

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

## HIGHLIGHTS

- Determination of a genetic scale of reading frame coding.
- Trinucleotide circular codes.
- Bijective genetic codes.
- Trinucleotide codes of amino acids.

## GRAPHICAL ABSTRACT



## ARTICLE INFO

*Article history:*
Received 6 November 2013
Received in revised form
18 March 2014
Accepted 18 March 2014
Available online 31 March 2014

*Keywords:*
Comma free-code
Circular code
Bijective genetic code
Random code

## 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. A first objective is the definition of a new and simple statistical parameter PrRFC for analysing the probability (efficiency) of reading frame coding (RFC) of any trinucleotide code. A second objective is to reveal different classes and subclasses of trinucleotide codes involved in reading frame coding: the circular codes of 20 trinucleotides and the bijective genetic codes of 20 trinucleotides coding the 20 amino acids. This approach allows us to propose a genetic scale of reading frame coding which ranges from 1/3 with the random codes (RFC probability identical in the three frames) to 1 with the comma-free circular codes (RFC probability maximal in the reading frame and null in the two shifted frames). This genetic scale shows, in particular, the reading frame coding probabilities of the 12,964,440 circular codes (PrRFC = 83.2% in average), the 216 $C^3$ self-complementary circular codes (PrRFC = 84.1% in average) including the code $X$ identified in eukaryotic and prokaryotic genes (PrRFC = 81.3%) and the 339,738,624 bijective genetic codes (PrRFC = 61.5% in average) including the 52 codes without permuted trinucleotides (PrRFC = 66.0% in average). Otherwise, the reading frame coding probabilities of each trinucleotide code coding an amino acid with the universal genetic code are also determined. The four amino acids Gly, Lys, Phe and Pro are coded by codes (not circular) with RFC probabilities equal to 2/3, 1/2, 1/2 and 2/3, respectively. The amino acid Leu is coded by a circular code (not comma-free) with a RFC probability equal to 18/19. The 15 other amino acids are coded by comma-free circular codes, i.e. with RFC probabilities equal to 1. The identification of coding properties in some classes of trinucleotide codes studied here may bring new insights in the origin and evolution of the genetic code.

## 1. Introduction

The reading frame coding (RFC) of a code (set) of trinucleotides, e.g. the genetic code, is a fascinating and open problem. It is also an old problem. Almost 60 years ago (in 1957), before the discovery of

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

http://dx.doi.org/10.1016/j.jtbi.2014.03.029

the genetic code, a class of trinucleotide codes, called comma-free codes (or codes without commas) was proposed by Crick et al. (1957) for explaining how the reading of a series of trinucleotides could code amino acids. The two questions of interest were: why are there more trinucleotides than amino acids and, how does one choose the reading frame? Crick et al. (1957) proposed that only 20 trinucleotides among 64 code the 20 amino acids. Such a bijective code implies that the coding trinucleotides are found only in one frame. The determination of a set of 20 trinucleotides forming a comma-free code has several constraints:

(i) A periodic permuted trinucleotide, i.e. a trinucleotide with identical nucleotides, must be excluded from such a code. Indeed, the concatenation of AAA with itself, for instance, does not allow the (original) reading frame to be retrieved as there are three possible decompositions: …AAA · AAA · AAA… (original frame), …A · AAA · AAA · AA…, and …AA · AAA · AAA · A…, the concatenation operator " · " showing the adopted decomposition.
(ii) Two non-periodic permuted trinucleotides, i.e. two trinucleotides related to the circular permutation map, e.g. ACG and CGA, must also be excluded from such a code. Indeed, the concatenation of ACG with itself, for instance, does not allow the (original) reading frame to be retrieved as there are two possible decompositions: …ACG · ACG · ACG… (original frame) and …A · CGA · CGA · CG…

Therefore, 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 identical to the amino acid number, thus leading to a code assigning one trinucleotide per amino acid without ambiguity. A few combinatorial results on trinucleotide comma-free codes were obtained by Golomb et al. (1958a, 1958b). However, no trinucleotide comma-free code was identified in genes statistically. Furthermore, in the late 1950s, 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 a comma-free code over the alphabet {A, C, G, T}. For several biological reasons, in particular the interaction between mRNA and tRNA, this concept was again taken up later over the purine/pyrimidine alphabet $\{R, Y\}$ ($R = \{A, G\}$, $Y = \{C, T\}$) with two trinucleotide comma-free codes for primitive genes: RRY (Crick et al., 1976) and RNY ($N = \{R, Y\}$) (Eigen and Schuster, 1978).

In 1996, a statistical analysis of occurrence frequencies of the 64 trinucleotides {AAA, …, TTT} in the three frames 0, 1 and 2 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 {AAA, CCC, GGG, TTT} 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): eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686 sequences, 4,709,758 trinucleotides) (Arquès and Michel, 1996). This set $X$ contains the 20 following trinucleotides:

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

The two sets $X_1$ and $X_2$, of 20 trinucleotides each, in the shifted frames 1 and 2 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 of the circular code $X$ (words, not necessarily unique, in two or three frames), this window needs a length of 13 nucleotides with $X$ (Fig. 3 in Michel, 2012). A window of 13 nucleotide length is the largest window of $X$, i.e. it allows to retrieve the reading frame for all the ambiguous words of $X$. Gonzalez et al. (2011), by defining a statistical function analysing the covering capability of a circular code, have recently showed on a gene data set from 13 classes of proteins that the 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., 2008a). A review of this 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 decoding center which recognizes the codon–anticodon helix in A-tRNA (Michel, 2012). A 3D visualization of $X$ motifs in the ribosome (crystal structure 3I8G, Jenner et al., 2010) shows several spatial configurations involving mRNA $X$ motifs, A-tRNA and E-tRNA $X$ motifs, and four 16S rRNA $X$ motifs. These results led to the concept of a possible translation (framing) code based on circular code (Michel, 2012).

Comma-free and circular codes have two different definitions in combinatorics (Definitions 5 and 10 below). In fact, these two classes of codes should not be considered as different. Indeed, it was proved recently that a comma-free code is a particular circular code (Proposition 3 in Michel et al., 2008a). Precisely, a hierarchy of circular codes is closed by the strongest ones which are comma-free and the weakest ones which are circular with large "necklaces" (Proposition 4 and Remark 4 in Michel et al., 2008a). Furthermore, it exists as circular codes even stronger than the comma-free codes and called strong circular codes (Michel and Pirillo, 2011). There are 12,964,440 (maximal, i.e. of 20 trinucleotide length) trinucleotide circular codes (Table 2(d) in Arquès and Michel, 1996; growth function in Table 1 in Michel and Pirillo, 2010) which include the 408 comma-free codes (growth function in Table 2a and list in Table 2b in Michel et al., 2008b).

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 5*LDCN* (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. Comma-free codes have their trinucleotides only in reading frame, thus short necklaces, while circular codes (not comma-free) have their trinucleotides in reading frame but also in the two shifted frames 1 and 2, thus large necklaces (see the most general hierarchy given in Proposition 4.1 in Michel and Pirillo, 2011).

The first objective of this study is to define a new and simple statistical parameter PrRFC for analysing the probability (efficiency) of reading frame coding (RFC) of any trinucleotide code $C$.

The second objective is to reveal different classes of trinucleotide codes $C$ involved in reading frame coding, almost all of them have never been studied. These trinucleotide codes are important from a fundamental point of view, e.g. for understanding the properties of the genetic code and its origin and evolution. From a practical point of view, the motifs generated by such codes have important genetic properties, e.g. reading frame retrieval, complementary, permutation, etc., and may be involved in different biological processes, e.g. translation (Michel, 2012, 2013). Two main classes of trinucleotide codes $C$ are analysed here:

(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 the complementarity and circular permutation maps. It includes the $C^3$ self-complementary circular codes. All these codes $CC$ are issued from our combinatorial work during the last 20 years.

(ii) The 339,738,624 bijective genetic codes $BGC$ of 20 trinucleotides coding the 20 amino acids. These codes $BGC$ are not circular as no circular code among the 12,964,440 ones codes 20 amino acids (see Introduction in Michel and Pirillo, 2013). Several subclasses are identified here. To our knowledge, these codes $BGC$ have been never studied. They have a great number of combinatorial and biological properties which remain to be discovered.

The approach developed here leads to the determination of a genetic scale of reading frame coding for various classes of trinucleotide codes, including the codes $CC$ and $BGC$ and their subclasses.

Finally, this method allows the determination of the reading frame coding probabilities of each trinucleotide code coding an amino acid with the universal genetic code.

## 2. Method

### 2.1. Definitions of codes

We briefly recall a few classical definitions of codes in order to understand the different classes of codes studied for the property of reading frame coding.

**Notation 1.** The letters (or nucleotides or bases) define the genetic alphabet $A_4 = \{A, C, G, T\}$. The set of non-empty words (words resp.) over $A_4$ is denoted by $A_4^+$ ($A_4^*$ resp.). The set of the 64 words of length 3 (trinucleotides or triletters) on $A_4$ is denoted by $A_4^3 = \{AAA, \ldots, TTT\}$. 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.** By convention here, the reading frame established by a start codon {ATG, GTG, TTG} is the frame 0, and the frames 1 and 2 are the reading frame 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 $\mathcal{C}(A) = T$, $\mathcal{C}(C) = G$, $\mathcal{C}(G) = C$, $\mathcal{C}(T) = A$. 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. $\mathcal{C}(ACG) = CGT$. By extension to a trinucleotide set $S$, the *set complementarity map* $\mathcal{C} : S \to S$ is defined by $\mathcal{C}(S) = \{v | \ u, v \in A_4^3, u \in S, v = \mathcal{C}(u)\}$, i.e. a complementary trinucleotide set $\mathcal{C}(S)$ is obtained by applying the complementarity map $\mathcal{C}$ to all its trinucleotides, e.g. $\mathcal{C}(\{ACG, AGT\}) = \{ACT, CGT\}$.

**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. $\mathcal{P}(ACG) = CGA$. The 2nd iterate of $\mathcal{P}$ is denoted $\mathcal{P}^2$, e.g. $\mathcal{P}^2(ACG) = GAC$. By extension to a trinucleotide set $S$, the *set circular permutation map* $\mathcal{P} : S \to S$ is defined by $\mathcal{P}(S) = \{v | \ u, v \in A_4^3, u \in S, v = \mathcal{P}(u)\}$, 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 $\mathcal{P}^2(\{ACG, AGT\}) = \{GAC, TAG\}$.

**Remark 1.** There are two classes of permuted trinucleotides: periodic permuted trinucleotides $PPT$ such that $t = \mathcal{P}(t) = \mathcal{P}^2(t)$ for a trinucleotide $t \in A_4^3$, i.e. $PPT = \{AAA, CCC, GGG, TTT\}$, and non-periodic permuted trinucleotides $NPPT = \{t, \mathcal{P}(t), \mathcal{P}^2(t)\}$ for a trinucleotide $t \in A_4^3 \setminus PPT$.

**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 $s x_2 \cdots x_n r = y_1 \cdots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word) and $x_i = y_i$ for $i = 1, \ldots, n$.

**Remark 2.** A trinucleotide code $C$ containing either one periodic permuted trinucleotide $PPT$ or two non-periodic permuted trinucleotides $NPPT$, i.e. $\{t, \mathcal{P}(t)\}$ or $\{t, \mathcal{P}^2(t)\}$, cannot be circular (details, e.g., in Michel, 2008). Obviously, a trinucleotide code $C$ containing three non-periodic permuted trinucleotides $NPPT$, i.e. $\{t, \mathcal{P}(t), \mathcal{P}^2(t)\}$, is not circular. Thus, the two trinucleotide codes $A_4^3$ and $A_4^3 \setminus PPT$ are not circular.

**Remark 3.** 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 (see Fig. 1 for a graphical representation of the circular code definition and Fig. 2 for an example in Michel, 2012). The reading frame in a sequence (a 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 (see Fig. 3 for an example in Michel, 2012).

**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 permuted trinucleotide sets $CC_1 = \mathcal{P}(CC)$ and $CC_2 = \mathcal{P}^2(CC)$ are trinucleotide circular codes.

**Remark 4.** A trinucleotide circular code $CC \subset A_4^3$ does not necessarily imply that $CC_1 = \mathcal{P}(CC)$ and $CC_2 = \mathcal{P}^2(CC)$ are also 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).

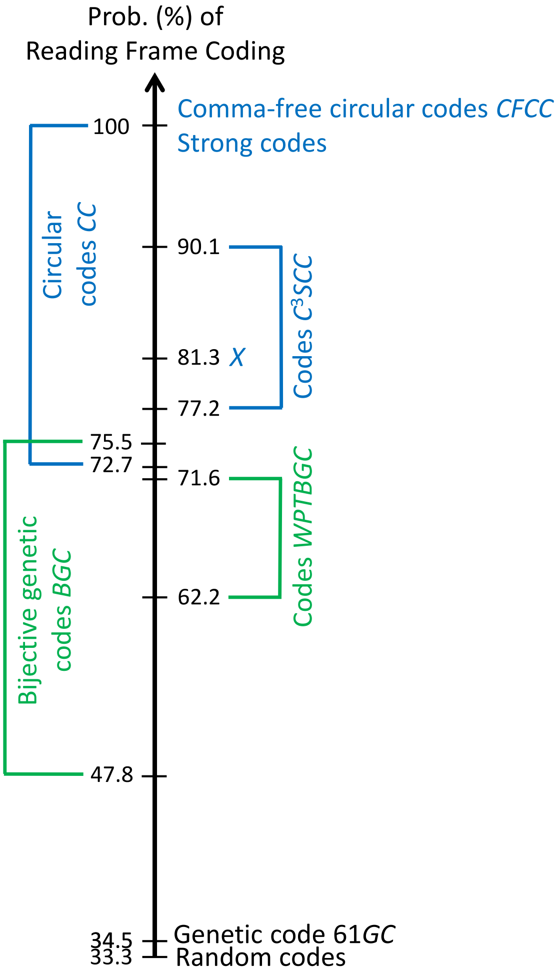


**Fig. 1.** A genetic scale of reading frame coding of trinucleotide codes. The RFC probabilities of reading frame coding (Eq. (6)) of the two main classes of trinucleotide codes are represented: (i) the 12,964,440 circular codes $CC$, where the reading frame is always retrieved with more or less efficiency, 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 $BGC$, where the reading frame is not always retrieved with more or less efficiency, including the 52 bijective genetic codes $WPTBGC$ without permuted trinucleotides ($WPT$), i.e. without the periodic permuted trinucleotides $PPT=\{\text{AAA}, \text{CCC}, \text{GGG}, \text{TTT}\}$ and without the non-periodic permuted trinucleotides $NPPT=\{t, \mathcal{P}(t), \mathcal{P}^2(t)\}$ (details in Remarks 1 and 2). The RFC probability scale ranges from 1/3 with the random codes (one chance out of three to retrieve the reading frame among the three possible frames in genes) to 1 with the comma-free codes and the strong codes (the reading frame is always retrieved). There is a non-empty RFC probability intersection of 2.8% between the circular codes $CC$ and the bijective genetic codes $BGC$: $\text{PrRFC}(CC \cap BGC) \in [72.7, 75.5]$ (%) (Table 3a). The $C^3$ self-complementary circular code $X$ (1) identified in eukaryotic and prokaryotic genes has a RFC probability equal to 81.3%.

**Definition 9.** A trinucleotide circular code $CC \subset A_4^3$ is maximal if for each $y \in A_4^3$, $y \notin CC$, $CC \cup \{y\}$ is not a trinucleotide circular code.

Any trinucleotide circular code $CC$ with 20 trinucleotides is maximal (proof obvious, details in Lemma 1 and Remark 3 in Michel et al., 2012). In other words, any code containing more than 20 trinucleotides cannot be circular. This number 20 operates with the maximum number of trinucleotides in a circular code and the number of amino acids coded by the genetic code.

The trinucleotide set $X = X_0$ (1) coding the reading frames (frames 0) in eukaryotic and prokaryotic genes is a maximal $C^3$ self-complementary trinucleotide 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:

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

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

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

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 a trinucleotide code

#### 2.2.1. Definition

A simple quantitative definition is proposed for measuring the probability $\text{PrRFC}(C)$ of reading frame coding (RFC) for any code (set) $C$ of trinucleotides $t$ with assigned probabilities. Remember that there are $2^{64}-1 \approx 1.8 \times 10^{19}$ trinucleotide codes from length 1 (64 codes at one trinucleotide) to 64 (one code at 64 trinucleotides). The RFC probability $\text{PrRFC}(C)$ 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. It is important to stress that the proposed definition satisfies the combinatorial properties of trinucleotides codes. In particular, the RFC probability is equal to 1 with the comma-free codes where the trinucleotides are only in the reading frame 0, i.e. the reading frame 0 is coded maximally, and the RFC probability is equal to 1/3 with the random codes where the trinucleotides are found in the three frames 0, 1 and 2 equiprobably, i.e. the reading frame 0 is "not coded".

Let $t = l_0l_1l_2$, $l_0, l_1, l_2 \in A_4$, be a trinucleotide of a code $C$. Each trinucleotide $t$ of $C$ is assigned to a probability $\Pr(t)$ such that the probability $\Pr(C)$ of the code $C$ is equal to

$$\Pr(C) = \sum_{t \in C} \Pr(t) = 1. \tag{2}$$

For example, the reading frame coding measure of a code $C$ of 20 trinucleotides can take the simplest hypothesis that each trinucleotide $t$ of $C$ occurs with the same probability, i.e. $\Pr(t) = 1/20$ for all $t \in C$. The method proposed here can be applied to any trinucleotide probabilities $\Pr(t)$ as long as its sum is equal to 1 for satisfying Eq. (2).

By convention, the reading frame $f = 0$ is established by the letter $l_0$ of $t = l_0l_1l_2$, $l_0, l_1, l_2 \in A_4$. The frames $f = 1$ and 2 start with the letters $l_1$ and $l_2$ of $t$, respectively. Let the di-trinucleotide $w$ be a concatenation of two trinucleotides $t' = l'_0l'_1l'_2 \in C$ and $t'' = l''_0l''_1l''_2 \in C$, i.e. $w = t't'' \in C^2$. We denote by $t_0(w) = l'_0l'_1l'_2 \in A_4^3$, $t_1(w) = l'_1l'_2l''_0 \in A_4^3$ and $t_2(w) = l'_2l''_0l''_1 \in A_4^3$ the trinucleotides in frames 0, 1 and 2, respectively, of a di-trinucleotide $w \in C^2$. The concatenation of the two trinucleotides $t' \in C$ and $t'' \in C$ may yield a trinucleotide $t_f(w)$ in a shifted frame $f \in \{1,2\}$ belonging to $C$. For example, with the code $C = \{\text{AAA}, \text{AAC}\}$, the concatenation of the trinucleotides $t' = \text{AAA} \in C$ and $t'' = \text{AAC} \in C$, i.e. $w = \text{AAAAAC}$, leads to the trinucleotides $t_1(w) = \text{AAA} \in C$ and $t_2(w) = \text{AAA} \in C$, thus AAA occurs in frame 0, obviously as it belongs to $C$, but also in frames 1 and 2.

The probability of a trinucleotide $t$ to occur in a shifted frame is first determined for a given di-trinucleotide. The probability $\Pr(t_f(w), f)$ of a trinucleotide $t_f(w) \in A_4^3$ in a frame $f \in \{1,2\}$ of a di-trinucleotide $w = t't'' \in C^2$ is equal to the product of probabilities $\Pr(t')$ and $\Pr(t'')$ (with the simplest hypothesis of independent events)

$$\Pr(t_f(w), f) = \Pr(t') \times \Pr(t''). \tag{3}$$

Then, the probability $\text{PrFrame}(t,f)$ of a trinucleotide $t \in C$ in a frame $f \in \{1,2\}$ in all the di-trinucleotides $w = t't'' \in C^2$ is equal

$$\text{PrFrame}(t,f) = \sum_{w \in C^2 \mid t = t_f(w) \in C} \Pr(t_f(w),f). \tag{4}$$

Then, the probability $\text{PrFrame}(C,f)$ of a code $C$ in a frame $f \in \{1,2\}$ is equal

$$\text{PrFrame}(C,f) = \sum_{t \in C} \text{PrFrame}(t,f). \tag{5}$$

Finally, the reading frame coding probability $\text{PrRFC}(C)$ (efficiency) of a code $C$ is equal to

$$\begin{aligned}\text{PrRFC}(C) &= \frac{\Pr(C)}{\Pr(C) + \sum_{f=1}^{2} \text{PrFrame}(C,f)} \\ &= \frac{1}{1 + \sum_{f=1}^{2} \text{PrFrame}(C,f)}.\end{aligned} \tag{6}$$

**Property 1.** $\frac{1}{3} \le \text{PrRFC}(C) \le 1$ according to Eq. (6).

The more the RFC probability $\text{PrRFC}(C)$ value is raised, the more the code $C$ has efficiency for coding the reading frame 0.

For the bound 1, the RCF probability can, in addition to the property of coding measure, be associated to the property of reading frame retrieval. If $\text{PrRFC}(C) = 1$, i.e. $\text{PrFrame}(C,1) = \text{PrFrame}(C,2) = 0$, then the code $C$ only codes the reading frame 0 and always retrieves the reading frame.

If $\text{PrRFC}(C) = 1/3$, i.e. $\Pr(C) = \text{PrFrame}(C,1) = \text{PrFrame}(C,2) = 1$, then the code $C$ codes the three frames 0, 1 and 2 with the same efficiency and the reading frame 0 is retrieved randomly.

This RFC probability (Eq. (6)) can very easily be programmed in a computer language or in a spreadsheet.

#### 2.2.2. Explained examples

(i) The code $C = \{\text{AAA}\}$ which is a periodic permuted trinucleotide $PPT$ (see Remark 1) (with $\Pr(\text{AAA}) = 1$), has a reading frame coding probability $\text{PrRFC}(C) = 1/3$ (detailed calculus in Table 1a).

(ii) The code $C = \{\text{AAC}\}$ (with $\Pr(\text{AAC}) = 1$) has a RFC probability $\text{PrRFC}(C) = 1$ (detailed calculus in Table 1b).

(iii) The code $C = \{\text{AAA}, \text{AAC}, \text{ACA}\}$ with, for example, the following probabilities $\Pr(\text{AAA}) = 1/6$, $\Pr(\text{AAC}) = 1/3$ and $\Pr(\text{ACA}) = 1/2$, has a RFC probability $\text{PrRFC}(C) = 6/13$ (detailed calculus in Table 1c).

(iv) The code $C = PPT = \{\text{AAA}, \text{CCC}, \text{GGG}, \text{TTT}\}$ (see Remark 1) with equiprobability, i.e. $\Pr(\text{AAA}) = \Pr(\text{CCC}) = \Pr(\text{GGG}) = \Pr(\text{TTT}) = 1/4$, has a RFC probability $\text{PrRFC}(C) = 2/3$.

(v) The genetic code $61GC$ with 61 equiprobable codons without the three stop codons {TAA, TAG, TGA} has a RFC probability $\text{PrRFC}(61GC) = 3721/10,779 \approx 34.5\%$.

(vi) The $C^3$ self-complementary circular code $X$ has a RFC probability $\text{PrRFC}(X) = 100/123 \approx 81.3\%$. Indeed, by considering the 400 pairs $X^2$ of trinucleotides of $X$, the number of trinucleotides belonging to $X$, $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ in the frames 1 and 2 of $X^2$ are:

(via) In frame 1: 11.50% of trinucleotides belong to $X$, 72.50% to $X_1$, 12.25% to $X_2$, and 3.75% to $PPT = \{\text{AAA}, \text{CCC}, \text{GGG}, \text{TTT}\}$ with no GGG (1.25% of AAA, 1.50% of CCC, 1.00% of TTT) (by a computer analysis done here). Note that $11.50 + 12.25 + 3.75 = 27.5\%$ (Section 3.7 and Fig. 3 in Arquès and Michel, 1996). If the four trinucleotides $PPT$ are not considered in frame 1 then 11.95% of trinucleotides belong to $X$, 75.32% to $X_1$, and 12.73% to $X_2$ (Section 3.3 in Arquès et al., 1997).

(vib) In frame 2: 11.50% of trinucleotides belong to $X$, 12.25% to $X_1$, 72.50% to $X_2$, and 3.75% to $PPT = \{\text{AAA}, \text{CCC}, \text{GGG}, \text{TTT}\}$ with no CCC (1.00% of AAA, 1.50% of GGG, 1.25% of TTT). These numbers in frame 2 can be deduced from the numbers in frame 1 by self-complementarity of $X$. Thus, $\text{PrRFC}(X) = 1/(1 + 2 \times 11.5/100) = 100/123$.

### 2.3. Classes of trinucleotide codes analysed

Two main classes of trinucleotide codes are studied for their properties of reading frame coding.

#### 2.3.1. Trinucleotide circular codes

There are 12,964,440 (maximal) trinucleotide circular codes $CC$ (Table 2(d) in Arquès and Michel, 1996; growth function in Table 1 in Michel and Pirillo, 2010).

These circular codes $CC$ can be divided into seven necklace subclasses according to a hierarchy based on the necklace which is a combinatorial measure of the nucleotide length for retrieving the reading frame. These seven necklace subclasses, given in inclusion sets (Proposition 8 and, Tables 2 and 3 in Michel et al., 2012) are explicitly identified here (number and list of trinucleotides codes). The notation used here is based on the notation in Michel et al. (2012).

(i) 408 comma-free circular codes $CFCC$ ($I^2$ in Table 3 in Michel et al., 2012);

(ii) 2352 circular codes $J^3CCC$ ($J^3C = I^3C \setminus I^2$ where $I^3C$ ($C$ meaning Continued) and $I^2$ are defined in Table 3 in Michel et al., 2012);

(iii) 294,312 circular codes $J^3CC$ ($J^3 = I^3 \setminus I^3C$);

(iv) 252,960 circular codes $J^4CCC$ ($J^4C = I^4C \setminus I^3$);

(v) 4,566,696 circular codes $J^4CC$ ($J^4 = I^4 \setminus I^4C$);

(vi) 823,920 circular codes $J^5CCC$ ($J^5C = I^5C \setminus I^4$);

(vii) 7,023,792 circular codes $J^5CC$ ($J^5 = I^5 \setminus I^5C$).

The seven necklace subclasses of circular codes $CC$ are pairwise disjoint with a sum equal to 12,964,440. They are ranged from the

**Table 1a**
Example: the trinucleotide code $C = \{\text{AAA}\}$ with $\Pr(\text{AAA}) = 1$ has a probability $\text{PrRFC}(C)$ of reading frame coding (Eq. (6)) equal to 1/3 (in bold).

| | | | | | | | Frame $f=0$ | Frame $f=0$ | Frame $f=1$ | Frame $f=2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | $t'$ | | | $t''$ | | | $\Pr(t')$ | $\Pr(t'')$ | $\Pr(t_1(w),f) = \Pr(t') \times \Pr(t'')$ Eq. (3) | $\Pr(t_2(w),f) = \Pr(t') \times \Pr(t'')$ Eq. (3) |
| $w$ | A | A | A | A | A | A | 1 | 1 | | |
| | | A | A | A | | | | | 1 ($t_1(w) = \text{AAA} \in C$) | |
| | | | A | A | A | | | | | 1 ($t_2(w) = \text{AAA} \in C$) |

| | Frame $f=0$ | Frame $f=1$ | Frame $f=2$ | | |
|---|---|---|---|---|---|
| $t \in C$ | $\Pr(t)$ | $\text{PrFrame}(t,1)$ Eq. (4) | $\text{PrFrame}(t,2)$ Eq. (4) | | |
| AAA | 1 | 1 | 1 | | |
| $\text{PrFrame}(C,f)$ Eq. (5) | 1 | 1 | 1 | $1/(1+1+1) =$ **1/3** | $\text{PrRFC}(C)$ Eq. (6) |

**Table 1b**
Example: the trinucleotide code $C=\{\text{AAC}\}$ with $\Pr(\text{AAC})=1$ has a probability $\text{PrRFC}(C)$ of reading frame coding (Eq. (6)) equal to 1 (in bold).

| | | | | | | | Frame $f=0$ | Frame $f=0$ | Frame $f=1$ | Frame $f=2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | t' | | | t'' | | | $\Pr(t')$ | $Pr(t'')$ | $\Pr(t_1(w),f)=Pr(t')\times Pr(t'')$ Eq. (3) | $\Pr(t_2(w),f)=Pr(t')\times Pr(t'')$ Eq. (3) |
| $w$ | A | A | C | A | A | C | 1 | 1 | | |
| | | A | C | A | | | | | 1 ($t_1(w)=\text{ACA}\notin C$) | |
| | | | C | A | A | | | | | 1 ($t_2(w)=\text{CAA}\notin C$) |

| | Frame $f=0$ | Frame $f=1$ | Frame $f=2$ | | |
|---|---|---|---|---|---|
| $t\in C$ | $\Pr(t)$ | $\text{PrFrame}(t,1)$ Eq. (4) | $\text{PrFrame}(t,2)$ Eq. (4) | | |
| AAC | 1 | 0 | 0 | | |
| $\text{PrFrame}(C,f)$ Eq. (5) | 1 | 0 | 0 | $1/(1+0+0)$ = **1** | $\text{PrRFC}(C)$ Eq. (6) |

**Table 1c**
Example: the trinucleotide code $C=\{\text{AAA, AAC, ACA}\}$ with $\Pr(\text{AAA})=1/6$, $\Pr(\text{AAC})=1/3$ and $\Pr(\text{ACA})=1/2$ has a probability $\text{PrRFC}(C)$ of reading frame coding (Eq. (6)) equal to 6/13 (in bold).

| | | | | | | | Frame $f=0$ | Frame $f=0$ | Frame $f=1$ | Frame $f=2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | t' | | | t'' | | | $\Pr(t')$ | $\Pr(t'')$ | $\Pr(t_1(w),f)=\Pr(t')\times\Pr(t'')$ Eq. (3) | $\Pr(t_2(w),f)=\Pr(t')\times\Pr(t'')$ Eq. (3) |
| $w$ | A | A | A | A | A | A | 1/6 | 1/6 | | |
| | | A | A | A | | | | | 1/36 ($t_1(w)=\text{AAA}\in C$) | |
| | | | A | A | A | | | | | 1/36 ($t_2(w)=\text{AAA}\in C$) |
| $w$ | A | A | A | A | A | C | 1/6 | 1/3 | | |
| | | A | A | A | | | | | 1/18 ($t_1(w)=\text{AAA}\in C$) | |
| | | | A | A | A | | | | | 1/18 ($t_2(w)=\text{AAA}\in C$) |
| $w$ | A | A | A | A | C | A | 1/6 | 1/2 | | |
| | | A | A | A | | | | | 1/12 ($t_1(w)=\text{AAA}\in C$) | |
| | | | A | A | C | | | | | 1/12 ($t_2(w)=\text{AAC}\in C$) |
| $w$ | A | A | C | A | A | A | 1/3 | 1/6 | | |
| | | A | C | A | | | | | 1/18 ($t_1(w)=\text{ACA}\in C$) | |
| | | | C | A | A | | | | | 1/18 ($t_2(w)=\text{CAA}\notin C$) |
| $w$ | A | A | C | A | A | C | 1/3 | 1/3 | | |
| | | A | C | A | | | | | 1/9 ($t_1(w)=\text{ACA}\in C$) | |
| | | | C | A | A | | | | | 1/9 ($t_2(w)=\text{CAA}\notin C$) |
| $w$ | A | A | C | A | C | A | 1/3 | 1/2 | | |
| | | A | C | A | | | | | 1/6 ($t_1(w)=\text{ACA}\in C$) | |
| | | | C | A | C | | | | | 1/6 ($t_2(w)=\text{CAC}\notin C$) |
| $w$ | A | C | A | A | A | A | 1/2 | 1/6 | | |
| | | C | A | A | | | | | 1/12 ($t_1(w)=\text{CAA}\notin C$) | |
| | | | A | A | A | | | | | 1/12 ($t_2(w)=\text{AAA}\in C$) |
| $w$ | A | C | A | A | A | C | 1/2 | 1/3 | | |
| | | C | A | A | | | | | 1/6 ($t_1(w)=\text{CAA}\notin C$) | |
| | | | A | A | A | | | | | 1/6 ($t_2(w)=\text{AAA}\in C$) |
| $w$ | A | C | A | A | C | A | 1/2 | 1/2 | | |
| | | C | A | A | | | | | 1/4 ($t_1(w)=\text{CAA}\notin C$) | |
| | | | A | A | C | | | | | 1/4 ($t_2(w)=\text{AAC}\in C$) |

| | Frame $f=0$ | Frame $f=1$ | Frame $f=2$ | | |
|---|---|---|---|---|---|
| $t\in C$ | $\Pr(t)$ | $\text{PrFrame}(t,1)$ Eq. (4) | $\text{PrFrame}(t,2)$ Eq. (4) | | |
| AAA | 1/6 | $1/36+1/18+1/12=1/6$ | $1/36+1/18+1/12+1/6=1/3$ | | |
| AAC | 1/3 | 0 | $1/12+1/4=1/3$ | | |
| ACA | 1/2 | $1/18+1/9+1/6=1/3$ | 0 | | |
| $\text{PrFrame}(C,f)$ Eq. (5) | 1 | 1/2 | 2/3 | $1/(1+1/2+2/3)=$ **6/13** | $\text{PrRFC}(C)$ Eq. (6) |

shortest necklace with the codes *CFCC* to the largest necklace with the codes $J^5CC$.

The class of the 408 comma-free circular codes *CFCC* is known (Golomb et al., 1958a, 1958b; growth function in Table 2a and list in Table 2b in Michel et al., 2008b). Its classical definition is recalled in order to stress two remarks.

**Definition 10.** A trinucleotide code $C\subset A_4^3$ is *comma-free* if for each $x\in C$ and $u,v\in A_4^*$ such that $u\cdot x\cdot v=y_1\cdots y_n$ with $y_1,\ldots,y_n\in C$, $n\geq 1$, it holds that $u,v\in C^*$.

Definition 10 of a trinucleotide comma-free code differs from Definition 5 of a trinucleotide circular code. However, a comma-free code is always circular and belongs to the class of circular codes *CC* (Tables 2 and 3 in Michel et al., 2012). Thus, it is called here comma-free circular codes (*CFCC*). Furthermore, Definition 10 implies that all the trinucleotides of a comma-free circular code *CFCC* are always in reading frame. Thus, the reading frame coding probability (Eq. (6)) is also a combinatorial proof to demonstrate that a trinucleotide code is comma-free.

**Proposition 1.** *The following conditions are equivalent*:

(i) *the trinucleotide code C is comma-free*;
(ii) *the trinucleotide code C has a RFC probability* $\text{PrRFC}(C)=1$ (Eq. (6)).

The circular codes *CC* can also be divided in three map subclasses according to the complementarity and circular permutation maps (Definitions 1 and 2):

(i) 221,328 $C^3$ circular codes $C^3CC$;
(ii) 312 self-complementary circular codes $SCC$;
(iii) 216 $C^3$ self-complementary circular codes $C^3SCC$ (Definition 8; Table 2(d) in Arquès and Michel, 1996) including the code $X$ (1) identified in eukaryotic and prokaryotic genes.

The three map subclasses of circular codes $CC$ are pairwise disjoint.

The two subclasses of circular codes $C^3CC$ and $SCC$ are newly studied here.

The reading frame coding probability PrRFC($CC$) of a circular code $CC$ is determined by applying Eq. (6) where each trinucleotide $t$ of $CC$ is assigned to a probability $\Pr(t) = 1/20$ (by taking the simplest hypothesis of equiprobability in order to have a reference value for the reading frame coding).

### 2.3.2. Bijective genetic codes

In the universal genetic code, 61 codons code 20 amino acids as there are three stop codons. As two amino acids are encoded by a single codon, nine amino acids, by two codons, one amino acid, by three codons, five amino acids, by four codons and three amino acids, by six codons, there are $2^9 \times 3 \times 4^5 \times 6^3 = 339,738,624$ bijective genetic codes $BGC$ of 20 codons coding the 20 amino acids, i.e. with a bijective map. To our knowledge, these codes $BGC$, a subset of all possible bijective codes $64!/(44!20!) \approx 2 \times 10^{16}$ were never studied.

There is no self-complementary bijective genetic code $BGC$, i.e. a code $BGC$ such that 10 trinucleotides are complementary to the 10 other trinucleotides (by a computer analysis done here).

There is no bijective genetic code $BGC$ which is circular as none circular code among the 12,964,440 ones codes 20 amino acids (detailed in Introduction in Michel and Pirillo, 2013).

**Table 2**
The 52 bijective genetic codes $WPTBGC$ without permuted trinucleotides ($WPT$), i.e. without the periodic permuted trinucleotides $PPT = \{\text{AAA}, \text{CCC}, \text{GGG}, \text{TTT}\}$ and without the non-periodic permuted trinucleotides $NPPT = \{t, \mathcal{P}(t), \mathcal{P}^2(t)\}$ (details in Remarks 1 and 2), coding the 20 amino acids. They are not circular codes but they have the necessary condition to be circular codes.

| |
|---|
| {AAG,AAT,ACT,ATC,ATG,CAA,CAC,CCG,CGT,GAC,GAG,GCA,GGC,GTA,TAT,TCC,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACT,AGT,ATC,ATG,CAA,CAC,CCG,CGT,CTC,GAC,GAG,GCA,GGC,GTT,TAT,TGC,TGG,TTC} |
| {AAG,AAT,ACT,ATC,ATG,CAA,CAC,CCT,CGC,GAC,GAG,GCA,GGC,GTA,TAT,TCG,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACT,AGT,ATC,ATG,CAA,CAC,CCT,CGC,GAC,GAG,GCA,GGC,GTC,TAT,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACC,ATG,ATT,CAA,CAT,CCG,CGT,GAC,GAG,GCA,GGC,GTA,TAC,TCC,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACC,AGT,ATG,ATT,CAA,CAT,CCG,CGT,CTC,GAC,GAG,GCA,GGC,GTT,TAC,TGC,TGG,TTC} |
| {AAG,AAT,ACC,ATG,ATT,CAA,CAT,CCT,CGC,GAC,GAG,GCA,GGC,GTA,TAC,TCG,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACC,AGT,ATG,ATT,CAA,CAT,CCT,CGC,GAC,GAG,GCA,GGC,GTC,TAC,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACT,ATC,ATG,CAA,CAC,CCG,CGT,CTG,GAC,GAG,GCA,GGC,GTA,TAT,TCC,TGG,TGT,TTC} |
| {AAG,AAT,ACT,ATC,ATG,CAA,CAC,CCT,CGC,CTG,GAC,GAG,GCA,GGC,GTA,TAT,TCG,TGG,TGT,TTC} |
| {AAG,AAT,ACT,AGT,ATC,ATG,CAA,CAC,CCT,CGC,CTG,GAC,GAG,GCA,GGC,GTC,TAT,TGG,TGT,TTC} |
| {AAG,AAT,ACC,ATG,ATT,CAA,CAT,CCG,CGT,CTG,GAC,GAG,GCA,GGC,GTA,TAC,TCC,TGG,TGT,TTC} |
| {AAG,AAT,ACC,ATG,ATT,CAA,CAT,CCT,CGC,CTG,GAC,GAG,GCA,GGC,GTA,TAC,TCG,TGG,TGT,TTC} |
| {AAG,AAT,ACC,AGT,ATG,ATT,CAA,CAT,CCT,CGC,CTG,GAC,GAG,GCA,GGC,GTC,TAC,TGG,TGT,TTC} |
| {AAC,AAG,ACT,ATA,ATG,CAC,CAG,CCT,CGT,GAC,GAG,GCC,GGC,GTA,TAT,TCA,TGC,TGG,TTC,TTG} |
| {AAC,AAG,ACT,ATA,ATG,CAG,CAT,CCA,CGT,GAC,GAG,GCC,GGC,GTA,TAT,TCC,TGC,TGG,TTC,TTG} |
| {AAC,AAG,ACT,AGT,ATA,ATG,CAG,CAT,CCA,CGT,CTC,GAC,GAG,GCC,GGC,GTT,TAT,TGC,TGG,TTC} |
| {AAC,AAG,ACC,AGT,ATA,ATG,CAG,CAT,CCT,CGT,GAC,GAG,GCC,GGC,GTT,TAC,TGC,TGG,TTA,TTC} |
| {AAC,AAG,ACC,AGT,ATA,ATG,CAG,CAT,CCT,CGT,CTA,GAC,GAG,GCC,GGC,GTT,TAT,TGC,TGG,TTC} |
| {AAC,AAG,ACT,ATA,ATG,CAC,CAG,CCT,CGT,CTG,GAC,GAG,GCC,GGC,GTA,TAT,TCA,TGG,TGT,TTC} |
| {AAC,AAG,ACT,ATA,ATG,CAG,CAT,CCA,CGT,CTG,GAC,GAG,GCC,GGC,GTA,TAT,TCC,TGG,TGT,TTC} |
| {AAG,AAT,ACT,AGC,ATC,ATG,CAA,CAC,CCT,CGT,GAC,GAG,GCC,GGC,GTA,TAT,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACC,AGC,ATG,ATT,CAA,CAT,CCT,CGT,GAC,GAG,GCC,GGC,GTA,TAC,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACA,AGT,ATC,ATG,CAC,CAG,CCT,CGT,GAC,GAG,GCC,GGC,GTT,TAC,TGC,TGG,TTA,TTC} |
| {AAG,AAT,ACA,AGT,ATC,ATG,CAC,CAG,CCT,CGT,CTA,GAC,GAG,GCC,GGC,GTT,TAT,TGC,TGG,TTC} |
| {AAG,AAT,ACA,ATG,ATT,CAC,CAG,CCT,CGT,GAC,GAG,GCC,GGC,GTA,TAC,TCA,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACA,ATG,ATT,CAG,CAT,CCA,CGT,GAC,GAG,GCC,GGC,GTA,TAC,TCC,TGC,TGG,TTC,TTG} |
| {AAG,AAT,ACA,AGT,ATG,ATT,CAG,CAT,CCA,CGT,CTC,GAC,GAG,GCC,GGC,GTT,TAC,TGC,TGG,TTC} |
| {AAG,AAT,ACT,AGC,ATC,ATG,CAA,CAC,CCT,CGT,CTG,GAC,GAG,GCC,GGC,GTA,TAT,TGG,TGT,TTC} |
| {AAG,AAT,ACC,AGC,ATG,ATT,CAA,CAT,CCT,CGT,CTG,GAC,GAG,GCC,GGC,GTA,TAC,TGG,TGT,TTC} |
| {AAG,AAT,ACA,ATG,ATT,CAC,CAG,CCT,CGT,CTG,GAC,GAG,GCC,GGC,GTA,TAC,TCA,TGG,TGT,TTC} |
| {AAG,AAT,ACA,ATG,ATT,CAG,CAT,CCA,CGT,CTG,GAC,GAG,GCC,GGC,GTA,TAC,TCC,TGG,TGT,TTC} |
| {AAC,AAG,ACT,ATA,ATG,CAC,CAG,CCG,CGT,CTC,GAC,GAG,GCT,GGC,GTA,TAT,TCA,TGG,TGT,TTC} |
| {AAC,AAG,ACT,ATA,ATG,CAG,CAT,CCA,CGC,CTC,GAC,GAG,GCT,GGC,GTA,TAT,TCG,TGG,TGT,TTC} |
| {AAC,AAG,ACT,AGT,ATA,ATG,CAG,CAT,CCA,CGC,CTC,GAC,GAG,GCT,GGC,GTC,TAT,TGG,TGT,TTC} |
| {AAC,AAG,ACC,ATA,ATG,CAG,CAT,CCG,CGT,GAC,GAG,GCT,GGC,GTA,TAC,TCC,TGG,TGT,TTA,TTC} |
| {AAC,AAG,ACC,ATA,ATG,CAG,CAT,CCG,CGT,CTA,GAC,GAG,GCT,GGC,GTA,TAT,TCC,TGG,TGT,TTC} |
| {AAC,AAG,ACC,ATA,ATG,CAG,CAT,CCT,CGC,GAC,GAG,GCT,GGC,GTA,TAC,TCG,TGG,TGT,TTA,TTC} |
| {AAC,AAG,ACC,AGT,ATA,ATG,CAG,CAT,CCT,CGC,GAC,GAG,GCT,GGC,GTC,TAC,TGG,TGT,TTA,TTC} |
| {AAC,AAG,ACC,ATA,ATG,CAG,CAT,CCT,CGC,CTA,GAC,GAG,GCT,GGC,GTA,TAT,TCG,TGG,TGT,TTC} |
| {AAC,AAG,ACC,AGT,ATA,ATG,CAG,CAT,CCT,CGC,CTA,GAC,GAG,GCT,GGC,GTC,TAT,TGG,TGT,TTC} |
| {AAG,AAT,ACT,AGC,ATC,ATG,CAA,CAC,CCG,CGT,CTC,GAC,GAG,GCT,GGC,GTA,TAT,TGG,TGT,TTC} |
| {AAG,AAT,ACC,AGC,ATG,ATT,CAA,CAT,CCG,CGT,CTC,GAC,GAG,GCT,GGC,GTA,TAC,TGG,TGT,TTC} |
| {AAG,AAT,ACA,ATC,ATG,CAC,CAG,CCG,CGT,GAC,GAG,GCT,GGC,GTA,TAC,TCC,TGG,TGT,TTA,TTC} |
| {AAG,AAT,ACA,ATC,ATG,CAC,CAG,CCG,CGT,CTA,GAC,GAG,GCT,GGC,GTA,TAT,TCC,TGG,TGT,TTC} |
| {AAG,AAT,ACA,ATC,ATG,CAC,CAG,CCT,CGC,GAC,GAG,GCT,GGC,GTA,TAC,TCG,TGG,TGT,TTA,TTC} |
| {AAG,AAT,ACA,AGT,ATC,ATG,CAC,CAG,CCT,CGC,GAC,GAG,GCT,GGC,GTC,TAC,TGG,TGT,TTA,TTC} |
| {AAG,AAT,ACA,ATC,ATG,CAC,CAG,CCT,CGC,CTA,GAC,GAG,GCT,GGC,GTA,TAT,TCG,TGG,TGT,TTC} |
| {AAG,AAT,ACA,AGT,ATC,ATG,CAC,CAG,CCT,CGC,CTA,GAC,GAG,GCT,GGC,GTC,TAT,TGG,TGT,TTC} |
| {AAG,AAT,ACA,ATG,ATT,CAC,CAG,CCG,CGT,CTC,GAC,GAG,GCT,GGC,GTA,TAC,TCA,TGG,TGT,TTC} |
| {AAG,AAT,ACA,ATG,ATT,CAG,CAT,CCA,CGC,CTC,GAC,GAG,GCT,GGC,GTA,TAC,TCG,TGG,TGT,TTC} |
| {AAG,AAT,ACA,AGT,ATG,ATT,CAG,CAT,CCA,CGC,CTC,GAC,GAG,GCT,GGC,GTC,TAC,TGG,TGT,TTC} |

By computer analysis, three subclasses of bijective genetic codes *BGC* are newly identified here according to the existence or not of permuted trinucleotides. Indeed, the absence of permuted trinucleotides in a code is a necessary condition, but not sufficient, for a code to be circular (Remark 2). So, some combinatorial properties may be identified between the class of bijective genetic codes *BGC* and the class of circular codes *CC* in future.

(i) The 1st subclass contains 52 bijective genetic codes *WPTBGC* without permuted trinucleotides (*WPT*), i.e. without the periodic permuted trinucleotides $PPT=\{\text{AAA},\text{CCC},\text{GGG},\text{TTT}\}$ and without the non-periodic permuted trinucleotides $NPPT=\{t,\mathcal{P}(t),\mathcal{P}^2(t)\}$ (details in Remarks 1 and 2). These codes listed in Table 2 are important for further investigation as they are potential circular codes but they lost some combinatorial properties to be circular codes, properties which are unknown so far.

(ii) The 2nd subclass, less constraint than the previous one, has 36,328 bijective genetic codes *WNPPTBGC* without the non-periodic permuted trinucleotides *NPPT* (*WNPPT*).

(iii) The 3rd subclass, the less constrained, has 47,775,744 bijective genetic codes *WPPTBGC* without the periodic permuted trinucleotides *PPT* (*WPPT*).

The reading frame coding probability PrRFC(*BGC*) of a bijective genetic code *BGC* is determined by applying Eq. (6) where each trinucleotide $t$ of *BGC* is assigned to a probability $\Pr(t)=1/20$ (by taking, as with the circular codes *CC*, the simplest hypothesis of equiprobability in order to have a reference value for the reading frame coding).

### 2.3.3. *Random trinucleotide codes*

There are several random trinucleotide codes *RC* with a probability PrRFC(*RC*) of reading frame coding equal to 1/3. There are four random codes of length 1 for each periodic permuted trinucleotide *PPT* (example (i) in Section 2.2.2). There is a random code of length 64 with the trinucleotide code $A_4^3=\{\text{AAA},\ldots,\text{TTT}\}$ where the 64 trinucleotides are associated to any probabilities, i.e. equiprobable or not (see Eq. (6)) but with a sum is equal to 1.

## 3. Results

### 3.1. *Reading frame coding of trinucleotide circular codes*

A trinucleotide circular code *CC* always retrieves the reading frame (by definition, see also Remark 3). However, the efficiency of reading frame retrieval depends on the code *CC* which is measured by its RFC probability PrRFC(*CC*).

(i) The 12,964,440 circular codes *CC* have an average RFC probability PrRFC(*CC*) equal to 83.2% in the range [72.7, 100] (Table 3a). Among these 12,964,440 codes *CC*, ten codes *CC* (list not given) code a maximum number of 18 amino acids (Table 4). All these 10 codes *CC* code the following set of 17 amino acids {Arg, Asn, Asp, Cys, Gln, Glu, Gly, Ile, Leu, Lys, Met, Phe, Pro, Ser, Trp, Tyr, Val} and for the 18th amino acid, six codes code His and four codes, Thr. Thus, Ala is never coded by these 10 codes *CC*. Seventy-three codes *CC* (list not given) code a minimum number of 5 amino acids (Table 4; list of amino acids not given).

(ii) The 408 comma-free circular codes *CFCC* have a RFC probability PrRFC(*CFCC*) equal to 1 (by Definition 10; see also Table 3b). This class has the highest RFC efficiency. Four codes *CFCC* code a maximum number of 13 amino acids (Table 4):
  - the code *CFCC* {AAC, AAG, ATA, CAC, GAC, CTA, CAG, GAG, GTA, ATC, ATG, TTA, CCG, CTC, GGC, GTC, CTG, TTC, GTG, TTG} codes {Asn, Asp, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Val};
  - the code *CFCC* {AAC, AAG, ATA, CAC, GAC, CTA, CAG, GAG, GTA, ATC, ATG, TTA, CCG, CTC, GCG, GTC, CTG, TTC, GTG, TTG} codes {Ala, Asn, Asp, Gln, Glu, His, Ile, Leu, Lys, Met, Phe, Pro, Val};
  - the code *CFCC* {AAC, AAG, ATA, CAC, GAC, CTA, CAG, GAG, GTA, ATC, ATG, TTA, CGC, CTC, GGC, GTC, CTG, TTC, GTG, TTG} codes {Arg, Asn, Asp, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Val};
  - the code *CFCC* {AAC, AAG, ATA, CAC, GAC, CTA, CAG, GAG, GTA, ATC, ATG, TTA, GCC, CTC, CGG, GTC, CTG, TTC, GTG, TTG} codes {Ala, Arg, Asn, Asp, Gln, Glu, His, Ile, Leu, Lys, Met, Phe, Val}.

(iii) The six other necklace classes of codes $J^3CCC, J^3CC, J^4CCC, J^4CC, J^5CCC$ and $J^5CC$ representing 12,964,032 codes *CC* have expected decreasing RFC probabilities (Table 3b) according to the necklace combinatorial property (Michel et al., 2012). Their amino acid properties are given in Table 4. Their combinatorial and genetic properties remain to be investigated.

(iv) The 221,328 $C^3$ circular codes $C^3CC$ have an average RFC probability $\mathrm{PrRFC}(C^3CC)$ equal to 85.5% in the range [74.9, 100] (Table 3b). The upper bound 100% of this class was totally unexpected. It proves that some codes $C^3CC$ are comma-free in frame 0 (by Proposition 1). However, no code is $C^3$ comma-free, i.e. comma-free in frames 0, 1 and 2 simultaneously (Table 4a in Michel et al., 2008b). Thus, by computer analysis, a new class of 192 trinucleotide codes is identified here which is comma-free in frame 0 and circular in frames 1 and 2 but not comma-free in frames 1 and 2.

(v) The 216 $C^3$ self-complementary circular codes $C^3SCC$ have an average RFC probability $\mathrm{PrRFC}(C^3SCC)$ equal to 84.1% in the range [77.2, 90.1] (Table 3b). Four codes $C^3SCC$ code a maximum number of 14 amino acids (Table 4):
  - the code $C^3SCC$ {AAC, AAG, AAT, ACC, ACG, ACT, AGT, ATG, ATT, CAG, CAT, CCG, CGG, CGT, CTC, CTG, CTT, GAG, GGT, GTT} codes {Arg, Asn, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Pro, Ser, Thr, Val};
  - the code $C^3SCC$ {AAC, AAG, AAT, ACT, AGT, ATG, ATT, CAC, CAG, CAT, CCG, CGG, CTC, CTG, CTT, GAC, GAG, GTC, GTG, GTT} codes {Arg, Asn, Asp, Gln, Glu, His, Ile, Leu, Lys, Met, Pro, Ser, Thr, Val};
  - the code $C^3SCC$ {AAG, AAT, ACA, ACG, ACT, AGT, ATG, ATT, CAG, CAT, CCA, CCG, CGG, CGT, CTC, CTG, CTT, GAG, TGG, TGT} codes {Arg, Asn, Cys, Gln, Glu, His, Ile, Leu, Lys, Met, Pro, Ser, Thr, Trp};
  - the code $C^3SCC$ {AAG, AAT, ACT, AGT, ATG, ATT, CAA, CAC, CAG, CAT, CCG, CGG, CTC, CTG, CTT, GAC, GAG, GTC, GTG,

**Table 3a**
Probability PrRFC(*C*) (%) of reading frame coding (Eq. (6)) of the two main classes of trinucleotide codes *C*: the 12,964,440 circular codes *CC* and the 339,738,624 bijective genetic codes *BGC*. The genetic code 61*GC* with 61 equiprobable codons without the three stop codons {TAA, TAG, TGA} (Section 2.2.2.(v)) and the random codes *RC* (Section 2.3.3) are also represented. For the sets of trinucleotide codes *C*, the probability PrRFC(*C*) (%) is also given for the 25th and 75th percentiles, and for the minimum and maximum values.

| Trinucleotide codes *C* | PrRFC(*C*) (%) | | | | |
|---|---|---|---|---|---|
| | Mean | 25th %ile | 75th %ile | Min | Max |
| Circular codes *CC* (12,964,440) | 83.2 | 81.3 | 85.1 | 72.7 | 100 |
| Bijective genetic codes *BGC* (339,738,624) | 61.5 | 59.7 | 63.2 | 47.8 | 75.5 |
| Genetic code 61*GC* | 34.5 | | | | |
| Random codes *RC* | 33.3 | | | | |

**Table 3b**
Probability PrRFC($C$) (%) of reading frame coding (Eq. (6)) of subclasses of the two main classes of trinucleotide codes $C$: (i) the class of the 12,964,440 circular codes $CC$ with (ia) the seven necklace subclasses; (ib) the three map subclasses based on the complementarity and circular permutation maps; (ic) the $C^3SCC$ code $X$ identified in eukaryotic and prokaryotic genes; (ii) the class of the 339,738,624 bijective genetic codes $BGC$ with (iia) the bijective genetic codes $WPTBGC$ without permuted trinucleotides ($WPT$), i.e. without the periodic permuted trinucleotides $PPT = \{\text{AAA}, \text{CCC}, \text{GGG}, \text{TTT}\}$ and without the non-periodic permuted trinucleotides $NPPT = \{t, \mathcal{P}(t), \mathcal{P}^2(t)\}$ (details in Remarks 1 and 2); (iib) the bijective genetic codes $WNPPTBGC$ without $NPPT$ ($WNPPT$); (iic) the bijective genetic codes $WPPTBGC$ without $PPT$ ($WPPT$). For the sets of trinucleotide codes $C$, the probability PrRFC($C$) (%) is also given for the 25th and 75th percentiles, and for the minimum and maximum values.

| Trinucleotide codes $C$ | PrRFC($C$) (%) | | | | |
|---|---|---|---|---|---|
| | Mean | 25th %ile | 75th %ile | Min | Max |
| *Circular codes CC* | | | | | |
| Necklace classes | | | | | |
| Comma-free circular codes $CFCC$ (408) | 100 | 100 | 100 | 100 | 100 |
| Circular codes $J^3CCC$ (2,352) | 91.6 | 89.7 | 93.2 | 88.1 | 96.6 |
| Circular codes $J^3CC$ (294,312) | 88.1 | 86.2 | 89.7 | 79.5 | 97.3 |
| Circular codes $J^4CCC$ (252,960) | 86.0 | 84.7 | 87.3 | 78.3 | 93.9 |
| Circular codes $J^4CC$ (4,566,696) | 84.5 | 83.0 | 86.2 | 74.1 | 94.6 |
| Circular codes $J^5CCC$ (823,920) | 83.2 | 81.8 | 84.6 | 75.6 | 90.9 |
| Circular codes $J^5CC$ (7,023,792) | 82.0 | 80.3 | 83.7 | 72.7 | 91.1 |
| Map classes | | | | | |
| $C^3$ circular codes $C^3CC$ (221,328) | 85.5 | 83.7 | 87.0 | 74.9 | 100 |
| Self-complementary circular codes $SCC$ (312) | 81.6 | 78.7 | 84.4 | 74.1 | 89.3 |
| $C^3$ self-complementary circular codes $C^3SCC$ (216) | 84.1 | 81.3 | 86.2 | 77.2 | 90.1 |
| $C^3SCC$ code $X$ | 81.3 | | | | |
| *Bijective genetic codes BGC* | | | | | |
| Bijective genetic codes $WPTBGC$ (52) | 66.0 | 64.3 | 67.0 | 62.2 | 71.6 |
| Bijective genetic codes $WNPPTBGC$ (36,328) | 63.7 | 62.0 | 65.3 | 55.9 | 71.6 |
| Bijective genetic codes $WPPTBGC$ (47,775,744) | 62.4 | 60.6 | 64.0 | 50.0 | 75.5 |

**Table 4**
Number of amino acids in the 12,964,440 circular codes $CC$, the seven necklace subclasses, the three map subclasses based on the complementarity and circular permutation maps and the $C^3SCC$ code $X$ identified in eukaryotic and prokaryotic genes. For the sets of trinucleotide codes $C$, the number of amino acids is also given for the 25th and 75th percentiles, and for the minimum and maximum values.

| | Number of amino acids | | | | |
|---|---|---|---|---|---|
| | Mean | 25th %ile | 75th %ile | Min | Max |
| Circular codes $CC$ (12,964,440) | 11.4 | 10 | 12 | 5 | 18 |
| *Necklace classes* | | | | | |
| Comma-free circular codes $CFCC$ (408) | 9.1 | 8 | 10 | 6 | 13 |
| Circular codes $J^3CCC$ (2,352) | 10.6 | 9 | 12 | 7 | 16 |
| Circular codes $J^3CC$ (294,312) | 10.7 | 10 | 12 | 5 | 17 |
| Circular codes $J^4CCC$ (252,960) | 11.2 | 10 | 12 | 5 | 17 |
| Circular codes $J^4CC$ (4,566,696) | 11.2 | 10 | 12 | 5 | 18 |
| Circular codes $J^5CCC$ (823,920) | 11.4 | 10 | 12 | 5 | 17 |
| Circular codes $J^5CC$ (7,023,792) | 11.6 | 11 | 13 | 5 | 18 |
| *Map classes* | | | | | |
| $C^3$ circular codes $C^3CC$ (221,328) | 10.7 | 10 | 12 | 5 | 17 |
| Self-complementary circular codes $SCC$ (312) | 11.6 | 10 | 13 | 7 | 15 |
| $C^3$ self-complementary circular codes $C^3SCC$ (216) | 10.8 | 10 | 12 | 6 | 14 |
| $C^3SCC$ code $X$ | 12 | | | | |

TTG} codes {Arg, Asn, Asp, Gln, Glu, His, Ile, Leu, Lys, Met, Pro, Ser, Thr, Val}.

### 3.2. Reading frame coding of bijective genetic codes

A bijective genetic code $BGC$ is not circular (Section 2.3.2). Thus, it does not always retrieve the reading frame. The efficiency of non-reading frame retrieval depends on the code $BGC$ which is measured by its RFC probability PrRFC($BGC$).

The 339,738,624 bijective genetic codes $BGC$ have an average RFC probability PrRFC($BGC$) equal to 61.5% in the range [47.8, 75.5] (Table 3a). The code $BGC_\text{L}$ coding the 20 amino acids with the lowest (L) RFC probability $\text{PrRFC}(BGC_\text{L}) = 47.8\%$ is the set {AAA, AAT, ACA, AGA, ATG, ATT, CAA, CAT, CCA, GAA, GAT, GCA, GGA, GTT, TAT, TCA, TGG, TGT, TTG, TTT}. Note that $BGC_\text{L}$ contains two periodic permuted trinucleotides $PPT$ (AAA and TTT). The code $BGC_\text{H}$ coding the 20 amino acids with the highest (H) RFC probability $\text{PrRFC}(BGC_\text{H}) = 75.5\%$ is the set {AAC, AAG, ACC, ATC, ATG, CAC, CAG, CCT, CGG, GAC, GAG, GCT, GGC, GTC, TAC, TCC, TGC, TGG, TTA, TTC}. Note that $BGC_\text{H}$ contains no periodic permuted trinucleotide $PPT$.

The three subclasses of codes $WPTBGC$, $WNPPTBGC$ and $WPPTBGC$ have expected decreasing RFC probabilities (Table 3b) according to the number of permuted trinucleotides (Remark 2). Precisely, a code $WPTBGC$ has no permuted trinucleotide. A code $WNPPTBGC$ can have a maximum of four permuted trinucleotides $PPT$ as AAA codes for Lys, CCC for Pro, GGG for Gly and TTT for Phe. A code $WPPTBGC$ can have 20 trinucleotides such that each trinucleotide $t$ of $WPPTBGC$ is associated to one or two permuted trinucleotides, i.e. with the following partitions $\{t, \mathcal{P}(t)\}$, $\{t, \mathcal{P}^2(t)\}$ and $\{t, \mathcal{P}(t), \mathcal{P}^2(t)\}$. By computer analysis, 33 codes $WPPTBGC$ among 47,775,744 are identified here such that their 20 trinucleotides are related each other by the permutation map (Table 5).

### 3.3. A genetic scale of reading frame coding of trinucleotide codes

The previous results allow a genetic scale of reading frame coding of trinucleotide codes to be defined. Fig. 1 represents this RFC probability scale which ranges from 1/3 with the random codes (one chance out of three to retrieve the reading frame among the three possible frames in genes) to 1 with the comma-free codes and the strong codes (the reading frame is always retrieved). The RFC probabilities of two fundamental classes of trinucleotide codes are represented: the 12,964,440 circular codes *CC* where the reading frame is always retrieved with more or less efficiency and the 339,738,624 bijective genetic codes *BGC* where the reading frame is not always retrieved with more or less efficiency. There is a non-empty RFC probability intersection of 2.8% between the circular codes *CC* and the bijective genetic codes *BGC*: $\text{PrRFC}(CC \cap BGC) \in [72.7, 75.5]$ (%) (Table 3a and Fig. 1).

### 3.4. Trinucleotide codes coding the 20 amino acids

Table 6 gives the 20 trinucleotide codes *C* coding the 20 amino acids with the universal genetic code. The probabilities $\Pr(t)$ of trinucleotides *t* in a code *C* are chosen uniform in order to have a reference value for the reading frame coding. Four trinucleotide codes *C* coding the four amino acids Gly, Lys, Phe and Pro are (obviously) not circular (existence of a periodic permuted trinucleotide *PPT*, see Remarks 1 and 2). The trinucleotide code coding the amino acid Leu is circular *CC* but not comma-free. As its RFC probability is less than 1, a proof in necessary to decide that it is circular (see Appendix). The 15 other trinucleotide codes coding the 15 other amino acids are comma-free circular codes *CFCC* (by Proposition 1 as RFC probabilities are equal to 1). Why the amino acid Leu coded by a trinucleotide code with a combinatorial property less constraint compared to the 15 other trinucleotide codes of amino acids (circular versus comma-free) remains open. It could be related to a particular chemical property or origin of Leu.

## 4. Conclusion

A genetic scale of reading frame coding of trinucleotide codes is determined here (Fig. 1). It ranges from 1/3 with the random codes to 1 with the comma-free circular codes and the strong codes. For the bound 1 of the genetic scale, the RFC probability can, in addition to the property of coding measure, be associated to the property of reading frame retrieval. Indeed, such trinucleotide

**Table 6**
The 20 trinucleotide codes *C* coding the 20 amino acids with the universal genetic code and their probability $\text{PrRFC}(C)$ of reading frame coding (Eq. (6)). The probabilities $\Pr(t)$ of trinucleotides *t* in a code *C* are chosen uniform in order to have a reference value for the reading frame coding (4th column). The four amino acids Gly, Lys, Phe and Pro are coded by codes *C* (not circular) with RFC probabilities equal to 2/3, 1/2, 1/2 and 2/3, respectively. The amino acid Leu is coded by a circular code *CC* (not comma-free) with a RFC probability equal to 18/19. The 15 other amino acids are coded by comma-free circular codes *CFCC*, i.e. with RFC probabilities equal to 1.

| Trinucleotide codes *C* | Amino acid | Class | $\Pr(t)$ | $\text{PrRFC}(C)$ |
|---|---|---|---|---|
| {ATG} | Met (M) | *CFCC* | 1 | 1 |
| {TGG} | Trp (W) | *CFCC* | 1 | 1 |
| {AAC, AAT} | Asn (N) | *CFCC* | 1/2 | 1 |
| {GAC, GAT} | Asp (D) | *CFCC* | 1/2 | 1 |
| {TGC, TGT} | Cys (C) | *CFCC* | 1/2 | 1 |
| {CAA, CAG} | Gln (Q) | *CFCC* | 1/2 | 1 |
| {GAA, GAG} | Glu (E) | *CFCC* | 1/2 | 1 |
| {CAC, CAT} | His (H) | *CFCC* | 1/2 | 1 |
| {AAA, AAG} | Lys (K) | *C* | 1/2 | 1/2 |
| {TTC, TTT} | Phe (F) | *C* | 1/2 | 1/2 |
| {TAC, TAT} | Tyr (Y) | *CFCC* | 1/2 | 1 |
| {ATA, ATC, ATT} | Ile (I) | *CFCC* | 1/3 | 1 |
| {GCA, GCC, GCG, GCT} | Ala (A) | *CFCC* | 1/4 | 1 |
| {GGA, GGC, GGG, GGT} | Gly (G) | *C* | 1/4 | 2/3 |
| {CCA, CCC, CCG, CCT} | Pro (P) | *C* | 1/4 | 2/3 |
| {ACA, ACC, ACG, ACT} | Thr (T) | *CFCC* | 1/4 | 1 |
| {GTA, GTC, GTG, GTT} | Val (V) | *CFCC* | 1/4 | 1 |
| {AGA, AGG, CGA, CGC, CGG, CGT} | Arg (R) | *CFCC* | 1/6 | 1 |
| {CTA, CTC, CTG, CTT, TTA, TTG} | Leu (L) | *CC* | 1/6 | 18/19 |
| {AGC, AGT, TCA, TCC, TCG, TCT} | Ser (S) | *CFCC* | 1/6 | 1 |

**Table 5**
The 33 bijective genetic codes *WPPTBGC* without the periodic permuted trinucleotides $PPT = \{\text{AAA}, \text{CCC}, \text{GGG}, \text{TTT}\}$ (*WPPT*) coding the 20 amino acids such that each trinucleotide *t* of *WPPTBGC* is associated to one or two permuted trinucleotides, i.e. with the following partitions $\{t, \mathcal{P}(t)\}$, $\{t, \mathcal{P}^2(t)\}$ and $\{t, \mathcal{P}(t), \mathcal{P}^2(t)\}$.

| |
|---|
| {AAC,AAG,ACA,AGA,ATG,ATT,CAC,CAG,CCA,CTT,GAA,GAT,GCA,GGT,GTT,TAT,TCT,TGG,TGT,TTC} |
| {AAC,AAG,ACA,AGA,ATG,ATT,CAC,CAG,CCA,GAA,GAT,GCA,GGT,GTT,TAT,TCT,TGG,TGT,TTC,TTG} |
| {AAG,AAT,ACT,AGA,ATA,ATG,CAC,CAG,CCA,CTT,GAA,GAT,GCA,GGT,GTT,TAC,TCT,TGG,TGT,TTC} |
| {AAG,AAT,ACT,AGA,ATA,ATG,CAC,CAG,CCA,GAA,GAT,GCA,GGT,GTT,TAC,TCT,TGG,TGT,TTC,TTG} |
| {AAC,AAG,ACA,ATG,ATT,CAA,CAC,CCA,CGC,CTT,GAA,GAT,GCC,GGT,GTT,TAT,TCT,TGG,TGT,TTC} |
| {AAC,AAG,ACA,ATG,ATT,CAA,CAC,CCA,CGC,GAA,GAT,GCC,GGT,GTT,TAT,TCT,TGG,TGT,TTC,TTG} |
| {AAC,AAG,ACC,ATG,ATT,CAA,CAC,CCA,CGC,CTT,GAA,GAT,GCC,GGT,GTT,TAT,TCT,TGG,TGT,TTC} |
| {AAC,AAG,ACC,ATG,ATT,CAA,CAC,CCA,CGC,GAA,GAT,GCC,GGT,GTT,TAT,TCT,TGG,TGT,TTC,TTG} |
| {AAC,AAG,ACC,AGA,ATG,ATT,CAA,CAC,CCG,CTT,GAA,GAT,GCC,GGT,GTT,TAT,TCT,TGG,TGT,TTC} |
| {AAC,AAG,ACC,AGA,ATG,ATT,CAA,CAC,CCG,GAA,GAT,GCC,GGT,GTT,TAT,TCT,TGG,TGT,TTC,TTG} |
| {AAC,AAG,ACC,ATG,ATT,CAA,CAC,CCG,CGC,CTT,GAA,GAT,GCC,GGT,GTT,TAT,TCT,TGG,TGT,TTC} |
| {AAC,AAG,ACC,ATG,ATT,CAA,CAC,CCG,CGC,GAA,GAT,GCC,GGT,GTT,TAT,TCT,TGG,TGT,TTC,TTG} |
| {AAC,AAG,ACT,AGA,ATC,ATG,CAA,CAT,CCG,CTT,GAA,GAT,GCC,GGT,GTT,TAC,TCT,TGG,TGT,TTC} |
| {AAC,AAG,ACT,AGA,ATC,ATG,CAA,CAT,CCG,GAA,GAT,GCC,GGT,GTT,TAC,TCT,TGG,TGT,TTC,TTG} |
| {AAC,AAG,ACT,ATC,ATG,CAA,CAT,CCG,CGC,CTT,GAA,GAT,GCC,GGT,GTT,TAC,TCT,TGG,TGT,TTC} |
| {AAC,AAG,ACT,ATC,ATG,CAA,CAT,CCG,CGC,GAA,GAT,GCC,GGT,GTT,TAC,TCT,TGG,TGT,TTC,TTG} |
| {AAC,AAG,ACA,ATG,ATT,CAA,CAC,CCA,CGG,CTT,GAA,GAT,GCG,GGT,GTT,TAT,TCT,TGG,TGT,TTC} |
| {AAC,AAG,ACA,ATG,ATT,CAA,CAC,CCA,CGG,GAA,GAT,GCG,GGT,GTT,TAT,TCT,TGG,TGT,TTC,TTG} |
| {AAC,AAG,ACC,ATG,ATT,CAA,CAC,CCA,CGG,CTT,GAA,GAT,GCG,GGT,GTT,TAT,TCT,TGG,TGT,TTC} |
| {AAC,AAG,ACC,ATG,ATT,CAA,CAC,CCA,CGG,GAA,GAT,GCG,GGT,GTT,TAT,TCT,TGG,TGT,TTC,TTG} |
| {AAC,AAG,ACA,AGG,ATG,ATT,CAA,CAC,CCA,CTT,GAA,GAT,GCT,GGA,GTG,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACC,AGG,ATG,ATT,CAA,CAC,CCA,CTT,GAA,GAT,GCT,GGA,GTG,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACA,ATG,ATT,CAA,CAC,CCA,CGG,CTT,GAA,GAT,GCT,GGC,GTG,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACC,ATG,ATT,CAA,CAC,CCA,CGG,CTT,GAA,GAT,GCT,GGC,GTG,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACA,ATG,ATT,CAA,CAC,CCA,CGT,CTT,GAA,GAT,GCT,GGT,GTC,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACA,AGA,ATG,ATT,CAA,CAC,CCA,CTT,GAA,GAT,GCT,GGT,GTG,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACC,ATG,ATT,CAA,CAC,CCA,CGT,CTT,GAA,GAT,GCT,GGT,GTC,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACC,AGA,ATG,ATT,CAA,CAC,CCA,CTT,GAA,GAT,GCT,GGT,GTG,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACG,ATG,ATT,CAA,CAC,CCA,CGA,CTT,GAA,GAT,GCT,GGT,GTG,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACC,ATG,ATT,CAA,CAC,CCG,CGC,CTT,GAA,GAT,GCT,GGT,GTG,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACT,ATC,ATG,CAA,CAT,CCG,CGC,CTT,GAA,GAT,GCT,GGT,GTG,TAC,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACA,AGA,ATG,ATT,CAA,CAC,CCA,CTT,GAG,GAT,GCT,GGA,GTG,TAT,TCT,TGC,TGG,TTC} |
| {AAC,AAG,ACC,AGA,ATG,ATT,CAA,CAC,CCA,CTT,GAG,GAT,GCT,GGA,GTG,TAT,TCT,TGC,TGG,TTC} |

codes only code for the reading frame 0, i.e. they always retrieve the reading frame. The RFC probabilities of reading frame coding of a huge number of trinucleotide codes are computed and represented in this scale: (i) the 12,964,440 circular codes $CC$ including in particular the 408 comma-free circular codes $CFCC$, the six other necklace classes of 12,964,032 codes $J^3CCC$, $J^3CC$, $J^4CCC$, $J^4CC$, $J^5CCC$ and $J^5CC$, and the 216 $C^3$ self-complementary circular codes $C^3SCC$ containing the code $X$ (1) identified in eukaryotic and prokaryotic genes; (ii) the 339,738,624 bijective genetic codes of 20 trinucleotides coding the 20 amino acids including the 52 bijective genetic codes $WPTBGC$ without permuted trinucleotides ($WPT$).

Comma-free codes are the most RFC efficient codes with a RFC probability of reading frame coding equal to 1 (Table 3b and Fig. 1) and a reading frame always retrieved (Definition 10). However, they are not statistically observed in today genes. At least two combinatorial properties may explain the fact that genetic information does not use comma-free codes for coding reading frames in genes: (i) there is no self-complementary comma-free code containing 20 trinucleotides (see the growth function in Table 3a in Michel et al., 2008b); (ii) there is no comma-free code $Y$ containing 20 trinucleotides such that the codes $Y_1=\mathcal{P}(Y)$ and $Y_2=\mathcal{P}^2(Y)$ obtained by circular permutation of $Y$, are also comma-free codes (see the growth function in Table 4a in Michel et al., 2008b). Self-complementary circular codes and $C^3$ circular codes are important for coding the six frames in the two complementary strands of the DNA double helix (the reading frame 0 and the two shifted frames 1 and 2 in each strand; detailed in Fig. 4 in Arquès and Michel, 1996).

The $C^3SCC$ code $X$ statistically observed in genes of prokaryotes and eukaryotes has a RFC probability of reading frame coding equal to 81.3% which is not the most RFC efficient of the 216 $C^3$ self-complementary circular codes $C^3SCC$ ranging up to 90.1% (Table 3b and Fig. 1). However, some combinatorial properties may be specific to the circular code $X$:

(i) $X$ has the maximum length, i.e. 13 nucleotides in the scale [5,7,9,13], of the minimal window of the 216 codes $C^3SCC$ ((i) in Section 3.7 in Arquès and Michel, 1996).
(ii) $X$ is one of the 216 codes $C^3SCC$ having the most trinucleotides misplaced in the shifted frames, precisely 27.5% in the scale [6.5,31.0] (Fig. 3 in Arquès and Michel, 1996).
(iii) $X$ can code the comma-free code $\text{RNY}=\{\text{RRY},\text{RYY}\}$ with $\text{R}=\{\text{A},\text{G}\}$, $\text{Y}=\{\text{C},\text{T}\}$, and $\text{N}=\{\text{R},\text{Y}\}$ (Table 3(a) in Arquès and Michel, 1996; Eigen and Schuster, 1978).
(iv) Transversion I ($\mathcal{T}_\text{I}(\text{A})=\text{T}$, $\mathcal{T}_\text{I}(\text{C})=\text{G}$, $\mathcal{T}_\text{I}(\text{G})=\text{C}$, $\mathcal{T}_\text{I}(\text{T})=\text{A}$) on the 2nd position of any subset of trinucleotides of $X$ generates trinucleotide circular codes which are always $C^3$ (Benard and Michel, 2013).

Thus, these combinatorial properties of $X$ as well as the existence of $X$ motifs in transfer and ribosomal RNAs (Michel, 2012, 2013) may explain why the $C^3SCC$ code $X$ is used for reading frame coding, retrieval and maintenance in genes. While some combinatorial properties of $X$ may still be identified, the experimental biological properties of $X$ remain unknown.

The $C^3SCC$ code $X$ is a weak statistical property at the sequence level in today genes. However, even if the genetic code uses 61 trinucleotides for coding the 20 amino acids in today genes, the 20 trinucleotides of $X$ may still have a biological function, even incomplete, of reading frame coding, retrieval and maintenance in genes. This concept is reinforced by the fact that the $X$ motifs identified in transfer and ribosomal RNAs may lead to a possible translation code (Michel, 2012, 2013). At the population level, the $C^3SCC$ code $X$ is a strong statistical property as it observed in a great number of genes from diverse taxonomic groups of prokaryotes and eukaryotes according to the data already available in the EMBL database in 1996. Precisely, with the hypothesis that gene evolution is mainly a random process then the law of large numbers asserts that the $C^3SCC$ code $X$ is a strong statistical property in primitive genes (before evolution). Thus, the $C^3SCC$ code $X$ could be the unique biological process of reading frame coding, retrieval and maintenance in primitive genes. Furthermore, its very simple structure only using trinucleotides on the genetic alphabet is compatible with an early stage of evolution, i.e. before the protein complexity of the translation process in today genes using two alphabets, the genetic and the protein alphabets.

Only the class of the 216 codes $C^3SCC$ has been significantly studied, even if a few combinatorial problems remain still open (Koch and Lehmann, 1997; Lacan and Michel, 2001; Gonzalez et al., 2011 who mentioned "…the issue on which codes can be obtained by relaxing the conditions established in Koch and Lehmann (1997) remains open"). During this research work here, three interesting trinucleotide codes are identified and should be investigated from a combinatorial and biological point of views in future. In the class of the 408 comma-free circular codes $CFCC$, there is a subclass of 192 codes $CFCC$ which are also circular in frames 1 and 2 (but not comma-free in frames 1 and 2). The 52 bijective genetic codes $WPTBGC$ without permuted trinucleotides ($WPT$) can code the 20 amino acids with the necessary condition to be circular codes. However, they lost some combinatorial properties to be circular codes. The 33 bijective genetic codes $WPPTBGC$ without the periodic permuted trinucleotides $PPT$ ($WPPT$) can code the 20 amino acids such that each trinucleotide $t$ of $WPPTBGC$ is associated to one or two permuted trinucleotides, i.e. with the following partitions $\{t,\mathcal{P}(t)\}$, $\{t,\mathcal{P}^2(t)\}$ and $\{t,\mathcal{P}(t),\mathcal{P}^2(t)\}$.

## Acknowledgments

I thank an anonymous reviewer for his advices and Denise Besch, Svetlana Gorchkova, Elisabeth Michel and Jean-Marc Vassards for their support.

## Appendix

We prove that the trinucleotide code $C_\text{Leu}=\{\text{CTA, CTC, CTG, CTT, TTA, TTG}\}$ coding the amino acid Leu (L) is a circular code but not comma-free.

Let $l_1,l_2,\ldots,l_{n-1},l_n,\ldots$ be letters in $A_4=\{\text{A},\text{C},\text{G},\text{T}\}$, $d_1,d_2,\ldots,d_{n-1},d_n,\ldots$ be diletters in $A_4^2=\{\text{AA},\ldots,\text{TT}\}$ and $n$ be an integer satisfying $n\geq 2$.

**Definition.** *Letter Diletter Continued Necklaces* ($LDCN$): We say that the ordered sequence $l_1,d_1,l_2,d_2,\ldots,d_{n-1},l_n,d_n,l_{n+1}$ is an $(n+1)LDCN$ for a subset $Y\subset A_4^3$ if

$$l_1d_1,l_2d_2,\ldots,l_nd_n\in Y$$

and

$$d_1l_2,d_2l_3,\ldots,d_{n-1}l_n,d_nl_{n+1}\in Y.$$

We use here the simplest proposition for a proof by hand.

**Proposition 2.** (Pirillo, 2003). *The following conditions are equivalent*:

(i) *$Y$ is a trinucleotide circular code.*
(ii) *$Y$ has no 5$LDCN$.*

**Proposition 3.** *The trinucleotide code $C_\text{Leu}$ is circular.*

**Proof.** By way of contradiction, suppose that $C_{\text{Leu}}$ admits a 5*LDCN*. As the letter $l_1 \in \{\text{C}, \text{T}\}$ (prefix of trinucleotides of $C_{\text{Leu}}$), it is enough to prove that each letter choice leads to a contradiction.

(i) If $l_1 = \text{C}$ then there are four possible diletters $d_1 \in \{\text{TA}, \text{TC}, \text{TG}, \text{TT}\}$ (suffix of trinucleotides of $C_{\text{Leu}}$).
  (ia) If $d_1 \in \{\text{TA}, \text{TC}, \text{TG}\}$ then there is a contradiction with these three cases as no trinucleotide of $C_{\text{Leu}}$ has such prefixes.
  (ib) If $d_1 \in \{\text{TT}\}$ then there are two possible letters $l_2 \in \{\text{A}, \text{G}\}$ (suffix of trinucleotides of $C_{\text{Leu}}$). If $l_2 \in \{\text{A}, \text{G}\}$ then there is a contradiction with these two cases as no trinucleotide of $C_{\text{Leu}}$ has such prefixes.
(ii) If $l_1 = \text{T}$ then there are two possible diletters $d_1 \in \{\text{TA}, \text{TG}\}$ (suffix of trinucleotides of $C_{\text{Leu}}$). These two cases lead to a contradiction (see (ia)).

As, for each letter, we cannot complete the assumed 5*LDCN* for $C_{\text{Leu}}$, we are in contradiction. Hence, $C_{\text{Leu}}$ is a circular code.

**Proposition 4.** *The trinucleotide circular code* $C_{\text{Leu}}$ *is not comma-free.*

**Proof.** It is enough to give a counterexample of Definition 10. Take $x = \text{CTT} \in C_{\text{Leu}}$, $y_1 = \text{CTC} \in C_{\text{Leu}}$ and $y_2 = \text{TTA} \in C_{\text{Leu}}$ then $u = \text{CT} \notin C_{\text{Leu}}$ and $v = \text{A} \notin C_{\text{Leu}}$.

## References

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

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

Benard, E., Michel, C.J., 2013. Transition and transversion on the common trinucleotide circular code. Comput. Biol. J., 1–10 (Article ID 795418).?

Berstel, J., Perrin, D., 1985. Theory of Codes. Pure and Applied Mathematics. vol. 117. Academic Press, London, UK.

Crick, F.H.C., Brenner, S., Klug, A., Pieczenik, G., 1976. A speculation on the origin of protein synthesis. Origins Life 7, 389–397.

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

Eigen, M., Schuster, P., 1978. The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. Naturwissenschaften 65, 341–369.

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

Golomb, S.W., Welch, L.R., Delbrück, M., 1958b. Construction and properties of comma-free codes. Biol. Medd. Dan. Vidensk. Selsk. 23, 1–34.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2011. Circular codes revisited: a statistical approach. *J. Theor. Biol.* 275, 21–28.

Jenner, L.B., Demeshkina, N., Yusupova, G., Yusupov, M., 2010. Structural aspects of messenger RNA reading frame maintenance by the ribosome. Nat. Struct. Mol. Biol. 17, 555–560.

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

Lassez, J.-L., 1976. Circular codes and synchronization. Int. J. Comput. Inf. Sci. 5, 201–208.

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

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., 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., 2013. A permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes. *J. Theor. Biol.* 319, 116–121.

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

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

Michel, C.J., Pirillo, G., Pirillo, M.A., 2012. A classification of 20-trinucleotide circular codes. Inf. Comput. 212, 55–63.

Nirenberg, M.W., Matthaei, J.H., 1961. The dependance of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc. Natl. Acad. Sci. USA 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.