Journal of Theoretical Biology 380 (2015) 156–177

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses

Christian J. Michel

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

HIGHLIGHTS

- The maximal C^3 self-complementary trinucleotide circular code X in genes.
- Genes of bacteria, eukaryotes, plasmids and viruses.
- Circular code asymmetries in the three frames of genes.
- Variant X codes in genes.

ARTICLE INFO

Article history:
Received 26 August 2014
Received in revised form
28 February 2015
Accepted 9 April 2015
Available online 29 April 2015

Keywords:
Circular code in genes
Genes of bacteria
Genes of eukaryotes
Genes of plasmids
Genes of viruses

ABSTRACT

In 1996, a set X of 20 trinucleotides is identified in genes of both prokaryotes and eukaryotes which has in average the highest occurrence in reading frame compared to the two shifted frames (Arquès and Michel, 1996). Furthermore, this set X has an interesting mathematical property as X is a maximal C^3 self-complementary trinucleotide circular code (Arquès and Michel, 1996). In 2014, the number of trinucleotides in prokaryotic genes has been multiplied by a factor of 527. Furthermore, two new gene kingdoms of plasmids and viruses contain enough trinucleotide data to be analysed. The approach used in 1996 for identifying a preferential frame for a trinucleotide is quantified here with a new definition analysing the occurrence probability of a complementary/permutation (CP) trinucleotide set in a gene kingdom. Furthermore, in order to increase the statistical significance of results compared to those of 1996, the circular code X is studied on several gene taxonomic groups in a kingdom. Based on this new statistical approach, the circular code X is strengthened in genes of prokaryotes and eukaryotes, and now also identified in genes of plasmids. A subset of X with 18 or 16 trinucleotides is identified in genes of viruses. Furthermore, a simple probabilistic model based on the independent occurrence of trinucleotides in reading frame of genes explains the circular code frequencies and asymmetries observed in the shifted frames in all studied gene kingdoms. Finally, the developed approach allows to identify variant X codes in genes, i.e. trinucleotide codes which differ from X. In genes of bacteria, eukaryotes and plasmids, 14 among the 47 studied gene taxonomic groups (about 30%) have variant X codes. Seven variant X codes are identified with at least 16 trinucleotides of X. Two variant X codes X_A in cyanobacteria and plasmids of cyanobacteria, and X_D in birds are self-complementary, without permuted trinucleotides but non-circular. Five variant X codes X_B in deinococcus, plasmids of chloroflexi and deinococcus, mammals and kinetoplasts, X_C in elusimicrobia and apicomplexans, X_E in fishes, X_F in insects, and X_G in basidiomycetes and plasmids of spirochaetes are C^3 self-complementary circular. In genes of viruses, no variant X code is found.

1. Introduction

The trinucleotide codes, e.g. the genetic code, constitute 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, and 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

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

http://dx.doi.org/10.1016/j.jtbi.2015.04.009
0022-5193/

per amino acid without ambiguity. However, statistically, no trinucleotide comma-free code was identified in genes. Furthermore, in the beginning sixties, 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 {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): prokaryotes (13,686 sequences, 4,708,758 trinucleotides) and eukaryotes (26,757 sequences, 11,397,678 trinucleotides) (Arquès and Michel, 1996). This set X contains the 20 following trinucleotides:

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

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 maximal 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. 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 nucleotides allows to retrieve the reading frame for all the ambiguous words generated with X. Such a window explains that the circular codes are less constrained than the comma-free codes. 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). Several X motifs 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, 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 the circular code (Michel, 2012).

During the last 20 years, several classes of methods were developed for searching circular codes in genes, in particular:

(i) trinucleotide frequency per frame (Arquès and Michel, 1996);
(ii) correlation function per frame (Arquès and Michel, 1997);
(iii) frame permuted trinucleotide frequency (Frey and Michel, 2003, 2006);
(iv) 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).

The approach used in 1996 for identifying a preferential frame for a trinucleotide is quantified here with a new definition analysing the occurrence probability of a complementary/permutation (CP) trinucleotide set in a gene kingdom. Furthermore, in order to increase the statistical significance of results compared to those of 1996, the circular code X is studied on gene taxonomic groups of a kingdom. The statistical analysis here is carried on 2,481,566,882 trinucleotides of prokaryotic genes, a trinucleotide number multiplied by a factor of 527 compared to 1996, and on 824,825,761 trinucleotides of eukaryotic genes, a trinucleotide number also multiplied by a significant factor of 72 compared to 1996. Furthermore, two new gene kingdoms of plasmids and viruses contain enough trinucleotide data to be analysed for a search of the circular code X. The proposed approach strengthened the circular code X in genes of prokaryotes and eukaryotes. Furthermore, it identifies the circular code X in genes of plasmids and a subset of X in genes of viruses. The development of a simple probabilistic model based on the independent occurrence of trinucleotides in reading frame of genes will explain the circular code frequencies and asymmetries observed in the shifted frames in all studied gene kingdoms. Finally, the developed approach also allows to identify variant X codes in genes, i.e. trinucleotide codes which differ from X. In genes of bacteria, eukaryotes and plasmids, 14 among the 47 studied gene taxonomic groups (about 30%) have variant X codes. Seven variant X codes are identified with at least 16 trinucleotides of X and which are either circular or non-circular. In genes of viruses, no variant X code is found.

2. Method

2.1. Definitions

A few classical definitions are briefly recalled in order to understand the main properties of the trinucleotide circular code X identified in genes of prokaryotes and eukaryotes (Arquès and Michel, 1996).

Notation 1. The letters (or nucleotides or bases) define the genetic alphabet $A_4=\{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$. 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 $A_4^3=\{\mathrm{AAA},\ldots,\mathrm{TTT}\}$. 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 {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 $\mathcal{C}(\mathrm{A})=\mathrm{T}$, $\mathcal{C}(\mathrm{C})=\mathrm{G}$, $\mathcal{C}(\mathrm{G})=\mathrm{C}$, $\mathcal{C}(\mathrm{T})=\mathrm{A}$. 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. $\mathcal{C}(\mathrm{ACG})=\mathrm{CGT}$. By extension to a trinucleotide set S, the *set complementarity map* $\mathcal{C}: \mathbb{P}\left(A_4^3\right) \to \mathbb{P}\left(A_4^3\right)$, $\mathbb{P}$ being the set of all subsets of A_4^3, is defined by $\mathcal{C}(S)=\left\{v \;\middle|\; u,v\in A_4^3, u\in S, v=\mathcal{C}(u)\right\}$, i.e. a complementary trinucleotide

set $\mathcal{C}(S)$ is obtained by applying the complementarity map $\mathcal{C}$ to all its trinucleotides, e.g. $\mathcal{C}(\{\mathrm{ACG}, \mathrm{AGT}\}) = \{\mathrm{ACT}, \mathrm{CGT}\}$.

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

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 $A_4^3 = \{\mathrm{AAA}, \ldots, \mathrm{TTT}\}$ 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 1. A trinucleotide code C containing either one periodic permuted trinucleotide $PPT = \{\mathrm{AAA}, \mathrm{CCC}, \mathrm{GGG}, \mathrm{TTT}\}$ 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, 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 associated to an unidirectional axis such as the $5'-3'$ direction (Michel, 2012). The circular code $X_1 = \mathcal{P}(X)$ contains the 20 following trinucleotides:

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

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

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

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

2.2. Kingdoms and taxonomic groups of genes

Kingdoms K and taxonomic groups G of genes belonging to complete genomes of bacteria, (nuclear) eukaryotes, (bacterial) plasmids and viruses are extracted from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, May 2014) (Table 1). Usual preliminary tests exclude genes with nucleotides different from A_4, without start codons $\{\mathrm{ATG}, \mathrm{GTG}, \mathrm{TTG}\}$ (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=t), without stop codons $\{\mathrm{TAA}, \mathrm{TAG}, \mathrm{TGA}\}$ and with nucleotide lengths non-modulo 3. In a given kingdom, taxonomic groups containing less than 500 genes are eliminated in this study. Note also that the rare start codon CTG in the standard code is very often associated to inaccurate sequences and it has been removed from the set of start codons in order to (slightly) increase the speed of data acquisition.

The kingdom of bacteria $K = B$ contains 7,851,762 genes (2,481,566,882 trinucleotides) and is divided into 25 taxonomic groups $G = B_i$ listed and characterized in Table 1. The kingdom of eukaryotes $K = E$ has 1,662,579 genes (824,825,761 trinucleotides) and is divided into 11 taxonomic groups $G = E_i$ (Table 1). The kingdom of plasmids $K = P$ contains 237,486 genes (68,244,356 trinucleotides) and is divided into 11 taxonomic groups $G = P_i$ (Table 1). The kingdom of viruses $K = V$ has 184,344 genes (45,688,798 trinucleotides) and is divided into six taxonomic groups $G = V_i$ (Table 1). The two gene kingdoms of plasmids and viruses analysed for the first time here have numbers of genes and trinucleotides significantly less than those in the gene kingdoms of prokaryotes and eukaryotes. The statistical analysis developed here considers each gene taxonomic group in a kingdom with the same weight, i.e. with the same biological importance, whatever its numbers of genes and trinucleotides (see below Eqs. (4) and (12)). Furthermore, the trinucleotide frequencies are computed here on large gene taxonomic groups, i.e. greater than 500 genes, in order to have stable and average values for the analysis of the circular code X.

2.3. Preferential frame of a trinucleotide in a gene taxonomic group or a gene kingdom

A (very) simple parameter is defined for determining a preferential frame for a trinucleotide in a gene taxonomic group or a gene kingdom. It quantifies the approach proposed in 1996 which allowed the identification of the circular code X in genes of prokaryotes and eukaryotes (Arquès and Michel, 1996).

Table 1
Kingdoms K and their taxonomic groups G of genes extracted from the GenBank database with their number of genes and trinucleotides.

Kingdom K	Group G	Symbol	Nb of genes	Nb of trinucleotides
Bacteria B	Actinobacteria	B_{ACT}	1,089,730	357,508,023
	Aquificae	B_{AQU}	20,280	6,165,531
	Armatimonadetes	B_{ARM}	2809	1,017,309
	Bacteroidetes	B_{BAC}	318,160	110,404,488
	Caldiserica	B_{CAL}	1581	481,735
	Chlamydiae	B_{CHA}	127,066	43,201,999
	Chloroflexi	B_{CHO}	51,703	17,418,781
	Chrysiogenetes	B_{CHR}	2571	858,437
	Cyanobacteria	B_{CYA}	283,213	87,868,703
	Deferribacteres	B_{DEF}	9387	3,041,832
	Deinococcus	B_{DEI}	50,870	15,707,697
	Dictyoglomi	B_{DIC}	3654	1,177,023
	Elusimicrobia	B_{ELU}	1529	494,013
	Fibrobacteres	B_{FIB}	41,927	15,301,826
	Firmicutes	B_{FIR}	1,692,429	502,060,799
	Fusobacteria	B_{FUS}	15,867	4,998,875
	Gemmatimonadetes	B_{GEM}	3935	1,420,091
	Nitrospirae	B_{NIT}	11,182	3,380,667
	Planctomycetes	B_{PLA}	27,692	10,270,528
	Proteobacteria	B_{PRO}	3,873,667	1,225,804,951
	Spirochaetes	B_{SPI}	114,930	38,196,569
	Synergistetes	B_{SYN}	8903	2,837,236
	Tenericutes	B_{TEN}	63,690	20,717,741
	Thermodesulfobacteria	B_{THD}	3791	1,199,562
	Thermotogae	B_{THG}	31,196	10,032,466
	Sum		7,851,762	2,481,566,882
Eukaryotes E	Birds	E_{BIR}	79,171	46,172,760
	Fishes	E_{FIS}	76,602	43,256,890
	Insects	E_{INS}	30,062	16,365,302
	Mammals	E_{MAM}	700,461	372,644,961
	Roundworms	E_{RWO}	42,418	17,007,286
	Ascomycetes	E_{ASC}	171,318	81,422,894
	Basidiomycetes	E_{BAS}	19,618	10,421,444
	Green_Algae	E_{GAL}	25,537	10,010,687
	Land_Plants	E_{LPL}	428,348	172,311,552
	Apicomplexans	E_{API}	42,190	27,166,377
	Kinetoplasts	E_{KIN}	46,854	28,045,608
	Sum		1,662,579	824,825,761
Plasmids P	Actinobacteria	P_{ACT}	17,531	4,963,407
	Bacteroidetes	P_{BAC}	3773	1,241,432
	Chlamydiae	P_{CMD}	617	172,493
	Chloroflexi	P_{CRF}	1239	384,468
	Cyanobacteria	P_{CYA}	12,540	3,768,101
	Deinococcus	P_{DEI}	6633	1,937,528
	Fibrobacteres	P_{FIB}	844	323,211
	Firmicutes	P_{FIR}	29,628	7,258,266
	Fusobacteria	P_{FUS}	1083	322,669
	Proteobacteria	P_{PRO}	157,869	46,615,252
	Spirochaetes	P_{SPI}	5729	1,257,529
	Sum		237,486	68,244,356
Viruses V	dsDNA	V_{DSD}	172,198	39,934,299
	dsRNA	V_{DSR}	973	654,931
	Retro-transcribing	V_{RTR}	559	269,070
	ssDNA	V_{SSD}	3562	796,401
	ssRNA	V_{SSR}	4492	3,510,773
	Phages	V_{PHA}	2560	523,324
	Sum		184,344	45,688,798

Let the (protein coding) gene family $\mathcal{F}$ be a gene taxonomic group G or a gene kingdom K. Let $PrFr_f(t,\mathcal{F})$ be the occurrence frequency of a trinucleotide $t \in A_4^3$ in a frame $f \in \{0,1,2\}$ of a gene family $\mathcal{F}$. Let $MdPrFr_f(t,\mathcal{F})$ be the median occurrence frequency of a trinucleotide $t \in A_4^3$ in a frame $f \in \{0,1,2\}$ of a gene family $\mathcal{F}$. Thus, there are $3 \times 64 = 192$ trinucleotide occurrence frequencies $PrFr_f(t,\mathcal{F})$ ($MdPrFr_f(t,\mathcal{F})$) in the three frames f of a gene family $\mathcal{F}$. Then, the preferential frame $PrefFr(t,\mathcal{F}) \in \{0,1,2\}$ ($MdPrefFr(t,\mathcal{F}) \in \{0,1,2\}$, respectively) of a trinucleotide $t \in A_4^3$ in a gene family $\mathcal{F}$ is defined by the frame having the maximal occurrence frequency $PrFr_f(t,\mathcal{F})$ ($MdPrFr_f(t,\mathcal{F})$, respectively) among the three frames $f \in \{0,1,2\}$ of genes in $\mathcal{F}$

$$\begin{cases} PrefFr(t,\mathcal{F}) = \arg\max_{f \in \{0,1,2\}} PrFr_f(t,\mathcal{F}) \\ MdPrefFr(t,\mathcal{F}) = \arg\max_{f \in \{0,1,2\}} MdPrFr_f(t,\mathcal{F}). \end{cases} \quad (2)$$

Remark 3. With the large gene taxonomic groups G studied (Section 2.2), the three trinucleotide occurrence frequencies $PrFr_f(t,G)$ in the three frames f have always different values.

Table 2a gives the mean occurrence frequencies $PrFr_f(t,K)$ of the 64 trinucleotides t in the three frames $f \in \{0,1,2\}$ of genes in bacteria $K=B$ (7,851,762 genes, 2,481,566,882 trinucleotides), eukaryotes $K=E$ (1,662,579 genes, 824,825,761 trinucleotides), plasmids $K=P$ (237,486 genes, 68,244,356 trinucleotides) and viruses $K=V$ (184,344 genes, 45,688,798 trinucleotides) (Table 1). The codon usage is given by $PrFr_0(t,K)$ in reading frame $f=0$ of genes. These 64 trinucleotide frequencies $PrFr_f(t,K)$ in a gene kingdom K are obtained from the sum of trinucleotide occurrence numbers in the gene groups G belonging to K. By defining the occurrence number $NbFr_f(t,G)$ of a trinucleotide $t \in A_4^3$ in a frame $f \in \{0,1,2\}$ of a group G, then the mean occurrence frequency $PrFr_0(\mathrm{AAA},V)$ of AAA in frame 0 of the viral kingdom V, for example, is equal to $PrFr_0(\mathrm{AAA},V) = \frac{1}{45,688,798}[NbFr_0(\mathrm{AAA},\mathrm{dsDNA}) + NbFr_0(\mathrm{AAA},\mathrm{dsRNA}) + NbFr_0(\mathrm{AAA},\mathrm{Retro{-}transcribing}) + NbFr_0(\mathrm{AAA},\mathrm{ssDNA}) + NbFr_0(\mathrm{AAA},\mathrm{ssRNA}) + NbFr_0(\mathrm{AAA},\mathrm{Phages})] = \frac{1,647,554}{45,688,798} = 3.61\%$ which is given in Table 2a. These mean trinucleotide frequencies can be considered as reference values for the codon usage in bacteria, eukaryotes, plasmids and viruses. However, some statistical aspects should not be ignored, in particular (i) some gene taxonomic groups are overrepresented or underrepresented (experimental difficulty or not to sequence some genomes, large number of species or not in a today's taxonomic group, etc.); (ii) an overrepresented gene taxonomic group may reflect or not the trinucleotide distribution in its kingdom.

Table 2b gives the median occurrence frequencies $MdPrFr_f(t,K)$ of the 64 trinucleotides t in the three frames $f \in \{0,1,2\}$ of the 25 gene taxonomic groups $G=B_i$ in bacteria $K=B$ (7,851,762 genes, 2,481,566,882 trinucleotides), the 11 taxonomic groups $G=E_i$ in eukaryotes $K=E$ (1,662,579 genes, 824,825,761 trinucleotides), the 11 taxonomic groups $G=P_i$ in plasmids $K=P$ (237,486 genes, 68,244,356 trinucleotides) and the six taxonomic groups $G=V_i$ in viruses $K=V$ (184,344 genes, 45,688,798 trinucleotides) (Table 1).

2.4. Preferential frame occurrence probability of a trinucleotide in a frame of a gene kingdom

Let a gene kingdom K with $\mathrm{Card}(K)$ gene taxonomic groups G, i.e. $K = \{G_1,\ldots,G_{\mathrm{Card}(K)}\}$. Let the indicator function $\delta_f(PrefFr(t,G))$ ($\delta_f(MdPrefFr(t,G))$, respectively) be equal to 1 if the preferential frame $PrefFr(t,G) \in \{0,1,2\}$ ($MdPrefFr(t,G) \in \{0,1,2\}$, respectively) of the trinucleotide $t \in A_4^3$ is equal to a given frame $f \in \{0,1,2\}$ of a gene taxonomic group G, 0 otherwise

$$\begin{cases} \delta_f(PrefFr(t,G)) = \begin{cases} 1 & \text{if } PrefFr(t,G) = f, \\ 0 & \text{otherwise} \end{cases} \\ \delta_f(MdPrefFr(t,G)) = \begin{cases} 1 & \text{if } MdPrefFr(t,G) = f, \\ 0 & \text{otherwise} \end{cases} \end{cases} \quad (3)$$

where $PrefFr(t,G)$ and $MdPrefFr(t,G)$ are defined in Eq. (2).

Then, the preferential frame occurrence probability $PrPrefFr_f(t,K)$ of a trinucleotide $t \in A_4^3$ in a frame $f \in \{0,1,2\}$ of a gene kingdom K is

Table 2a
Mean occurrence frequencies $PrFr_f(t,K)$ of the 64 trinucleotides t in the three frames $f \in \{0,1,2\}$ of genes in bacteria $K=B$ (7,851,762 genes, 2,481,566,882 trinucleotides), eukaryotes $K=E$ (1,662,579 genes, 824,825,761 trinucleotides), plasmids $K=P$ (237,486 genes, 68,244,356 trinucleotides) and viruses $K=V$ (184,344 genes, 45,688,798 trinucleotides) (Table 1). The codon usage is given by $PrFr_0(t,K)$ in reading frame $f=0$ of genes.

	Bacteria B			Eukaryotes E			Plasmids P			Viruses V		
Frame t	$f=0$ $PrFr_0$	$f=1$ $PrFr_1$	$f=2$ $PrFr_2$	$f=0$ $PrFr_0$	$f=1$ $PrFr_1$	$f=2$ $PrFr_2$	$f=0$ $PrFr_0$	$f=1$ $PrFr_1$	$f=2$ $PrFr_2$	$f=0$ $PrFr_0$	$f=1$ $PrFr_1$	$f=2$ $PrFr_2$
AAA	2.87	2.49	2.22	2.73	2.23	2.32	2.15	1.93	1.75	3.61	2.77	2.83
AAC	1.79	1.46	1.14	2.00	1.38	1.14	1.77	1.35	0.99	2.46	1.61	1.51
AAG	1.97	2.74	0.75	3.21	2.72	1.24	2.16	2.10	0.68	2.58	2.92	0.93
AAT	1.93	1.26	1.60	2.11	1.19	1.43	1.69	1.09	1.24	2.71	1.70	1.96
ACA	1.00	1.66	1.03	1.55	2.52	1.35	0.93	1.76	0.90	1.65	2.51	1.36
ACC	2.12	1.65	0.73	1.56	1.76	1.10	2.22	1.74	0.71	1.62	1.61	1.00
ACG	1.39	2.46	0.67	0.78	1.40	0.47	1.66	2.56	0.71	1.20	2.67	0.67
ACT	0.90	1.02	1.07	1.44	1.49	1.14	0.75	0.96	1.03	1.64	1.41	1.10
AGA	0.54	1.55	2.25	1.36	2.94	3.09	0.52	1.75	1.78	1.08	2.00	3.02
AGC	1.39	1.97	1.53	1.70	2.36	1.66	1.49	2.07	1.33	1.07	1.54	1.50
AGG	0.32	2.34	1.27	1.17	2.98	1.45	0.45	2.67	1.03	0.56	2.11	1.34
AGT	0.81	0.92	1.11	1.33	1.42	1.34	0.73	0.89	0.84	1.09	1.33	1.37
ATA	0.89	1.87	0.58	0.99	1.56	0.57	0.77	1.50	0.51	1.43	2.64	1.00
ATC	2.71	1.61	0.73	1.89	1.40	1.13	3.03	1.73	0.73	2.17	1.52	1.11
ATG	2.34	2.53	0.39	2.24	3.07	0.60	2.29	2.38	0.43	2.45	2.96	0.66
ATT	2.43	1.40	1.49	1.88	1.32	1.30	1.72	1.19	1.16	2.50	1.82	1.84
CAA	1.61	1.20	2.53	1.59	1.68	3.09	1.33	1.08	2.65	1.83	1.75	2.99
CAC	1.05	0.86	1.76	1.32	1.27	1.90	1.09	0.89	1.96	1.07	1.04	1.75
CAG	2.18	1.45	1.11	2.80	2.40	1.97	2.35	1.33	1.28	1.72	1.78	1.05
CAT	1.06	0.71	2.24	1.20	1.12	2.40	1.07	0.70	2.42	0.98	1.13	2.08
CCA	0.77	1.88	1.51	1.69	2.53	1.93	0.70	1.99	1.63	1.10	1.66	1.27
CCC	1.08	1.38	1.08	1.47	1.58	1.50	1.29	1.56	1.14	0.92	0.95	1.06
CCG	1.88	3.29	1.57	0.83	1.29	1.00	2.10	3.81	1.94	1.21	2.03	1.03
CCT	0.81	1.18	1.91	1.66	1.47	2.05	0.71	1.40	2.21	1.13	0.83	1.36
CGA	0.42	1.52	3.87	0.65	0.70	1.70	0.57	1.83	4.44	0.60	1.22	3.23
CGC	2.25	2.38	3.17	0.95	0.92	1.13	2.67	2.78	3.72	1.36	1.30	2.04
CGG	1.08	2.67	3.00	0.88	1.18	1.08	1.47	3.09	3.37	0.66	1.63	1.89
CGT	1.10	0.86	2.11	0.59	0.57	1.05	0.95	0.86	2.39	1.08	0.96	1.67
CTA	0.56	0.93	1.05	0.78	1.15	0.94	0.51	0.69	1.16	0.84	1.71	0.98
CTC	1.73	0.74	0.99	1.75	1.63	1.68	2.19	0.72	1.09	1.28	1.11	1.00
CTG	3.66	1.53	0.93	2.85	2.81	1.19	3.92	1.39	1.10	1.91	2.19	0.79
CTT	1.28	0.96	1.66	1.55	1.31	1.94	1.31	0.74	1.72	1.27	1.11	1.44
GAA	3.47	0.72	1.87	3.21	1.17	2.93	2.99	0.69	1.96	3.58	0.96	2.29
GAC	2.63	0.44	1.44	2.40	0.67	1.33	2.90	0.51	1.76	2.92	0.57	1.37
GAG	2.62	0.65	0.65	3.68	1.33	1.52	2.88	0.64	0.77	2.67	0.94	0.81
GAT	2.80	0.39	2.36	2.73	0.58	1.66	2.56	0.42	2.53	3.22	0.64	1.82
GCA	1.69	1.86	2.02	1.75	1.84	1.94	1.66	2.12	2.09	1.79	1.47	1.45
GCC	3.54	1.61	1.63	2.25	1.40	1.53	4.17	1.89	1.86	2.04	0.96	1.12
GCG	3.06	2.87	1.83	1.03	1.08	0.91	3.43	3.20	2.17	1.72	1.83	1.13
GCT	1.60	1.20	3.00	2.04	1.22	2.12	1.39	1.30	3.29	1.92	0.82	1.54
GGA	1.23	0.83	2.63	1.75	1.42	3.55	1.17	0.98	2.81	1.43	0.93	2.41
GGC	3.35	1.27	3.00	1.85	1.11	2.08	3.69	1.60	3.47	1.95	0.80	1.94
GGG	1.22	1.24	1.70	1.35	1.37	1.49	1.33	1.44	1.79	0.95	1.06	1.24
GGT	1.76	0.54	2.33	1.40	0.67	1.62	1.38	0.61	2.36	2.11	0.71	1.57
GTA	1.08	0.95	0.59	0.82	0.81	0.67	0.80	0.69	0.52	1.41	1.41	0.91
GTC	2.04	0.79	0.86	1.43	0.86	1.18	2.50	0.77	0.99	1.54	0.89	1.05
GTG	2.58	1.45	0.51	2.34	1.83	0.80	2.46	1.25	0.54	1.81	1.80	0.77
GTT	1.52	0.86	1.67	1.55	0.78	1.40	1.25	0.67	1.53	1.96	0.96	1.73
TAA	0.00	1.29	1.94	0.00	1.06	1.70	0.00	0.95	1.42	0.00	1.75	3.24
TAC	1.32	0.75	1.07	1.45	0.75	0.96	1.26	0.60	0.86	1.76	0.91	1.47
TAG	0.00	1.32	0.55	0.00	1.09	0.82	0.00	0.91	0.46	0.00	1.60	1.01
TAT	1.62	0.81	1.92	1.31	0.71	1.32	1.47	0.62	1.32	2.04	1.13	2.34
TCA	0.77	2.23	1.35	1.48	2.47	1.68	0.70	2.45	1.22	1.16	2.22	1.53
TCC	0.99	1.43	1.08	1.59	1.75	1.52	1.08	1.74	1.09	0.90	1.21	1.19
TCG	1.10	3.53	0.78	0.73	1.19	0.68	1.36	4.34	0.85	0.97	2.30	0.86
TCT	0.86	1.23	1.26	1.74	1.57	1.61	0.69	1.41	1.40	1.42	1.14	1.29
TGA	0.01	2.43	2.78	0.00	2.38	3.68	0.00	2.46	2.30	0.00	2.14	3.73
TGC	0.56	2.87	2.18	1.03	2.11	2.20	0.65	2.95	2.13	0.70	1.60	1.99
TGG	1.25	3.41	1.59	1.18	3.21	2.33	1.38	3.24	1.38	1.34	2.36	1.97
TGT	0.38	1.31	1.63	0.96	1.39	2.14	0.33	1.22	1.37	0.78	1.45	2.06
TTA	1.64	1.72	0.92	1.02	1.29	0.70	1.01	1.18	0.70	1.71	2.30	1.16
TTC	1.94	1.33	1.15	1.91	1.61	1.56	2.23	1.34	1.02	1.87	1.35	1.30
TTG	1.44	2.66	0.49	1.65	2.65	0.66	1.21	2.15	0.47	1.39	2.79	0.75
TTT	2.01	1.54	2.19	1.86	1.31	1.80	1.52	1.13	1.54	2.18	1.47	2.14

Table 2b
Median occurrence frequencies $MdPrFr_f(t,K)$ of the 64 trinucleotides t in the three frames $f \in \{0,1,2\}$ of the 25 gene taxonomic groups $G=B_i$ in bacteria $K=B$ (7,851,762 genes, 2,481,566,882 trinucleotides), the 11 taxonomic groups $G=E_i$ in eukaryotes $K=E$ (1,662,579 genes, 824,825,761 trinucleotides), the 11 taxonomic groups $G=P_i$ in plasmids $K=P$ (237,486 genes, 68,244,356 trinucleotides) and the six taxonomic groups $G=V_i$ in viruses $K=V$ (184,344 genes, 45,688,798 trinucleotides) (Table 1).

	Bacteria B			Eukaryotes E			Plasmids P			Viruses V		
Frame t	$f=0$ $MdPrFr_0$	$f=1$ $MdPrFr_1$	$f=2$ $MdPrFr_2$	$f=0$ $MdPrFr_0$	$f=1$ $MdPrFr_1$	$f=2$ $MdPrFr_2$	$f=0$ $MdPrFr_0$	$f=1$ $MdPrFr_1$	$f=2$ $MdPrFr_2$	$f=0$ $MdPrFr_0$	$f=1$ $MdPrFr_1$	$f=2$ $MdPrFr_2$
AAA	3.94	3.43	2.94	2.78	2.07	2.06	3.84	3.62	2.91	3.39	2.56	2.78
AAC	1.83	1.50	1.49	2.05	1.43	1.14	1.74	1.43	1.84	2.19	1.65	1.69
AAG	2.06	3.50	1.15	3.21	2.76	1.15	2.36	3.65	1.06	2.68	2.79	1.19
AAT	2.34	1.54	2.32	1.82	1.11	1.30	3.06	1.77	2.28	2.77	1.75	2.09
ACA	1.24	1.64	1.06	1.54	2.39	1.35	1.49	1.80	1.08	1.88	2.42	1.55
ACC	1.76	1.26	0.82	1.51	1.43	1.11	1.67	1.47	0.77	1.38	1.46	1.32
ACG	1.15	1.98	0.63	0.96	2.06	0.52	1.04	1.86	0.70	1.17	2.01	0.74
ACT	1.08	1.02	1.25	1.39	1.50	1.11	1.34	1.05	1.31	1.81	1.38	1.23
AGA	1.02	1.67	3.27	1.38	2.53	2.97	0.94	1.66	3.51	1.54	2.11	2.93
AGC	1.12	1.46	1.70	1.57	2.39	1.62	1.19	1.55	1.75	1.05	1.50	1.80
AGG	0.40	2.30	1.70	1.12	3.13	1.48	0.45	2.16	1.52	1.03	2.05	1.57
AGT	0.90	0.95	1.40	1.23	1.44	1.29	1.27	0.97	1.37	1.20	1.31	1.67
ATA	1.63	2.17	0.69	0.88	1.37	0.49	1.14	2.49	0.70	1.58	2.43	1.01
ATC	2.11	1.70	0.92	2.02	1.36	1.14	1.96	1.75	0.91	1.80	1.71	1.30
ATG	2.19	2.51	0.49	2.28	2.92	0.53	1.88	2.51	0.49	2.45	3.28	0.76
ATT	2.68	1.91	1.65	1.69	1.09	1.17	2.85	2.00	2.11	2.41	1.91	1.92
CAA	1.43	1.41	2.20	1.70	1.53	3.10	1.27	1.76	2.44	2.11	2.18	2.84
CAC	0.75	0.82	1.21	1.46	1.28	1.98	0.86	1.00	1.28	1.03	1.32	1.61
CAG	1.73	1.66	1.16	1.80	1.90	1.36	1.88	1.66	1.06	1.80	2.19	1.14
CAT	0.99	0.69	2.07	1.14	0.97	2.25	1.03	0.86	2.07	1.33	1.34	2.03
CCA	0.83	1.83	1.13	1.64	2.24	1.53	0.89	1.70	1.52	1.65	1.77	1.31
CCC	1.08	1.11	0.95	1.43	1.24	1.25	1.20	1.34	0.90	0.95	0.98	1.10
CCG	0.91	2.00	0.97	0.93	1.29	0.93	0.69	1.64	0.96	0.88	1.26	0.72
CCT	1.13	1.21	1.51	1.34	1.30	1.67	0.94	1.05	1.32	1.45	0.91	1.24
CGA	0.43	0.95	2.48	0.66	1.15	2.48	0.61	1.00	2.89	0.63	1.17	2.07
CGC	1.02	0.95	1.41	0.88	1.09	1.37	1.08	0.96	1.07	0.83	0.92	1.38
CGG	0.77	1.73	1.84	0.72	1.26	1.17	0.76	1.41	1.19	0.52	1.16	1.13
CGT	0.84	0.56	1.52	0.69	0.83	1.17	0.92	0.58	1.19	1.16	0.84	1.24
CTA	0.60	1.02	0.98	0.73	0.97	0.89	0.95	1.45	1.09	1.00	1.87	0.87
CTC	1.53	0.91	0.85	1.65	1.66	1.64	1.28	0.87	1.01	1.25	1.40	1.19
CTG	1.98	1.81	0.65	1.98	2.54	0.72	1.68	1.71	0.67	1.40	2.44	0.80
CTT	2.05	1.36	1.69	1.31	1.17	1.82	1.57	1.41	1.71	1.41	1.36	1.48
GAA	4.32	1.22	1.96	3.28	1.10	2.96	4.14	1.08	1.97	3.34	1.14	2.44
GAC	2.01	0.55	1.00	2.50	0.70	1.39	2.09	0.54	1.02	2.33	0.71	1.33
GAG	2.58	1.11	0.73	3.97	1.34	1.45	2.48	1.08	0.76	2.48	1.24	1.04
GAT	2.99	0.61	2.02	2.67	0.59	1.69	3.14	0.54	1.91	3.23	0.77	1.78
GCA	1.88	1.41	1.49	1.64	1.64	2.06	1.91	1.46	1.63	1.93	1.38	1.42
GCC	2.46	0.88	0.99	2.07	0.99	1.48	2.07	0.86	0.92	1.55	0.83	1.10
GCG	1.53	1.36	0.86	1.16	1.24	0.83	1.13	1.21	0.90	1.24	1.02	0.84
GCT	1.84	0.84	1.70	1.99	0.93	2.36	1.92	0.95	1.71	2.12	0.81	1.29
GGA	1.76	0.94	2.60	1.72	1.34	3.89	1.76	0.98	2.16	1.67	1.23	2.38
GGC	1.85	0.85	1.97	1.75	0.93	1.93	1.75	0.84	1.89	1.25	0.78	1.52
GGG	1.19	1.19	1.56	1.23	1.22	1.51	1.16	1.27	1.36	1.08	1.09	1.13
GGT	1.76	0.56	1.71	1.36	0.61	1.57	1.36	0.63	1.60	1.92	0.76	1.41
GTA	1.47	1.03	0.59	0.81	0.77	0.70	1.38	1.10	0.65	1.41	1.36	0.88
GTC	1.24	0.69	0.60	1.58	0.87	1.23	1.18	0.75	0.65	1.39	0.85	1.04
GTG	1.88	1.39	0.40	2.33	1.73	0.74	1.74	1.47	0.41	1.67	1.87	0.70
GTT	2.01	0.88	1.48	1.36	0.73	1.42	1.92	0.80	1.61	2.05	1.02	1.67
TAA	0.00	1.82	2.64	0.00	0.99	1.42	0.00	1.97	3.07	0.00	1.79	2.88
TAC	1.34	0.91	1.32	1.59	0.68	0.88	1.22	0.98	1.16	1.62	1.09	1.52
TAG	0.00	1.99	0.75	0.00	0.91	0.69	0.00	1.91	0.96	0.00	1.78	1.20
TAT	1.92	1.18	2.43	1.16	0.67	1.17	2.00	1.18	2.13	2.09	1.15	2.11
TCA	0.90	1.97	1.35	1.38	2.53	1.60	1.09	2.08	1.46	1.54	2.22	1.74
TCC	0.96	1.16	1.35	1.58	1.38	1.48	0.81	1.02	1.25	1.06	1.24	1.38
TCG	0.71	1.90	0.64	1.08	1.90	0.73	0.83	1.60	0.87	0.79	1.62	0.88
TCT	1.25	1.23	1.59	1.67	1.47	1.51	1.25	1.34	1.48	1.53	1.21	1.41
TGA	0.00	2.27	3.29	0.00	2.35	3.55	0.00	2.07	3.05	0.00	2.16	4.01
TGC	0.46	1.80	2.23	1.13	2.24	2.22	0.46	1.82	2.15	0.77	1.49	2.08
TGG	1.11	2.69	1.78	1.15	3.02	2.02	1.36	2.47	1.73	1.35	2.21	2.12
TGT	0.47	1.03	2.00	0.98	1.36	2.04	0.58	1.04	1.75	0.90	1.42	2.21
TTA	2.06	2.25	1.21	0.81	1.13	0.62	2.06	2.51	1.30	1.86	2.25	1.20
TTC	1.75	1.69	1.55	1.96	1.45	1.36	1.68	1.55	1.49	1.88	1.52	1.44
TTG	1.44	3.22	0.50	1.63	2.24	0.56	1.42	2.92	0.45	1.31	2.85	0.83
TTT	2.98	2.37	3.33	1.72	1.13	1.55	2.83	2.10	3.04	2.20	1.46	2.24

simply defined by

$$PrPrefFr_f(t,K)=\frac{1}{\mathrm{Card}(K)}\sum_{i=1}^{\mathrm{Card}(K)}\delta_f(PrefFr(t,G_i)). \quad (4)$$

Remark 4. $\sum_{f=0,1,2} PrPrefFr_f(t,K)=1.$

Proposition 1. *If $PrPrefFr_f(t,K)=1/3$ for the three frames $f\in\{0,1,2\}$ then the trinucleotide t has no preferential frame in the gene kingdom K, i.e. t occurs in the three frames equiprobably in K.*

2.5. Occurrence probability of a complementary/permutation trinucleotide set in a gene kingdom

The class of C^3 self-complementary trinucleotide circular codes C^3SCC (Definition 8) is included in a larger class of codes C^3SC by relaxing the circularity property. Precisely, a new definition of a class of codes is given here on which the developed statistical approach will be based.

Definition 9. A trinucleotide code $C\subset A_4^3$ is C^3 *self-complementary* and called C^3SC if C, $C_1=\mathcal{P}(C)$ and $C_2=\mathcal{P}^2(C)$ are trinucleotide codes satisfying the following properties $C=\mathcal{C}(C)$ (self-complementary), $\mathcal{C}(C_1)=C_2$ and $\mathcal{C}(C_2)=C_1$ (C_1 and C_2 are complementary).

In order to study the C^3 self-complementary codes C^3SC including the class of circular codes C^3SCC, Eq. (4) defined for a trinucleotide is extended to a set T of six trinucleotides related to the complementarity and permutation maps $\mathcal{C}$ and $\mathcal{P}$ simultaneously, precisely $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ with $t\in A_4^3\backslash PPT=\{\mathrm{AAA},\mathrm{CCC},\mathrm{GGG},\mathrm{TTT}\}$ and $\{t,\mathcal{C}(t)\}$ in frame 0, $\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\}$ in frame 1 and $\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}$ in frame 2.

Remark 5. $\mathcal{P}(t)=\mathcal{C}(\mathcal{P}^2(\mathcal{C}(t)))$ and $\mathcal{P}^2(t)=\mathcal{C}(\mathcal{P}(\mathcal{C}(t)))$.

Then, the occurrence probability $PrCP(T,K)$ of a complementary and permutation (CP) trinucleotide set $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ in a gene kingdom K is equal to

$$\begin{aligned}PrCP(T,K)=\frac{1}{6}[&PrPrefFr_0(t,K)+PrPrefFr_0(\mathcal{C}(t),K)+PrPrefFr_1(\mathcal{P}(t),K)\\&+PrPrefFr_1(\mathcal{P}(\mathcal{C}(t)),K)+PrPrefFr_2(\mathcal{P}^2(t),K)\\&+PrPrefFr_2(\mathcal{P}^2(\mathcal{C}(t)),K)]\end{aligned} \quad (5)$$

where $PrPrefFr_f(t,K)$ is defined in Eq. (4).

Proposition 2. *If $PrCP(T,K)=1/3$ then the complementary and permutation (CP) trinucleotide set T has no preferential frame in the gene kingdom K.*

When the trinucleotide t is given then the trinucleotide $\mathcal{C}(t)$ is also known. Thus, there are $60/2=30$ CP trinucleotide sets noted $T_1,\ldots,T_{30}$ where $T_i=\{\{t,\mathcal{C}(t)\}_i,\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\}_i,\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}_i\}$ with $\{t,\mathcal{C}(t)\}_i$ in frame 0, $\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\}_i$ in frame 1 and $\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}_i$ in frame 2. A maximal (20 trinucleotides) C^3 self-complementary code C^3SC is identified with the 10 first values of the occurrence probabilities $PrCP(T_1,K),\ldots,PrCP(T_{10},K)$. Precisely, the code C^3SC has 20 trinucleotides $C^3SC=C^3SC_0=\{\{t,\mathcal{C}(t)\}_1,\ldots,\{t,\mathcal{C}(t)\}_{10}\}$ in frame 0, 20 trinucleotides $C^3SC_1=\mathcal{P}(C^3SC)=\{\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\}_1,\ldots,\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\}_{10}\}$ in frame 1 and 20 trinucleotides $C^3SC_2=\mathcal{P}^2(C^3SC)=\{\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}_1,\ldots,\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}_{10}\}$ in frame 2 with $C^3SC=\mathcal{C}(C^3SC)$, $\mathcal{C}(C^3SC_1)=C^3SC_2$ and $\mathcal{C}(C^3SC_2)=C^3SC_1$. Only 216 C^3 self-complementary trinucleotide codes C^3SC among $\binom{30}{10}=30,045,015$ are circular C^3SCC.

Notation 3. A complementary and permutation (CP) trinucleotide set $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ is said to belong to the C^3 self-complementary trinucleotide circular code X, i.e. $T\in X$, if $\{t,\mathcal{C}(t)\}\cap X\neq\varnothing$. Ten among the 30 CP trinucleotide sets T belong to the C^3 circular code X, i.e. such that $\{t,\mathcal{C}(t)\}\in X$, $\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\}\in\mathcal{P}(X)=X_1$ and $\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\in\mathcal{P}^2(X)=X_2$.

2.6. A probabilistic model for estimating a code probability in a shifted frame of a gene kingdom

The probabilistic model is applied here to the four codes $X=X_0$, $X_1=\mathcal{P}(X)$, $X_2=\mathcal{P}^2(X)$ and $PPT=\{\mathrm{AAA},\mathrm{CCC},\mathrm{GGG},\mathrm{TTT}\}$. The four studied codes C can be investigated by the preferential frame approach developed here. The preferential frame occurrence probability $PrPrefFr_f(C,K)$ of a trinucleotide code C in a frame $f\in\{0,1,2\}$ of a gene kingdom K is obviously equal to

$$PrPrefFr_f(C,K)=\frac{1}{\mathrm{Card}(C)}\sum_{t\in C}PrPrefFr_f(t,K) \quad (6)$$

where $PrPrefFr_f(t,K)$ is defined in Eq. (4).

These four studied codes C can also be analysed by the frequency approach. The occurrence probability $PrFr_f(C,K)$ of a trinucleotide code C in a frame $f\in\{0,1,2\}$ of a gene kingdom K is obviously equal to

$$PrFr_f(C,K)=\frac{1}{\mathrm{Card}(C)}\sum_{t\in C}PrFr_f(t,K) \quad (7)$$

where $PrFr_f(t,K)$ is defined in Eq. (2).

Eqs. (6) and (7) rely on the same frame concept.

A simple probabilistic model based on the independent occurrence of trinucleotides in reading frame $f=0$ can estimate the real circular code probabilities $PrPrefFr_f(C,K)$ (Eq. (6)) and $PrFr_f(C,K)$ (Eq. (7)) observed in the shifted frames $f\in\{1,2\}$ of a gene kingdom K. Indeed, the estimated trinucleotide probabilities $\widehat{PrFr}_f(t,K)$ for the two shifted frames $f\in\{1,2\}$ of a gene kingdom K are obtained from the product of two trinucleotide probabilities $PrFr_0(t',K)$ and $PrFr_0(t'',K)$ in frame 0 with the $64\times64=4096$ di-trinucleotides $t't''\in A_4^3\times A_4^3$ and with the simplest hypothesis of independent events. The method here is very similar to the method developed in Michel (2014). Let $t=l_0l_1l_2\in A_4^3$ be a trinucleotide occurring with a frequency $PrFr_0(t,K)$ in frame $f=0$ of a gene kingdom K such that $\sum_{t\in A_4^3}PrFr_0(t,K)=1$. By convention, the reading frame $f=0$ is established by the letter l_0 of $t=l_0l_1l_2$. The frames $f=1$ and $f=2$ start with the letters l_1 and l_2, respectively, of t. Let the di-trinucleotide w be a concatenation of two trinucleotides $t'=l_0'l_1'l_2'\in A_4^3$ and $t''=l_0''l_1''l_2''\in A_4^3$, i.e. $w=t't''\in A_4^3\times A_4^3$. We denote by $t_1(w)=l_1'l_2'l_0''\in A_4^3$ and $t_2(w)=l_2'l_0''l_1''\in A_4^3$ the trinucleotides in frames 1 and 2, respectively, of a di-trinucleotide w. The concatenation of the two trinucleotides t' and t'' yield a trinucleotide $t_f(w)$ in a shifted frame $f\in\{1,2\}$. For example, the concatenation of the trinucleotides $t'=$ ACG and $t''=$ TAC, i.e. $w=$ ACGTAC, leads to the trinucleotides $t_1(w)=$ CGT in frame $f=1$ and $t_2(w)=$ GTA in frame $f=2$.

The estimated probability $\widehat{PrFr}_f(t_f(w),K)$ of a trinucleotide $t_f(w)\in A_4^3$ in a frame $f\in\{1,2\}$ of a di-trinucleotide $w=t't''\in A_4^3\times$

A_4^3 in a gene kingdom K is equal to the product of probabilities $PrFr_0(t',K)$ and $PrFr_0(t'',K)$ in frame 0 (with the simplest hypothesis of independent events)

$$\widehat{PrFr}_f\left(t_f(w),K\right)=PrFr_0(t',K)\times PrFr_0(t'',K).$$

Then, the estimated probability $\widehat{PrFr}_f(t,K)$ of a trinucleotide $t\in A_4^3$ in a frame $f\in\{1,2\}$ of a gene kingdom K is equal to the probability sum obtained with all the di-trinucleotides $w=t't''\in A_4^3\times A_4^3$

$$\widehat{PrFr}_f(t,K)=\sum_{w\in A_4^3\times A_4^3\ |\ t_f(w)=t}\widehat{PrFr}_f\left(t_f(w),K\right).$$

Finally, the estimated probability $\widehat{PrFr}_f(C,K)$ of a trinucleotide code C in a frame $f\in\{1,2\}$ of a gene kingdom K is equal to

$$\widehat{PrFr}_f(C,K)=\sum_{t\in C}\widehat{PrFr}_f(t,K). \tag{8}$$

2.7. Identification of variant X codes

Section 3 will show that the C^3 self-complementary circular code X is again identified in average in genes of prokaryotes, eukaryotes, plasmids and almost completely in genes of viruses.

However, as already observed in 1996, a few trinucleotides among the $60\times 3=180$ trinucleotides are poorly assigned in some frames of genes, mainly GTG and TGG in prokaryotes and AAG, GTG, TGC and TGG in eukaryotes (Arquès and Michel, 1996, Tables 1a and 1b, respectively). Gonzalez et al. (2011) have also observed this variability of the circular code X by analysing its covering capability among the 216 C^3 self-complementary trinucleotide circular codes C^3SCC. A third approach based on the probability (efficiency) of reading frame coding (RFC) of usage of the circular code X also confirmed this variability of X (Michel, 2015). Indeed, the highest RFC probabilities of usage of X are identified in bacterial plasmids and bacteria (about 49.0%), then, by decreasing values, in viruses (45.4%) and nuclear eukaryotes (42.8%) (Michel, 2015).

The statistical approach developed in Michel (2015) allows to quantify the usage of X (RFC probability) in genes. It does not propose another code, i.e. a subset of X which is always circular or a code Y different from X which can be circular or not, in the case where a low usage of X is observed in some genes, i.e. in the case where some trinucleotides of X are less significant for reading frame coding. The approach developed here allows to identify variant X codes in genes, i.e. trinucleotide codes which differ from the circular code X.

For each gene taxonomic group G in a given kingdom K, the number of correctly assigned trinucleotides (CAT) with respect to the frame is counted in a complementary and permutation (CP) trinucleotide set $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ (Section 2.5), precisely $\{t,\mathcal{C}(t)\}$ are in frame 0, $\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\}$ in frame 1 and $\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}$ in frame 2. Precisely, the numbers $NbCAT(T,G)$ and $MdNbCAT(T,G)$ of correctly assigned trinucleotides (CAT) with respect to the frame in a CP trinucleotide set $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ of a gene taxonomic group G are equal to

$$\begin{cases} NbCAT(T,G)=\delta_0(PrefFr(t,G))+\delta_0(PrefFr(\mathcal{C}(t),G))\\ \quad+\delta_1(PrefFr(\mathcal{P}(t),G))+\delta_1(PrefFr(\mathcal{P}(\mathcal{C}(t)),G))\\ \quad+\delta_2\left(PrefFr\left(\mathcal{P}^2(t),G\right)\right)+\delta_2\left(PrefFr\left(\mathcal{P}^2(\mathcal{C}(t)),G\right)\right)\\ MdNbCAT(T,G)=\delta_0(MdPrefFr(t,G))+\delta_0(MdPrefFr(\mathcal{C}(t),G))\\ \quad+\delta_1(MdPrefFr(\mathcal{P}(t),G))+\delta_1(MdPrefFr(\mathcal{P}(\mathcal{C}(t)),G))\\ \quad+\delta_2\left(MdPrefFr\left(\mathcal{P}^2(t),G\right)\right)+\delta_2\left(MdPrefFr\left(\mathcal{P}^2(\mathcal{C}(t)),G\right)\right)\end{cases} \tag{9}$$

where $\delta_f(PrefFr(t,G))$ and $\delta_f(MdPrefFr(t,G))$ are defined in Eq. (3).

Proposition 3. $0\le NbCAT(T,G)\le 6$ *and* $0\le MdNbCAT(T,G)\le 6$ *according to Eq. (9).*

Proposition 4.

$$\mathcal{P}(T)=\{\{\mathcal{P}(t),\mathcal{C}(\mathcal{P}(t))\},\{\mathcal{P}^2(t),\mathcal{P}(\mathcal{C}(\mathcal{P}(t)))\},\{t,\mathcal{P}^2(\mathcal{C}(\mathcal{P}(t)))\}\}$$
$$=\{\{\mathcal{P}(t),\mathcal{C}(\mathcal{P}(t))\},\{\mathcal{P}^2(t),\mathcal{C}(t)\},\{t,\mathcal{P}(\mathcal{C}(t))\}\}$$

and

$$\mathcal{P}^2(T)=\{\{\mathcal{P}^2(t),\mathcal{C}(\mathcal{P}^2(t))\},\{t,\mathcal{P}(\mathcal{C}(\mathcal{P}^2(t)))\},\{\mathcal{P}(t),\mathcal{P}^2(\mathcal{C}(\mathcal{P}^2(t)))\}\}$$
$$=\{\{\mathcal{P}^2(t),\mathcal{C}(\mathcal{P}^2(t))\},\{t,\mathcal{P}^2(\mathcal{C}(t))\},\{\mathcal{P}(t),\mathcal{C}(t)\}\}.$$

Proof. (i) For $\mathcal{P}(T)$, as $\mathcal{P}(t)=\mathcal{C}(\mathcal{P}^2(\mathcal{C}(t)))$ (see Remark 5), $\mathcal{P}(\mathcal{C}(\mathcal{P}(t)))=\mathcal{P}(\mathcal{C}(\mathcal{C}(\mathcal{P}^2(\mathcal{C}(t)))))=\mathcal{P}(\mathcal{P}^2(\mathcal{C}(t)))=\mathcal{C}(t)$ and $\mathcal{P}^2(\mathcal{C}(\mathcal{P}(t)))=\mathcal{P}^2(\mathcal{C}(\mathcal{C}(\mathcal{P}^2(\mathcal{C}(t)))))=\mathcal{P}^2(\mathcal{P}^2(\mathcal{C}(t)))=\mathcal{P}(\mathcal{C}(t))$.

(ii) For $\mathcal{P}^2(T)$, as $\mathcal{P}^2(t)=\mathcal{C}(\mathcal{P}(\mathcal{C}(t)))$ (see Remark 5), $\mathcal{P}(\mathcal{C}(\mathcal{P}^2(t)))=\mathcal{P}(\mathcal{C}(\mathcal{C}(\mathcal{P}(\mathcal{C}(t)))))=\mathcal{P}^2(\mathcal{C}(t))$ and $\mathcal{P}^2(\mathcal{C}(\mathcal{P}^2(t)))=\mathcal{P}^2(\mathcal{C}(\mathcal{C}(\mathcal{P}(\mathcal{C}(t)))))=\mathcal{C}(t)$.

Proposition 5.

$NbCAT(T,G)+NbCAT(\mathcal{P}(T),G)+NbCAT\left(\mathcal{P}^2(T),G\right)=6$ *by definition.*

A complementary trinucleotide pair $\{t,\mathcal{C}(t)\}$ of the circular code X is replaced by the complementary trinucleotide pair $\{\mathcal{P}(t),\mathcal{C}(\mathcal{P}(t))\}\notin X$ if

$$NbCAT(\mathcal{P}(T),G)\ge 4. \tag{10}$$

Similarly, a complementary trinucleotide pair $\{t,\mathcal{C}(t)\}$ of the circular code X is replaced by the complementary trinucleotide pair $\{\mathcal{P}^2(t),\mathcal{C}(\mathcal{P}^2(t))\}\notin X$ if

$$NbCAT\left(\mathcal{P}^2(T),G\right)\ge 4. \tag{11}$$

The chosen minimal number 4 means that four trinucleotides among six, i.e. a number strictly greater than the mean number 3, in a CP trinucleotide set T are correctly assigned with respect to the frame. If inequality (10) or (11) is verified then $NbCAT(T,G)<3$ according to Proposition 5.

Finally, the mean numbers $\overline{NbCAT}(T,K)$ and $\overline{MdNbCAT}(T,K)$ of correctly assigned trinucleotides (CAT) with respect to the frame in a CP trinucleotide set $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ of a gene kingdom K are equal to

$$\begin{cases}\overline{NbCAT}(T,K)=\frac{1}{\mathrm{Card}(K)}\sum_{i=1}^{\mathrm{Card}(K)}NbCAT(T,G_i)\\ \overline{MdNbCAT}(T,K)=\frac{1}{\mathrm{Card}(K)}\sum_{i=1}^{\mathrm{Card}(K)}MdNbCAT(T,G_i)\end{cases} \tag{12}$$

where $NbCAT(T,G_i)$ and $MdNbCAT(T,G_i)$ are defined in Eq. (9).

2.8. Explained example of the statistical approach developed

Table 3 gives an explained example of the statistical approach developed. Let a gene kingdom K with five groups G, i.e. $K=\{G_1,\ldots,G_5\}$. Let the CP trinucleotide set $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}=\{\{\mathrm{CAG},\mathrm{CTG}\},\{\mathrm{AGC},\mathrm{TGC}\},\{\mathrm{GCA},\mathrm{GCT}\}\}$ with $\{\mathrm{CAG},\mathrm{CTG}\}$ in frame 0, $\{\mathrm{AGC},\mathrm{TGC}\}$ in frame 1 and $\{\mathrm{GCA},\mathrm{GCT}\}$ in frame 2. Then, the CP occurrence probability $PrCP(T,K)$ of T in K is obtained as follows:

(i) Computation of the three occurrence frequencies $PrFr_f(t,G)$ of the six trinucleotides $t,\mathcal{C}(t),\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t)),\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\in T$ in the three frames $f\in\{0,1,2\}$ of each group G in kingdom K.

(ii) Determination with Eq. (2) of the preferential frame $PrefFr(t,G)\in\{0,1,2\}$ of the six trinucleotides t belonging to T in each group G of kingdom K. For example, AGC in G_1 has a preferential frame $PrefFr(\mathrm{AGC},G_1)=1$ as $PrFr_1(\mathrm{AGC},G_1)=1.96>\mathrm{Max}\{PrFr_0(\mathrm{AGC},G_1),PrFr_2(\mathrm{AGC},G_1)\}=1.53$ (%).

(iii) Determination with Eq. (4) of the three preferential frame occurrence probabilities $PrPrefFr_f(t,K)$ of the six trinucleotides t

Table 3

Explained example of the statistical approach developed. Gene kingdom K with five groups G, i.e. $K=\{G_1,\ldots,G_5\}$, and complementary and permutation (CP) trinucleotide set $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ $=\{\{\text{CAG},\text{CTG}\},\{\text{AGC},\text{TGC}\},\{\text{GCA},\text{GCT}\}\}$ with {CAG, CTG} in frame 0, {AGC, TGC} in frame 1 and {GCA, GCT} in frame 2. $PrFr_f(t,G)$ is the occurrence frequency (%) of a trinucleotide t belonging to T in a frame $f\in\{0,1,2\}$ of a group G in the kingdom K. $PrefFr(t,G)\in\{0,1,2\}$ (Eq. (2)) is the preferential frame of a trinucleotide t belonging to T in a group G of the kingdom K. $PrPrefFr_f(t,K)$ (Eq. (4)) is the preferential frame occurrence probability (%) of a trinucleotide t belonging to T in a frame $f\in\{0,1,2\}$ of the kingdom K. $PrCP(T,K)$ (Eq. (5)) is the occurrence probability (%) of the CP trinucleotide set T in the kingdom K.

K	Group G_1			Group G_2			Group G_3			Group G_4			Group G_5		
Frame	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$
t	$PrFr_0(t,G_1)$	$PrFr_1(t,G_1)$	$PrFr_2(t,G_1)$	$PrFr_0(t,G_2)$	$PrFr_1(t,G_2)$	$PrFr_2(t,G_2)$	$PrFr_0(t,G_3)$	$PrFr_1(t,G_3)$	$PrFr_2(t,G_3)$	$PrFr_0(t,G_4)$	$PrFr_1(t,G_4)$	$PrFr_2(t,G_4)$	$PrFr_0(t,G_5)$	$PrFr_1(t,G_5)$	$PrFr_2(t,G_5)$
AGC	1.39	1.96	1.53	1.74	2.39	1.68	1.18	1.50	1.58	1.07	1.53	1.50	1.49	2.07	1.33
CAG	2.18	1.45	1.11	2.88	2.44	2.04	1.45	1.81	1.26	1.72	1.78	1.04	2.35	1.33	1.28
CTG	3.66	1.53	0.93	2.94	2.81	1.25	1.85	1.80	0.77	1.91	2.19	0.78	3.91	1.39	1.10
GCA	1.69	1.86	2.02	1.74	1.87	1.96	2.02	1.15	0.96	1.79	1.47	1.45	1.66	2.11	2.08
GCT	1.60	1.20	3.00	2.02	1.25	2.12	1.59	0.81	2.16	1.92	0.82	1.54	1.39	1.30	3.28
TGC	0.56	2.87	2.18	1.06	2.13	2.17	0.45	1.31	1.87	0.70	1.60	1.98	0.65	2.95	2.13

K	Group G_1	Group G_2	Group G_3	Group G_4	Group G_5
t	Preferential frame $PrefFr(t,G_1)$	Preferential frame $PrefFr(t,G_2)$	Preferential frame $PrefFr(t,G_3)$	Preferential frame $PrefFr(t,G_4)$	Preferential frame $PrefFr(t,G_5)$
AGC	1	1	2	1	1
CAG	0	0	1	1	0
CTG	0	0	0	1	0
GCA	2	2	0	0	1
GCT	2	2	2	0	2
TGC	1	2	2	2	1

Kingdom K

Frame	$f=0$	$f=1$	$f=2$
t	$PrPrefFr_0(t,K)$	$PrPrefFr_1(t,K)$	$PrPrefFr_2(t,K)$
AGC	0	80	20
CAG	60	40	0
CTG	80	20	0
GCA	40	20	40
GCT	20	0	80
TGC	0	40	60

$PrPrefFr_0(t,K)$	$PrPrefFr_0(\mathcal{C}(t),K)$	$PrPrefFr_1(\mathcal{P}(t),K)$	$PrPrefFr_1(\mathcal{P}(\mathcal{C}(t)),K)$	$PrPrefFr_2(\mathcal{P}^2(t),K)$	$PrPrefFr_2(\mathcal{P}^2(\mathcal{C}(t)),K)$	
CAG	CTG	AGC	TGC	GCA	GCT	$PrCP(T,K)$
60	80	80	40	40	80	$=380/6\approx 63$

belonging to T in the three frames $f \in \{0,1,2\}$ of kingdom K. For example, AGC in K has the preferential frame occurrence probabilities $PrPrefFr_0(\mathrm{AGC},K)=0$ in frame 0 (zero preferential frame $PrefFr(\mathrm{AGC},K)=0$ in K), $PrPrefFr_1(\mathrm{AGC},K)=80\%$ in frame 1 (four preferential frames $PrefFr(\mathrm{AGC},K)=1$ among 5 in K) and $PrPrefFr_2(\mathrm{AGC},K)=20\%$ in frame 2 (one preferential frame $PrefFr(\mathrm{AGC},K)=2$ among 5).

(iv) Computation with Eq. (5) of the CP occurrence probability $PrCP(T,K)=\frac{1}{6}(PrPrefFr_0(\mathrm{CAG},K)+PrPrefFr_0(\mathrm{CTG},K)+PrPrefFr_1(\mathrm{AGC},K)+PrPrefFr_1(\mathrm{TGC},K)+PrPrefFr_2(\mathrm{GCA},K)+PrPrefFr_2(\mathrm{GCT},K))=\frac{1}{6}(60+80+80+40+40+80)=\frac{380}{6}\approx 63\%$.

The number $NbCAT(T,G)$ of correctly assigned trinucleotides (CAT) with respect to the frame in the above chosen CP trinucleotide set $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}=\{\{\mathrm{CAG},\mathrm{CTG}\},\{\mathrm{AGC},\mathrm{TGC}\},\{\mathrm{GCA},\mathrm{GCT}\}\}$ is given for the selected group G_3: $NbCAT(T,G_3)=\delta_0(PrefFr(\mathrm{CAG},G_3))+\delta_0(PrefFr(\mathrm{CTG},G_3))+\delta_1(PrefFr(\mathrm{AGC},G_3))+\delta_1(PrefFr(\mathrm{TGC},G_3))+\delta_2(PrefFr(\mathrm{GCA},G_3))+\delta_2(PrefFr(\mathrm{GCT},G_3))=0+1+0+0+0+1=2<4$. Thus, the complementary pair $\{t,\mathcal{C}(t)\}=\{\mathrm{CAG},\mathrm{CTG}\}$ is replaced either by the complementary pair $\{\mathcal{P}(t),\mathcal{C}(\mathcal{P}(t))\}=\{\mathrm{AGC},\mathrm{GCT}\}$ or $\{\mathcal{P}^2(t),\mathcal{C}(\mathcal{P}^2(t))\}=\{\mathrm{GCA},\mathrm{TGC}\}$. The number $NbCAT(\mathcal{P}(T),G_3)$ in the CP trinucleotide set $\mathcal{P}(T)=\{\mathcal{P}(t),\mathcal{C}(\mathcal{P}(t))\},\{\mathcal{P}^2(t),\mathcal{C}(t)\},\{t,\mathcal{P}(\mathcal{C}(t))\}\}=\{\{\mathrm{AGC},\mathrm{GCT}\},\{\mathrm{GCA},\mathrm{CTG}\},\{\mathrm{CAG},\mathrm{TGC}\}\}$ (Proposition 4; trinucleotides {GCA, CTG} in non-lexicographical order for easier reading) in G_3 is equal to $NbCAT(\mathcal{P}(T),G_3)=\delta_0(PrefFr(\mathrm{AGC},G_3))+\delta_0(PrefFr(\mathrm{GCT},G_3))+\delta_1(PrefFr(\mathrm{GCA},G_3))+\delta_1(PrefFr(\mathrm{CTG},G_3))+\delta_2(PrefFr(\mathrm{CAG},G_3))+\delta_2(PrefFr(\mathrm{TGC},G_3))=0+0+0+0+0+1=1<4$. The number $NbCAT(\mathcal{P}^2(T),G_3)$ in the CP trinucleotide set $\mathcal{P}^2(T)=\{\{\mathcal{P}^2(t),\mathcal{P}(\mathcal{C}(t))\},\{t,\mathcal{P}^2(\mathcal{C}(t))\},\{\mathcal{P}(t),\mathcal{C}(t)\}\}=\{\{\mathrm{GCA},\mathrm{TGC}\},\{\mathrm{CAG},\mathrm{GCT}\},\{\mathrm{AGC},\mathrm{CTG}\}\}$ (Proposition 4) in G_3 is equal to $NbCAT(\mathcal{P}^2(T),G_3)=\delta_0(PrefFr(\mathrm{GCA},G_3))+\delta_0(PrefFr(\mathrm{TGC},G_3))+\delta_1(PrefFr(\mathrm{CAG},G_3))+\delta_1(PrefFr(\mathrm{GCT},G_3))+\delta_2(PrefFr(\mathrm{AGC},G_3))+\delta_2(PrefFr(\mathrm{CTG},G_3))=1+0+1+0+1+0=3<4$. Even if $NbCAT(\mathcal{P}^2(T),G_3)=3>\mathrm{Max}\{NbCAT(T,G_3),NbCAT(\mathcal{P}(T),G_3)\}=2$, the number $NbCAT(\mathcal{P}^2(T),G_3)$ of correctly assigned trinucleotides (CAT) with respect to the frame is considered by this approach as being not significant (not strictly greater than the mean number 3). In this example, the complementary pair $\{t,\mathcal{C}(t)\}=\{\mathrm{CAG},\mathrm{CTG}\}$ is removed from the trinucleotide code and not replaced, i.e. leading to a subset of the code and not to a variant code.

3. Results

3.1. C^3 self-complementary circular code X in genes of bacteria

The two indicators $PrefFr(t,B)$ and $MdPrefFr(t,B)$ (Eq. (2)) show very close results for the trinucleotide assignment of a preferential frame in the bacterial gene kingdom $K=B$ (Tables 2a and 2b). Indeed, 54 trinucleotides among 64, i.e. about 84%, have the same preferential frame with $PrefFr(t,B)$ and $MdPrefFr(t,B)$. The 10 trinucleotides with a different preferential frame in B are AGC, CTA, CTT, GCA, GCT, **GGC**, **GGT**, **GTT**, TCC and TGC (the trinucleotides of X being in bold). Seventeen trinucleotides of X are assigned to the frame 0 with both indicators $PrefFr(t,B)$ and $MdPrefFr(t,B)$. The three trinucleotides of X having a different preferential frame assignment in B are **GGC** in frame 0 with $PrefFr(t,B)$ and in frame 2 with $MdPrefFr(t,B)$, **GGT** in frame 2 with $PrefFr(t,B)$ and in frame 0 with $MdPrefFr(t,B)$, and **GTT** in frame 2 with $PrefFr(t,B)$ and in frame 0 with $MdPrefFr(t,B)$. The indicator $MdPrefFr(t,B)$ identifies directly 19 trinucleotides of X among 20 in frame 0.

Table 4a gives the preferential frame occurrence probabilities $PrPrefFr_f(t,B)$ (Eq. (4)) of the 60 trinucleotides (without $PPT=\{\mathrm{AAA},\mathrm{CCC},\mathrm{GGG},\mathrm{TTT}\}$) in the three frames $f\in\{0,1,2\}$ of genes in bacteria B (Table 1). As in 1996, the three sets of trinucleotides X, $X_1=\mathcal{P}(X)$ and $X_2=\mathcal{P}^2(X)$ are almost directly identified. Indeed, only by inspection, i.e. without any statistical tool, a set S_0 of 24 trinucleotides occurs in frame $f=0$ such that its trinucleotides have preferential frame occurrence probabilities $PrPrefFr_0(t,B)>4/10$ (Table 4a), the value 4/10 being the ceiling of the random value $PrPrefFr_0(t,B)=1/3$ (Proposition 1). By decreasing values of $PrPrefFr_0(t,B)$, the set S_0 (ordered by $>$ and with a notation given in parenthesis) is (**ATT**, **GAA**, **GAC**, **GCC**, **GAG**, **GTC**, **GAT**, **GTA**, *GTG*, **AAT**, **ACC**, **GTT**, **ATC**, **CTC**, **AAC**, ***CTT***, ***GCA***, *GCG*, ***GCT***, **GGT**, **TAC**, **CTG**, **CAG**, **TTC**) (Table 4a) where the trinucleotides of X are in bold, the trinucleotides of X_1 are in italics and the trinucleotides of X_2 are both in bold and italics. S_0 contains 19 trinucleotides of X. Similarly, a set S_1 of 21 trinucleotides occurs in frame $f=1$ such that its trinucleotides have probabilities $PrPrefFr_1(t,B)>4/10$ (Table 4a). By decreasing values of $PrPrefFr_1(t,B)$, the ordered set S_1 is (*ACG*, *TAG*, *CCG*, *TCG*, *TTG*, *AAG*, *ATA*, *CCA*, ***TGG***, *TTA*, *ATG*, *ACA*, *TCA*, *AGG*, ***CTA***, **CAG**, **CTG**, *AGC*, *GCG*, *TGC*, *TCC*) (Table 4a). S_1 contains 17 trinucleotides of X_1. A set S_2 of 24 trinucleotides occurs in frame $f=2$ such that its trinucleotides have probabilities $PrPrefFr_2(t,B)>4/10$ (Table 4a). By decreasing values of $PrPrefFr_2(t,B)$, the ordered set S_2 is {***CGA***, ***CAT***, ***CAC***, ***TGT***, ***CGC***, ***CGT***, ***AGA***, ***TAT***, ***GGA***, ***TAA***, ***TGA***, ***AGT***, ***CGG***, **GGC**, ***CCT***, *AGC*, ***CAA***, *TCC*, *TCT*, *TGC*, **TAC**, ***CTA***, ***GCT***, **GGT**} (Table 4a). S_2 contains 17 trinucleotides of X_2. Thus, only by inspection, three sets of trinucleotides are identified, one set per frame. In addition, these three sets are related by the complementarity and the circular permutation maps simultaneously. The complementarity map (Definition 1) can be revealed by inspection (i) within the trinucleotides in S_0 (self-complementarity of S_0); and (ii) among the trinucleotides S_1 and S_2 (complementarity between S_1 and S_2). Furthermore, as the complementarity map of a trinucleotide set has a biophysical basis with the complementary and antiparallel double helix, its identification in S_0, S_1 and S_2 is obvious. The circular permutation map of a trinucleotide set can also be revealed by inspection of the trinucleotides in S_0, S_1 and S_2. However, in contrast to the complementarity map, the permutation map (Definition 2) has no biophysical basis and is mainly related to a "mathematical" property of codes. All subsequent works on trinucleotide circular codes after 1996 have shown its importance in biology, in particular for coding the shifted frames in frameshift genes, e.g. Ahmed and Michel, 2011.

The identification of three sets related to the complementarity and permutation maps such that each set has 20 trinucleotides per frame needs a minor statistical investigation. Indeed, a problem arises with a few trinucleotides having preferential frame occurrence probabilities close to the random value $PrPrefFr_0(t,B)=1/3$ (Proposition 1) (Table 4a). In this case, a preferential frame for such trinucleotides cannot be assigned easily. In addition, there are also trinucleotides with a preferential frame but also occurring in the other frames. For example, by restricting to the trinucleotides having a preferential frame in frame 0, AAC also occurs in frames 1 and 2 with equiprobability ($PrPrefFr_0(\mathrm{AAC},B)=60\%$, $PrPrefFr_1(\mathrm{AAC},B)=20\%$, $PrPrefFr_2(\mathrm{AAC},B)=20\%$, Table 4a), AAT also occurs in frame 2 but not in frame 1 ($PrPrefFr_0(\mathrm{AAT},B)=68\%$, $PrPrefFr_1(\mathrm{AAT},B)=0$, $PrPrefFr_2(\mathrm{AAT},B)=32\%$, Table 4a), ATC also occurs in frame 1 but rarely in frame 2 ($PrPrefFr_0(\mathrm{ATC},B)=64\%$, $PrPrefFr_1(\mathrm{ATC},B)=32\%$, $PrPrefFr_2(\mathrm{ATC},B)=4\%$, Table 4a), etc. A similar observation exists with the trinucleotides having a preferential frame in frames 1 and 2 (Table 4a).

The 10 complementary and permutation (CP) trinucleotide sets $T\in X$, i.e. $\{t,\mathcal{C}(t)\}\in X$, $\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\}\in X_1$ and $\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\in X_2$, belong to the 10 highest values $PrCP(T,B)$ (Eq. (5)) among 30 in genes of bacteria (Table 4a), leading to the 20 trinucleotides of X in frame 0, 20 trinucleotides of X_1 in frame 1 and 20 trinucleotides of X_2 in frame 2. The 10th value $PrCP(T,B)=39\%$ is greater than the random value 1/3 (Proposition 2). This result confirms the

Table 4a
Preferential frame occurrence probabilities $PrPrefFr_f(t,B)$ (Eq. (4)) (rounded %) of the 60 trinucleotides t (without {AAA, CCC, GGG, TTT}) in the three frames $f \in \{0,1,2\}$ of genes in bacteria B (Table 1). Occurrence probabilities $PrCP(T,B)$ (Eq. (5)) (rounded %) of the 30 complementary and permutation (CP) trinucleotide sets $T = \{\{t, \mathcal{C}(t)\}, \{\mathcal{P}(t), \mathcal{P}(\mathcal{C}(t))\}, \{\mathcal{P}^2(t), \mathcal{P}^2(\mathcal{C}(t))\}\}$ in bacteria B. The CP trinucleotide sets T are ranged according to the decreasing values of $PrCP(T,B)$. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1 = \mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2 = \mathcal{P}^2(X)$ are both in bold and italics. The 10 CP trinucleotide sets T such that the complementary pairs $\{t, \mathcal{C}(t)\}$ belong to the circular code X are in bold.

t	Frame $f=0$	Frame $f=1$	Frame $f=2$	Frame $f=0$		Frame $f=1$		Frame $f=2$		$PrCP$
	$PrPrefFr_0$	$PrPrefFr_1$	$PrPrefFr_2$	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$	
AAC	60	20	20	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	93
AAG	24	76	0	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	77
AAT	68	0	32	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	73
ACA	28	64	8	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	70
ACC	68	8	24	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	69
ACG	0	100	0	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	65
ACT	40	24	36	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	60
AGA	0	24	76	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	59
AGC	0	44	56	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	57
AGG	0	60	40	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	39
AGT	0	32	68	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	33
ATA	24	76	0	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	33
ATC	64	32	4	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	32
ATG	32	68	0	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	29
ATT	100	0	0	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	27
CAA	12	32	56	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	25
CAC	0	16	84	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	23
CAG	44	52	4	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	21
CAT	0	8	92	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	19
CCA	24	76	0	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	17
CCG	4	96	0	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	15
CCT	36	0	64	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	14
CGA	0	0	100	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	13
CGC	16	4	80	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	11
CGG	0	32	68	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	9
CGT	20	0	80	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	7
CTA	0	56	44	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	5
CTC	64	32	4	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	3
CTG	48	52	0	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	2
CTT	56	8	36	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	1
GAA	100	0	0							
GAC	100	0	0							
GAG	92	8	0							
GAT	76	0	24							
GCA	56	32	12							
GCC	96	0	4							
GCG	56	44	0							
GCT	56	0	44							
GGA	28	0	72							
GGC	32	0	68							
GGT	56	0	44							
GTA	72	8	20							
GTC	80	16	4							
GTG	72	28	0							
GTT	68	0	32							
TAA	0	28	72							
TAC	52	0	48							
TAG	0	100	0							
TAT	24	0	76							
TCA	16	64	20							
TCC	4	40	56							
TCG	0	96	4							
TCT	20	24	56							
TGA	0	28	72							
TGC	0	44	56							
TGG	0	72	28							
TGT	0	16	84							
TTA	20	72	8							
TTC	44	32	24							
TTG	12	88	0							

existence of the circular code X in genes of bacteria observed in 1996. Furthermore, its statistical significance is strongly increased. Indeed, the quantitative approach is based here on 25 taxonomic groups of bacterial genes and the number of trinucleotides is multiplied by a factor of 527 (4,708,758 trinucleotides in 1996, 2,481,566,882 trinucleotides here).

3.2. C^3 self-complementary circular code X in genes of eukaryotes

The two indicators $PrefFr(t,E)$ and $MdPrefFr(t,E)$ (Eq. (2)) also show, as in the bacterial gene kingdom, very close results for the trinucleotide assignment of a preferential frame in the eukaryotic gene kingdom $K=E$ (Tables 2a and 2b). Indeed, 56 trinucleotides among 64, i.e. about 88%, have the same preferential frame with $PrefFr(t,E)$ and $MdPrefFr(t,E)$. The eight trinucleotides with a different preferential frame in E are **ACC**, **CAG**, CCC, **CTC**, **CTG**, **GTT**, TCC and TGC. Thirteen trinucleotides of X are assigned to the frame 0 with both indicators $PrefFr(t,E)$ and $MdPrefFr(t,E)$, and **GGC** and **GGT** to the frame 2. The five trinucleotides of X having a different preferential frame assignment in E are **ACC** in frame 1 with $PrefFr(t,E)$ and in frame 0 with $MdPrefFr(t,E)$, **CAG** in frame 0 with $PrefFr(t,E)$ and in frame 1 with $MdPrefFr(t,E)$, **CTC** in frame 0 with $PrefFr(t,E)$ and in frame 1 with $MdPrefFr(t,E)$, **CTG** in frame 0 with $PrefFr(t,E)$ and in frame 1 with $MdPrefFr(t,E)$, and **GTT** in frame 0 with $PrefFr(t,E)$ and in frame 2 with $MdPrefFr(t,E)$. The indicator $PrefFr(t,E)$ identifies directly 17 trinucleotides of X among 20 in frame 0. The indicator $MdPrefFr(t,E)$ is less significant compared to $PrefFr(t,E)$ in E containing a small number of gene taxonomic groups $G=E_i$ (11 versus 25 in B). Note also that some trinucleotides of X have very close median occurrence frequencies $MdPrFr_f(t,E)$ (%) in the three frames, i.e. $MdPrFr_0(\mathrm{ACC},E)=1.51$, $MdPrFr_1(\mathrm{ACC},E)=1.43$, $MdPrFr_0(\mathrm{CAG},E)=1.80$, $MdPrFr_1(\mathrm{CAG},E)=1.90$, $MdPrFr_0(\mathrm{CTC},E)=1.65$, $MdPrFr_1(\mathrm{CTC},E)=1.66$, $MdPrFr_2(\mathrm{CTC},E)=1.64$, $MdPrFr_0(\mathrm{GTT},E)=1.36$ and $MdPrFr_2(\mathrm{GTT},E)=1.42$ (Table 2b).

Table 4b gives the preferential frame occurrence probabilities $PrPrefFr_f(t,E)$ (Eq. (4)) of the 60 trinucleotides in the three frames $f\in\{0,1,2\}$ of genes in eukaryotes E (Table 1).

As in genes of bacteria, there are trinucleotides with preferential frame occurrence probabilities close to the random value $PrPrefFr_f(t,E)=1/3$ (Proposition 1) or with a preferential frame but also occurring in the other frames (Table 4b). Furthermore, the combinatorial complexity is increased as some trinucleotides may have the same preferential frame in bacteria and eukaryotes while other trinucleotides have a different statistical behavior in these two gene kingdoms. For example, ATT has a preferential frame in frame 0 in genes of both bacteria ($PrPrefFr_0(\mathrm{ATT},B)=100\%$, Table 4a) and eukaryotes ($PrPrefFr_0(\mathrm{ATT},E)=100\%$, Table 4b), AAT has a preferential frame in frame 0 but also occurs in frame 2 of bacterial genes ($PrPrefFr_0(\mathrm{AAT},B)=68\%$, $PrPrefFr_1(\mathrm{AAT},B)=0$, $PrPrefFr_2(\mathrm{AAT},B)=32\%$, Table 4a) while it has a preferential frame only in frame 0 in eukaryotic genes ($PrPrefFr_0(\mathrm{AAT},E)=100\%$, Table 4b), etc. However, this trinucleotide variability within eukaryotic genes and among prokaryotic genes leads to the same circular code X.

Indeed, the 10 CP trinucleotide sets $T\in X$ belong to the 10 highest values $PrCP(T,E)$ (Eq. (5)) among 30 in genes of eukaryotes (Table 4b). The 10th value $PrCP(T,E)=50\%$ is greater than the random value 1/3 (Proposition 2). This result again confirms the existence of the circular code X in genes of eukaryotes observed in 1996. Furthermore, its statistical significance is strongly increased. Indeed, the quantitative approach is based here on 11 taxonomic groups of eukaryotic genes and the number of trinucleotides is multiplied by a factor of 72 (11,397,678 trinucleotides in 1996, 824,825,761 trinucleotides here).

3.3. Identification of the C^3 self-complementary circular code X in genes of plasmids

Genes of plasmids is a kingdom studied for the first time. The two indicators $PrefFr(t,P)$ and $MdPrefFr(t,P)$ (Eq. (2)) show close results for the trinucleotide assignment of a preferential frame in the plasmid gene kingdom $K=P$ (Table 2a and 2b). Indeed, 45 trinucleotides among 64, i.e. about 70%, have the same preferential frame with $PrefFr(t,P)$ and $MdPrefFr(t,P)$. The 19 trinucleotides with a different preferential frame in P are **AAC**, AAG, ACT, AGC, AGT, CGC, CGG, CTA, **CTG**, GCA, GCG, GCT, **GGC**, **GTT**, TAT, TCC, TCT, TGA and TGC. Most of trinucleotides with a different preferential frame assignment concern the circular codes X_1 in frame 1 and X_2 in frame 2. Fifteen trinucleotides of X are assigned to the frame 0 with both indicators $PrefFr(t,P)$ and $MdPrefFr(t,P)$, and **GGT** to the frame 2. The four trinucleotides of X having a different preferential frame assignment are **AAC** in frame 0 with $PrefFr(t,P)$ and in frame 2 with $MdPrefFr(t,P)$, **CTG** in frame 0 with $PrefFr(t,P)$ and in frame 1 with $MdPrefFr(t,P)$, **GGC** in frame 0 with $PrefFr(t,P)$ and in frame 2 with $MdPrefFr(t,P)$ and **GTT** in frame 2 with $PrefFr(t,P)$ and in frame 0 with $MdPrefFr(t,P)$. The indicator $PrefFr(t,P)$ identifies directly 18 trinucleotides of X among 20 in frame 0.

Table 4c gives the preferential frame occurrence probabilities $PrPrefFr_f(t,P)$ (Eq. (4)) of the 60 trinucleotides in the three frames $f\in\{0,1,2\}$ of genes in plasmids P (Table 1). The 10 CP trinucleotide sets $T\in X$ belong to the 10 highest values $PrCP(T,P)$ (Eq. (5)) among 30 in genes of plasmids (Table 4c). The 10th value $PrCP(T,P)=41\%$ is greater than the random value 1/3 (Proposition 2). Thus, the C^3 self-complementary circular code X is now also identified in genes of plasmids.

3.4. Identification of a subset of the C^3 self-complementary circular code X in genes of viruses

Genes of viruses is also a kingdom studied for the first time. The two indicators $PrefFr(t,V)$ and $MdPrefFr(t,V)$ (Eq. (2)) show very close results for the trinucleotide assignment of a preferential frame in the viral gene kingdom $K=V$ (Tables 2a and b). Indeed, 54 trinucleotides among 64, i.e. about 84%, have the same preferential frame with $PrefFr(t,V)$ and $MdPrefFr(t,V)$. The 10 trinucleotides with a different preferential frame in V are **ACC**, AGC, CCT, CGG, **CTC**, GCG, **GGC**, GTG, TCC and TTT. Fifteen trinucleotides of X are assigned to the frame 0 with both indicators $PrefFr(t,V)$ and $MdPrefFr(t,V)$, and **CAG** and **CTG** to the frame 1. The three trinucleotides of X having a different preferential frame assignment in V are **ACC** in frame 0 with $PrefFr(t,V)$ and in frame 1 with $MdPrefFr(t,V)$, **CTC** in frame 0 with $PrefFr(t,V)$ and in frame 1 with $MdPrefFr(t,V)$ and **GGC** in frame 0 with $PrefFr(t,V)$ and in frame 2 with $MdPrefFr(t,V)$. The indicator $PrefFr(t,V)$ identifies directly 18 trinucleotides of X among 20 in frame 0.

Table 4d gives the preferential frame occurrence probabilities $PrPrefFr_f(t,V)$ (Eq. (4)) of the 60 trinucleotides in the three frames $f\in\{0,1,2\}$ of genes in viruses V (Table 1). Nine CP trinucleotide sets $T\in X$ belong to the nine highest values $PrCP(T,V)$ (Eq. (5)) among 30 in genes of viruses (Table 4d). The 10th set $T=\{\{\mathrm{CAG},\mathrm{CTG}\},\{\mathrm{AGC},\mathrm{TGC}\},\{\mathrm{GCA},\mathrm{GCT}\}\}\in X$ has a probability $PrCP(T,V)=8\%$ which is significantly less than the random value 1/3 (Proposition 2), meaning that the two trinucleotides CAG and $\mathcal{C}(\mathrm{CAG})=\mathrm{CTG}$ do not occur preferentially in frame 0 of viral genes and must be excluded from the code X. Thus, a subset of X which is a non-maximal C^3 self-complementary circular code is identified in genes of viruses. A search of variant X codes in viral genes will confirm this result (see below Section 3.7.4).

3.5. Circular code asymmetries of the C^3 self-complementary circular code X, $X_1=P(X)$ and $X_2=P^2(X)$

3.5.1. Circular code asymmetries in genes of bacteria, eukaryotes, plasmids and viruses

Table 5a gives the occurrence probabilities $PrPrefFr_f(C,K)$ (Eq. (6)) of preferential frame of the C^3 self-complementary circular codes $C=X$, $C=X_1=\mathcal{P}(X)$ and $C=X_2=\mathcal{P}^2(X)$ which are computed in the three frames $f\in\{0,1,2\}$ of genes in bacteria B,

Table 4b
Preferential frame occurrence probabilities $PrPrefFr_f(t,E)$ (Eq. (4)) (rounded %) of the 60 trinucleotides t (without {AAA, CCC, GGG, TTT}) in the three frames $f \in \{0,1,2\}$ of genes in eukaryotes E (Table 1). Occurrence probabilities $PrCP(T,E)$ (Eq. (5)) (rounded %) of the 30 complementary and permutation (CP) trinucleotide sets $T = \{\{t, \mathcal{C}(t)\}, \{\mathcal{P}(t), \mathcal{P}(\mathcal{C}(t))\}, \{\mathcal{P}^2(t), \mathcal{P}^2(\mathcal{C}(t))\}\}$ in eukaryotes E. The CP trinucleotide sets T are ranged according to the decreasing values of $PrCP(T,E)$. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1 = \mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2 = \mathcal{P}^2(X)$ are both in bold and italics. The 10 CP trinucleotide sets T such that the complementary pairs $\{t, \mathcal{C}(t)\}$ belong to the circular code X are in bold.

t	Frame $f=0$	Frame $f=1$	Frame $f=2$	Frame $f=0$		Frame $f=1$		Frame $f=2$		$PrCP$
	$PrPrefFr_0$	$PrPrefFr_1$	$PrPrefFr_2$	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$	
AAC	91	9	0	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	97
AAG	91	9	0	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	92
AAT	100	0	0	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	86
ACA	0	100	0	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	83
ACC	36	55	9	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	73
ACG	0	100	0	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	59
ACT	36	64	0	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	58
AGA	0	36	64	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	56
AGC	9	73	18	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	56
AGG	0	100	0	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	50
AGT	9	73	18	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	39
ATA	0	100	0	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	38
ATC	73	27	0	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	38
ATG	18	82	0	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	26
ATT	100	0	0	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	23
CAA	0	9	91	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	18
CAC	0	9	91	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	17
CAG	64	36	0	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	17
CAT	0	9	91	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	15
CCA	9	91	0	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	12
CCG	0	82	18	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	12
CCT	36	0	64	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	11
CGA	0	0	100	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	6
CGC	18	9	73	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	5
CGG	9	55	36	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	5
CGT	9	0	91	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	3
CTA	0	45	55	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	2
CTC	45	36	18	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	2
CTG	55	45	0	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	2
CTT	27	0	73	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	2
GAA	64	0	36							
GAC	100	0	0							
GAG	100	0	0							
GAT	82	0	18							
GCA	45	9	45							
GCC	100	0	0							
GCG	45	36	18							
GCT	45	0	55							
GGA	9	0	91							
GGC	9	0	91							
GGT	45	0	55							
GTA	27	36	36							
GTC	91	0	9							
GTG	55	45	0							
GTT	55	0	45							
TAA	0	0	100							
TAC	91	0	9							
TAG	0	100	0							
TAT	27	0	73							
TCA	0	100	0							
TCC	36	36	27							
TCG	0	100	0							
TCT	45	27	27							
TGA	0	27	73							
TGC	0	55	45							
TGG	0	73	27							
TGT	0	9	91							
TTA	18	82	0							
TTC	64	9	27							
TTG	9	91	0							

eukaryotes E and plasmids P (Table 1). The viral gene kingdom with only six taxonomic groups is excluded from this computation. Table 5b gives the occurrence probabilities $PrFr_f(C,K)$ (Eq. (7)) of the codes $C = X$, $C = X_1$, $C = X_2$ and $C = PPT = \{AAA, CCC, GGG, TTT\}$ which are also computed in the three frames $f \in \{0,1,2\}$ of genes in the three previous kingdoms B, E and P, and also in viruses V (Table 1).

The two real probabilities $PrPrefFr_f(C,K)$ (Eq. (6)) and $PrFr_f(C,K)$ (Eq. (7)) retrieve the classical circular code asymmetry in frame 0 in all studied gene kingdoms K which has already been observed in

Table 4c
Preferential frame occurrence probabilities $PrPrefFr_f(t,P)$ (Eq. (4)) (rounded %) of the 60 trinucleotides t (without {AAA, CCC, GGG, TTT}) in the three frames $f \in \{0,1,2\}$ of genes in plasmids P (Table 1). Occurrence probabilities $PrCP(T,P)$ (Eq. (5)) (rounded %) of the 30 complementary and permutation (CP) trinucleotide sets $T = \{\{t, \mathcal{C}(t)\}, \{\mathcal{P}(t), \mathcal{P}(\mathcal{C}(t))\}, \{\mathcal{P}^2(t), \mathcal{P}^2(\mathcal{C}(t))\}\}$ in plasmids P. The CP trinucleotide sets T are ranged according to the decreasing values of $PrCP(T,P)$. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1 = \mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2 = \mathcal{P}^2(X)$ are both in bold and italics. The 10 CP trinucleotide sets T such that the complementary pairs $\{t, \mathcal{C}(t)\}$ belong to the circular code X are in bold.

t	Frame $f=0$	Frame $f=1$	Frame $f=2$	Frame $f=0$		Frame $f=1$		Frame $f=2$		$PrCP$
	$PrPrefFr_0$	$PrPrefFr_1$	$PrPrefFr_2$	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$	
AAC	45	18	36	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	88
AAG	36	64	0	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	80
AAT	73	9	18	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	62
ACA	36	55	9	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	62
ACC	55	18	27	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	61
ACG	9	91	0	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	61
ACT	45	27	27	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	59
AGA	0	45	55	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	55
AGC	0	45	55	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	55
AGG	0	55	45	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	41
AGT	9	45	45	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	39
ATA	9	91	0	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	33
ATC	55	36	9	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	33
ATG	27	73	0	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	32
ATT	100	0	0	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	29
CAA	27	18	55	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	27
CAC	0	27	73	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	27
CAG	45	55	0	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	24
CAT	9	9	82	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	24
CCA	18	64	18	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	18
CCG	0	100	0	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	15
CCT	45	0	55	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	14
CGA	0	0	100	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	14
CGC	18	0	82	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	12
CGG	0	64	36	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	12
CGT	36	0	64	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	11
CTA	0	55	45	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	6
CTC	55	45	0	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	5
CTG	45	55	0	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	2
CTT	36	9	55	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	0
GAA	91	0	9							
GAC	91	0	9							
GAG	100	0	0							
GAT	55	0	45							
GCA	55	27	18							
GCC	100	0	0							
GCG	73	27	0							
GCT	55	0	45							
GGA	36	0	64							
GGC	18	0	82							
GGT	55	0	45							
GTA	64	0	36							
GTC	82	9	9							
GTG	55	45	0							
GTT	55	0	45							
TAA	0	9	91							
TAC	45	0	55							
TAG	0	100	0							
TAT	45	0	55							
TCA	9	55	36							
TCC	0	45	55							
TCG	0	100	0							
TCT	36	45	18							
TGA	0	45	55							
TGC	0	45	55							
TGG	0	64	36							
TGT	0	27	73							
TTA	27	73	0							
TTC	45	27	27							
TTG	18	82	0							

Table 4d

Preferential frame occurrence probabilities $PrPrefFr_f(t,V)$ (Eq. (4)) (rounded %) of the 60 trinucleotides t (without {AAA, CCC, GGG, TTT}) in the three frames $f \in \{0,1,2\}$ of genes in viruses V (Table 1). Occurrence probabilities $PrCP(T,V)$ (Eq. (5)) (rounded %) of the 30 complementary and permutation (CP) trinucleotide sets $T = \{\{t, \mathcal{C}(t)\}, \{\mathcal{P}(t), \mathcal{P}(\mathcal{C}(t))\}, \{\mathcal{P}^2(t), \mathcal{P}^2(\mathcal{C}(t))\}\}$ in viruses V. The CP trinucleotide sets T are ranged according to the decreasing values of $PrCP(T,V)$. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1 = \mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2 = \mathcal{P}^2(X)$ are both in bold and italics. The 10 CP trinucleotide sets T such that the complementary pairs $\{t, \mathcal{C}(t)\}$ belong to the circular code X are in bold.

t	Frame $f=0$	Frame $f=1$	Frame $f=2$	Frame $f=0$		Frame $f=1$		Frame $f=2$		$PrCP$
	$PrPrefFr_0$	$PrPrefFr_1$	$PrPrefFr_2$	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$	
AAC	83	17	0	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	94
AAG	33	67	0	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	92
AAT	100	0	0	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	92
ACA	0	83	17	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	92
ACC	33	33	33	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	72
ACG	0	100	0	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	67
ACT	100	0	0	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	64
AGA	0	0	100	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	64
AGC	0	17	83	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	50
AGG	0	100	0	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	50
AGT	0	33	67	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	44
ATA	0	100	0	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	42
ATC	67	33	0	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	36
ATG	0	100	0	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	36
ATT	83	17	0	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	22
CAA	17	0	83	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	14
CAC	0	0	100	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	14
CAG	33	67	0	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	11
CAT	0	0	100	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	8
CCA	33	67	0	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	8
CCG	0	100	0	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	8
CCT	83	0	17	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	6
CGA	0	0	100	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	6
CGC	0	0	100	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	6
CGG	0	50	50	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	3
CGT	33	0	67	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	0
CTA	0	100	0	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	0
CTC	33	67	0	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	0
CTG	0	100	0	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	0
CTT	17	0	83	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	0
GAA	100	0	0							
GAC	100	0	0							
GAG	100	0	0							
GAT	100	0	0							
GCA	100	0	0							
GCC	100	0	0							
GCG	83	17	0							
GCT	100	0	0							
GGA	0	0	100							
GGC	17	0	83							
GGT	83	0	17							
GTA	67	33	0							
GTC	100	0	0							
GTG	33	67	0							
GTT	100	0	0							
TAA	0	0	100							
TAC	67	0	33							
TAG	0	100	0							
TAT	17	0	83							
TCA	0	83	17							
TCC	0	33	67							
TCG	0	100	0							
TCT	83	0	17							
TGA	0	0	100							
TGC	0	0	100							
TGG	0	50	50							
TGT	0	0	100							
TTA	17	83	0							
TTC	83	17	0							
TTG	0	100	0							

Table 5a
Circular code asymmetries in genes of bacteria, eukaryotes and plasmids. Occurrence probabilities $PrPrefFr_f(X,K)$, $PrPrefFr_f(X_1,K)$ and $PrPrefFr_f(X_2,K)$ (Eq. (6)) (%) of preferential frame of the C^3 self-complementary circular code X, $X_1=\mathcal{P}(X)$ and $X_2=\mathcal{P}^2(X)$, respectively, in the three frames $f\in\{0,1,2\}$ of genes in bacteria $K=B$ (7,851,762 genes, 2,481,566,882 trinucleotides), eukaryotes $K=E$ (1,662,579 genes, 824,825,761 trinucleotides) and plasmids $K=P$ (237,486 genes, 68,244,356 trinucleotides) (Table 1).

	Bacteria B			Eukaryotes E			Plasmids P		
	Frame $f=0$ $PrPrefFr_0$	Frame $f=1$ $PrPrefFr_1$	Frame $f=2$ $PrPrefFr_2$	Frame $f=0$ $PrPrefFr_0$	Frame $f=1$ $PrPrefFr_1$	Frame $f=2$ $PrPrefFr_2$	Frame $f=0$ $PrPrefFr_0$	Frame $f=1$ $PrPrefFr_1$	Frame $f=2$ $PrPrefFr_2$
X	69.2	13.0	17.8	69.5	12.7	17.7	63.6	13.6	22.7
X_1	18.8	64.2	17.0	18.6	73.6	7.7	20.0	64.1	15.9
X_2	15.2	19.4	65.4	11.8	18.2	70.0	18.6	22.3	59.1

Table 5b
Circular code asymmetries in genes of bacteria, eukaryotes, plasmids and viruses. Occurrence probabilities $PrFr_f(X,K)$, $PrFr_f(X_1,K)$, $PrFr_f(X_2,K)$ and $PrFr_f(PPT,K)$ (Eq. (7)) (%) of the C^3 self-complementary circular code X, $X_1=\mathcal{P}(X)$, $X_2=\mathcal{P}^2(X)$ and $PPT=\{AAA,CCC,GGG,TTT\}$, respectively, in the three frames $f\in\{0,1,2\}$ of genes in bacteria $K=B$ (7,851,762 genes, 2,481,566,882 trinucleotides), eukaryotes $K=E$ (1,662,579 genes, 824,825,761 trinucleotides), plasmids $K=P$ (237,486 genes, 68,244,356 trinucleotides) and viruses $K=V$ (184,344 genes, 45,688,798 trinucleotides) (Table 1). Estimated probabilities $\widehat{PrFr}_f(X,K)$, $\widehat{PrFr}_f(X_1,K)$, $\widehat{PrFr}_f(X_2,K)$ and $\widehat{PrFr}_f(PPT,K)$ (Eq. (8)) (%) of the codes X, X_1, X_2 and PPT, respectively, for the two shifted frames $f\in\{1,2\}$ in the four previous gene kingdoms B, E, P and V. The eight estimated probabilities $\widehat{PrFr}_f(C,K)$ are very close to the eight real probabilities $PrFr_f(C,K)$ in the shifted frames $f\in\{1,2\}$ in all gene kingdoms K: correlation coefficients $r\left(PrFr_f(C,K),\widehat{PrFr}_f(C,K)\right)$ are equal to 0.999 in bacterial genes, 1.000 in eukaryotic genes, 0.999 in viral genes and 1.000 in plasmid genes.

	Bacteria B					Eukaryotes E					Plasmids P					Viruses V				
	Frame $f=0$	Frame $f=1$		Frame $f=2$		Frame $f=0$	Frame $f=1$		Frame $f=2$		Frame $f=0$	Frame $f=1$		Frame $f=2$		Frame $f=0$	Frame $f=1$		Frame $f=2$	
	$PrFr_0$	$PrFr_1$	$\widehat{PrFr}_1$	$PrFr_2$	$\widehat{PrFr}_2$	$PrFr_0$	$PrFr_1$	$\widehat{PrFr}_1$	$PrFr_2$	$\widehat{PrFr}_2$	$PrFr_0$	$PrFr_1$	$\widehat{PrFr}_1$	$PrFr_2$	$\widehat{PrFr}_2$	$PrFr_0$	$PrFr_1$	$\widehat{PrFr}_1$	$PrFr_2$	$\widehat{PrFr}_2$
X	46.6	21.4	22.0	27.4	27.4	41.5	25.6	25.5	29.4	29.4	47.5	21.0	20.9	27.9	27.5	43.4	24.5	24.6	27.6	27.7
X_1	25.8	43.1	42.4	21.3	22.7	28.2	39.0	39.0	23.0	23.7	25.9	43.4	43.0	21.2	22.1	26.7	41.2	40.7	22.5	23.3
X_2	20.3	28.9	29.5	44.1	43.4	22.8	28.9	29.1	40.5	40.0	20.3	29.6	29.8	44.6	44.1	22.3	28.0	28.4	42.5	41.8
PPT	7.2	6.6	6.3	7.2	6.7	7.4	6.5	6.3	7.1	7.0	6.3	6.1	6.3	6.2	6.3	7.7	6.3	6.3	7.3	7.2

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)

$$\begin{cases} PrPrefFr_0(X_1,K) > PrPrefFr_0(X_2,K) \ \text{(Table 5a)} \\ PrFr_0(X_1,K) > PrFr_0(X_2,K) \ \text{(Table 5b)} \end{cases}. \quad (13)$$

In frame 0, the circular code X_1 occurs with a frequency higher than the circular code X_2.

These probabilities $PrPrefFr_f(C,K)$ (Eq. (6)) and $PrFr_f(C,K)$ (Eq. (7)) also identify other circular code asymmetries in frames 1 and 2 in all studied gene kingdoms K (Table 5a and 5b), given here with $PrFr_f(C,K)$

$$\begin{cases} PrFr_2(X,K) > PrFr_1(X,K) \\ PrFr_1(X_2,K) > PrFr_1(X,K) \\ PrFr_2(X,K) > PrFr_2(X_1,K) \end{cases}. \quad (14)$$

Thus, in particular, the circular code X occurs with a frequency in frame 2 higher than in frame 1. This circular code asymmetry may be related to the biological observation that there are more putative overlapping genes in the frame 2 than the frame 1 of mitochondrial genes of primates (Seligmann, 2011), Drosophila (Seligmann, 2012a) and turtles (Seligmann, 2012b).

3.5.2. A probabilistic model for the circular code asymmetries

The estimated probabilities $\widehat{PrFr}_f(C,K)$ (Eq. (8)) of the codes $C=X$, $C=X_1$, $C=X_2$ and $C=PPT$ are determined for the two shifted frames $f\in\{1,2\}$ in the four previous gene kingdoms B, E, P and V (Table 5b). Very surprisingly, the eight estimated probabilities $\widehat{PrFr}_f(C,K)$ are very close to the eight real probabilities $PrFr_f(C,K)$ (Eq. (7)) in the shifted frames $f\in\{1,2\}$ in all studied gene kingdoms K: correlation coefficients $r\left(PrFr_f(C,B),\widehat{PrFr}_f(C,B)\right)=r\left(PrFr_f(C,V),\widehat{PrFr}_f(C,V)\right)=0.999$ in genes of bacteria B and viruses V and $r\left(PrFr_f(C,E),\widehat{PrFr}_f(C,E)\right)=r\left(PrFr_f(C,P),\widehat{PrFr}_f(C,P)\right)=1.000$ in genes of eukaryotes E and plasmids P. The estimated probabilities $\widehat{PrFr}_f(C,K)$ retrieve the set of inequalities (13) and (14). Thus, the probabilities of the circular codes X, $X_1=\mathcal{P}(X)$, $X_2=\mathcal{P}^2(X)$ in frame 0 explain the circular code probabilities and asymmetries in the two shifted frames $f\in\{1,2\}$ of genes in bacteria, eukaryotes, plasmids and viruses.

3.6. C^3 self-complementary circular code X in genes of bacteria, eukaryotes, plasmids and viruses

In the large class of $\binom{30}{10}=30,045,015$ C^3 self-complementary trinucleotide codes, the C^3 self-complementary circular code X occurs preferentially in genes of bacteria B, eukaryotes E, plasmids P and viruses V by considering both the occurrence frequencies $PrFr_f(t,K)$ and the median occurrence frequencies $MdPrFr_f(t,K)$ of trinucleotides $t\in A_4^3$ in the three frames $f\in\{0,1,2\}$ in these four gene kingdoms K.

Indeed, for the frequency $PrFr_f(t,K)$, the nine CP trinucleotide sets $T_1, T_2, T_4, \ldots, T_{10}\in X$ have each at least three values $NbCAT(T,K)\geq 4$ (Eq. (9)) among four values $NbCAT(T,K)$ and mean numbers $\overline{NbCAT}(T,BEPV)\geq 4.0$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame (Table 6a). The

Table 6a

Identification of the C^3 self-complementary circular code X in the four gene kingdoms K of bacteria B, eukaryotes E, plasmids P and viruses V (Table 1). Number $NbCAT(T,K)$ (Eq. (9)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 complementary and permutation (CP) trinucleotide sets $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ of the four gene kingdoms K. Mean number $\overline{NbCAT}(T,BEPV)$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 CP trinucleotide sets T of bacteria B, eukaryotes E, plasmids P and viruses V. The values $NbCAT(T,K)\geq 4$ are in bold. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1=\mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2=\mathcal{P}^2(X)$ are both in bold and italics. The first 10 CP trinucleotide sets $T_1,\ldots,T_{10}$ belonging to the C^3 self-complementary circular code X have complementary pairs $\{t,\mathcal{C}(t)\}$ in bold.

T	Frame $f=0$		Frame $f=1$		Frame $f=2$		B	E	P	V	$\overline{NbCAT}$
	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$					
T_1	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	**5**	**6**	**5**	**6**	5.5
T_2	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	**6**	**6**	**5**	**6**	5.8
T_3	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	3	2	3	**4**	3.0
T_4	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	**6**	**6**	**5**	**6**	5.8
T_5	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	**6**	**5**	**5**	1	4.3
T_6	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	**6**	**6**	**6**	**6**	6.0
T_7	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	**5**	**4**	**5**	**5**	4.8
T_8	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	**6**	**6**	**6**	**6**	6.0
T_9	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	**5**	**4**	**5**	**6**	5.0
T_{10}	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	**5**	**4**	**4**	3	4.0
T_{11}	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	1	1	1	0	0.8
T_{12}	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	0	0	0	0	0.0
T_{13}	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	0	0	0	0	0.0
T_{14}	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	0	1	0	2	0.8
T_{15}	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	0	1	1	3	1.3
T_{16}	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	0	0	0	0	0.0
T_{17}	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	0	0	1	0	0.3
T_{18}	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	0	0	1	0	0.3
T_{19}	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	0	0	0	0	0.0
T_{20}	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	0	0	0	0	0.0
T_{21}	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	1	2	1	0	1.0
T_{22}	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	3	**4**	3	2	3.0
T_{23}	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	1	1	2	0	1.0
T_{24}	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	0	0	0	0	0.0
T_{25}	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	0	0	0	0	0.0
T_{26}	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	0	0	0	0	0.0
T_{27}	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	0	1	0	1	0.5
T_{28}	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	0	0	0	2	0.5
T_{29}	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	0	0	0	0	0.0
T_{30}	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	1	0	1	0	0.5

Table 6b

Identification of the C^3 self-complementary circular code X in the four gene kingdoms K of bacteria B, eukaryotes E, plasmids P and viruses V (Table 1). Number $MdNbCAT(T,K)$ (Eq. (9)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 complementary and permutation (CP) trinucleotide sets $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ of the four gene kingdoms K. Mean number $\overline{MdNbCAT}(T,BEPV)$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 CP trinucleotide sets T of bacteria B, eukaryotes E, plasmids P and viruses V. The values $MdNbCAT(T,K)\geq 4$ are in bold. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1=\mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2=\mathcal{P}^2(X)$ are both in bold and italics. The first 10 CP trinucleotide sets $T_1,\ldots,T_{10}$ belonging to the C^3 self-complementary circular code X have complementary pairs $\{t,\mathcal{C}(t)\}$ in bold.

T	Frame $f=0$		Frame $f=1$		Frame $f=2$		B	E	P	V	$\overline{MdNbCAT}$
	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$					
T_1	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	**6**	**5**	**5**	**6**	5.5
T_2	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	**6**	**6**	**6**	**6**	6.0
T_3	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	**4**	3	3	**4**	3.5
T_4	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	**6**	**6**	**6**	**6**	6.0
T_5	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	2	**4**	1	0	1.8
T_6	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	**5**	**4**	**5**	3	4.3
T_7	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	**4**	**4**	**5**	**5**	4.5
T_8	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	**6**	**6**	**6**	**6**	6.0
T_9	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	**4**	**4**	3	3	3.5
T_{10}	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	**4**	**4**	**4**	**4**	4.0
T_{11}	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	2	1	1	0	1.0
T_{12}	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	0	0	1	0	0.3
T_{13}	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	0	0	0	0	0.0
T_{14}	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	1	1	2	2	1.5
T_{15}	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	2	1	3	3	2.3
T_{16}	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	1	1	1	3	1.5
T_{17}	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	0	0	0	0	0.0
T_{18}	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	0	0	0	0	0.0
T_{19}	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	0	0	0	0	0.0
T_{20}	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	0	0	0	0	0.0
T_{21}	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	2	2	3	3	2.5
T_{22}	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	2	3	3	2	2.5
T_{23}	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	1	1	0	0	0.5
T_{24}	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	0	0	0	0	0.0
T_{25}	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	0	1	0	0	0.3
T_{26}	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	0	0	0	0	0.0
T_{27}	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	0	1	0	1	0.5
T_{28}	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	2	1	2	3	2.0
T_{29}	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	0	0	0	0	0.0
T_{30}	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	0	1	0	0	0.3

two CP trinucleotide sets $T_3\in X$ and $T_{22}\notin X$ have each no value $NbCAT(T,K)>4$, only one value $NbCAT(T,K)=4$ and mean numbers $\overline{NbCAT}(T,BEPV)=3.0$ (Table 6a). The 19 remaining CP trinucleotide sets $T_{11},\ldots,T_{21}$, $T_{23},\ldots,T_{30}\notin X$ have all the values $NbCAT(T,K)<4$ and mean numbers $\overline{NbCAT}(T,BEPV)\leq 1.3$ (Table 6a).

For the median frequency $MdPrFr_f(t,K)$, the nine CP trinucleotide sets $T_1,\ldots,T_4$, $T_6,\ldots,T_{10}\in X$ have each at least two values $MdNbCAT(T,K)\geq 4$ (Eq. (9)) among four values $MdNbCAT(T,K)$ and mean numbers $\overline{MdNbCAT}(T,BEPV)\geq 3.5$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame (Table 6b). The CP trinucleotide set $T_5\in X$ has one value $MdNbCAT(T_5,K)=4$ but a very low value $\overline{MdNbCAT}(T_5,BEPV)=1.8$ (Table 6b). The 20 CP trinucleotide sets $T_{11},\ldots,T_{30}\notin X$ have all the values $MdNbCAT(T,K)<4$ and mean numbers $\overline{MdNbCAT}(T,BEPV)\leq 2.5$ (Table 6b).

The partition $T_1,\ldots,T_{10}\in X$ and $T_{11},\ldots,T_{30}\notin X$ observed with $NbCAT(T,K)$, $\overline{NbCAT}(T,BEPV)$, $MdNbCAT(T,K)$ and $\overline{MdNbCAT}(T,BEPV)$ confirms that the average code in genes of bacteria B, eukaryotes E, plasmids P and viruses V is X. Furthermore, this partition is also retrieved by considering the gene taxonomic groups of bacteria (see the results with $\overline{NbCAT}(T,B)$ in Section 3.7.1), eukaryotes (see the results with $\overline{NbCAT}(T,E)$ in Section 3.7.2), plasmids (see the results with $\overline{NbCAT}(T,P)$ in Section 3.7.3) and viruses (see the results with $\overline{NbCAT}(T,V)$ in Section 3.7.4).

3.7. Variant X codes

3.7.1. Variant X codes in genes of bacteria

Three variant X codes, i.e. trinucleotide codes which differ from the C^3 self-complementary circular code X, are identified in cyanobacteria B_{CYA}, deinococcus B_{DEI} and elusimicrobia B_{ELU} among the 25 gene taxonomic groups B_G of bacteria B (Table 7a), i.e. $3/25=12\%$. Indeed, among the $20\times 25=500$ values $NbCAT(T_i,B_{G_j})$ (Eq. (9)) with $i\in\{11,\ldots,30\}$ and $j\in\{1,\ldots,25\}$, only three values $NbCAT(T_i,B_{G_j})$ are greater or equal to 4, i.e. $3/(500/2)\approx 1\%$ as 10 CP trinucleotide sets are exclusive from 10 other CP trinucleotide sets with $NbCAT(T_i,B_{G_j})\geq 4$. These three values are $NbCAT(T_{21},B_{CYA})=NbCAT(T_{22},B_{DEI})=NbCAT(T_{15},B_{ELU})=4$ (Table 7a). Furthermore, these three variant X codes only differ by one complementary trinucleotide pair with respect to X. Otherwise, all the CP trinucleotide sets $T_1,\ldots,T_{10}\in X$ have mean numbers $\overline{NbCAT}(T,B)\geq 2.4$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame while all the CP trinucleotide sets $T_{11},\ldots,T_{30}\notin X$ have values $\overline{NbCAT}(T,B)\leq 1.9$ (Table 7a last column). The partition $T_1,\ldots,T_{10}\in X$ and

Table 7a

Variant X codes in genes of the 25 taxonomic groups B_G in bacteria B (Table 1). Number $NbCAT(T, B_G)$ (Eq. (9)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 complementary and permutation (CP) trinucleotide sets $T = \{\{t, \mathcal{C}(t)\}, \{\mathcal{P}(t), \mathcal{P}(\mathcal{C}(t))\}, \{\mathcal{P}^2(t), \mathcal{P}^2(\mathcal{C}(t))\}\}$ of the 25 gene taxonomic groups B_G. Mean number $\overline{NbCAT}(T, B)$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 CP trinucleotide sets T of bacteria B. The values $NbCAT(T, B_G) \geq 4$ are in bold. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1 = \mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2 = \mathcal{P}^2(X)$ are both in bold and italics. The first 10 CP trinucleotide sets $T_1, \ldots, T_{10}$ belonging to the C^3 self-complementary circular code X have complementary pairs $\{t, \mathcal{C}(t)\}$ in bold.

T	Frame $f=0$		Frame $f=1$		Frame $f=2$		B_{ACT}	B_{AQU}	B_{ARM}	B_{BAC}	B_{CAL}	B_{CHA}	B_{CHO}	B_{CHR}	B_{CYA}	B_{DEF}	B_{DEI}	B_{DIC}	B_{ELU}	B_{FIB}	B_{FIR}	B_{FUS}	B_{GEM}	B_{NIT}	B_{PLA}	B_{PRO}	B_{SPI}	B_{SYN}	B_{TEN}	B_{THD}	B_{THG}	$\overline{NbCAT}$
	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$																										
T_1	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	3	**5**	**5**	**6**	3	3	**4**	**6**	3	**4**	3	3	**5**	**5**	3	3	3	**5**	**5**	**4**	**6**	**6**	3	**4**	**5**	4.3
T_2	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	**5**	3	**6**	**6**	**4**	**5**	**6**	**6**	**5**	**6**	**5**	3	**4**	**5**	**5**	**4**	**4**	**4**	**4**	**6**	**6**	3	**5**	3	3	4.3
T_3	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	3	3	3	**4**	3	**5**	3	3	**4**	**4**	2	**4**	**4**	3	**4**	3	3	3	3	3	**4**	3	3	**6**	2	3.4
T_4	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	3	**5**	**6**	**6**	**4**	**4**	**6**	**5**	**4**	**6**	3	**4**	3	3	**5**	3	**5**	**5**	3	**4**	**6**	3	3	**4**	**6**	4.3
T_5	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	**5**	0	**6**	1	0	0	**5**	**5**	0	0	**5**	0	1	**5**	0	0	**5**	**4**	**6**	**6**	0	**5**	0	0	0	2.4
T_6	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	**6**	**5**	**5**	**5**	1	1	**6**	**6**	**4**	2	**6**	1	2	**6**	2	0	**6**	**6**	**6**	**6**	**5**	**6**	0	1	**4**	3.8
T_7	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	**4**	3	3	**4**	3	3	**5**	**4**	**4**	3	**4**	3	3	**4**	3	3	3	3	**5**	**5**	3	**4**	3	3	**4**	3.5
T_8	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	**6**	**6**	**6**	**6**	**6**	**5**	**6**	**6**	**5**	**6**	**6**	**5**	**5**	**6**	**6**	**4**	**6**	**6**	**6**	**6**	**6**	**6**	**4**	3	**6**	5.4
T_9	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	**5**	**5**	3	**5**	**4**	3	**4**	**5**	2	**5**	3	3	**6**	**6**	**5**	**4**	3	3	**4**	**5**	**4**	3	**6**	**4**	**4**	4.1
T_{10}	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	**4**	**4**	**5**	3	3	3	**5**	**4**	3	3	**4**	3	3	**4**	3	3	3	**5**	3	3	3	**6**	3	3	**4**	3.5
T_{11}	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	2	3	3	1	3	0	1	2	2	2	2	3	1	2	3	1	3	3	1	1	2	2	0	3	2	1.9
T_{12}	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	0	1	0	0	3	0	0	0	0	2	0	3	1	0	2	3	0	0	0	0	0	0	3	1	1	0.9
T_{13}	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	2	0	0	0	0	0	0	2	3	0	0.6
T_{14}	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	0	1	0	3	2	3	0	1	3	2	0	3	3	0	3	3	0	0	0	1	3	0	3	3	1	1.5
T_{15}	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	1	3	0	2	3	3	1	1	3	3	1	3	**4**	1	3	3	1	1	0	0	3	1	3	3	3	1.9
T_{16}	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	0	1	1	1	3	3	0	0	1	3	0	3	1	0	3	3	0	0	0	0	1	0	3	3	1	1.3
T_{17}	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	1	3	0	0	2	0	0	0	0	0	1	3	2	1	0	1	2	2	2	0	0	3	0	3	3	1.5
T_{18}	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	3	1	0	0	1	2	0	1	2	0	3	2	1	3	1	1	1	1	3	2	0	3	1	2	0	1.4
T_{19}	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	0	0	0	0	3	0	0	0	0	1	0	2	0	0	2	3	0	0	0	0	0	0	3	0	2	0.9
T_{20}	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0.1
T_{21}	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	1	1	3	1	0	3	2	1	**4**	1	3	3	0	0	1	1	3	3	2	1	2	3	0	2	2	1.8
T_{22}	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	3	3	3	2	0	1	3	3	2	1	**4**	0	2	3	0	0	3	3	3	3	2	3	0	0	2	1.7
T_{23}	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	2	1	1	0	1	0	1	1	0	1	2	0	0	2	0	0	3	1	3	2	0	0	0	0	1	0.9
T_{24}	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	2	0	0	0	0	0	0	2	0	0	0.4
T_{25}	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	0	0	0	0	2	2	0	0	1	1	0	2	3	0	1	3	0	0	0	0	0	0	3	2	1	0.9
T_{26}	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0.0
T_{27}	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	0	0	0	1	0	3	0	0	0	1	0	0	2	0	0	2	0	0	0	0	1	0	3	0	0	0.5
T_{28}	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	0	3	0	3	3	3	0	0	3	3	0	3	1	0	3	3	0	1	0	0	3	0	3	3	3	1.7
T_{29}	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0.3
T_{30}	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	3	0	1	0	0	3	2	0	3	0	3	0	0	1	1	0	3	1	1	2	0	0	0	1	0	0.8

$T_{11}, \ldots, T_{30} \notin X$ confirms again that the average code in bacterial genes is X (see Sections 3.1 and 3.6).

In cyanobacteria B_{CYA}, the complementary trinucleotide pair {GCC, GGC} of X is replaced by {CGC, GCG} (T_{21}) leading to the variant X code

$$X_A = \{\text{AAC, AAT, ACC, ATC, ATT, CAG, CGC, CTC, CTG, GAA, GAC, GAG, GAT, GCG, GGT, GTA, GTC, GTT, TAC, TTC}\}. \quad (15)$$

In deinococcus B_{DEI}, the complementary trinucleotide pair {ACC, GGT} of X is replaced by {CAC, GTG} (T_{22}) leading to the variant X code

$$X_B = \{\text{AAC, AAT, ATC, ATT, CAC, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GTA, GTC, GTG, GTT, TAC, TTC}\}. \quad (16)$$

In elusimicrobia B_{ELU}, the complementary trinucleotide pair {CAG, CTG} of X is replaced by {AGC, GCT} (T_{15}) leading to the variant X code

$$X_C = \{\text{AAC, AAT, ACC, AGC, ATC, ATT, CTC, GAA, GAC, GAG, GAT, GCC, GCT, GGC, GGT, GTA, GTC, GTT, TAC, TTC}\}. \quad (17)$$

3.7.2. Variant X codes in genes of eukaryotes

Seven variant X codes are identified in birds E_{BIR}, fishes E_{FIS}, insects E_{INS}, mammals E_{MAM}, basidiomycetes E_{BAS}, apicomplexans E_{API} and kinetoplasts E_{KIN} among the 11 gene taxonomic groups E_G of eukaryotes E (Table 7b), i.e. $7/11 \approx 64\%$. Among the $20 \times 11 = 220$ values $NbCAT\left(T_i, E_{G_j}\right)$ (Eq. (9)) where $i \in \{11, \ldots, 30\}$ and $j \in \{1, \ldots, 11\}$, only nine values $NbCAT\left(T_i, E_{G_j}\right)$ are greater or equal to 4, i.e. $9/(220/2) \approx 8\%$. These nine values are $NbCAT(T_{21}, E_{BIR}) = NbCAT(T_{22}, E_{BIR}) = NbCAT(T_{11}, E_{FIS}) = NbCAT(T_{22}, E_{FIS}) = NbCAT(T_{11}, E_{INS}) = NbCAT(T_{22}, E_{MAM}) = NbCAT(T_{14}, E_{BAS}) = NbCAT(T_{15}, E_{API}) = NbCAT(T_{22}, E_{KIN}) = 4$ (Table 7b). Furthermore, these seven variant X codes only differ by two complementary trinucleotide pairs with respect to X in E_{BIR} and E_{FIS}, and by one complementary trinucleotide pair with respect to X in E_{INS}, E_{MAM}, E_{BAS}, E_{API} and E_{KIN}. All the CP trinucleotide sets $T_1, \ldots, T_{10} \in X$ have mean numbers $\overline{NbCAT}(T, E) \geq 3.0$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame while all the CP trinucleotide sets $T_{11}, \ldots, T_{30} \notin X$ have values $\overline{NbCAT}(T, E) \leq 2.4$ (Table 7b last column). The partition $T_1, \ldots, T_{10} \in X$ and $T_{11}, \ldots, T_{30} \notin X$ confirms again that the average code in eukaryotic genes is X (see Sections 3.2 and 3.6).

In birds E_{BIR}, the two complementary trinucleotide pairs {{GCC, GGC}, {ACC, GGT}} of X are replaced by {{CGC, GCG}, {CAC, GTG}} (T_{21}, T_{22}) leading to the variant X code

$$X_D = \{\text{AAC, AAT, ATC, ATT, CAC, CAG, CGC, CTC, CTG, GAA, GAC, GAG, GAT, GCG, GTA, GTC, GTG, GTT, TAC, TTC}\}. \quad (18)$$

In fishes E_{FIS}, the two complementary trinucleotide pairs {{GAA, TTC}, {ACC, GGT}} of X are replaced by {{AAG, CTT}, {CAC, GTG}} (T_{11}, T_{22}) leading to the variant X code

$$X_E = \{\text{AAC, AAG, AAT, ATC, ATT, CAC, CAG, CTC, CTG, CTT, GAC, GAG, GAT, GCC, GGC, GTA, GTC, GTG, GTT, TAC}\}. \quad (19)$$

In insects E_{INS}, the complementary trinucleotide pair {GAA, TTC} of X is replaced by {AAG, CTT} (T_{11}) leading to the variant X code

$$X_F = \{\text{AAC, AAG, AAT, ACC, ATC, ATT, CAG, CTC, CTG, CTT, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC}\}. \quad (20)$$

In mammals E_{MAM} and kinetoplasts E_{KIN}, the complementary trinucleotide pair {ACC, GGT} of X is replaced by {CAC, GTG} (T_{22}) leading to the variant X code X_B (Eq. (16)). In particular, the code $X_B \backslash \{\text{GCC, GGC}\}$ of 18 trinucleotides is observed in the human genes (102,788 genes, 62,777,956 trinucleotides according to the data acquisition in GenBank).

In basidiomycetes E_{BAS}, the complementary trinucleotide pair {GTA, TAC} of X is replaced by {ACT, AGT} (T_{14}) leading to the variant X code

$$X_G = \{\text{AAC, AAT, ACC, ACT, AGT, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTC, GTT, TTC}\}. \quad (21)$$

In apicomplexans E_{API}, the complementary trinucleotide pair {CAG, CTG} of X is replaced by {AGC, GCT} (T_{15}) leading to the variant X code X_C (Eq. (17)).

3.7.3. Variant X codes in genes of plasmids

Four variant X codes are identified in plasmids of chloroflexi P_{CRF}, cyanobacteria P_{CYA}, deinococcus P_{DEI} and spirochaetes P_{SPI} among the 11 gene taxonomic groups P_G in plasmids P (Table 7c), i.e. $4/11 \approx 36\%$. Indeed, among the $20 \times 11 = 220$ values $NbCAT\left(T_i, P_{G_j}\right)$ (Eq. (9)) where $i \in \{11, \ldots, 30\}$ and $j \in \{1, \ldots, 11\}$, only four values $NbCAT\left(T_i, P_{G_j}\right)$ are greater or equal to 4, i.e. $4/(220/2) \approx 4\%$. These four values are $NbCAT(T_{22}, P_{CRF}) = NbCAT(T_{21}, P_{CYA}) = NbCAT(T_{22}, P_{DEI}) = NbCAT(T_{14}, P_{SPI}) = 4$ (Table 7c). Furthermore, these four variant X codes only differ by one complementary trinucleotide pair with respect to X. All the CP trinucleotide sets $T_1, \ldots, T_{10} \in X$ have mean numbers $\overline{NbCAT}(T, P) \geq 2.5$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame while all the CP trinucleotide sets $T_{11}, \ldots, T_{30} \notin X$ have values $\overline{NbCAT}(T, P) \leq 2.4$ (Table 7c last column). The partition $T_1, \ldots, T_{10} \in X$ and $T_{11}, \ldots, T_{30} \notin X$ confirms again that the average code in plasmid genes is X (see Sections 3.3 and 3.6).

In plasmids of chloroflexi P_{CRF} and deinococcus P_{DEI}, the complementary trinucleotide pair {ACC, GGT} of X is replaced by {CAC, GTG} (T_{22}) leading to the variant X code X_B (Eq. (16)).

In plasmids of cyanobacteria P_{CYA}, the complementary trinucleotide pair {GCC, GGC} of X is replaced by {CGC, GCG} (T_{21}) leading to the variant X code X_A (Eq. (15)).

In plasmids of spirochaetes P_{SPI}, the complementary trinucleotide pair {GTA, TAC} of X is replaced by {ACT, AGT} (T_{14}) leading to the variant X code X_G (Eq. (21)).

3.7.4. Subsets of the circular code X in genes of viruses

For the CP trinucleotide set T_5, all the numbers $NbCAT(T_5, V_{G_j})$ (Eq. (9)) for the six viral gene taxonomic groups V_{G_j} where $j \in \{1, \ldots, 6\}$ are less or equal to 2 and their mean number $\overline{NbCAT}(T_5, V) = 0.5$ (Eq. (12)) (Table 7d), confirming that the complementary trinucleotide pair {CAG, CTG} does not belong to X (see Section 3.4). For the CP trinucleotide set T_{10}, five numbers $NbCAT(T_{10}, V_{G_i})$ among 6 are less or equal to 3 and their mean number $\overline{NbCAT}(T_{10}, V) = 3.0$ (Table 7d) suggesting that the complementary trinucleotide pair {GTA, TAC} may also not belong to X. All the CP trinucleotide sets $T_1, \ldots, T_4, T_6, \ldots, T_9 \in X$ have mean numbers $\overline{NbCAT}(T, V) \geq 3.8$ of correctly assigned trinucleotides (CAT) with respect to the frame while all the CP trinucleotide sets $T_{11}, \ldots, T_{30} \notin X$ have values $\overline{NbCAT}(T, V) \leq 3.0$ (Table 7d last column). Furthermore, no CP trinucleotide set $T_{11}, \ldots, T_{30} \notin X$ has a value $NbCAT\left(T_i, V_{G_j}\right) \geq 4$ with the six viral groups V_G (Table 7d). Thus, no variant X code is identified in genes of viruses, only a subset of the C^3 self-complementary circular code X which may have either 18 trinucleotides $X \backslash \{\text{CAG, CTG}\}$ leading to the non-maximal C^3 self-complementary circular code

$$X_{18} = \{\text{AAC, AAT, ACC, ATC, ATT, CTC, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC}\} \quad (22)$$

or 16 trinucleotides $X \backslash \{\{\text{CAG, CTG}\}, \{\text{GTA, TAC}\}\}$ leading to the non-maximal C^3 self-complementary circular code

$$X_{16} = \{\text{AAC, AAT, ACC, ATC, ATT, CTC, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTC, GTT, TTC}\}. \quad (23)$$

Table 7b
Variant X codes in genes of the 11 taxonomic groups E_G in eukaryotes E (Table 1). Number $NbCAT(T,E_G)$ (Eq. (9)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 complementary and permutation (CP) trinucleotide sets $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ of the 11 gene taxonomic groups E_G. Mean number $\overline{NbCAT}(T,E)$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 CP trinucleotide sets T of eukaryotes E. The values $NbCAT(T,E_G)\geq 4$ are in bold. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1=\mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2=\mathcal{P}^2(X)$ are both in bold and italics. The first 10 CP trinucleotide sets $T_1,\ldots,T_{10}$ belonging to the C^3 self-complementary circular code X have complementary pairs $\{t,\mathcal{C}(t)\}$ in bold.

T	Frame $f=0$		Frame $f=1$		Frame $f=2$		E_{BIR}	E_{FIS}	E_{INS}	E_{MAM}	E_{RWO}	E_{ASC}	E_{BAS}	E_{GAL}	E_{LPL}	E_{API}	E_{KIN}	$\overline{NbCAT}$
	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$												
T_1	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	**6**	**5**	**5**	**5**	**4**	**6**	**6**	3	**6**	**6**	**5**	5.2
T_2	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	**5**	**5**	**6**	**5**	**6**	**5**	**6**	**5**	**6**	**6**	**6**	5.5
T_3	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	2	2	3	2	3	**5**	**6**	3	**5**	**4**	2	3.4
T_4	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	**6**	**6**	**5**	**6**	**4**	**6**	**6**	3	**5**	**5**	3	5.0
T_5	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	**6**	**5**	**6**	**6**	0	2	1	**6**	0	0	**6**	3.5
T_6	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	**5**	**5**	**5**	**6**	2	3	**4**	**5**	3	**5**	**5**	4.4
T_7	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	**4**	2	2	**5**	3	**4**	2	3	2	3	3	3.0
T_8	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	**6**	**6**	**6**	**6**	**5**	**6**	**6**	**6**	**6**	**6**	**5**	5.8
T_9	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	1	**5**	**4**	1	**4**	**5**	**4**	3	3	**4**	3	3.4
T_{10}	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	**5**	**4**	**4**	**5**	**4**	2	1	**4**	3	3	**4**	3.5
T_{11}	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	1	**4**	**4**	1	3	1	2	3	2	1	3	2.3
T_{12}	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	0	0	0	0	1	0	0	0	0	0	0	0.1
T_{13}	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	0	0	0	0	1	0	0	0	0	0	0	0.1
T_{14}	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	0	0	0	0	2	3	**4**	0	3	3	0	1.4
T_{15}	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	0	1	0	0	3	3	3	0	3	**4**	0	1.5
T_{16}	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	0	1	0	0	3	2	2	0	3	0	0	1.0
T_{17}	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	1	1	0	1	0	0	0	0	0	0	0	0.3
T_{18}	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	0	0	1	0	1	0	0	3	1	1	3	0.9
T_{19}	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	0	0	0	0	3	0	0	0	0	0	0	0.3
T_{20}	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	1	0	1	2	0	0	0	0	0	0	0	0.4
T_{21}	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	**4**	1	1	3	2	1	2	3	3	2	3	2.3
T_{22}	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	**4**	**4**	3	**4**	0	1	0	3	1	2	**4**	2.4
T_{23}	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	1	2	2	1	0	1	1	2	0	0	2	1.1
T_{24}	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	0	0	0	0	1	0	0	0	0	0	0	0.1
T_{25}	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	1	0	1	0	1	1	0	1	0	1	1	0.6
T_{26}	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	0	0	0	0	0	0	0	0	0	0	1	0.1
T_{27}	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	1	0	0	0	0	1	2	0	2	2	0	0.7
T_{28}	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	0	0	0	0	3	1	2	0	3	2	0	1.0
T_{29}	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	0	0	0	0	0	1	0	1	0	0	0	0.2
T_{30}	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	0	1	1	1	1	0	0	3	0	0	1	0.7

The statistical approach performed here in the viral kingdom containing only six gene taxonomic groups of small size (see Table 1) leaves open three hypotheses for the identification of a circular code in viral genes: the maximal C^3 self-complementary circular code X or the non-maximal C^3 self-complementary circular codes X_{18} or X_{16}. Additional statistical studies together with an increase of viral gene data should solve this problem in future.

3.7.5. *Combinatorial properties of the variant X codes*

The variant X codes X_A (Eq. (15)) in cyanobacteria B_{CYA} and plasmids of cyanobacteria P_{CYA}, and X_D (Eq. (18)) in birds E_{BIR} are self-complementary, without permuted trinucleotides but non-circular.

The variant X codes X_B (Eq. (16)) in deinococcus B_{DEI}, plasmids of chloroflexi P_{CRF} and deinococcus P_{DEI}, mammals E_{MAM} and kinetoplasts E_{KIN}, X_C (Eq. (17)) in elusimicrobia B_{ELU} and apicomplexans E_{API}, X_E (Eq. (19)) in fishes E_{FIS}, X_F (Eq. (20)) in insects E_{INS} and X_G (Eq. (21)) in basidiomycetes E_{BAS} and plasmids of spirochaetes P_{SPI} are maximal C^3 self-complementary circular. Furthermore, the maximal C^3 self-complementary circular codes X_B, X_C and X_E (but not X_F and X_G) belong to the class of 88 circular codes generated by the nucleotide frequency (NF) method (Lacan and Michel, 2001; Koch and Lehmann, 1997; Fimmel et al., 2014), X_B, X_C and X_E being the 19th, 17th and 35th codes, respectively, in Table 3 in Lacan and Michel (2001).

The circular code X codes 12 amino acids $AA=\{$Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr, Val$\}$ according to the standard genetic code (Arquès and Michel, 1996, Table 4a). The code X_A codes 13 amino acids $AA\cup$Arg, the code X_B codes 12 amino acids $\{AA\backslash$Thr$\}\cup$His, the code X_C codes 12 amino acids $\{AA\backslash$Gln$\}\cup$Ser, the code X_D codes 12 amino acids $\{AA\backslash\{$Gly, Thr$\}\}\cup\{$Arg, His$\}$, the code X_E codes 12 amino acids $\{AA\backslash\{$Phe, Thr$\}\}\cup\{$His, Lys$\}$, the code X_F codes 12 amino acids $\{AA\backslash$Phe$\}\cup$Lys and the code X_G codes 12 amino acids $\{AA\backslash$Tyr$\}\cup$Ser.

4. Conclusion

The statistical approach developed here quantifies the concept used in 1996 for determining a preferential frame for the trinucleotides among the three possible frames in genes. Based on the occurrence probability $PrCP(T,K)$ (Eq. (5)) of complementary and permutation (CP) trinucleotide sets T in gene kingdoms K, it confirms the C^3 self-complementary circular code X in genes of bacteria and eukaryotes. It also identifies this circular code X in genes of plasmids and a subset of X in genes of viruses. Note that, for an order of magnitude, the probability to retrieve the same circular code X in three independent gene kingdoms is equal to $1/\binom{30}{10}^3\approx 4\times 10^{-23}$.

There are some significant differences between the methods developed in 1996 and here: (i) in 1996, the trinucleotide frequencies in the three frames were studied by inspection, here, as mentioned above, they are analysed by the quantitative parameter $PrCP(T,K)$; (ii) in 1996, only single gene populations of bacteria and eukaryotes were available, here the approach uses large gene taxonomic groups in four

Table 7c
Variant X codes in genes of the 11 taxonomic groups P_G in plasmids P (Table 1). Number $NbCAT(T,P_G)$ (Eq. (9)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 complementary and permutation (CP) trinucleotide sets $T=\{\{t,\mathcal{C}(t)\},\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\},\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}\}$ of the 11 gene taxonomic groups P_G. Mean number $\overline{NbCAT}(T,P)$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 CP trinucleotide sets T of plasmids P. The values $NbCAT(T,P_G)\geq 4$ are in bold. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1=\mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2=\mathcal{P}^2(X)$ are both in bold and italics. The first 10 CP trinucleotide sets $T_1,\ldots,T_{10}$ belonging to the C^3 self-complementary circular code X have complementary pairs $\{t,\mathcal{C}(t)\}$ in bold.

T	Frame $f=0$		Frame $f=1$		Frame $f=2$		P_{ACT}	P_{BAC}	P_{CMD}	P_{CRF}	P_{CYA}	P_{DEI}	P_{FIB}	P_{FIR}	P_{FUS}	P_{PRO}	P_{SPI}	$\overline{NbCAT}$
	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$												
T_1	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	3	**6**	3	3	3	**4**	**5**	3	3	**4**	3	3.6
T_2	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	**4**	**6**	**4**	**5**	**5**	**5**	**5**	**5**	3	**5**	**6**	4.8
T_3	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	3	**4**	**4**	2	**6**	2	3	3	3	3	3	3.3
T_4	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	3	**6**	**4**	**4**	**4**	3	**4**	3	**4**	3	3	3.7
T_5	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	**5**	0	0	**6**	0	**6**	**5**	0	0	**5**	0	2.5
T_6	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	**6**	**5**	1	**6**	2	**6**	**6**	1	1	**6**	1	3.7
T_7	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	**4**	3	3	**5**	**4**	**4**	3	3	3	**4**	3	3.5
T_8	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	**6**	**6**	**4**	**4**	**5**	**6**	**6**	**5**	**6**	**6**	**4**	5.3
T_9	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	**4**	3	3	3	2	3	**5**	3	**5**	**5**	**4**	3.6
T_{10}	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	**4**	3	3	3	3	**4**	**4**	3	3	**4**	2	3.3
T_{11}	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	2	3	0	1	2	2	3	2	1	2	1	1.7
T_{12}	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	0	0	2	0	0	0	0	2	3	0	3	0.9
T_{13}	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	0	0	2	2	1	0	0	1	0	0	1	0.6
T_{14}	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	0	2	3	0	3	0	0	3	3	0	**4**	1.6
T_{15}	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	1	3	3	0	3	0	1	3	3	1	3	1.9
T_{16}	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	0	1	3	0	3	0	0	3	3	0	3	1.5
T_{17}	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	1	0	1	1	0	1	1	0	3	1	0	0.8
T_{18}	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	3	0	1	2	2	3	2	3	1	3	2	2.0
T_{19}	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	0	0	0	0	0	0	0	3	2	0	3	0.7
T_{20}	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	0	0	0	0	0	0	0	0	0	0	0	0.0
T_{21}	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	2	3	3	3	**4**	3	1	3	1	1	2	2.4
T_{22}	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	3	2	2	**4**	0	**4**	3	0	1	3	0	2.0
T_{23}	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	2	1	0	3	0	2	2	0	0	2	0	1.1
T_{24}	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	0	0	1	0	0	0	0	0	1	0	1	0.3
T_{25}	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	0	0	2	0	1	0	0	2	2	0	2	0.8
T_{26}	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	0	0	0	0	0	0	0	0	0	0	1	0.1
T_{27}	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	0	0	3	0	0	0	0	1	2	0	2	0.7
T_{28}	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	0	3	3	0	3	0	0	3	3	0	3	1.6
T_{29}	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	1	0	1	0	1	0	0	1	0	0	0	0.4
T_{30}	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	3	0	1	3	3	2	1	1	0	2	0	1.5

kingdoms; and (iii) the amount of gene data has increased considerably, e.g. by a factor of 527 for bacterial genes.

The method developed by Gonzalez et al. (2011) also demonstrated that the circular code X has on average the best covering capability (CC). This CC method is based on a simple definition of a statistical function, in the same line of the method proposed here. It should be stressed that the definition of a simple statistical parameter in this coding research field, obvious a posteriori, is not immediate. This also explains the intermediate development of more elaborate statistical methods for searching circular codes in genes, e.g. the correlation function per frame (Arquès and Michel, 1997) or the frame permuted trinucleotide frequency methods (Frey and Michel, 2003, 2006). However, all these code search methods rely on the same principle with a study of the three frames in genes. Nevertheless, there are some complementary aspects between the CC method and the method proposed here: (i) the CC method uses a gene data set finely selected from 13 classes of proteins, here large gene taxonomic groups of bacteria, eukaryotes, plasmids and viruses are investigated; and (ii) the CC method explores the circular code X among the class of the 216 C^3 self-complementary circular codes, here the circular code X is analysed in the large class of $\binom{30}{10}=30,045,015$ C^3 self-complementary trinucleotide codes which contains in particular these 216 circular codes.

Several circular code asymmetries of the C^3 self-complementary circular code X, $X_1=\mathcal{P}(X)$ and $X_2=\mathcal{P}^2(X)$ are identified in the three frames of genes in bacteria, eukaryotes, plasmids and viruses. In particular, (i) in frame 0, the circular code X_1 occurs with a frequency higher than the circular code X_2; and (ii) the circular code X occurs with a frequency in frame 2 higher than in frame 1. The development of a simple probabilistic model based on the independent occurrence of trinucleotides in reading frame (frame 0) of genes can estimate the probabilities and asymmetries of the circular codes X, X_1 and X_2 in the two shifted frames $f\in\{1,2\}$ of genes in bacteria, eukaryotes, plasmids and viruses.

The developed approach also allows the identification of variant X codes in each gene taxonomic group, i.e. trinucleotide codes which differ from the C^3 self-complementary circular code X. In genes of bacteria, eukaryotes and plasmids, 14 among the 47 studied gene taxonomic groups (about 30%) have variant trinucleotide codes close to X, i.e. containing at least 16 trinucleotides of X. Seven variant X codes are identified. Two variant X codes X_A in cyanobacteria and plasmids of cyanobacteria, and X_D in birds are self-complementary, without permuted trinucleotides but non-circular. Five variant X codes X_B in deinococcus, plasmids of chloroflexi and deinococcus, mammals and kinetoplasts, X_C in elusimicrobia and apicomplexans, X_E in fishes, X_F in insects, and X_G in basidiomycetes and plasmids of spirochaetes are C^3 self-complementary circular. In genes of viruses, no variant X code is observed but a subset of X which may have 18 or 16 trinucleotides according to the viral gene data acquired. The evolution of the circular code X to a variant X code is an open problem which needs several investigations, from a theoretical point of view (combinatorics, statistics) and biological point of view, in particular in relation to the genetic code (amino acid coding).

In summary, the proposed quantitative statistical approach based on massive gene data shows that the maximal C^3 self-complementary trinucleotide circular code X is a common

Table 7d
No variant X codes in genes of the six taxonomic groups V_G in viruses V (Table 1). Number $NbCAT(T, V_G)$ (Eq. (9)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 complementary and permutation (CP) trinucleotide sets $T = \{\{t, \mathcal{C}(t)\}, \{\mathcal{P}(t), \mathcal{P}(\mathcal{C}(t))\}, \{\mathcal{P}^2(t), \mathcal{P}^2(\mathcal{C}(t))\}\}$ of the six gene taxonomic groups V_G. Mean number $\overline{NbCAT}(T, V)$ (Eq. (12)) of correctly assigned trinucleotides (CAT) with respect to the frame in the 30 CP trinucleotide sets T of viruses V. The values $NbCAT(T, V_G) \geq 4$ are in bold. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of $X_1 = \mathcal{P}(X)$ are in italics and the 20 trinucleotides of $X_2 = \mathcal{P}^2(X)$ are both in bold and italics. The first 10 CP trinucleotide sets $T_1, \ldots, T_{10}$ belonging to the C^3 self-complementary circular code X have complementary pairs $\{t, \mathcal{C}(t)\}$ in bold.

T	Frame $f=0$		Frame $f=1$		Frame $f=2$		V_{DSD}	V_{DSR}	V_{RTR}	V_{SSD}	V_{SSR}	V_{PHA}	$\overline{NbCAT}$
	t	$\mathcal{C}(t)$	$\mathcal{P}(t)$	$\mathcal{P}(\mathcal{C}(t))$	$\mathcal{P}^2(t)$	$\mathcal{P}^2(\mathcal{C}(t))$							
T_1	**AAC**	**GTT**	*ACA*	*TTG*	***CAA***	***TGT***	**6**	**6**	3	**6**	**6**	**6**	5.5
T_2	**AAT**	**ATT**	*ATA*	*TTA*	***TAA***	***TAT***	**6**	**6**	**5**	**4**	**6**	**6**	5.5
T_3	**ACC**	**GGT**	*CCA*	*GTG*	***CAC***	***TGG***	**4**	3	1	**5**	**6**	**5**	4.0
T_4	**ATC**	**GAT**	*TCA*	*ATG*	***CAT***	***TGA***	**6**	**5**	**6**	**4**	**6**	**6**	5.5
T_5	**CAG**	**CTG**	*AGC*	*TGC*	***GCA***	***GCT***	2	0	0	1	0	0	0.5
T_6	**CTC**	**GAG**	*TCC*	*AGG*	***CCT***	***GGA***	**6**	3	**4**	3	**4**	3	3.8
T_7	**GAA**	**TTC**	*AAG*	*TCT*	***AGA***	***CTT***	**5**	**5**	**4**	**4**	**4**	**4**	4.3
T_8	**GAC**	**GTC**	*ACG*	*TCG*	***CGA***	***CGT***	**6**	**6**	**6**	**5**	**6**	**5**	5.7
T_9	**GCC**	**GGC**	*CCG*	*GCG*	***CGC***	***CGG***	**6**	3	3	3	**4**	**4**	3.8
T_{10}	**GTA**	**TAC**	*TAG*	*ACT*	***AGT***	***CTA***	3	**4**	3	2	3	3	3.0
T_{11}	*AAG*	***CTT***	***AGA***	**TTC**	**GAA**	*TCT*	0	0	2	1	1	1	0.8
T_{12}	*ACA*	***TGT***	***CAA***	**GTT**	**AAC**	*TTG*	0	0	0	0	0	0	0.0
T_{13}	*ACG*	***CGT***	***CGA***	**GTC**	**GAC**	*TCG*	0	0	0	1	0	1	0.3
T_{14}	*ACT*	***AGT***	***CTA***	**GTA**	**TAC**	*TAG*	2	2	3	3	3	3	2.7
T_{15}	*AGC*	***GCT***	***GCA***	**CTG**	**CAG**	*TGC*	3	3	3	3	3	3	3.0
T_{16}	*AGG*	***CCT***	***GGA***	**CTC**	**GAG**	*TCC*	0	3	2	3	2	3	2.2
T_{17}	*ATA*	***TAT***	***TAA***	**ATT**	**AAT**	*TTA*	0	0	0	2	0	0	0.3
T_{18}	*ATG*	***CAT***	***TGA***	**ATC**	**GAT**	*TCA*	0	1	0	2	0	0	0.5
T_{19}	*CCA*	***TGG***	***CAC***	**GGT**	**ACC**	*GTG*	0	2	2	0	0	0	0.7
T_{20}	*CCG*	***CGG***	***CGC***	**GGC**	**GCC**	*GCG*	0	0	0	0	0	0	0.0
T_{21}	*GCG*	***CGC***	***CGG***	**GCC**	**GGC**	*CCG*	0	3	3	3	2	2	2.2
T_{22}	*GTG*	***CAC***	***TGG***	**ACC**	**GGT**	*CCA*	2	1	3	1	0	1	1.3
T_{23}	*TAG*	***CTA***	***AGT***	**TAC**	**GTA**	*ACT*	1	0	0	1	0	0	0.3
T_{24}	*TCA*	***TGA***	***CAT***	**GAT**	**ATC**	*ATG*	0	0	0	0	0	0	0.0
T_{25}	*TCC*	***GGA***	***CCT***	**GAG**	**CTC**	*AGG*	0	0	0	0	0	0	0.0
T_{26}	*TCG*	***CGA***	***CGT***	**GAC**	**GTC**	*ACG*	0	0	0	0	0	0	0.0
T_{27}	*TCT*	***AGA***	***CTT***	**GAA**	**TTC**	*AAG*	1	1	0	1	1	1	0.8
T_{28}	*TGC*	***GCA***	***GCT***	**CAG**	**CTG**	*AGC*	1	3	3	2	3	3	2.5
T_{29}	*TTA*	***TAA***	***TAT***	**AAT**	**ATT**	*ATA*	0	0	1	0	0	0	0.2
T_{30}	*TTG*	***CAA***	***TGT***	**AAC**	**GTT**	*ACA*	0	0	3	0	0	0	0.5

(average) property in genes of bacteria, eukaryotes, plasmids and viruses.

Acknowledgment

I thank the three reviewers for their advice, and Denise Besch, Svetlana Gorchkova, Elisabeth Michel and Jean-Marc Vassards for their support.

References

Ahmed, A., Michel, C.J., 2011. Circular code signal in frameshift genes. J. Comput. Sci. Syst. Biol. 4, 7–15.
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.
Arquès, D.G., Michel, C.J., 1997. A code in the protein coding genes. Biosystems 44, 107–134.
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.
Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Natl. Acad. Sci. USA 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. 70, 1623–1644. http://dx.doi.org/10.1007/s00285-014-0806-7.
Frey, G., Michel, C.J., 2003. Circular codes in archaeal genomes. J. Theor. Biol. 223, 413–431.
Frey, G., Michel, C.J., 2006. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Comput. Biol. Chem. 30, 87–101.
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.
Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.
Michel, C.J., 2012. Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Comput. Biol. Chem. 37, 24–37.
Michel, C.J., 2013. Circular code motifs in transfer RNAs. Comput. Biol. Chem. 45, 17–29.
Michel, C.J., 2014. A genetic scale of reading frame coding. J. Theor. Biol. 355, 83–94.
Michel, C.J., 2015. An extended genetic scale of reading frame coding. J. Theor. Biol. 365, 164–174.
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.
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.
Seligmann, H., 2011. Two genetic codes, one genome: frameshifted primate mitochondrial genes code for additional proteins in presence of antisense antitermination tRNAs. Biosystems 105, 271–285.
Seligmann, H., 2012a. An overlapping genetic code for frameshifted overlapping genes in *Drosophila mitochondria*: antisense antitermination tRNAs UAR insert serine. J. Theor. Biol. 298, 51–76.
Seligmann, H., 2012b. Overlapping genetic codes for overlapping frameshifted genes in Testudines, and *Lepidochelys olivacea* as special case. Comput. Biol. Chem. 41, 18–34.