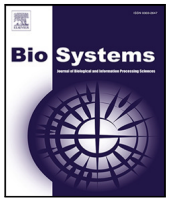
BioSystems 244 (2024) 105308

Contents lists available at ScienceDirect

BioSystems

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



Circular code identified by the codon usage

Christian J. Michel

Theoretical bioinformatics, ICube, University of Strasbourg, C.N.R.S., 300 Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Keywords:
Statistical 1-frame method
Trinucleotide codes
Trinucleotide circular codes
Codon usage
Genes
Bacteria
Archaea
Eukaryotes

ABSTRACT

Since 1996, circular codes in genes have been identified thanks to the development of 6 statistical approaches: trinucleotide frequencies per frame (Arquès and Michel, 1996), correlation functions per frame (Arquès and Michel, 1997), frame permuted trinucleotide frequencies (Frey and Michel, 2003, 2006), advanced statistical functions at the gene population level (Michel, 2015) and at the gene level (Michel, 2017). All these 3-frame statistical methods analyse the trinucleotide information in the 3 frames of genes: the reading frame and the 2 shifted frames. Notably, codon usage does not allow for the identification of circular codes (Michel, 2020). This has been a long-standing problem since 1996, hindering biologists' access to circular code theory.

By considering circular code conditions resulting from code theory, particularly the concept of permutation class, and building upon previous statistical work, a new statistical approach based solely on the codon usage, i.e. a 1-frame statistical method, surprisingly reveals the maximal C^3 self-complementary trinucleotide circular code X in bacterial genes and in average (bacterial, archaeal, eukaryotic) genes, and almost in archaeal genes. Additionally, a new parameter definition indicates that bacterial and archaeal genes exhibit codon usage dispersion of the same order of magnitude, but significantly higher than that observed in eukaryotic genes. This statistical finding may explain the greater variability of codes in eukaryotic genes compared to bacterial and archaeal genes, an issue that has been open for many years. Finally, biologists can now search for new (variant) circular codes at both the genome level (across all genes in a given genome) and the gene level using only codon usage, without the need for analysing the shifted frames.

1. Introduction

The circular code theory provides a mathematical structure to the genetic code and its sub-codes, allowing simultaneous study of amino acid encoding and reading frame synchronization properties during the translation process. In 1996, a maximal C^3 self-complementary trinucleotide circular code X was discovered in the genes (Arquès and Michel, 1996)

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

This circular code X is able to retrieve the correct reading frame during translation - a property called *circularity* that generalizes the Crick's classical notion of comma-freeness. The historical context of this result is described in the article by Michel (2020).

From 1996 to 2017, six statistical approaches applied to an exponentially growing number of genes were developed to confirm the circular code X (1) in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Michel, 2017). In 2017, the analysed kingdom of bacteria contained 15,735,053 genes and 5,222,267,667 trinucleotides (13,688 genes and 4,708,758 trinucleotides in 1996), i.e. a bacterial trinucleotide number multiplied by 1109, and the analysed kingdom of eukaryotes contained 4,356,391 genes and 2,406,844,838 trinucleotides (26,757 genes and 11,397,678 trinucleotides in 1996), i.e. an eukaryotic trinucleotide number multiplied by 211. All the developed 3-frame statistical methods analyse the trinucleotide information in the 3 frames of genes (the reading frame 0 and the 2 shifted frames 1 and 2): trinucleotide frequencies per frame (Arquès and Michel, 1996), correlation functions per frame (Arquès and Michel, 1997), frame permuted trinucleotide frequencies (Frey and Michel, 2003, 2006), advanced statistical functions at the gene population level (Michel, 2015) and at the gene level (Michel, 2017). However, this genetic information is not available in any biological software. Furthermore, codon usage, the most widely used statistical method in genetics, does not allow for the identification of circular codes (Michel, 2020). In this work, we investigate this open problem since 1996.

By considering circular code conditions resulting from code theory and building upon previous statistical work, a new developed statistical approach based solely on the codon usage allows determining codes that are potentially circular without needing to analyse the shifted frames 1 and 2. Surprisingly, this 1-frame statistical method reveals

E-mail address: c.michel@unistra.fr.
URL: https://dpt-info.di.unistra.fr/~c.michel/.

https://doi.org/10.1016/j.biosystems.2024.105308
Received 16 June 2024; Received in revised form 4 August 2024; Accepted 13 August 2024
Available online 17 August 2024

the circular code X (1) in bacterial genes and in average (bacterial, archaeal, eukaryotic) genes. A new parameter definition based on the codon permutation class will show that the genes of bacteria and archaea have a codon usage dispersion of the same order of magnitude, but significantly higher than the one in the genes of eukaryotes. This new statistical result could be the fundamental explanation for the variability of codes, whether circular or not, in eukaryotic genes.

This article is organized as follows. Section 2 is divided into 6 parts. The necessary definitions and notation of code, circular code, complementary map and permutation map are gathered in Section 2.1. Section 2.2 describes the acquisition of codon usage for the genomes of bacteria, archaea and eukaryotes from the codon statistics database (CSD) (Subramanian et al., 2022). The codon usage of bacterial, archaeal and eukaryotic genes is computed in Section 2.3. Section 2.4 states the circular code conditions applied to the new 1-frame statistical method for identifying trinucleotide codes potentially circular, based solely on codon usage. Section 2.5 describes two procedures to identify a trinucleotide code and a self-complementary trinucleotide code of cardinality 20. Section 2.6 proposes a new definition related to the dispersion of codon usage in a permutation class.

Section 3 is divided into 5 parts. Sections 3.1, 3.2, 3.3 identify trinucleotide codes and self-complementary trinucleotide codes of cardinality 20 in the genes of bacteria, archaea and eukaryotes, respectively. Section 3.4 identifies a trinucleotide code and a self-complementary trinucleotide code of cardinality 20 on average in genes of bacteria, archaea and eukaryotes. Section 3.5 demonstrates a lowest codon dispersion in eukaryotic genes compared to bacterial and archaeal genes.

2. Method

2.1. Definitions

In this section, only a few basic definitions are necessary to be recalled. Let $\mathcal{B} = \{A, C, G, T\}$ be the genetic alphabet where the nucleotides (also called bases) A stands for Adenine, C for Cytosine, G for Guanine and T for Thymine. Triletter words $N = N_1N_2N_3 \in \mathcal{B}^3$ are usually called trinucleotides or codons. There are 64 codons in total and each of them encodes an amino acid or the stop signal. This encoding map forms the standard genetic code that is present in all living organisms nowadays (see Woese, 1965, for more details on the biological background).

2.1.1. Code

A code Y is a set of words (perhaps of different lengths) over some alphabet Σ such that any concatenation of codewords from Y can be read in a unique way on a line. Any subset of $\mathcal{B}^3$ is a code.

2.1.2. Circular code (Arquès and Michel, 1996; Fimmel et al., 2016)

A code Y is circular over some alphabet Σ such that any concatenation of codewords from Y can be read in a unique way on a circle.

The circular code X (1) in genes belongs to the (most) important class of the 216 maximal C^3 self-complementary trinucleotide circular codes over the genetic alphabet (Arquès and Michel, 1996) whose list is explicitly given in Michel et al. (2008). Interestingly, there are (at least) 3 possible partitions of the 216 circular codes: (i) 27 classes (27×8) by letter invariance with respect to self-complementarity (Fimmel et al., 2014); (ii) 3 classes ($2 \times 56 + 104$) based on the necklace definition (Michel et al., 2008); and (iii) 2 classes ($88+128$) based on probabilities of nucleotides at the three codon sites (Koch and Lehmann, 1997; Lacan and Michel, 2001).

The properties and mathematical proofs that enable us to determine whether a code is circular or not, are outside the scope of this article. We refer the reader to the reviews (Michel, 2008; Fimmel and Strüngmann, 2018) for the biological context and the basic mathematical properties of circular codes.

2.1.3. Complementary map

The nucleotide complementary map $C : \mathcal{B} \to \mathcal{B}$ of a base $B \in \mathcal{B}$ is defined by $C(A) := T$, $C(T) := A$, $C(C) := G$ and $C(G) := C$. The trinucleotide complementary map $\mathcal{C} : \mathcal{B}^3 \to \mathcal{B}^3$ of a trinucleotide $N = N_1N_2N_3 \in \mathcal{B}^3$ is defined by $\mathcal{C}(N) := C(N_3)C(N_2)C(N_1)$. The trinucleotide set complementary map $\mathcal{C} : \mathcal{B}^3 \to \mathcal{B}^3$ of a trinucleotide code Y is defined by $\mathcal{C}(Y) = \{M : N, M \in \mathcal{B}^3, N \in Y, M = \mathcal{C}(N)\}$. A trinucleotide code $Y \subseteq \mathcal{B}^3$ is self-complementary if for every trinucleotide $N \in Y$, the complementary trinucleotide $\mathcal{C}(N)$ also belongs to Y, i.e. $Y = \mathcal{C}(Y) := \{\mathcal{C}(N) : N \in Y\}$. A self-complementary trinucleotide Y has an even cardinality.

Example 1. $\mathcal{C}(\{ACG, AGT\}) = \{ACT, CGT\}$.

Our method will involve the 30 following self-complementary codon pairs **SC**:

$$\begin{aligned}
&\{AAC, GTT\}, \{AAG, CTT\}, \{AAT, ATT\},\\
&\{ACA, TGT\}, \{ACC, GGT\}, \{ACG, CGT\},\\
&\{ACT, AGT\}, \{AGC, GCT\}, \{AGG, CCT\},\\
&\{ATA, TAT\}, \{ATC, GAT\}, \{ATG, CAT\},\\
&\{CAG, CTG\}, \{CCA, TGG\}, \{CCG, CGG\},\\
&\{CTC, GAG\}, \{GAA, TTC\}, \{GAC, GTC\},\\
&\{GCC, GGC\}, \{GCG, CGC\}, \{GTA, TAC\},\\
&\{GTG, CAC\}, \{TAG, CTA\}, \{TCA, TGA\},\\
&\{TCC, GGA\}, \{TCG, CGA\}, \{TCT, AGA\},\\
&\{TGC, GCA\}, \{TTA, TAA\}, \{TTG, CAA\}.
\end{aligned} \tag{2}$$

2.1.4. Permutation map

The trinucleotide permutation map $\mathcal{P} : \mathcal{B}^3 \to \mathcal{B}^3$ of a trinucleotide $N = N_1N_2N_3 \in \mathcal{B}^3$ is defined by $\mathcal{P}(N) := N_2N_3N_1$. The 2nd iterate is $\mathcal{P}^2(N) := N_3N_1N_2$. The trinucleotide set permutation map $\mathcal{P} : \mathcal{B}^3 \to \mathcal{B}^3$ of a trinucleotide code Y is defined by $\mathcal{P}(Y) = \{M : N, M \in \mathcal{B}^3, N \in Y, M = \mathcal{P}(N)\}$.

Example 2. $\mathcal{P}(\{ACG, AGT\}) = \{CGA, GTA\}$ and $\mathcal{P}^2(\{ACG, AGT\}) = \{GAC, TAG\}$.

Our method will involve the 20 following codon permutation classes **P**:

$$\begin{aligned}
&\{AAC, ACA, CAA\}, \{AAG, AGA, GAA\},\\
&\{AAT, ATA, TAA\}, \{ACC, CCA, CAC\},\\
&\{ACG, CGA, GAC\}, \{ACT, CTA, TAC\},\\
&\{AGC, GCA, CAG\}, \{AGG, GGA, GAG\},\\
&\{AGT, GTA, TAG\}, \{ATC, TCA, CAT\},\\
&\{ATG, TGA, GAT\}, \{ATT, TTA, TAT\},\\
&\{CCG, CGC, GCC\}, \{CCT, CTC, TCC\},\\
&\{CGG, GGC, GCG\}, \{CGT, GTC, TCG\},\\
&\{CTG, TGC, GCT\}, \{CTT, TTC, TCT\},\\
&\{GGT, GTG, TGG\}, \{GTT, TTG, TGT\}.
\end{aligned} \tag{3}$$

The complementary and permutation maps are the two most important maps related to the complementary DNA double helix. While the complementarity map is classic and well-known among biologists, the permutation map introduced in the theory of circular code in 1996 (Arquès and Michel, 1996) is equally important concerning gene coding. Indeed, these two maps are related to each other according to 2 relations. Without loss of generality, take a trinucleotide t, then (Remark 5 in Michel, 2015):

$$\mathcal{P}(t) = \mathcal{C}(\mathcal{P}^2(\mathcal{C}(t))) \quad \text{and} \quad \mathcal{P}^2(t) = \mathcal{C}(\mathcal{P}(\mathcal{C}(t))). \tag{4}$$

Furthermore, without loss of generality, take a self-complementary trinucleotide Y, i.e. $Y = C(Y)$, circular or not, then (Arquès and Michel, 1996; Remark 5 in Michel, 2021):

$$\mathcal{P}(X) = C(\mathcal{P}^2(X)) \quad \text{and} \quad \mathcal{P}^2(X) = C(\mathcal{P}(X)). \tag{5}$$

Relation (5) means the following biological property with the two complementary strands s_1 and $s_2 = C(s_1)$ of DNA. A trinucleotide t in frame 0 in s_1 is paired with its complementary trinucleotide $C(t)$ in frame 0 in s_2, and a trinucleotide t in frame 1 (2, respectively) in s_1 is paired with its complementary trinucleotide $C(t)$ in frame 2 (1, respectively) in s_2.

The definition of codon permutation class **P** is necessary to obtain a trinucleotide circular code. Indeed, a necessary condition for a trinucleotide code to be circular is that it contains no permuted trinucleotide. Thus, a circular code includes only one codon in each permutation class. As there are 20 permutation classes, the maximum cardinality of a trinucleotide circular code is 20 (with codewords of length 3 on a 4-letter alphabet).

The definition of self-complementary codon pair **SC** is necessary to obtain a self-complementary trinucleotide code, circular or not. A self-complementary trinucleotide code must contain both a codon and its complementary codon.

The two definitions **P** and **SC** are necessary to obtain a self-complementary trinucleotide circular code like X (1) identified in genes.

2.2. Data

The codon statistics database (CSD), recently developed, provides the codon usage for all the species with reference or representative genomes in RefSeq (Subramanian et al., 2022). It is free to access without registration at http://codonstatsdb.unr.edu. From this CSD, we extract (July 2022) the codon usage of genomes of three kingdoms: bacteria, archaea and eukaryota. The few exceptional genomes in which the codon usage of the stop codons is not given, are not considered.

The Archaea (Id 2157) can be directly extracted leading to 432 archaeal genomes Arc with 1,280,890 genes representing 367,937,932 codons.

The Bacteria and Eukaryota cannot be directly obtained from CSD, which is restricted to taxa for which the genetic code is homogeneous, i.e. species with the same genetic code. For example, the bacterial Mycoplasmatales uses a different genetic code with only two stop codons TAA and TAG, TGA coding Trp.

Thus, the Bacteria are constructed from the union of the 22 following bacterial classes: Acidobacteria (Id 57723), Actinobacteria (Id 201174), Aquificae (Id 187857), Bacteroidetes (Id 976), Balneolia (Id 1853221), Chlamydiia (Id 204429), Chloroflexi (Id 200795), Cyanobacteria (Id 1117), Deferribacteres (Id 68337), Deinococcus-Thermus (Id 1297), Epsilonproteobacteria (Id 29547), Firmicutes (Id 1239), Fusobacteria (Id 32066), Mycoplasmatales (Id 2085), Nitrospirae (Id 40117), Planctomycetes (Id 203682), Pseudomonadales (Id 72274), Spirochaetes (Id 203691), Synergistetes (Id 508458), Thermodesulfobacteria (Id 200940), Thermotogae (Id 200918) and Verrucomicrobia (Id 74201). This data construction leads to 8345 bacterial genomes Bac with 34,020,997 genes representing 11,087,876,805 codons.

In a similar way, the Eukaryota are constructed from the union of the 4 following eukaryotic classes: Metazoa (animals, Id 33208), Plants (Embryophyta, land plants, Id 3193; Chlorophyta, green algae, Id 3041; Rhodophyta, red algae, Id 2763), Fungi (Agaricomycotina Id 5302; Pezizomycotina, Id 147538; Saccharomyces, Id 4930; Ustilaginomycotina, Id 452284; Basidiomycota, Id 5204) and Protists (Apicomplexa, Id 5794; Kinetoplastea, Id 5653). This data construction leads to 1150 eukaryotic genomes Euk with 20,206,058 genes representing 10,374,305,634 codons.

2.3. Codon usage of bacterial, archaeal and eukaryotic genes

In Table 1, the codon usage frequency f (%) is computed in 34,020,997 genes of 8345 bacterial genomes Bac, 1,280,890 genes of 432 archaeal genomes Arc and 20,206,058 genes of 1150 eukaryotic genomes Euk obtained from the codon statistics database (CSD) (see Section 2.2). The massive statistics of genes in the three kingdoms of bacteria, archaea and eukaryotes given in Table 1, also enable the reader to carry out other research studies.

2.4. Circular code conditions

We develop a 1-frame statistical method based on circular code conditions, resulting from code theory and previous statistical work, in order to identify potential circular codes in genes (in reading frame 0) from the codon usage only, i.e. without having to consider genetic information in the shifted frames 1 and 2, i.e. without having to compute trinucleotides in the frames 1 and 2 as with all 3-frame statistical methods used to date.

(i) The 60 codons $B^3 \setminus \{AAA, CCC, GGG, TTT\}$ are considered, i.e. the 4 periodic codons are excluded.
(ii) The cardinality of trinucleotide codes to be identified potentially circular, is at most 20.
(iii) The permuted codons do not belong to trinucleotide codes to be identified potentially circular.
(iv) From Conditions (i), (ii) and (iii), the 20 codon permutation classes **P** (3) are considered.
(v) The codon usage frequency f is normalized in each permutation class **P** (3), the normalized frequency is named g. It is the most important idea that enables to neutralize the effect of codon usage in different genomes depending on a great number of biological factors.
(vi) The trinucleotide codes X_1 and X_2 in frames 1 and 2, respectively, are deduced from the trinucleotide code X in reading frame of genes as follows: $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$.

2.5. Statistical method

The procedure P for identifying a trinucleotide code Y of cardinality 20 in genes (in reading frame) is divided into 3 steps.

(P1) Compute the normalized frequency g of the 3 codons in each permutation class **P** (3) among 20. Precisely, let a permutation class $\{c_1, c_2, c_3\}$ of 3 codons c_1, c_2 and c_3 be associated with the codon usage frequencies f_1, f_2 and f_3, respectively, from Table 1. Then, a normalized frequency g_i of a codon c_i of frequency f_i in its permutation class is (obviously) equal to

$$g_i = \frac{f_i}{\sum_{i=1}^{3} f_i}. \tag{6}$$

Note that $\sum_{i=1}^{3} g_i = 1$. In the tables below, g will be given in %.

(P2) For each permutation class **P** (3), identify the codon of highest normalized frequency $g_{\max}$ such that

$$g_{\max} = \max\left\{g_i \ : \ i \in \{1,2,3\}\right\}. \tag{7}$$

(P3) Deduce the trinucleotide code Y of cardinality 20 with the 20 codons of highest normalized frequencies $g_{\max}$ in the 20 codon permutation classes **P** (3).

Remark 1. The normalized frequency g of a codon c is not necessary to find the maximum value (Procedure (P2)) but it is important for Procedure (P3) that ranks the codons from 1st to 20th place. Remark 7 in Section 3.1 gives an example with bacterial genes.

Table 1
Codon usage frequency f (%) of 34,020,997 genes of 8345 bacterial genomes Bac (11,087,876,805 codons), 1,280,890 genes of 432 archaeal genomes Arc (367,937,932 codons) and 20,206,058 genes of 1150 eukaryotic genomes Euk (10,374,305,634 codons) obtained from the codon statistics database (CSD) (see Section 2.2).

	Bac	Arc	Euk		Bac	Arc	Euk		Bac	Arc	Euk		Bac	Arc	Euk
AAA	2.37	1.74	2.60	*CAA*	1.20	0.59	1.61	*GAA*	2.91	3.04	2.96	*TAA*	0.12	0.11	0.06
AAC	1.80	2.01	2.12	*CAC*	1.27	1.40	1.37	*GAC*	3.42	5.25	2.57	*TAC*	1.56	2.13	1.55
AAG	2.00	1.88	3.12	*CAG*	2.16	1.78	2.75	*GAG*	3.22	5.07	3.69	*TAG*	0.06	0.08	0.05
AAT	1.64	1.05	1.95	*CAT*	0.78	0.46	1.17	*GAT*	2.25	2.02	2.64	*TAT*	1.35	0.96	1.24
ACA	0.93	0.85	1.57	*CCA*	0.63	0.67	1.61	*GCA*	1.43	1.48	1.76	*TCA*	0.64	0.68	1.38
ACC	2.56	2.16	1.63	*CCC*	1.39	1.35	1.49	*GCC*	4.32	3.49	2.23	*TCC*	1.25	1.08	1.60
ACG	1.51	2.20	0.94	*CCG*	2.14	1.91	0.97	*GCG*	3.22	3.42	1.12	*TCG*	1.23	1.62	0.90
ACT	0.76	0.73	1.40	*CCT*	0.67	0.52	1.58	*GCT*	1.30	1.05	2.02	*TCT*	0.67	0.56	1.60
AGA	0.54	0.69	1.27	*CGA*	0.43	0.72	0.71	*GGA*	1.39	1.52	1.76	*TGA*	0.15	0.19	0.08
AGC	1.28	1.27	1.72	*CGC*	2.26	1.80	1.03	*GGC*	3.63	3.40	1.92	*TGC*	0.53	0.41	1.05
AGG	0.38	0.81	1.13	*CGG*	1.64	1.51	0.82	*GGG*	1.43	1.78	1.26	*TGG*	1.31	1.08	1.22
AGT	0.69	0.63	1.27	*CGT*	0.80	0.50	0.68	*GGT*	1.54	1.34	1.43	*TGT*	0.29	0.42	0.95
ATA	0.82	1.33	1.02	*CTA*	0.45	0.47	0.77	*GTA*	1.05	0.86	0.84	*TTA*	1.25	0.75	0.92
ATC	2.76	2.90	2.04	*CTC*	2.23	3.86	1.83	*GTC*	2.66	4.30	1.58	*TTC*	2.11	2.58	2.01
ATG	2.15	2.00	2.26	*CTG*	3.86	2.25	2.87	*GTG*	2.66	1.88	2.28	*TTG*	1.21	0.68	1.58
ATT	1.91	1.31	1.75	*CTT*	1.01	1.09	1.46	*GTT*	1.20	1.36	1.51	*TTT*	1.61	1.00	1.73

Remark 2. In a codon usage (with real numbers), the codon associated with $g_{\max}$ is unique. In case there are 2 or 3 codons with the same value $g_{\max}$, any codon with $g_{\max}$ can be chosen or its permutation class can be discarded.

The procedure Q for identifying a self-complementary trinucleotide code Z of cardinality 20 in genes (in reading frame) is divided into 2 steps.

(Q1) Compute the mean normalized frequency $\overline{g}$ of the 2 codons in each self-complementary pair **SC** (2) among 30. Precisely, let a self-complementary pair $\{c, C(c)\}$ of 2 codons $c_1 = c$ and $c_2 = C(c)$ be associated with the normalized frequencies g_1 and g_2. Then, a mean normalized frequency $\overline{g}$ of 2 codons in its self-complementary pair is (obviously) equal to

$$\overline{g} = \frac{g_1 + g_2}{2} \tag{8}$$

where g_i (6) is the normalized frequency of a codon c_i in the self-complementary codon pair $\{c_1, c_2\}$.

(Q2) Deduce the self-complementary trinucleotide code Z of cardinality 20 with the 20 codons of highest mean normalized frequencies $\overline{g}$ in the 10 self-complementary codon pairs **SC** (2).

The procedures P and Q can be easily programmed in a programming language or spreadsheet program. Several examples of computation of g and $\overline{g}$ will be given in Section 3 to make it easier for the reader to implement the software.

Remark 3. The procedure P forces maximality and the procedure Q forces both maximality and self-complementarity. These two properties appear naturally and are totally unexpected in the two original 3-frames methods (Arquès and Michel, 1996, 1997).

Remark 4. The trinucleotide code Y and the self-complementary trinucleotide code Z obtained by the 2 procedures P and Q verify the necessary conditions of maximal cardinality 20 and the absence of permuted codons, but they are not necessarily circular which require additional mathematical properties (not in the scope of this paper, see e.g. Arquès and Michel, 1996; Fimmel et al., 2016).

Remark 5. The 2 procedures P and Q can obviously identify trinucleotide codes Y and self-complementary trinucleotide codes Z of cardinality less than 20 by considering a subset of codons among the 20 highest normalized frequencies $g_{\max}$ or a subset of self-complementary codon pairs among the 10 mean normalized frequencies $\overline{g}$. This approach is particularly interesting for identifying non-maximal circular codes when a code of cardinality 20 (maximal) is not circular.

Remark 6. As the shifted frames 1 and 2 are not analysed, it is impossible to determine the trinucleotides of the codes X_1 and X_2 in frames 1 and 2, respectively. Condition (vi) must be applied for this case.

2.6. Codon usage dispersion in a permutation class

We extend here the definition of codon usage dispersion introduced in a previous work (Michel and Sereni, 2023) to a codon permutation class **P** (3). A codon usage in a codon permutation class $\{c_1, c_2, c_3\} \in \mathbf{P}$ of cardinality 3 is uniform if the 3 codons c_1, c_2 and c_3 have the same occurrence frequency $\frac{1}{3}$. We can now define the dispersion of codon usage in a permutation class.

Definition 1 (*Codon Usage Dispersion in a Codon Permutation Class*)**.** The dispersion d of codon usage in a codon permutation class **P** (3) is given by

$$d = \sum_{i=1}^{3} \left| g_i - \frac{1}{3} \right| \tag{9}$$

where g_i (6) is the normalized frequency of a codon c_i in the codon permutation class $\{c_1, c_2, c_3\}$.

This new parameter definition d will show that the genes of bacteria and archaea have a codon usage dispersion of the same order of magnitude, but significantly higher than the one in the genes of eukaryotes. This result could be the fundamental explanation for the greater variability of codes in eukaryotic genes, compared to bacterial and archaeal genes.

3. Results

The two procedures P and Q will be applied to the codon usage of genes of bacteria, archaea and eukaryotes in order to identify trinucleotide codes which are potentially circular.

3.1. Trinucleotide codes in bacterial genes

In Table 2, the codon usage frequency f (%) of 34,020,997 genes of 8345 bacterial genomes Bac is copied from the codon usage Table 1 into each codon permutation class **P** (3) and Procedure (P1) computes the normalized frequency g (6) given in %.

For the reader's convenience, an example is given for computing g (6).

Table 2
Frequency f (codon usage from Table 1 in %) and normalized frequency g (6) (%) of 34,020,997 genes of 8345 bacterial genomes Bac (11,087,876,805 codons) for each codon permutation class **P** (3) (frequency sum equal to 100%; Procedure P). The codons are ordered according to the 20 permutation classes and the numbers are rounded to the 2nd decimal place.

	f	g		f	g		f	g		f	g
AAC	1.80	45.83	ACT	0.76	27.41	ATG	2.15	47.25	CGT	0.80	17.07
ACA	0.93	23.61	CTA	0.45	16.16	TGA	0.15	3.26	GTC	2.66	56.68
CAA	1.20	30.57	TAC	1.56	56.43	GAT	2.25	49.49	TCG	1.23	26.25
Total	3.92	100.00	Total	2.76	100.00	Total	4.55	100.00	Total	4.70	100.00
AAG	2.00	36.69	AGC	1.28	26.25	ATT	1.91	42.41	CTG	3.86	67.81
AGA	0.54	9.93	GCA	1.43	29.38	TTA	1.25	27.68	TGC	0.53	9.34
GAA	2.91	53.39	CAG	2.16	44.37	TAT	1.35	29.91	GCT	1.30	22.85
Total	5.45	100.00	Total	4.87	100.00	Total	4.51	100.00	Total	5.70	100.00
AAT	1.64	63.75	AGG	0.38	7.64	CCG	2.14	24.54	CTT	1.01	26.54
ATA	0.82	31.67	GGA	1.39	27.77	CGC	2.26	25.89	TTC	2.11	55.72
TAA	0.12	4.57	GAG	3.22	64.59	GCC	4.32	49.57	TCT	0.67	17.73
Total	2.57	100.00	Total	4.99	100.00	Total	8.72	100.00	Total	3.79	100.00
ACC	2.56	57.46	AGT	0.69	38.31	CCT	0.67	16.09	GGT	1.54	27.92
CCA	0.63	14.05	GTA	1.05	58.22	CTC	2.23	53.86	GTG	2.66	48.24
CAC	1.27	28.49	TAG	0.06	3.47	TCC	1.25	30.05	TGG	1.31	23.84
Total	4.46	100.00	Total	1.80	100.00	Total	4.15	100.00	Total	5.51	100.00
ACG	1.51	28.24	ATC	2.76	65.96	CGG	1.64	19.33	GTT	1.20	44.54
CGA	0.43	8.01	TCA	0.64	15.31	GGC	3.63	42.75	TTG	1.21	44.70
GAC	3.42	63.75	CAT	0.78	18.73	GCG	3.22	37.91	TGT	0.29	10.76
Total	5.37	100.00	Total	4.18	100.00	Total	8.49	100.00	Total	2.70	100.00

Example 3. The normalized frequency g for AAC in Table 2 needs the frequencies $f = 0.0180$ for AAC, $f = 0.0093$ for ACA and $f = 0.0120$ for CAA

$$g = \frac{0.0180}{0.0180 + 0.0093 + 0.0120} = \frac{0.0180}{0.0392} \approx 0.4583 = 45.83\%.$$

The denominator $0.0392 = 3.92\%$ and the frequency $g = 45.83\%$ are given in Table 2.

In Table 3, Procedures (P2) and (P3) identify the bacterial trinucleotide code Y_{Bac} in reading frame of genes, which is ordered from the highest to the lowest frequency

$$Y_{\text{Bac}} = \{CTG, ATC, GAG, AAT, GAC, GTA, ACC, GTC, TAC, TTC, CTC, GAA, GCC, GAT, GTG, AAC, TTG, CAG, GGC, ATT\}. \quad (10)$$

The 14 codons of highest frequencies in Y_{Bac} belong to the circular code X (1) in genes and Y_{Bac} contains 18 codons of X, i.e. $|Y_{\text{Bac}} \cap X| = 18$, except $\{GTG, TTG\} \notin X$.

Remark 7. As already mentioned in Remark 1, the frequencies f and g identify the same trinucleotide code Y_{Bac} of cardinality 20. But, the occurrence frequencies of selected trinucleotides are completely different in Y_{Bac}. For example in Y_{Bac}, the codon CTG has the highest rank with the normalized frequency $g = 67.81\%$ (see Tables 2 and 3) whereas it is the codon GCC which has the highest rank with the frequency $f = 4.32\%$ (see Tables 1 and 2).

For the reader's convenience, an example is given for computing the mean normalized frequency $\overline{g}$ (8) in a self-complementary codon pair.

Example 4. The mean normalized frequency $\overline{g}$ for the self-complementary codon pair $\{AAC, GTT\}$ needs the frequencies $g = 0.4583$ for AAC (see Table 2 and Example 3) and $g = 0.4454$ for GTT (see Table 2)

$$\overline{g} = \frac{0.4583 + 0.4454}{2} \approx 0.4518 = 45.18\%.$$

The pair $\{AAC, GTT\}$ and its frequency $\overline{g} = 45.18\%$ are in the 9th rank in Table 3.

In Table 3, Procedures (Q1) and (Q2) identify the bacterial self-complementary trinucleotide code Z_{Bac} in reading frame, which is ordered by self-complementary codon pairs from the highest to the lowest frequency

$$Z_{\text{Bac}} = \{GAC, GTC, CTC, GAG, ATC, GAT, GTA, TAC, CAG, CTG, GAA, TTC, AAT, ATT, GCC, GGC, AAC, GTT, ACC, GGT\}. \quad (11)$$

Very interestingly and surprisingly, the bacterial self-complementary trinucleotide code Z_{Bac} identified by this 1-frame method (codon usage), is the circular code X (1) in genes

$$Z_{\text{Bac}} = X. \quad (12)$$

In order to have a statistical order of magnitude, let n be the total number of self-complementary codon pairs **SC** (2) and let m be the number of self-complementary codon pairs of a trinucleotide code Z, usually $m < n$. Then, the probability p that the m self-complementary codon pairs of Z occurs in the m first ranks among n is equal to

$$p = \frac{m!(n-m)!}{n!}. \quad (13)$$

If Z is the circular code X (1), i.e. $m = 10$ among $n = 30$, then $p \approx 3 \times 10^{-8}$.

Table 8 in Appendix A shows that the circular code X (1) observed in genes cannot be identified in 34,020,997 genes of 8345 bacterial genomes Bac without normalization of the frequency f in the self-complementary codon pairs **SC** (2). Indeed, if the mean frequency $\overline{f} = (f_1 + f_2)/2$ (defined in the same way as $\overline{g}$ (8)) computed from Table 1 in % is used instead of $\overline{g}$, then the code Z_{Bac} contains only 14 codons of X and is not circular for two reasons: (i) the pair $\{GCC, GGC\}$ in the 1st rank and the pair $\{CCG, CGG\}$ in the 10th rank have permuted codons; and (ii) the pair $\{GCG, CGC\}$ in the 4th rank is not circular (not detailed).

Furthermore, several random genetic codes are generated according to a uniform distribution of the 64 codon frequencies ($f = 1/64$ for every codon). Then, the two procedures P and Q are applied for computing the frequencies g (6) and $\overline{g}$ (8) in these random genes. The 20 codons and the 10 self-complementary codon pairs of the circular code X (1) in genes always have different ranks in each random genetic code. Furthermore, an example of random genetic code Rand is given in Table 9 in Appendix B with random trinucleotide frequencies f. Table 10 in Appendix B gives the trinucleotide code Y_{Rand} and the self-complementary trinucleotide code Z_{Rand}. In Y_{Rand}, 8 codons among 20

Table 3
Identification of a code Y_{Bac} (Procedure P) and a self-complementary code Z_{Bac} (Procedure Q) of cardinality 20 in 34,020,997 genes of 8345 bacterial genomes Bac (11,087,876,805 codons). The codons and the self-complementary codon pairs are given in descending order of frequencies g (6) and $\overline{g}$ (8), respectively. The codons in bold belong to the circular code X (1) observed in genes. The normalized frequency g (from Table 2 in %) identifies the codon of preferential occurrence (highest frequency) in each of the 20 codon permutation classes **P** (3) leading to a code Y_{Bac} of cardinality 20 with a rank from 1 to 20. The mean normalized frequency $\overline{g}$ (%) is computed for the 30 self-complementary codon pairs **SC** (2) and identifies the self-complementary codon pairs of preferential occurrence (highest frequency) leading to a self-complementary code Z_{Bac} of cardinality 20 with a rank from 1 to 10. The code Y_{Bac} contains 18 codons of the circular code X (1) in genes and the codes Z_{Bac} and X are identical.

Genes of bacteria Bac					
Code of cardinality 20			Self-complementary code of cardinality 20		
Rank	Y_{Bac}	g	Rank	Z_{Bac}	$\overline{g}$
1	$\boldsymbol{CTG}$	67.81	1	$\{\boldsymbol{GAC},\boldsymbol{GTC}\}$	60.22
2	$\boldsymbol{ATC}$	65.96	2	$\{\boldsymbol{CTC},\boldsymbol{GAG}\}$	59.23
3	$\boldsymbol{GAG}$	64.59	3	$\{\boldsymbol{ATC},\boldsymbol{GAT}\}$	57.72
4	$\boldsymbol{AAT}$	63.75	4	$\{\boldsymbol{GTA},\boldsymbol{TAC}\}$	57.32
5	$\boldsymbol{GAC}$	63.75	5	$\{\boldsymbol{CAG},\boldsymbol{CTG}\}$	56.09
6	$\boldsymbol{GTA}$	58.22	6	$\{\boldsymbol{GAA},\boldsymbol{TTC}\}$	54.56
7	$\boldsymbol{ACC}$	57.46	7	$\{\boldsymbol{AAT},\boldsymbol{ATT}\}$	53.08
8	$\boldsymbol{GTC}$	56.68	8	$\{\boldsymbol{GCC},\boldsymbol{GGC}\}$	46.16
9	$\boldsymbol{TAC}$	56.43	9	$\{\boldsymbol{AAC},\boldsymbol{GTT}\}$	45.18
10	$\boldsymbol{TTC}$	55.72	10	$\{\boldsymbol{ACC},\boldsymbol{GGT}\}$	42.69
11	$\boldsymbol{CTC}$	53.86		$\{GTG, CAC\}$	38.36
12	$\boldsymbol{GAA}$	53.39		$\{TTG, CAA\}$	37.63
13	$\boldsymbol{GCC}$	49.57		$\{ATG, CAT\}$	32.99
14	$\boldsymbol{GAT}$	49.49		$\{ACT, AGT\}$	32.86
15	GTG	48.24		$\{GCG, CGC\}$	31.90
16	$\boldsymbol{AAC}$	45.83		$\{AAG, CTT\}$	31.62
17	TTG	44.70		$\{ATA, TAT\}$	30.79
18	$\boldsymbol{CAG}$	44.37		$\{TCC, GGA\}$	28.91
19	$\boldsymbol{GGC}$	42.75		$\{AGC, GCT\}$	24.55
20	$\boldsymbol{ATT}$	42.41		$\{ACG, CGT\}$	22.65
				$\{CCG, CGG\}$	21.94
				$\{TGC, GCA\}$	19.36
				$\{CCA, TGG\}$	18.95
				$\{ACA, TGT\}$	17.18
				$\{TCG, CGA\}$	17.13
				$\{TTA, TAA\}$	16.13
				$\{TCT, AGA\}$	13.83
				$\{AGG, CCT\}$	11.86
				$\{TAG, CTA\}$	9.81
				$\{TCA, TGA\}$	9.28

belong to the circular code X (1) in genes. In Z_{Rand}, 6 codons among 20 belong to X as the 10 self-complementary codon pairs of X occur in the ranks 2, 4, 5, 12, 13, 18, 23, 26, 28 and 30, thus only 3 pairs have a rank ≤ 10.

Remark 8. In the limit case, if the 64 codons all have an exactly frequency $f = 1/64$ then no preferential codon can be identified with the procedures P and Q.

In this random context, we are investigating the "inverse" theoretical problem with the relation between a uniform amino acid code and the preferential codons of X (1) occurring in the codon permutation class **P** (3). Two cases are considered whether every amino acid has the same probability: (i) $f = 1/20$ and $f = 0$ for the 3 stop codons $\{TAA, TAG, TGA\}$; and (ii) $f' = 1/21$ including the 3 stop codons (Table 11 in Appendix C).

Example 5. The codon $\{ATG\}$ coding Met has the frequencies $f = 1/20 = 5\%$ and $f' = 1/21 = 4.76\%$. Each codon in $\{GCA, GCC, GCG, GCT\}$ coding Ala has the frequencies $f = 1/80 = 1.25\%$ and $f' = 1/84 = 1.19\%$.

Table 12 in Appendix C identifies 3 cases for the codons of X occurring in the codon permutation class **P** (3):

(i) 7 codons (8 codons with GTA) of X have a preferential occurrence (highest frequency) in **P**:

$\{AAT, CAG, GAC, GAG, GTC, TAC, TTC\}$ and GTA (with f).

(ii) 4 codons of X have an indeterminate preferential occurrence (2 codons with highest frequencies) in **P**:

$\{AAC, GAA, GCC, GGC\}$.

(iii) 8 codons (9 codons with GTA) of X have no preferential occurrence (not the highest frequency) in **P**:

$\{ACC, ATC, ATT, CTC, CTG, GAT, GGT, GTT\}$ and GTA (with f').

The frequencies f and f' lead to the same classification except for GTA. As the random genetic code, the uniform amino acid code cannot construct the circular code X (1) observed in genes.

3.2. Trinucleotide codes in archaeal genes

We apply the same approach in 1,280,890 genes of 432 archaeal genomes Arc: the codon usage frequency f (from Table 1 in %) and the normalized frequency g (6) (%) computed by Procedure (P1) are given for each codon permutation class **P** (3) in Table 13 in Appendix D.

In Table 4, Procedures (P2) and (P3) identify the archaeal trinucleotide code Y_{Arc} in reading frame of genes, which is ordered from the highest to the lowest frequency

$$Y_{\text{Arc}} = \{ATC, CTC, GAG, GTC, GAC, TAC, TTC, CTG, AAC, GTT, GTA, GAA, ATA, ACC, GCC, GAT, GTG, ATT, GCG, CAG\}. \quad (14)$$

The 12 codons of highest frequencies in Y_{Arc} belong to the circular code X (1) in genes and Y_{Arc} contains 17 codons of X, i.e. $|Y_{\text{Arc}} \cap X| = 17$, except $\{ATA, GTG, GCG\} \notin X$.

In Table 4, Procedures (Q1) and (Q2) identify the archaeal self-complementary trinucleotide code Z_{Arc} in reading frame, which is ordered by self-complementary codon pairs from the highest to the lowest frequency

$$Z_{\text{Arc}} = \{CTC, GAG, GAC, GTC, ATC, GAT, GTA, TAC, GAA, TTC, AAC, GTT, CAG, CTG, GCC, GGC, AAT, ATT, ATA, TAT\}. \quad (15)$$

The 9 self-complementary codon pairs of highest frequencies in Z_{Arc} belong to the circular code X (1). The code Z_{Arc} is not circular as $\{AAT, ATT\}$ and $\{ATA, TAT\}$ are permuted. However, if the 10th codon pair $\{ATA, TAT\}$ is replaced by the 11th codon pair $\{ACC, GGT\}$ of close frequencies (42.55% and 41.04%, respectively), then $Z_{\text{Arc}} = X$, result summarized as follows

$$Z_{\text{Arc}} = \{X : \{ATA, TAT\} \in Z_{\text{Arc}} \rightarrow \{ACC, GGT\} \in X\}. \quad (16)$$

Remark 9. The number 367,937,932 of archaeal codons represents only $\approx 3\%$ of the number 11,087,876,805 of bacterial codons (see Section 2.2), leading to data of very different sizes and less stability with the non-highest codon frequencies.

3.3. Trinucleotide codes in eukaryotic genes

We repeat the approach in 20,206,058 genes of 1150 eukaryotic genomes Euk: the codon usage frequency f (from Table 1 in %) and the normalized frequency g (6) (%) computed by Procedures (P1) are given for each codon permutation class **P** (3) in Table 14 in Appendix E.

Table 4
Identification of a code Y_{Arc} (Procedure P) and a self-complementary code Z_{Arc} (Procedure Q) of cardinality 20 in 1,280,890 genes of 432 archaeal genomes Arc (367,937,932 codons). The codons and the self-complementary codon pairs are given in descending order of frequencies g (6) and $\overline{g}$ (8), respectively. The codons in bold belong to the circular code X (1) observed in genes. The normalized frequency g (from Table 13 in %) identifies the codon of preferential occurrence (highest frequency) in each of the 20 codon permutation classes **P** (3) leading to a code Y_{Arc} of cardinality 20 with a rank from 1 to 20. The mean normalized frequency $\overline{g}$ (%) is computed for the 30 self-complementary codon pairs **SC** (2) and identifies the self-complementary codon pairs of preferential occurrence (highest frequency) leading to a self-complementary code Z_{Arc} of cardinality 20 with a rank from 1 to 10. The code Y_{Arc} contains 17 codons of the circular code X (1) in genes and the code Z_{Arc} contains 18 codons of X.

Genes of archaea Arc					
Code of cardinality 20			Self-complementary code of cardinality 20		
Rank	Y_{Arc}	g	Rank	Z_{Arc}	$\overline{g}$
1	$\boldsymbol{ATC}$	71.80	1	$\{\boldsymbol{CTC,GAG}\}$	69.56
2	$\boldsymbol{CTC}$	70.67	2	$\{\boldsymbol{GAC,GTC}\}$	65.61
3	$\boldsymbol{GAG}$	68.46	3	$\{\boldsymbol{ATC,GAT}\}$	59.94
4	$\boldsymbol{GTC}$	66.96	4	$\{\boldsymbol{GTA,TAC}\}$	59.45
5	$\boldsymbol{GAC}$	64.26	5	$\{\boldsymbol{GAA,TTC}\}$	57.62
6	$\boldsymbol{TAC}$	64.01	6	$\{\boldsymbol{AAC,GTT}\}$	56.72
7	$\boldsymbol{TTC}$	61.02	7	$\{\boldsymbol{CAG,CTG}\}$	50.00
8	$\boldsymbol{CTG}$	60.57	8	$\{\boldsymbol{GCC,GGC}\}$	44.64
9	$\boldsymbol{AAC}$	58.15	9	$\{\boldsymbol{AAT,ATT}\}$	42.86
10	$\boldsymbol{GTT}$	55.29	10	$\{ATA,TAT\}$	42.55
11	$\boldsymbol{GTA}$	54.88		$\{\boldsymbol{ACC,GGT}\}$	41.04
12	$\boldsymbol{GAA}$	54.22		$\{GTG,CAC\}$	38.43
13	ATA	53.40		$\{GCG,CGC\}$	33.07
14	$\boldsymbol{ACC}$	51.02		$\{ACT,AGT\}$	31.09
15	$\boldsymbol{GCC}$	48.48		$\{AAG,CTT\}$	29.64
16	$\boldsymbol{GAT}$	48.08		$\{ATG,CAT\}$	29.40
17	GTG	43.73		$\{AGC,GCT\}$	28.16
18	$\boldsymbol{ATT}$	43.50		$\{TTG,CAA\}$	22.34
19	GCG	41.11		$\{CCG,CGG\}$	22.29
20	$\boldsymbol{CAG}$	39.42		$\{TGC,GCA\}$	21.84
				$\{ACA,TGT\}$	20.95
				$\{CCA,TGG\}$	20.53
				$\{TCC,GGA\}$	20.17
				$\{ACG,CGT\}$	17.35
				$\{TCG,CGA\}$	17.04
				$\{TTA,TAA\}$	14.58
				$\{TCT,AGA\}$	12.74
				$\{TCA,TGA\}$	10.66
				$\{AGG,CCT\}$	10.27
				$\{TAG,CTA\}$	9.46

Table 5
Identification of a code Y_{Euk} (Procedure P) and a self-complementary code Z_{Euk} (Procedure Q) of cardinality 20 in 20,206,058 genes of 1150 eukaryotic genomes Euk (10,374,305,634 codons). The codons and the self-complementary codon pairs are given in descending order of frequencies g (6) and $\overline{g}$ (8), respectively. The codons in bold belong to the circular code X (1) observed in genes. The normalized frequency g (from Table 14 in %) identifies the codon of preferential occurrence (highest frequency) in each of the 20 codon permutation classes **P** (3) leading to a code Y_{Euk} of cardinality 20 with a rank from 1 to 20. The mean normalized frequency $\overline{g}$ (%) is computed for the 30 self-complementary codon pairs **SC** (2) and identifies the self-complementary codon pairs of preferential occurrence (highest frequency) leading to a self-complementary code Z_{Euk} of cardinality 20 with a rank from 1 to 10. The code Y_{Euk} contains 16 codons of the circular code X (1) in genes and the code Z_{Euk} contains 18 codons of X.

Genes of eukaryotes Euk					
Code of cardinality 20			Self-complementary code of cardinality 20		
Rank	Y_{Euk}	g	Rank	Z_{Euk}	$\overline{g}$
1	$\boldsymbol{AAT}$	64.35	1	$\{\boldsymbol{GAC,GTC}\}$	55.40
2	$\boldsymbol{GAC}$	60.87	2	$\{\boldsymbol{AAT,ATT}\}$	54.52
3	AGT	58.93	3	$\{\boldsymbol{GCC,GGC}\}$	51.26
4	$\boldsymbol{GAG}$	56.10	4	$\{\boldsymbol{ATC,GAT}\}$	48.68
5	$\boldsymbol{GAT}$	52.92	5	$\{ACT,AGT\}$	48.22
6	$\boldsymbol{GCC}$	52.72	6	$\{\boldsymbol{CTC,GAG}\}$	46.31
7	$\boldsymbol{GTC}$	49.94	7	$\{\boldsymbol{CAG,CTG}\}$	46.22
8	$\boldsymbol{GGC}$	49.80	8	$\{\boldsymbol{GTA,TAC}\}$	40.28
9	$\boldsymbol{CTG}$	48.29	9	$\{\boldsymbol{GAA,TTC}\}$	39.99
10	GTG	46.23	10	$\{\boldsymbol{AAC,GTT}\}$	38.67
11	$\boldsymbol{ATT}$	44.70		$\{GTG,CAC\}$	37.96
12	$\boldsymbol{ATC}$	44.44		$\{AAG,CTT\}$	35.61
13	$\boldsymbol{CAG}$	44.15		$\{ATG,CAT\}$	35.40
14	AAG	42.43		$\{TTG,CAA\}$	34.72
15	$\boldsymbol{TAC}$	41.68		$\{ATA,TAT\}$	32.69
16	$\boldsymbol{AAC}$	39.98		$\{\boldsymbol{ACC,GGT}\}$	32.19
17	$\boldsymbol{TTC}$	39.65		$\{AGC,GCT\}$	30.82
18	TTG	39.08		$\{CCA,TGG\}$	29.85
19	$\boldsymbol{CTC}$	36.51		$\{TCC,GGA\}$	29.29
20	$\boldsymbol{ACC}$	35.38		$\{GCG,CGC\}$	26.64
				$\{ACA,TGT\}$	26.61
				$\{AGG,CCT\}$	24.41
				$\{TCT,AGA\}$	24.40
				$\{TGC,GCA\}$	22.95
				$\{TCG,CGA\}$	22.62
				$\{CCG,CGG\}$	22.10
				$\{ACG,CGT\}$	21.97
				$\{TCA,TGA\}$	15.93
				$\{TTA,TAA\}$	12.78
				$\{TAG,CTA\}$	11.50

In Table 5, Procedures (P2) and (P3) identify the eukaryotic trinucleotide code Y_{Euk} in reading frame of genes, which is ordered from the highest to the lowest frequency

$$Y_{\text{Euk}} = \{AAT, GAC, AGT, GAG, GAT, GCC, GTC, GGC, CTG, GTG, ATT, ATC, CAG, AAG, TAC, AAC, TTC, TTG, CTC, ACC\}. \quad (17)$$

The code Y_{Euk} contains 16 codons of the circular code X (1) in genes, i.e. $|Y_{\text{Euk}} \cap X| = 16$, except $\{AGT, GTG, AAG, TTG\} \notin X$.

In Table 5, Procedures (Q1) and (Q2) identify the eukaryotic self-complementary trinucleotide code Z_{Euk} in reading frame, which is ordered by self-complementary codon pairs from the highest to the lowest frequency

$$Z_{\text{Euk}} = \{GAC, GTC, AAT, ATT, GCC, GGC, ATC, GAT, ACT, AGT, CTC, GAG, CAG, CTG, GTA, TAC, GAA, TTC, AAC, GTT\}. \quad (18)$$

In Z_{Euk}, the codon pairs $\{ACT, AGT\} \notin X$ in 5th rank and $\{GTA, TAC\} \in X$ in 8th rank are permuted. In order to identify a code that could be maximal and circular, the codon pair $\{GTA, TAC\}$ must be removed and another codon pair must be added. By using the codon pair $\{GTG, CAC\} \notin X$ in 11th rank (a permuted codon pair of $\{ACC, GGT\} \in X$), the trinucleotide code V_{Euk} is

$$V_{\text{Euk}} = \{GAC, GTC, AAT, ATT, GCC, GGC, ATC, GAT, ACT, AGT, CTC, GAG, CAG, CTG, GAA, TTC, AAC, GTT, GTG, CAC\}. \quad (19)$$

Both codes Z_{Euk} and V_{Euk} contains 16 codons of X. The code Z_{Euk} is not circular as there are permuted codons, but the code V_{Euk} is circular (proof not shown), precisely V_{Euk} is a maximal C^3 self-complementary trinucleotide circular code, as the class of X (1) in genes.

These eukaryotic statistical results lead to several observations.

- The statistical approach by codon usage only is less effective with genes of eukaryotes than with the genes of bacteria and archaea, in particular because permuted codon pairs can occur in the top 10 ranks. A reason is the lack of statistical analysis of trinucleotides in frames 1 and 2.
- In general terms, there is a greater statistical variability with eukaryotic genes for identifying circular codes, variability observed since 1996 with the 6 statistical methods using the 3 frames.
- From a biological point of view, this variability may be associated with a greater variety with the eukaryotic genes. For example,

Table 6
Identification of a self-complementary code Z_{Genes} (Procedure Q) of cardinality 20 in the 3 kingdoms of bacterial, archaeal and eukaryotic genes. The self-complementary codon pairs are given in descending order of frequencies $\overline{\overline{g}}$ (20). The codons in bold belong to the circular code X (1) observed in genes. The mean gene frequency $\overline{\overline{g}}$ (%) is computed for the 30 self-complementary codon pairs **SC** (2) and identifies the self-complementary codon pairs of preferential occurrence (highest frequency) leading to a self-complementary code Z_{Genes} of cardinality 20 with a rank from 1 to 10. The code Z_{Genes} is identical to the circular code X (1) in genes.

Genes (Bac, Arc, Euk)		
Self-complementary code of cardinality 20		
Rank	Z_{Genes}	$\overline{\overline{g}}$
1	$\{\boldsymbol{GAC,GTC}\}$	60.41
2	$\{\boldsymbol{CTC,GAG}\}$	58.37
3	$\{\boldsymbol{ATC,GAT}\}$	55.45
4	$\{\boldsymbol{GTA,TAC}\}$	52.35
5	$\{\boldsymbol{CAG,CTG}\}$	50.77
6	$\{\boldsymbol{GAA,TTC}\}$	50.72
7	$\{\boldsymbol{AAT,ATT}\}$	50.16
8	$\{\boldsymbol{GCC,GGC}\}$	47.35
9	$\{\boldsymbol{AAC,GTT}\}$	46.86
10	$\{\boldsymbol{ACC,GGT}\}$	38.64
	$\{GTG,CAC\}$	38.25
	$\{ACT,AGT\}$	37.39
	$\{ATA,TAT\}$	35.35
	$\{ATG,CAT\}$	32.60
	$\{AAG,CTT\}$	32.29
	$\{TTG,CAA\}$	31.56
	$\{GCG,CGC\}$	30.54
	$\{AGC,GCT\}$	27.85
	$\{TCC,GGA\}$	26.12
	$\{CCA,TGG\}$	23.11
	$\{CCG,CGG\}$	22.11
	$\{ACA,TGT\}$	21.58
	$\{TGC,GCA\}$	21.38
	$\{ACG,CGT\}$	20.66
	$\{TCG,CGA\}$	18.93
	$\{TCT,AGA\}$	16.99
	$\{AGG,CCT\}$	15.51
	$\{TTA,TAA\}$	14.50
	$\{TCA,TGA\}$	11.96
	$\{TAG,CTA\}$	10.26

Table 7
Codon usage dispersion d (9) in the 20 codon permutation classes **P** (3) of 34,020,997 genes of 8345 bacterial genomes Bac (11,087,876,805 codons), 1,280,890 genes of 432 archaeal genomes Arc (367,937,932 codons) and 20,206,058 genes of 1150 eukaryotic genomes Euk (10,374,305,634 codons). The last row gives the dispersion mean $\overline{d}$. The numbers are rounded to the 2nd decimal place.

	Bac	Arc	Euk
$\{AAC,ACA,CAA\}$	24.99	49.63	13.29
$\{AAG,AGA,GAA\}$	46.81	42.20	32.18
$\{AAT,ATA,TAA\}$	60.84	57.92	62.52
$\{ACC,CCA,CAC\}$	48.26	35.38	7.28
$\{ACG,CGA,GAC\}$	60.84	61.86	55.08
$\{ACT,CTA,TAC\}$	46.20	61.35	25.06
$\{AGC,GCA,CAG\}$	22.07	12.18	21.64
$\{AGG,GGA,GAG\}$	62.52	70.25	45.54
$\{AGT,GTA,TAG\}$	59.72	57.02	62.28
$\{ATC,TCA,CAT\}$	65.25	76.93	22.20
$\{ATG,TGA,GAT\}$	60.15	57.86	63.30
$\{ATT,TTA,TAT\}$	18.15	20.34	22.72
$\{CCG,CGC,GCC\}$	32.48	30.29	38.77
$\{CCT,CTC,TCC\}$	41.06	74.68	6.35
$\{CGG,GGC,GCG\}$	28.00	30.47	32.94
$\{CGT,GTC,TCG\}$	46.70	67.25	33.21
$\{CTG,TGC,GCT\}$	68.96	54.48	31.32
$\{CTT,TTC,TCT\}$	44.78	55.38	12.64
$\{GGT,GTG,TGG\}$	29.81	20.78	25.80
$\{GTT,TTG,TGT\}$	45.15	43.91	19.56
Dispersion mean $\overline{d}$	45.64	49.01	31.68

Table 8
Self-complementary code Z'_{Bac} of cardinality 20 in 34,020,997 genes of 8345 bacterial genomes Bac (11,087,876,805 codons) without normalization of the frequency f. The self-complementary codon pairs are given in descending order of mean frequencies $\overline{f}$ (defined in the same way as $\overline{\overline{g}}$ (8)) computed from Table 1 in %. The codons in bold belong to the circular code X (1) observed in genes. The mean frequency $\overline{f}$ (%) is computed for the 30 self-complementary codon pairs **SC** (2) and identifies the self-complementary codon pairs of preferential occurrence (highest frequency) leading to a self-complementary code Z'_{Bac} of cardinality 20 with a rank from 1 to 10. The code Z'_{Bac} contains only 14 codons of X and is not circular (see Section 3.1).

Genes of bacteria Bac		
Self-complementary code of cardinality 20		
Rank	Z'_{Bac}	$\overline{f}$
1	$\{\boldsymbol{GCC,GGC}\}$	7.95
2	$\{\boldsymbol{GAC,GTC}\}$	6.08
3	$\{\boldsymbol{CAG,CTG}\}$	6.03
4	$\{GCG,CGC\}$	5.48
5	$\{\boldsymbol{CTC,GAG}\}$	5.46
6	$\{\boldsymbol{GAA,TTC}\}$	5.02
7	$\{\boldsymbol{ATC,GAT}\}$	5.01
8	$\{\boldsymbol{ACC,GGT}\}$	4.10
9	$\{GTG,CAC\}$	3.93
10	$\{CCG,CGG\}$	3.78
	$\{\boldsymbol{AAT,ATT}\}$	3.55
	$\{AAG,CTT\}$	3.01
	$\{\boldsymbol{AAC,GTT}\}$	3.00
	$\{ATG,CAT\}$	2.93
	$\{TCC,GGA\}$	2.63
	$\{\boldsymbol{GTA,TAC}\}$	2.61
	$\{AGC,GCT\}$	2.58
	$\{TTG,CAA\}$	2.41
	$\{ACG,CGT\}$	2.32
	$\{ATA,TAT\}$	2.16
	$\{TGC,GCA\}$	1.96
	$\{CCA,TGG\}$	1.94
	$\{TCG,CGA\}$	1.66
	$\{ACT,AGT\}$	1.45
	$\{TTA,TAA\}$	1.36
	$\{ACA,TGT\}$	1.22
	$\{TCT,AGA\}$	1.21
	$\{AGG,CCT\}$	1.05
	$\{TCA,TGA\}$	0.79
	$\{TAG,CTA\}$	0.51

some eukaryotic genes may use the circular code X and some other eukaryotic genes may change a few self-complementary codon pairs of X, while maintaining a circular code close to X, e.g. V_{Euk}.

The problem of variability in circular code with the eukaryotic genes has been open for several years in the circular code theory, and no statistical explanation could be found for this. In Section 3.5, we will investigate this question with the current (frequency) data and we will provide a statistical explanation by analysing the codon usage dispersion in the codon permutation classes.

3.4. Trinucleotide code in bacterial, archaeal and eukaryotic genes

The trinucleotide codes Z_{Bac} (11), Z_{Arc} (15) and Z_{Euk} (18) in genes of bacteria, archaea and eukaryotes, respectively, ordered from the highest to the lowest frequency, are recalled here.

$$Z_{\text{Bac}} = \{GAC,GTC,CTC,GAG,ATC,GAT,GTA,TAC,CAG,CTG,\\ GAA,TTC,AAT,ATT,GCC,GGC,AAC,GTT,ACC,GGT\},$$

$$Z_{\text{Arc}} = \{CTC,GAG,GAC,GTC,ATC,GAT,GTA,TAC,GAA,TTC,\\ AAC,GTT,CAG,CTG,GCC,GGC,AAT,ATT,ATA,TAT\},$$

$$Z_{\text{Euk}} = \{GAC,GTC,AAT,ATT,GCC,GGC,ATC,GAT,ACT,AGT,\\ CTC,GAG,CAG,CTG,GTA,TAC,GAA,TTC,AAC,GTT\}.$$

Table 9
Random frequency f (%) in a random genetic code Rand according to a uniform distribution of the 64 codon frequencies ($f = 1/64$ for every codon), and normalized frequency g (6) (%) for each codon permutation class **P** (3) (frequency sum equal to 100%; Procedure P). The codons are ordered according to the 20 permutation classes and the numbers are rounded to the 2nd decimal place.

	f	g		f	g		f	g		f	g
AAC	1.18	56.19	ACT	1.75	38.72	ATG	0.91	24.66	CGT	1.76	48.75
ACA	0.82	39.05	CTA	1.54	34.07	TGA	1.81	49.05	GTC	1.75	48.48
CAA	0.10	4.76	TAC	1.23	27.21	GAT	0.97	26.29	TCG	0.10	2.77
Total	2.10	100.00	Total	4.52	100.00	Total	3.69	100.00	Total	3.61	100.00
AAG	1.10	24.39	AGC	0.47	8.44	ATT	2.41	48.01	CTG	0.53	9.43
AGA	3.04	67.41	GCA	2.18	39.14	TTA	1.97	39.24	TGC	2.34	41.64
GAA	0.37	8.20	CAG	2.92	52.42	TAT	0.64	12.75	GCT	2.75	48.93
Total	4.51	100.00	Total	5.57	100.00	Total	5.02	100.00	Total	5.62	100.00
AAT	0.33	5.76	AGG	2.04	48.92	CCG	0.36	8.87	CTT	1.39	28.66
ATA	2.46	42.93	GGA	1.48	35.49	CGC	1.77	43.60	TTC	0.46	9.48
TAA	2.94	51.31	GAG	0.65	15.59	GCC	1.93	47.54	TCT	3.00	61.86
Total	5.73	100.00	Total	4.17	100.00	Total	4.06	100.00	Total	4.85	100.00
ACC	0.42	6.62	AGT	1.56	23.74	CCT	1.05	25.80	GGT	2.49	32.51
CCA	2.97	46.85	GTA	2.81	42.77	CTC	0.18	4.42	GTG	2.07	27.02
CAC	2.95	46.53	TAG	2.20	33.49	TCC	2.84	69.78	TGG	3.10	40.47
Total	6.34	100.00	Total	6.57	100.00	Total	4.07	100.00	Total	7.66	100.00
ACG	0.51	17.41	ATC	1.83	46.68	CGG	0.52	12.56	GTT	2.01	35.96
CGA	0.46	15.70	TCA	0.77	19.64	GGC	2.09	50.48	TTG	3.08	55.10
GAC	1.96	66.89	CAT	1.32	33.67	GCG	1.53	36.96	TGT	0.50	8.94
Total	2.93	100.00	Total	3.92	100.00	Total	4.14	100.00	Total	5.59	100.00

The order of the 10 self-complementary codon pairs (in the top 10 ranks) differ in these 3 gene kingdoms. By considering each kingdom with the same weight, a mean gene frequency $\overline{\overline{g}}$ of a self-complementary codon pair is defined as follows

$$\overline{\overline{g}} = \frac{\overline{g}_{Bac} + \overline{g}_{Arc} + \overline{g}_{Euk}}{3} \tag{20}$$

where $\overline{g}_{Bac}$, $\overline{g}_{Arc}$ and $\overline{g}_{Euk}$ are the mean normalized frequencies $\overline{g}$ (8) in genes of bacteria, archaea and eukaryotes, respectively.

Remark 10. As the number of archaeal codons represents only ≈3% of the number of bacterial codons (see Remark 9), all codons of all genes from all organisms of all kingdoms would have constituted a biased sample to calculate the mean gene frequency $\overline{\overline{g}}$ that was not representative of the biological classification into 3 kingdoms.

An example is given for computing the mean gene frequency $\overline{\overline{g}}$ (20) in a self-complementary codon pair.

Example 6. The mean gene frequency $\overline{\overline{g}}$ for the self-complementary codon pair $\{AAC, GTT\}$ needs the mean normalized frequencies $\overline{g}_{Bac} = 0.4518$ (see Table 3 and Example 4), $\overline{g}_{Arc} = 0.5672$ (see Table 4) and $\overline{g}_{Euk} = 0.3867$ (see Table 5)

$$\overline{\overline{g}} = \frac{0.4518 + 0.5672 + 0.3867}{3} \approx 0.4686 = 46.86\%.$$

The pair $\{AAC, GTT\}$ and its frequency $\overline{\overline{g}} = 46.86\%$ are in the 9th rank in Table 6.

In Table 6, the average self-complementary trinucleotide code Z_{Genes} identified by this 1-frame statistical method, is the circular code X (1) in genes

$$Z_{\text{Genes}} = X. \tag{21}$$

3.5. Lowest codon dispersion in eukaryotic genes compared to bacterial and archaeal genes

The problem of variability in circular code is investigated by analysing the codon usage dispersion d (9) in the codon permutation classes **P** (3).

For the reader's convenience, an example is given for computing the codon usage dispersion d in a permutation class.

Example 7. The codon usage dispersion d in the permutation class $\{c_1, c_2, c_3\} = \{AAC, ACA, CAA\} \in \mathbf{P}$ of bacteria Bac needs the normalized frequencies (6) $g_1 = 0.4583$ for $c_1 = AAC$, $g_2 = 0.2361$ for $c_2 = ACA$ and $g_3 = 0.3057$ for $c_3 = CAA$ (see Table 2).

$$d = \sum_{i=1}^{3} \left| g_i - \frac{1}{3} \right|$$
$$= \left(0.4583 - \frac{1}{3}\right) + \left(\frac{1}{3} - 0.2361\right) + \left(\frac{1}{3} - 0.3057\right) \approx 0.2499 = 24.99\%.$$

The result 24.99% of $\{AAC, ACA, CAA\}$ is given in Table 7 (1st row).

Very interestingly, the dispersion means $\overline{d} = 45.64$ in genes of bacteria Bac and $\overline{d} = 49.01$ in genes of archaea Arc differ significantly from $\overline{d} = 31.68$ in genes of eukaryotes Euk. We use the Wilcoxon signed-rank test for a statistical evaluation. It is a classical non-parametric rank test for statistical hypothesis testing used to compare the means of two populations using two matched samples. We first use a one-sided p-value to compare the 20 dispersion values of the variable Arc with the highest mean $\overline{d} = 49.01$ and the variable Bac with the 2nd highest mean $\overline{d} = 45.64$: $p = 0.29$ (the null hypothesis is not rejected), i.e. there is no dispersion difference between the genes of archaea and bacteria. Then, we continue with a one-sided p-value to compare the 20 dispersion values of the variable Bac with the 2nd highest mean $\overline{d} = 45.64$ and the variable Euk with the lowest mean $\overline{d} = 31.68$: $p = 0.002$ (the null hypothesis is rejected), i.e. there is a dispersion difference between the genes of bacteria and eukaryotes. Finally, a one-sided p-value compares the 20 dispersion values of the variable Arc ($\overline{d} = 49.01$) and the variable Euk ($\overline{d} = 31.68$): $p = 0.005$ (the null hypothesis is rejected), i.e. there is a dispersion difference between the genes of archaea and eukaryotes. Note the detail that $p = 0.002$ with Bac and Euk is less than $p = 0.005$ with Arc and Euk, while $\overline{d} = 45.64$ with Bac is less than $\overline{d} = 49.01$ with Arc.

In conclusion, the codon dispersion in eukaryotic genes is significantly lower than the codon dispersion in bacterial and archaeal genes. Thus, the codons of eukaryotes have, on average, frequencies that are closer to each other. This statistical result could be the main explanation for the fact that the circular codes in eukaryotic genes are more variable and more difficult to identity with any statistical method.

Table 10
Identification of a code Y_{Rand} (Procedure P) and a self-complementary code Z_{Rand} (Procedure Q) of cardinality 20 in a random genetic code Rand according to a uniform distribution of the 64 codon frequencies ($f = 1/64$ for every codon). The codons and the self-complementary codon pairs are given in descending order of frequencies g (6) and $\overline{g}$ (8), respectively. The codons in bold belong to the circular code X (1) observed in genes. The normalized frequency g (from Table 9 in %) identifies the codon of preferential occurrence (highest frequency) in each of the 20 codon permutation classes **P** (3) leading to a code Y_{Rand} of cardinality 20 with a rank from 1 to 20. The mean normalized frequency $\overline{g}$ (%) is computed for the 30 self-complementary codon pairs **SC** (2) and identifies the self-complementary codon pairs of preferential occurrence (highest frequency) leading to a self-complementary code Z_{Rand} of cardinality 20 with a rank from 1 to 10. The codes Y_{Rand} and Z_{Rand} contain 8 and 6 codons of X, respectively.

Random genes					
Code of cardinality 20			Self-complementary code of cardinality 20		
Rank	Y_{Rand}	g	Rank	Z_{Rand}	$\overline{g}$
1	TCC	69.78	1	$\{TCT, AGA\}$	64.63
2	AGA	67.41	2	$\{\boldsymbol{GAC, GTC}\}$	57.69
3	$\boldsymbol{GAC}$	66.89	3	$\{TCC, GGA\}$	52.64
4	TCT	61.86	4	$\{\boldsymbol{GCC, GGC}\}$	49.01
5	$\boldsymbol{AAC}$	56.19	5	$\{\boldsymbol{AAC, GTT}\}$	46.07
6	TTG	55.10	6	$\{TTA, TAA\}$	45.28
7	$\boldsymbol{CAG}$	52.42	7	$\{CCA, TGG\}$	43.66
8	TAA	51.31	8	$\{TGC, GCA\}$	40.39
9	$\boldsymbol{GGC}$	50.48	9	$\{GCG, CGC\}$	40.28
10	TGA	49.05	10	$\{AGG, CCT\}$	37.36
11	GCT	48.93		$\{GTG, CAC\}$	36.78
12	AGG	48.92		$\{\boldsymbol{ATC, GAT}\}$	36.49
13	CGT	48.75		$\{\boldsymbol{GTA, TAC}\}$	34.99
14	$\boldsymbol{ATT}$	48.01		$\{TCA, TGA\}$	34.35
15	$\boldsymbol{GCC}$	47.54		$\{TAG, CTA\}$	33.78
16	CCA	46.85		$\{ACG, CGT\}$	33.08
17	$\boldsymbol{ATC}$	46.68		$\{ACT, AGT\}$	31.23
18	$\boldsymbol{GTA}$	42.77		$\{\boldsymbol{CAG, CTG}\}$	30.93
19	TGG	40.47		$\{TTG, CAA\}$	29.93
20	ACT	38.72		$\{ATG, CAT\}$	29.17
				$\{AGC, GCT\}$	28.69
				$\{ATA, TAT\}$	27.84
				$\{\boldsymbol{AAT, ATT}\}$	26.88
				$\{AAG, CTT\}$	26.53
				$\{ACA, TGT\}$	24.00
				$\{\boldsymbol{ACC, GGT}\}$	19.57
				$\{CCG, CGG\}$	10.71
				$\{\boldsymbol{CTC, GAG}\}$	10.01
				$\{TCG, CGA\}$	9.23
				$\{\boldsymbol{GAA, TTC}\}$	8.84

4. Conclusion

In this work, we have developed a new statistical approach to identify trinucleotide codes and self-complementary trinucleotide codes in genes that are potentially circular, based solely on the reading frame of genes. Surprisingly, this 1-frame method reveals the circular code X (1) in bacterial genes, in average (bacterial, archaeal, eukaryotic) genes. Furthermore, it allows deducing the circular code X in archaeal genes and identifying 16 codons of X in eukaryotic genes.

This 1-frame method based on codon usage only, avoids an analysis of the genetic information in the shifted frames (1 and 2). This approach differs from the six 3-frame methods used in the past (Arquès and Michel, 1996, 1997; Frey and Michel, 2003, 2006; Michel, 2015, 2017). Indeed, the 1-frame method is based on two important concepts: codon dispersion and permutation class. Furthermore, a 3-frame method needs the gene sequence in the reading frame but also its 2 sequences in the 2 shifted frames (1 and 2). Indeed, the trinucleotide frequencies in the 2 shifted frames cannot be deduced from the trinucleotide frequencies in reading frame (codon usage).

The statistical relation between the 1-frame method and the 3-frame methods is unknown to date. It is also unexpected that 2 different methods identify the same circular code, i.e. X. Indeed, as each permutation class has 3 codons and as there are 20 permutation classes, the number of codon possibilities in the 20 permutation classes is 3^{20}. Thus, the probability to observe the circular code X in the 20 permutation classes is equal to $1/3^{20} \approx 10^{-10}$. Furthermore, the 1-frame method without the concept of permutation class cannot identify a maximal circular code (see Section 3.1 and Table 8 in Appendix A). Whether this direct approach can replace the 3-frames methods remains an open question, emphasizing that improvements remain possible in the future.

The trinucleotide code variability in eukaryotic genes, a long-standing problem since 1996, is investigated with a new parameter definition based on the dispersion in the codon permutation classes. The genes of bacteria and archaea have a codon usage dispersion of the same order of magnitude, but significantly higher than the one in the genes of eukaryotes where the codons have, on average, frequencies that are closer to each other. This new statistical result could be the fundamental explanation for the variability of codes, whether circular or not, in eukaryotic genes.

The 1-frame statistical method developed solely on codon usage, the most standard procedure in genetics, now allows biologists to access the circular code theory. It can be applied at the genome level, in a group of genes, e.g. orthologs, and even at the gene level. Variability in circular code at the gene level should increase drastically, but the property of retrieving the reading frame could still be retained, for example with non-maximal circular codes.

CRediT authorship contribution statement

Christian J. Michel: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The author reports no conflict of interest.

Acknowledgments

I thank Dr Svetlana A. MICHEL for her constant support.

Appendix A. Bacterial case without normalization

See Table 8.

Appendix B. Random genetic code

See Tables 9 and 10.

Appendix C. Uniform amino acid code

See Tables 11 and 12.

Appendix D. Archaeal codon usage per codon permutation class

See Table 13.

Appendix E. Eukaryotic codon usage per codon permutation class

See Table 14.

Table 11
Uniform amino acid code where every amino acid has the same probability: (i) $f = 1/20$ with $f = 0$ for the 3 stop codons $\{TAA, TAG, TGA\}$; and (ii) $f' = 1/21$ including the 3 stop codons.

	Codon	Number	$f = 1/20$	$f\%$	$f' = 1/21$	$f'\%$
Met	$\{ATG\}$	1	1/20	5.00	1/21	4.76
Trp	$\{TGG\}$	1	1/20	5.00	1/21	4.76
Asn	$\{AAC, AAT\}$	2	1/40	2.50	1/42	2.38
Asp	$\{GAC, GAT\}$	2	1/40	2.50	1/42	2.38
Cys	$\{TGC, TGT\}$	2	1/40	2.50	1/42	2.38
Gln	$\{CAA, CAG\}$	2	1/40	2.50	1/42	2.38
Glu	$\{GAA, GAG\}$	2	1/40	2.50	1/42	2.38
His	$\{CAC, CAT\}$	2	1/40	2.50	1/42	2.38
Lys	$\{AAA, AAG\}$	2	1/40	2.50	1/42	2.38
Phe	$\{TTC, TTT\}$	2	1/40	2.50	1/42	2.38
Tyr	$\{TAC, TAT\}$	2	1/40	2.50	1/42	2.38
Ile	$\{ATA, ATC, ATT\}$	3	1/60	1.67	1/63	1.59
$Stop$	$\{TAA, TAG, TGA\}$	3	0	0	1/63	1.59
Ala	$\{GCA, GCC, GCG, GCT\}$	4	1/80	1.25	1/84	1.19
Gly	$\{GGA, GGC, GGG, GGT\}$	4	1/80	1.25	1/84	1.19
Pro	$\{CCA, CCC, CCG, CCT\}$	4	1/80	1.25	1/84	1.19
Thr	$\{ACA, ACC, ACG, ACT\}$	4	1/80	1.25	1/84	1.19
Val	$\{GTA, GTC, GTG, GTT\}$	4	1/80	1.25	1/84	1.19
Arg	$\{AGA, AGG, CGA, CGC, CGG, CGT\}$	6	1/120	0.83	1/126	0.79
Leu	$\{CTA, CTC, CTG, CTT, TTA, TTG\}$	6	1/120	0.83	1/126	0.79
Ser	$\{AGC, AGT, TCA, TCC, TCG, TCT\}$	6	1/120	0.83	1/126	0.79

Table 12
Codon frequencies f and f' of a uniform amino acid code where every amino acid has the same probability: (i) $f = 1/20$ with $f = 0$ for the 3 stop codons $\{TAA, TAG, TGA\}$; and (ii) $f = 1/21$ including the 3 stop codons (from Table 11) for each codon permutation class **P** (3). The codons in bold belong to the circular code X (1) observed in genes.

	f	f'		f	f'		f	f'		f	f'
AAC	2.50	2.38	ACT	1.25	1.19	ATG	5.00	4.76	CGT	0.83	0.79
ACA	1.25	1.19	CTA	0.83	0.79	TGA	0.00	1.59	***GTC***	1.25	1.19
CAA	2.50	2.38	***TAC***	2.50	2.38	***GAT***	2.50	2.38	TCG	0.83	0.79
AAG	2.50	2.38	AGC	0.83	0.79	***ATT***	1.67	1.59	***CTG***	0.83	0.79
AGA	0.83	0.79	GCA	1.25	1.19	TTA	0.83	0.79	TGC	2.50	2.38
GAA	2.50	2.38	***CAG***	2.50	2.38	TAT	2.50	2.38	GCT	1.25	1.19
AAT	2.50	2.38	AGG	0.83	0.79	CCG	1.25	1.19	CTT	0.83	0.79
ATA	1.67	1.59	GGA	1.25	1.19	CGC	0.83	0.79	***TTC***	2.50	2.38
TAA	0.00	1.59	***GAG***	2.50	2.38	***GCC***	1.25	1.19	TCT	0.83	0.79
ACC	1.25	1.19	AGT	0.83	0.79	CCT	1.25	1.19	***GGT***	1.25	1.19
CCA	1.25	1.19	***GTA***	1.25	1.19	***CTC***	0.83	0.79	GTG	1.25	1.19
CAC	2.50	2.38	TAG	0.00	1.59	TCC	0.83	0.79	TGG	5.00	4.76
ACG	1.25	1.19	***ATC***	1.67	1.59	CGG	0.83	0.79	***GTT***	1.25	1.19
CGA	0.83	0.79	TCA	0.83	0.79	***GGC***	1.25	1.19	TTG	0.83	0.79
GAC	2.50	2.38	CAT	2.50	2.38	GCG	1.25	1.19	TGT	2.50	2.38

Table 13
Frequency f (codon usage from Table 1 in %) and normalized frequency g (6) (%) of 1,280,890 genes of 432 archaeal genomes Arc (367,937,932 codons) for each codon permutation class **P** (3) (frequency sum equal to 100%; Procedure P). The codons are ordered according to the 20 permutation classes and the numbers are rounded to the 2nd decimal place.

	f	g		f	g		f	g		f	g
AAC	2.01	58.15	ACT	0.73	21.89	ATG	2.00	47.52	CGT	0.50	7.77
ACA	0.85	24.71	CTA	0.47	14.11	TGA	0.19	4.40	GTC	4.30	66.96
CAA	0.59	17.15	TAC	2.13	64.01	GAT	2.02	48.08	TCG	1.62	25.27
Total	3.46	100.00	Total	3.33	100.00	Total	4.21	100.00	Total	6.42	100.00
AAG	1.88	33.55	AGC	1.27	27.98	ATT	1.31	43.50	CTG	2.25	60.57
AGA	0.69	12.23	GCA	1.48	32.60	TTA	0.75	24.79	TGC	0.41	11.07
GAA	3.04	54.22	CAG	1.78	39.42	TAT	0.96	31.71	GCT	1.05	28.35
Total	5.61	100.00	Total	4.53	100.00	Total	3.02	100.00	Total	3.71	100.00
AAT	1.05	42.23	AGG	0.81	10.99	CCG	1.91	26.49	CTT	1.09	25.74
ATA	1.33	53.40	GGA	1.52	20.55	CGC	1.80	25.03	TTC	2.58	61.02
TAA	0.11	4.38	GAG	5.07	68.46	GCC	3.49	48.48	TCT	0.56	13.24
Total	2.49	100.00	Total	7.40	100.00	Total	7.21	100.00	Total	4.22	100.00
ACC	2.16	51.02	AGT	0.63	40.29	CCT	0.52	9.54	GGT	1.34	31.06
CCA	0.67	15.84	GTA	0.86	54.88	CTC	3.86	70.67	GTG	1.88	43.73
CAC	1.40	33.13	TAG	0.08	4.82	TCC	1.08	19.79	TGG	1.08	25.21
Total	4.22	100.00	Total	1.56	100.00	Total	5.47	100.00	Total	4.30	100.00
ACG	2.20	26.93	ATC	2.90	71.80	CGG	1.51	18.10	GTT	1.36	55.29
CGA	0.72	8.81	TCA	0.68	16.92	GGC	3.40	40.80	TTG	0.68	27.52
GAC	5.25	64.26	CAT	0.46	11.28	GCG	3.42	41.11	TGT	0.42	17.19
Total	8.16	100.00	Total	4.04	100.00	Total	8.33	100.00	Total	2.46	100.00

Table 14
Frequency f (codon usage from Table 1 in %) and normalized frequency g (6) (%) of 20,206,058 genes of 1150 eukaryotic genomes Euk (10,374,305,634 codons) for each codon permutation class **P** (3) (frequency sum equal to 100%; Procedure P). The codons are ordered according to the 20 permutation classes and the numbers are rounded to the 2nd decimal place.

	f	g		f	g		f	g		f	g
AAC	2.12	39.98	ACT	1.40	37.52	ATG	2.26	45.40	CGT	0.68	21.56
ACA	1.57	29.66	CTA	0.77	20.81	TGA	0.08	1.69	GTC	1.58	49.94
CAA	1.61	30.36	TAC	1.55	41.68	GAT	2.64	52.92	TCG	0.90	28.50
Total	5.30	100.00	Total	3.72	100.00	Total	4.99	100.00	Total	3.17	100.00
AAG	3.12	42.43	AGC	1.72	27.61	ATT	1.75	44.70	CTG	2.87	48.29
AGA	1.27	17.24	GCA	1.76	28.23	TTA	0.92	23.49	TGC	1.05	17.68
GAA	2.96	40.33	CAG	2.75	44.15	TAT	1.24	31.81	GCT	2.02	34.03
Total	7.35	100.00	Total	6.22	100.00	Total	3.91	100.00	Total	5.94	100.00
AAT	1.95	64.35	AGG	1.13	17.19	CCG	0.97	23.01	CTT	1.46	28.79
ATA	1.02	33.57	GGA	1.76	26.71	CGC	1.03	24.27	TTC	2.01	39.65
TAA	0.06	2.07	GAG	3.69	56.10	GCC	2.23	52.72	TCT	1.60	31.56
Total	3.03	100.00	Total	6.58	100.00	Total	4.24	100.00	Total	5.07	100.00
ACC	1.63	35.38	AGT	1.27	58.93	CCT	1.58	31.62	GGT	1.43	28.99
CCA	1.61	34.92	GTA	0.84	38.88	CTC	1.83	36.51	GTG	2.28	46.23
CAC	1.37	29.69	TAG	0.05	2.19	TCC	1.60	31.87	TGG	1.22	24.77
Total	4.61	100.00	Total	2.16	100.00	Total	5.01	100.00	Total	4.92	100.00
ACG	0.94	22.38	ATC	2.04	44.44	CGG	0.82	21.18	GTT	1.51	37.37
CGA	0.71	16.75	TCA	1.38	30.17	GGC	1.92	49.80	TTG	1.58	39.08
GAC	2.57	60.87	CAT	1.17	25.40	GCG	1.12	29.01	TGT	0.95	23.55
Total	4.22	100.00	Total	4.59	100.00	Total	3.85	100.00	Total	4.04	100.00

References

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

Arquès, D.G., Michel, C.J., 1997. A code in the protein coding genes. Biosystems 44, 107–134.

Fimmel, E., Giannerini, S., Gonzalez, D., Strüngmann, L., 2014. Circular codes, symmetries and transformations. J. Math. Biol. 70 (7), 1623–1644.

Fimmel, E., Michel, C.J., Strüngmann, L., 2016. n-Nucleotide circular codes in graph theory. Philosophical transactions of the royal society a: Mathematical. Phys. Eng. Sci. A 374 (20150058), 1–19.

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

Frey, G., Michel, C.J., 2003. Circular codes in archaeal genomes. J. Theoret. 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.

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

Lacan, J., Michel, C.J., 2001. Analysis of a circular code model. J. Theoret. 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., 2015. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. J. Theoret. Biol. 380, 156–177.

Michel, C.J., 2017. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. Life 7 (20), 1–16.

Michel, C.J., 2020. The maximality of circular codes in genes statistically verified. Biosystems 197 (104201), 1–7.

Michel, C.J., 2021. Genes on the circular code alphabet. Biosystems 206 (104431), 1–12.

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

Michel, C.J., Sereni, J.-S., 2023. Reading frame retrieval of genes: a new parameter of codon usage based on the circular code theory. Bull. Math. Biol. 85 (24), 1–21.

Subramanian, K., Payne, B., Feyertag, F., Alvarez-Ponce, D., 2022. The codon statistics database: a database of codon usage bias. Mol. Biol. Evol. 39 (8), 1–3.

Woese, C.R., 1965. On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA 54, 1546–1552.