life

MDPI

Article

The Maximal C^3 Self-Complementary Trinucleotide Circular Code X in Genes of Bacteria, Archaea, Eukaryotes, Plasmids and Viruses

Christian J. Michel

Theoretical Bioinformatics, ICube, University of Strasbourg, CNRS, 300 Boulevard Sébastien Brant, 67400 Illkirch, France; c.michel@unistra.fr; Tel.: +33-368854462

Academic Editor: David Deamer
Received: 6 February 2017; Accepted: 31 March 2017; Published: 18 April 2017

Abstract: In 1996, a set X of 20 trinucleotides was identified in genes of both prokaryotes and eukaryotes which has on average the highest occurrence in reading frame compared to its two shifted frames. Furthermore, this set X has an interesting mathematical property as X is a maximal C^3 self-complementary trinucleotide circular code. In 2015, by quantifying the inspection approach used in 1996, the circular code X was confirmed in the genes of bacteria and eukaryotes and was also identified in the genes of plasmids and viruses. The method was based on the preferential occurrence of trinucleotides among the three frames at the gene population level. We extend here this definition at the gene level. This new statistical approach considers all the genes, i.e., of large and small lengths, with the same weight for searching the circular code X. As a consequence, the concept of circular code, in particular the reading frame retrieval, is directly associated to each gene. At the gene level, the circular code X is strengthened in the genes of bacteria, eukaryotes, plasmids, and viruses, and is now also identified in the genes of archaea. The genes of mitochondria and chloroplasts contain a subset of the circular code X. Finally, by studying viral genes, the circular code X was found in DNA genomes, RNA genomes, double-stranded genomes, and single-stranded genomes.

Keywords: circular code in genes; DNA genes; RNA genes; double-stranded genes; single-stranded genes

1. Introduction

Circular code is a mathematical structure of genes and genomes. This concept initially found for genes is extended for genomes (non-coding regions of eukaryotes) according to recent results. A circular code X is a set of words such that any motif from X, called X motif, allows it to retrieve, maintain, and synchronize the original (construction) frame.

The circular code X identified in the genes of bacteria, eukaryotes, plasmids, and viruses [1,2] contains the 20 following trinucleotides

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

which allows it to both retrieve the reading frame with a window of 13 nucleotides (Figure 3 in [3]) and to code the 12 following amino acids

$$\{\mathrm{Ala}, \mathrm{Asn}, \mathrm{Asp}, \mathrm{Gln}, \mathrm{Glu}, \mathrm{Gly}, \mathrm{Ile}, \mathrm{Leu}, \mathrm{Phe}, \mathrm{Thr}, \mathrm{Tyr}, \mathrm{Val}\}. \quad (2)$$

The current genetic code is not circular. Thus, it cannot retrieve the reading frame. The loss during evolution of this circular code property on the 4-letter alphabet $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ required a complex translation mechanism using 20 amino acids and proteins in current genomes.

X motifs from Equation (1) are identified in (i) genes "universally" [1,4]; (ii) tRNAs of prokaryotes and eukaryotes [3,5]; (iii) rRNAs of prokaryotes (16S) and eukaryotes (18S), in particular in the ribosome decoding center where the universally conserved nucleotides G530, A1492, and A1493 are included in the X motifs [3,6,7]; and (iv) genomes (non-coding regions of eukaryotes) [4,8].

The X motifs of maximal cardinality 20 (composition) in genes with the properties of the circular code, C^3 and complementary allow the two reading frames and the four shifted frames to be retrieved by pairing between DNAs-DNAs, DNAs-mRNAs, mRNAs-rRNAs, mRNAs-tRNAs, and rRNAs-tRNAs, as shown with a 3D visualization of the X motifs in the ribosome [3,6,7].

The X motifs in genomes have a different structure compared to the X motifs in genes [8]. Indeed, their cardinality is not maximal (less than 10 for an order of magnitude), their size is longer, and their structure contains repeated trinucleotides. Furthermore, the X motifs of minimal cardinality 1 generated with the 20 repeated trinucleotides t^n where $t \in X$ (Equation (1)) are very common in the genomes of eukaryotes (e.g., [8–10]). Their length n can be very large (e.g., $n > 6000$, see Figure 1). The repeated trinucleotides are very unstable with mutation rates up to 100,000 times higher than the genomic average mutation rate. Mutation in repeats increases its evolutionary stability.

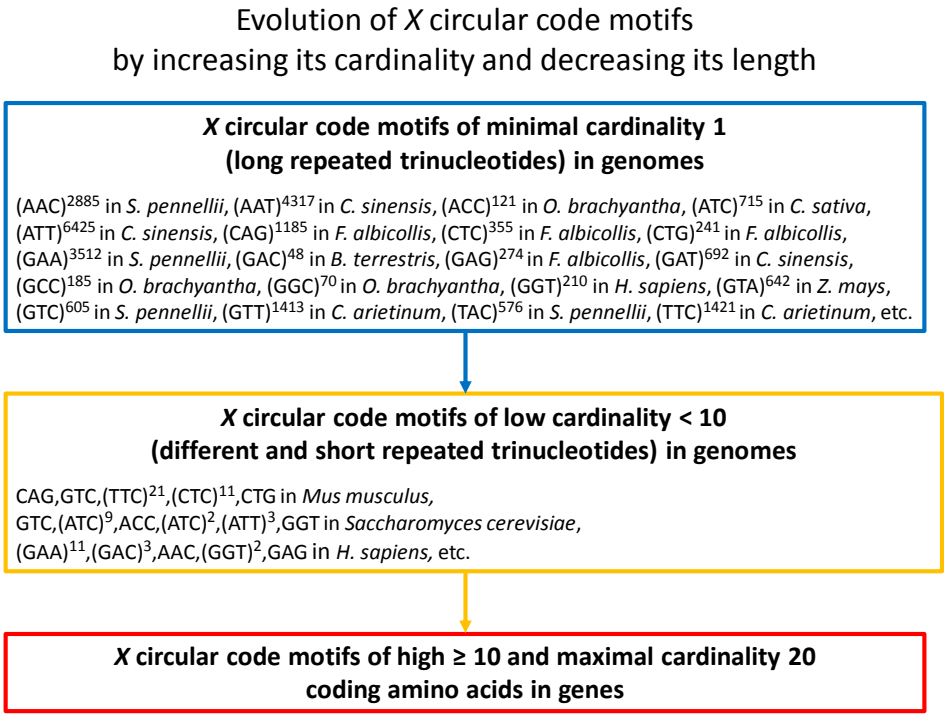

Figure 1. Model of evolution of the X circular code motifs (Equation (1)) by increasing its cardinality (composition) and decreasing its length. Evolution begins with X motifs of minimal cardinality 1 (long repeated trinucleotides) in genomes (the examples given are extracted from Table 2 in [8]). Then, the mutations in repeated trinucleotides lead to X motifs of low cardinality <10 (different and short repeated trinucleotides) in genomes (the examples given are extracted from Table 4 in [8]) up to X motifs of high ≥ 10 and maximal cardinality 20 coding the 12 amino acids (Equation (2)).

A model of evolution of the X motifs in genes and genomes can be proposed according to the previous works and the recent results [8]. It proposes that the X motifs of maximal cardinality 20 in genes have evolved from the X motifs of minimal cardinality 1 (repeated trinucleotides) in genomes (Figure 1). An X motif of minimal cardinality 1 which is unstable, mutates into an X motif of low cardinality <10 containing thus different repeated trinucleotides of short lengths. This evolutionary process continues by increasing the cardinality and decreasing the length of the X motifs up to generate the X motifs of high ≥ 10 and maximal cardinality 20 coding the 12 amino acids (Equation (2)) in genes.

The X motifs of high cardinality have acquired the protein coding function in addition to the reading frame retrieval. This model suggests that the property of reading frame retrieval has preceded the protein coding function.

Since 1996, all the statistical analyses studying the preferential occurrence of trinucleotides among the three frames were done at the gene population level (kingdoms, taxonomic groups, genomes). We extend here the method from [1] at the gene level. This new approach is important as all the genes, i.e., of large and small lengths, are now considered with the same weight in the statistical definition for searching the circular code X. As a consequence, the concept of circular code, in particular the reading frame retrieval, is directly associated to each gene. Thus, at the gene level, the circular code X is searched here in the genes of bacteria, archaea, eukaryotes, plasmids, viruses, and eukaryotic organelles, i.e., mitochondria and chloroplasts. Finally, genes of double-stranded DNA and RNA viruses, and single-stranded DNA and RNA viruses are also analysed with this approach in order to assign a genetic information unit (DNA or RNA, double-stranded or single-stranded) to the circular code X.

2. Method

2.1. Definitions

We recall a few definitions without detailed explanations (i.e., without figures and examples) for understanding the main properties of the trinucleotide circular code X identified in genes [1,2].

Notation 1. *Let us denote the nucleotide 4-letter alphabet $B = \{A, C, G, T\}$ where A stands for Adenine, C stands for Cytosine, G stands for Guanine, and T stands for Thymine. The trinucleotide set over B is denoted by $B^3 = \{AAA, \ldots, TTT\}$. The set of non-empty words (words, respectively) over B is denoted by B^+ (B^*, respectively).*

Notation 2. *Genes have three frames f. By convention here, the reading frame $f = 0$ is set up by a start trinucleotide $\{ATG, CTG, GTG, TTG\}$, 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 (to the right), respectively.*

Two biological maps are involved in gene coding.

Definition 1. *According to the complementary property of the DNA double helix, the nucleotide complementarity map $\mathcal{C} : B \to B$ is defined by $\mathcal{C}(A) = T$, $\mathcal{C}(C) = G$, $\mathcal{C}(G) = C$, and $\mathcal{C}(T) = A$. According to the complementary and antiparallel properties of the DNA double helix, the trinucleotide complementarity map $\mathcal{C} : B^3 \to B^3$ is defined by $\mathcal{C}(l_0 l_1 l_2) = \mathcal{C}(l_2)\mathcal{C}(l_1)\mathcal{C}(l_0)$ for all $l_0, l_1, l_2 \in B$. By extension to a trinucleotide set S, the set complementarity map $\mathcal{C} : \mathbb{P}(B^3) \to \mathbb{P}(B^3)$, $\mathbb{P}$ being the set of all subsets of B^3, is defined by $\mathcal{C}(S) = \{v : u, v \in B^3, u \in S, v = \mathcal{C}(u)\}$, e.g., $\mathcal{C}(\{CGA, GAT\}) = \{ATC, TCG\}$.*

Definition 2. *The trinucleotide circular permutation map $\mathcal{P} : B^3 \to B^3$ is defined by $\mathcal{P}(l_0 l_1 l_2) = l_1 l_2 l_0$ for all $l_0, l_1, l_2 \in B$. $\mathcal{P}^2$ denotes the 2nd iterate of $\mathcal{P}$. By extension to a trinucleotide set S, the set circular permutation map $\mathcal{P} : \mathbb{P}(B^3) \to \mathbb{P}(B^3)$ is defined by $\mathcal{P}(S) = \{v : u, v \in B^3, u \in S, v = \mathcal{P}(u)\}$, e.g., $\mathcal{P}(\{CGA, GAT\}) = \{ATG, GAC\}$ and $\mathcal{P}(\{CGA, GAT\}) = \{ACG, TGA\}$.*

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

Definition 4. *Any non-empty subset of the code B^3 is a code and called trinucleotide code C.*

Definition 5. *A trinucleotide code $C \subseteq B^3$ is self-complementary if, for each $t \in C$, $\mathcal{C}(t) \in C$, i.e., $C = \mathcal{C}(C)$.*

Definition 6. *A trinucleotide code $X \subseteq B^3$ is circular if, for each $x_1, \ldots, x_n, y_1, \ldots, y_m \in X$, $n, m \geq 1$, $r \in B^*$, $s \in B^+$, the conditions $sx_2 \ldots x_n r = y_1 \ldots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word), and $x_i = y_i$ for $i = 1, \ldots, n$.*

The proofs to decide whether a code is circular or not are based on the flower automaton [2], the necklace 5*LDCN* (Letter Diletter Continued Necklace) [11], the necklace *n*LDCCN (Letter Diletter Continued Closed Necklace) with $n \in \{2, 3, 4, 5\}$ [12], and the graph theory [13].

Definition 7. *A trinucleotide circular code $X \subseteq B^3$ is C^3 self-complementary if X, $X_1 = \mathcal{P}(X)$, and $X_2 = \mathcal{P}^2(X)$ are trinucleotide circular codes such that $X = \mathcal{C}(X)$ (self-complementary), $\mathcal{C}(X_1) = X_2$, and $\mathcal{C}(X_2) = X_1$ (X_1 and X_2 are complementary).*

The trinucleotide set $X = X_0$ (Equation (1)) coding the reading frame ($f = 0$) in genes is a maximal (20 trinucleotides) C^3 self-complementary trinucleotide circular code [2] where the circular code $X_1 = \mathcal{P}(X)$ coding the frame $f = 1$ contains the 20 following trinucleotides

$$X_1 = \{\text{AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG,GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG}\} \quad (3)$$

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

$$X_2 = \{\text{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT,CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT}\}. \quad (4)$$

The trinucleotide circular codes X_1 and X_2 are related by the permutation map, i.e., $X_2 = \mathcal{P}(X_1)$ and $X_1 = \mathcal{P}^2(X_2)$, and by the complementary map, i.e., $X_1 = \mathcal{C}(X_2)$ and $X_2 = \mathcal{C}(X_1)$ [14].

Several classes of methods were developed for identifying the circular code X in genes over the last 20 years: frequency methods [2,15,16], correlation function [17], covering capability function [18], and occurrence probability of a complementary/permutation (CP) trinucleotide set at the gene population level [1].

The class of the 216 C^3 self-complementary trinucleotide circular codes (Definition 7; [2]; list given in Tables 4a, 5a, and 6a in [19]; [20]) is included in a larger class of codes C by relaxing the circularity property which was defined in [1]:

Definition 8. *A trinucleotide code $C \subseteq B^3$ is C^3 self-complementary if $C = \mathcal{C}(C)$ (self-complementary), $\mathcal{C}(C_1) = C_2$, and $\mathcal{C}(C_2) = C_1$ (C_1 and C_2 are complementary) where $C_1 = \mathcal{P}(C)$ and $C_2 = \mathcal{P}^2(C)$.*

The statistical approach developed analyses the C^3 self-complementary codes (Definition 8) for searching the particular circular code X.

2.2. Gene Kingdoms

Gene kingdoms $\mathbb{K}$ of bacteria $\mathbb{B}$, archaea $\mathbb{A}$, plasmids $\mathbb{P}$, eukaryotes $\mathbb{E}$, chromosomes of eukaryotes $\mathbb{E}_{\text{chr}}$, mitochondria $\mathbb{M}$, chloroplasts $\mathbb{C}$, viruses $\mathbb{V}$, and its five taxonomic double-stranded DNA viruses $\mathbb{V}_{\text{dsDNA}}$, double-stranded RNA viruses $\mathbb{V}_{\text{dsRNA}}$, single-stranded DNA viruses $\mathbb{V}_{\text{ssDNA}}$, single-stranded RNA viruses $\mathbb{V}_{\text{ssRNA}}$, and retro-transcribing viruses $\mathbb{V}_{\text{rt}}$ are obtained from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, May 2016) (Table 1). Computer tests exclude genes when (i) their nucleotides do not belong to the alphabet B; (ii) they do not begin with a start trinucleotide $\{\text{ATG, CTG, GTG, TTG}\}$; (iii) they do not end with a stop trinucleotide $\{\text{TAA, TAG, TGA}\}$; and (iv) their lengths are not modulo 3. In order to have an order of magnitude of data acquisition (details in Table 1), the kingdom of bacteria $\mathbb{B}$ contains 15,735,053 genes and 5,222,267,667 trinucleotides (7,851,762 genes and 2,481,566,882 trinucleotides in [1]), i.e., a trinucleotide increase of about 110%, and the kingdom of eukaryotes $\mathbb{E}$ contains 4,356,391 genes and 2,406,844,838 trinucleotides (1,662,579 genes and 824,825,761 trinucleotides in [1]), i.e., a trinucleotide increase of about 192%. The gene kingdoms

$\mathbb{M}$, $\mathbb{C}$, $\mathbb{V}_{\text{dsRNA}}$, $\mathbb{V}_{\text{ssDNA}}$, and $\mathbb{V}_{\text{rt}}$ have gene and trinucleotide data that are significantly lower (less than 1 million trinucleotides) than the other gene kingdoms (Table 1).

Table 1. Kingdoms $\mathbb{K}$ of genes extracted from the GenBank database (http://www.ncbi.nlm.nih.gov/genome/browse/, May 2016) with their symbol and their numbers of genomes, genes, and trinucleotides.

Kingdom	$\mathbb{K}$ (Symbol)	Nb of Genomes	Nb of Genes	Nb of Trinucleotides
Bacteria	$\mathbb{B}$	7039	15,735,053	5,222,267,667
Archaea	$\mathbb{A}$	182	282,802	81,460,549
Plasmids	$\mathbb{P}$	2319	575,760	159,169,387
Eukaryotes	$\mathbb{E}$	190	4,356,391	2,406,844,838
Chromosomes of eukaryotes	$\mathbb{E}_{\text{chr}}$	2979	4,356,391	2,406,844,838
Mitochondria	$\mathbb{M}$	228	3347	862,327
Chloroplasts	$\mathbb{C}$	39	3192	925,303
Viruses	$\mathbb{V}$	5217	299,401	66,677,580
Double-stranded DNA viruses	$\mathbb{V}_{\text{dsDNA}} \subset \mathbb{V}$	2480	259,696	59,239,700
Double-stranded RNA viruses	$\mathbb{V}_{\text{dsRNA}} \subset \mathbb{V}$	211	1061	783,020
Single-stranded DNA viruses	$\mathbb{V}_{\text{ssDNA}} \subset \mathbb{V}$	715	3291	802,405
Single-stranded RNA viruses	$\mathbb{V}_{\text{ssRNA}} \subset \mathbb{V}$	1257	5093	4,406,365
Retro-transcribing viruses	$\mathbb{V}_{\text{rt}} \subset \mathbb{V}$	137	560	289,447

2.3. Preferential Frame of a Trinucleotide in a Gene

The method developed in [1] for identifying the circular code X in genes determined the preferential frame of trinucleotides at the gene population level (kingdoms, taxonomic groups, genomes), i.e., after summing the trinucleotide frequencies of all genes in a kingdom. We extend this method at the gene level, i.e., the preferential frame of trinucleotides among the three frames is determined for each gene. There is no sum of trinucleotide frequencies of all genes in a kingdom. Thus, all the genes, i.e., of large and small lengths, have the same weight in respect to the preferential frame.

Consider a gene kingdom $\mathbb{K}$ listed in Table 1. Let $Pr_f(t,g)$ be the occurrence frequency of a trinucleotide $t \in B^3$ in a frame $f \in \{0,1,2\}$ of a gene g belonging to a kingdom $\mathbb{K}$. Thus, there are $3 \times 64 = 192$ trinucleotide occurrence frequencies $Pr_f(t,g)$ in the three frames f of a gene g. Then, the preferential frame $F(t,g) \in \{0,1,2\}$ of a trinucleotide t in a gene g is the frame of maximal occurrence frequency $Pr_f(t,g)$ among the three frames f of g

$$F(t,g) = \arg \max_{f \in \{0,1,2\}} Pr_f(t,g). \tag{5}$$

The three frequencies of a given trinucleotide are computed in the three frames 0, 1, and 2 of a gene. Then, the preferential frame of the trinucleotide in this gene is the frame associated to its highest trinucleotide frequency.

Remark 1. *In* [1], *the three occurrence frequencies $Pr_f(t,\mathbb{K})$ of a trinucleotide t in the three frames f computed in a gene kingdom $\mathbb{K}$, always have different values, thus a unique preferential frame can be assigned to the trinucleotide. At the gene level, particularly for genes g of small lengths, a trinucleotide t may have an identical occurrence frequency $Pr_f(t,g)$ in two or three frames f. In this case, two or three preferential frames $F(t,g)$ are assigned to the trinucleotide t. If a trinucleotide t is absent in a gene g, mainly for genes g of very small lengths, then no preferential frame is attributed to t.*

The indicator function $\delta_f(F(t,g)) \in \{0,1\}$ is 1 if the preferential frame $F(t,g)$ of a trinucleotide t is equal to the frame f of a gene g, and 0 otherwise

$$\delta_f(F(t,g)) = \begin{cases} 1 \text{ if } F(t,g) = f \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

where $F(t,g)$ is defined in Equation (5).

2.4. Number of Preferential Frames of a Trinucleotide in a Gene Kingdom

The number $Nb_f(t,\mathbb{K}) \in \mathbb{N}$ of preferential frames of a trinucleotide $t \in B^3$ for each frame $f \in \{0,1,2\}$ in a gene kingdom $\mathbb{K}$ is simply obtained by summing for all genes in $\mathbb{K}$

$$Nb_f(t,\mathbb{K}) = \sum_{g \in \mathbb{K}} \delta_f(F(t,g)) \tag{7}$$

where $\delta_f(F(t,g))$ is defined in Equation (6).

2.5. Occurrence Probability of a Complementary/Permutation Trinucleotide Set in a Gene Kingdom

In order to study the C^3 self-complementary codes C (Definition 8) including the class of circular codes, and in particular the circular code X, Equation (7) for a trinucleotide t is expanded to a set T of six trinucleotides involving the complementarity map $\mathcal{C}$ and the permutation map $\mathcal{P}$ simultaneously, precisely $T = \{T^0, T^1, T^2\}$ with $T^0 = \{t, \mathcal{C}(t)\}$ in frame 0, $T^1 = \mathcal{P}(T^0) = \{\mathcal{P}(t), \mathcal{P}(\mathcal{C}(t))\}$ in frame 1, $T^2 = \mathcal{P}^2(T^0) = \{\mathcal{P}^2(t), \mathcal{P}^2(\mathcal{C}(t))\}$ in frame 2, and $t \in B^3 \backslash \{\mathrm{AAA}, \mathrm{CCC}, \mathrm{GGG}, \mathrm{TTT}\}$. T is called a complementary and permutation (CP) trinucleotide set and is completely defined by the trinucleotide t.

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

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 $T_i = \{T_i^0, T_i^1, T_i^2\}$ with $T_i^0 = \{t, \mathcal{C}(t)\}_i$ in frame 0, $T_i^1 = \mathcal{P}(T^0)_i = \{\mathcal{P}(t), \mathcal{P}(\mathcal{C}(t))\}_i$ in frame 1, and $T_i^2 = \mathcal{P}^2(T^0)_i = \{\mathcal{P}^2(t), \mathcal{P}^2(\mathcal{C}(t))\}_i$ in frame 2. A maximal (20 trinucleotides) C^3 self-complementary code C is identified with the first 10 values of the numbers $Nb(T_1,\mathbb{K}), \ldots, Nb(T_{10},\mathbb{K})$ (defined below). Precisely, the code C has 20 trinucleotides $C = C_0 = \{T_1^0, \ldots, T_{10}^0\}$ in frame 0, 20 trinucleotides $C_1 = \mathcal{P}(C) = \{T_1^1, \ldots, T_{10}^1\}$ in frame 1, and 20 trinucleotides $C_2 = \mathcal{P}^2(C) = \{T_1^2, \ldots, T_{10}^2\}$ in frame 2 with $C = \mathcal{C}(C)$ (self-complementary), $\mathcal{C}(C_1) = C_2$, and $\mathcal{C}(C_2) = C_1$ (C_1 and C_2 are complementary). There are $\begin{pmatrix} 30 \\ 10 \end{pmatrix} = 30,045,015$ C^3 self-complementary trinucleotide codes, and among them only 216 are circular [2,20].

Notation 3. *A CP trinucleotide set $T = \{T^0, T^1, T^2\}$ belongs to the C^3 self-complementary trinucleotide circular code X, i.e., $T \in X$, if $T^0 \cap X \neq \varnothing$, i.e., if the trinucleotide t and its complementary trinucleotide $\mathcal{C}(t)$ belong to X. Ten CP trinucleotide sets T among 30 belong to the C^3 circular code X, i.e., such that 10 sets $T^0 \in X$ with $T^1 = \mathcal{P}(T^0) \in \mathcal{P}(X) = X_1$ and $T^2 = \mathcal{P}^2(T^0) \in \mathcal{P}^2(X) = X_2$.*

Notation 4. *In order to facilitate the reading of* Table 2*, the 30 CP trinucleotide sets $T = \{T^0, T^1, T^2\}$ are presented in the following way (i) the first 10 sets $T_1, \ldots, T_{10}$ belong to the circular code X (with $T^0 = \{t, \mathcal{C}(t)\} \in X$, $T^1 \in X_1$ and $T^2 \in X_2$) and are in lexicographical order with respect to the trinucleotide $t \in X$ (in bold), and (ii) the 20 remaining sets $T_{11}, \ldots, T_{30}$ are in lexicographical order with respect to the trinucleotide $t \in X_1$ (in italics).*

The occurrence number $Nb(T,\mathbb{K})$ of a CP trinucleotide set $T = \{T^0, T^1, T^2\}$ in a gene kingdom $\mathbb{K}$ is equal to

$$\begin{aligned} Nb(T,\mathbb{K}) = {} & Nb_0(t,\mathbb{K}) + Nb_0(\mathcal{C}(t),\mathbb{K}) + Nb_1(\mathcal{P}(t),\mathbb{K}) + Nb_1(\mathcal{P}(\mathcal{C}(t)),\mathbb{K}) \\ & + Nb_2\left(\mathcal{P}^2(t),\mathbb{K}\right) + Nb_2\left(\mathcal{P}^2(\mathcal{C}(t)),\mathbb{K}\right) \end{aligned} \tag{8}$$

where $Nb_f(t,K)$ is defined in Equation (7).

In order to normalize the numbers $Nb(T,\mathbb{K})$ which depend on the numbers of genes in a kingdom $\mathbb{K}$, we simply define the occurrence probability $Pb(T,\mathbb{K})$ of a CP trinucleotide set $T=\{T^0,T^1,T^2\}$ in a gene kingdom $\mathbb{K}$ as follows

$$Pb(T,\mathbb{K})=\frac{Nb(T,\mathbb{K})}{\sum_{i=1}^{30}Nb(T_i,\mathbb{K})} \quad (9)$$

where $Nb(T,\mathbb{K})$ is defined in Equation (8).

The parameter $Rk(T,\mathbb{K})\in\{1,\ldots,30\}$ gives the rank of the values $Pb(T,\mathbb{K})$ among the 30 CP trinucleotide sets T, the 1st rank being associated to the highest value of $Pb(T,\mathbb{K})$ and the 30th rank, to the lowest value of $Pb(T,\mathbb{K})$.

2.6. A Statistical Test to Evaluate the Significance of the Obtained Ranks

In order to evaluate the statistical significance of the ranks $Rk(T,\mathbb{K})$ of the probabilities $Pb(T,\mathbb{K})$ (Equation (9)) of the 30 CP trinucleotide sets T in a given kingdom $\mathbb{K}$, we derive confidence intervals for $Pb(T,\mathbb{K})$. If the confidence interval for two probabilities $Pb(T,\mathbb{K})$ do not overlap, then their associated ranks $Rk(T,\mathbb{K})$ are assumed to be valid (in the population). The confidence interval for two probabilities $Pb(T,\mathbb{K})$ is evaluated by using the classical 2-sample z-test which is briefly recalled here.

Let $\mathcal{P}(T)$ and $\mathcal{P}(T')$ be the populations associated to the CP trinucleotide sets T and T' of probabilities $Pb(T,\mathcal{P})$ and $Pb(T',\mathcal{P})$, respectively. The probabilities $Pb(T,\mathbb{K})$ and $Pb(T',\mathbb{K})$ of T and T' are observed in a given gene kingdom $\mathbb{K}$ (sample) of size $n=\sum_{i=1}^{30}Nb(T_i,\mathbb{K})$ (defined from Equation (8)). The tests carried out in Section 3 are applied on large samples (the size of the smallest sample analysed being $n=10921$ with the archaea $\mathbb{A}$). Thus, the assumptions of normality for the variables and of the homogeneity for the variances in the two populations are not needed. The equality $H_0: Pb(T,\mathcal{P})\geq Pb(T',\mathcal{P})$ is tested against the alternative $H_1: Pb(T,\mathcal{P})<Pb(T',\mathcal{P})$ if they are not equal. Under H_0 and with large samples ($n>30$), $\min\{nPb(T,\mathcal{P}),n(1-Pb(T,\mathcal{P})),nPb(T',\mathcal{P}),n(1-Pb(T',\mathcal{P}))\}>5$ (always verified in the tests carried out in Section 3), and T and T' are independent events (realistic hypothesis with kingdoms $\mathbb{K}$ of large sizes), then

$$Z=\frac{Pb(T,\mathcal{P})-Pb(T',\mathcal{P})}{\sqrt{\left(\frac{nPb(T,\mathcal{P})+nPb(T',\mathcal{P})}{n+n}\right)\left(1-\frac{nPb(T,\mathcal{P})+nPb(T',\mathcal{P})}{n+n}\right)\left(\frac{1}{n}+\frac{1}{n}\right)}}=\frac{\sqrt{2}(Pb(T,\mathcal{P})-Pb(T',\mathcal{P}))}{\sqrt{-\frac{(Pb(T,\mathcal{P})+Pb(T',\mathcal{P})-2)(Pb(T,\mathcal{P})+Pb(T',\mathcal{P}))}{n}}}\sim\mathcal{N}(0,1).$$

The z-value and the p-value are given for each statistical test carried out in Section 3.

2.7. Explained Example of the Statistical Approach Developed

As an example, we explain the definition of the occurrence probability $Pb(T,\mathbb{K})$ (Equation (9)) which takes the value of 6.1% (see Table 2) with the CP trinucleotide set $T_1=\{T_1^0,T_1^1,T_1^2\}$ with $T_1^0=\{t,\mathcal{C}(t)\}=\{\text{AAC},\text{GTT}\}$ in frame 0, $T_1^1=\mathcal{P}(T^0)_1=\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\}=\{\text{ACA},\text{TTG}\}$ in frame 1 and $T_1^2=\mathcal{P}^2(T^0)_1=\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\}=\{\text{CAA},\text{TGT}\}$ in frame 2 in the gene kingdom of bacteria $\mathbb{K}=\mathbb{B}$ (Table 1).

The $3\times64=192$ occurrence frequencies $Pr_f(t,g)$ of the 64 trinucleotides t are computed in the three frames f of each gene g belonging to $\mathbb{B}$. Then, the preferential frame $F(t,g)$ of each trinucleotide t for each gene g in $\mathbb{B}$ is determined according to Equation (5). For example, with the trinucleotide $t=\text{AAC}$ in a gene g_1 of $\mathbb{B}$, if the frequency $Pr_0(\text{AAC},g_1)$ of AAC in frame $f=0$ (reading frame) is greater than the two frequencies $Pr_1(\text{AAC},g_1)$ and $Pr_2(\text{AAC},g_1)$ of AAC in frames $f=1$ and $f=2$, i.e., $Pr_0(\text{AAC},g_1)>Max\{Pr_1(\text{AAC},g_1),Pr_2(\text{AAC},g_1)\}$, then the preferential frame of AAC in g_1 is 0, i.e., $F(\text{AAC},g_1)=0$.

The indicator function $\delta_f(F(t,g))$ of each trinucleotide t for each gene g in $\mathbb{B}$ is obtained from Equation (6). With the previous example of AAC in the gene g_1 of $\mathbb{B}$, the indicator function is equal to $\delta_0(F(\text{AAC},g_1))=1$ for the frame $f=0$ and $\delta_1(F(\text{AAC},g_1))=\delta_2(F(\text{AAC},g_1))=0$ for the frames $f=1$ and $f=2$.

The number $Nb_f(t,\mathbb{B})$ of preferential frames of each trinucleotide t for each frame f in $\mathbb{B}$ is computed according to Equation (7). With the previous example of AAC in $\mathbb{B}$, the following numbers are obtained: $Nb_0(\mathrm{AAC},\mathbb{B}) = 3486$ for the frame $f = 0$, $Nb_1(\mathrm{AAC},\mathbb{B}) = 1742$ for the frame $f = 1$, and $Nb_2(\mathrm{AAC},\mathbb{B}) = 1819$ for the frame $f = 2$. Thus, the preferential frame of AAC in $\mathbb{B}$ is 0.

The occurrence number $Nb(T,\mathbb{B})$ of the 30 CP trinucleotide sets $T_i = \{T_i^0, T_i^1, T_i^2\}$ in $\mathbb{B}$ is determined according to Equation (8). With T_1 in $\mathbb{B}$, the following numbers are obtained: $Nb_0(\mathrm{GTT},\mathbb{B}) = 3765$ for the frame $f = 0$, $Nb_1(\mathrm{ACA},\mathbb{B}) = 4002$ and $Nb_1(\mathrm{TTG},\mathbb{B}) = 5650$ for the frame $f = 1$, and $Nb_2(\mathrm{CAA},\mathbb{B}) = 3999$ and $Nb_2(\mathrm{TGT},\mathbb{B}) = 4677$ for the frame $f = 2$. Then, the occurrence number of T_1 in $\mathbb{B}$ is equal to $Nb(T_1,\mathbb{B}) = Nb_0(\mathrm{AAC},\mathbb{B}) + Nb_0(\mathrm{GTT},\mathbb{B}) + Nb_1(\mathrm{ACA},\mathbb{B}) + Nb_1(\mathrm{TTG},\mathbb{B}) + Nb_2(\mathrm{CAA},\mathbb{B}) + Nb_2(\mathrm{TGT},\mathbb{B}) = 3486 + 3765 + 4002 + 5650 + 3999 + 4677 = 25579$.

Finally, the occurrence probability $Pb(T,\mathbb{B})$ of the 30 CP trinucleotide sets $T_i = \{T_i^0, T_i^1, T_i^2\}$ in $\mathbb{B}$ is deduced from Equation (9). With T_1 in $\mathbb{B}$, the occurrence probability of T_1 in $\mathbb{B}$ is equal to $Pb(T_1,\mathbb{B}) = \frac{Nb(T_1,\mathbb{B})}{\sum_{i=1}^{30} Nb(T_i,\mathbb{B})} = \frac{Nb(T_1,\mathbb{B})}{Nb(T_1,\mathbb{B})+\ldots+Nb(T_{30},\mathbb{B})} = \frac{25579}{25579+\ldots+11856} = \frac{25579}{422598} \approx 6.1\%$.

3. Results

3.1. Maximal C^3 Self-Complementary Circular Code X in Genes

This new statistical approach will show that the same set X of 20 trinucleotides among $\begin{pmatrix} 30 \\ 10 \end{pmatrix} = 30,045,015$ sets occurs preferentially in genes (reading frame) of bacteria $\mathbb{B}$, archaea $\mathbb{A}$, plasmids $\mathbb{P}$, eukaryotes $\mathbb{E}$, and viruses $\mathbb{V}$. This set X is the maximal C^3 self-complementary circular code defined in Equation (1).

3.1.1. Circular Code X in Genes of Bacteria

In the genes of bacteria $\mathbb{B}$, the 10 CP trinucleotide sets $T_1,\ldots,T_{10} \in X$ have occurrence probabilities $Pb(T,\mathbb{B})$ (Equation (9)) with the 10 highest ranks $Rk(T,\mathbb{B})$ among 30 (Table 2), i.e., $\{t,\mathcal{C}(t)\} \in X$, $\{\mathcal{P}(t),\mathcal{P}(\mathcal{C}(t))\} \in X_1$ and $\{\mathcal{P}^2(t),\mathcal{P}^2(\mathcal{C}(t))\} \in X_2$ 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 highest rank with $Pb(T_8,\mathbb{B}) = 8.2\%$ is related to the complementary pair $\{t,\mathcal{C}(t)\} = \{\mathrm{GAC},\mathrm{GTC}\} \in X$. The 10th rank with $Pb(T_5,\mathbb{B}) = 4.55\%$ is very significantly greater than the 11th rank with $Pb(T_{22},\mathbb{B}) = 3.31\%$ ($n = \sum_{i=1}^{30} Nb(T_i,\mathbb{B}) = 422598$, z-value = 29.33, p-value = 10^{-189}). The 20 trinucleotides of the circular code X are identified in the genes of bacteria:

$$X_{\mathbb{B}} = X. \tag{10}$$

The same result is obtained at the gene level and the gene population level [1].

3.1.2. Circular Code X in Genes of Archaea

In the genes of archaea $\mathbb{A}$, the eight CP trinucleotide sets $T_1, T_2, T_4, T_6, \ldots, T_{10} \in X$ (except T_3 and T_5) have occurrence probabilities $Pb(T,\mathbb{A})$ with the eight highest ranks $Rk(T,\mathbb{A})$ among 30 (Table 2). The highest rank with $Pb(T_8,\mathbb{A}) = 9.7\%$ is also related to the complementary pair $\{\mathrm{GAC},\mathrm{GTC}\} \in X$. The CP set $T_{22} \notin X$ with $Rk(T_{22},\mathbb{A}) = 9$ explains that the two complementary trinucleotides $\{t,\mathcal{C}(t)\} = \{\mathrm{ACC},\mathrm{GGT}\} \in X$ (T_3) do not occur preferentially in $\mathbb{A}$. As the CP set $T_5 \in X$ has a rank $Rk(T_5,\mathbb{A}) = 13$ with $Pb(T_5,\mathbb{A}) = 3.66\%$ greater than $Rk(T_{15},\mathbb{A}) = 14$ with $Pb(T_{15},\mathbb{A}) = 3.39\%$ and $Rk(T_{28},\mathbb{A}) = 15$ with $Pb(T_{28},\mathbb{A}) = 2.95\%$, the two complementary trinucleotides $\{t,\mathcal{C}(t)\} = \{\mathrm{CAG},\mathrm{CTG}\} \in X$ occur preferentially in $\mathbb{A}$ compared to $\{\mathrm{AGC},\mathrm{GCT}\}$ (T_{15}) and $\{\mathrm{GCA},\mathrm{TGC}\}$ (T_{28}), however the statistical significance between the ranks $Rk(T_5,\mathbb{A})$ and $Rk(T_{15},\mathbb{A})$ is not confirmed due to the lack of archaeal gene data (see Section 2.2) ($n = \sum_{i=1}^{30} Nb(T_i,\mathbb{A}) = 10921$, z-value = 1.08 , p-value = 0.14). Thus, a subset

of X of 18 trinucleotides (a non-maximal C^3 self-complementary circular code) is identified in the genes of archaea:

$$X_{\mathbb{A}} = X \backslash Y_{\mathbb{A}} \text{ with } Y_{\mathbb{A}} = \{\text{ACC}, \text{GGT}\}. \quad (11)$$

Note that the code $X_{\mathbb{A}} \cup \{\text{CAC}, \text{GTG}\}$ (T_{22}) is the variant X code observed in *Deinococcus* [1]. The circular code X retrieved in the genes of archaea is a new result which was not found in a study of variant X codes in archaeal genomes [15].

3.1.3. Circular Code X in Genes of Plasmids

In the genes of plasmids $\mathbb{P}$, the 10 CP trinucleotide sets $T_1, \ldots, T_{10} \in X$ have occurrence probabilities $Pb(T, \mathbb{P})$ with the 10 highest ranks $Rk(T, \mathbb{P})$ among 30 (Table 2). The highest rank with $Pb(T_8, \mathbb{P}) = 7.8\%$ is again related to the complementary pair $\{\text{GAC}, \text{GTC}\} \in X$. The 10th rank with $Pb(T_5, \mathbb{P}) = 3.93\%$ is very significantly greater than the 11th rank with $Pb(T_{21}, \mathbb{P}) = 3.43\%$ ($n = \sum_{i=1}^{30} Nb(T_i, \mathbb{P}) = 144366$, z-value = 7.14, p-value = 10^{-13}). The 20 trinucleotides of the circular code X are identified in the genes of plasmids:

$$X_{\mathbb{P}} = X. \quad (12)$$

The same result is obtained at the gene level and the gene population level [1].

3.1.4. Circular Code X in Genes of Eukaryotes

In the genes of eukaryotes $\mathbb{E}$, the 10 CP trinucleotide sets $T_1, \ldots, T_{10} \in X$ have occurrence probabilities $Pb(T, \mathbb{E})$ with the 10 highest ranks $Rk(T, \mathbb{E})$ among 30 (Table 2). The highest rank with $Pb(T_8, \mathbb{E}) = 9.0\%$ is again related to the complementary pair $\{\text{GAC}, \text{GTC}\} \in X$. The 10th rank with $Pb(T_5, \mathbb{E}) = 4.23\%$ is significantly greater than the 11th rank with $Pb(T_{22}, \mathbb{E}) = 3.82\%$ ($n = \sum_{i=1}^{30} Nb(T_i, \mathbb{E}) = 11401$, z-value = 1.57, p-value = 0.06). The 20 trinucleotides of the circular code X are identified in the genes of eukaryotes:

$$X_{\mathbb{E}} = X. \quad (13)$$

The same result is obtained at the gene level and the gene population level [1].

The subset $X_{\mathbb{E}_{\textit{Homo sapiens}}} = X \backslash \{\text{ACC}, \text{GCC}, \text{GGC}, \text{GGT}\}$ of X of 16 trinucleotides in the genes of *Homo sapiens* identified at the gene level is also identical to the subset found at the gene population level [1].

3.1.5. Circular Code X in Genes of Eukaryotic Chromosomes

The statistical analysis in Section 3.1.4 takes the eukaryotic genome as the genetic information unit. Indeed, Equation (7) with $g \in \mathbb{E}$ is achieved with $Card(\mathbb{E}) = 190$ eukaryotic genomes (see Table 1). We complete this classical approach by choosing the eukaryotic chromosome as the genetic information unit. Thus, Equation (7) with $g \in \mathbb{E}_{\text{chr}}$ is performed with $Card(\mathbb{E}_{\text{chr}}) = 2979$ eukaryotic chromosomes of $Card(\mathbb{E}) = 190$ genomes (see Table 1).

In the genes of eukaryotic chromosomes $\mathbb{E}_{\text{chr}}$, the 10 CP trinucleotide sets $T_1, \ldots, T_{10} \in X$ have occurrence probabilities $Pb(T, \mathbb{E}_{\text{chr}})$ with the 10 highest ranks $Rk(T, \mathbb{E}_{\text{chr}})$ among 30 (Table 2). The highest rank with $Pb(T_8, \mathbb{E}_{\text{chr}}) = 9.1\%$ is again related to the complementary pair $\{\text{GAC}, \text{GTC}\} \in X$. The 10th rank with $Pb(T_3, \mathbb{E}_{\text{chr}}) = 4.74\%$ is very significantly greater than the 11th rank with $Pb(T_{22}, \mathbb{E}_{\text{chr}}) = 4.47\%$ ($n = \sum_{i=1}^{30} Nb(T_i, \mathbb{E}_{\text{chr}}) = 179136$, z-value = 3.86, p-value = 10^{-5}). The 20 trinucleotides of the circular code X are identified in the genes of eukaryotic chromosomes:

$$X_{\mathbb{E}_{\text{chr}}} = X. \quad (14)$$

It is a new result which completes the statistical analysis of genes in eukaryotic genomes (Section 3.1.4).

3.1.6. Non-Maximal Circular Code X in Genes of Eukaryotic Organelles

The genes of eukaryotic organelles, i.e., mitochondria and chloroplasts, are investigated with this statistical approach. It should also be stressed that the available data have an order of magnitude very significantly lower than the other gene kingdoms studied (less than 1 million trinucleotides for each class of organelles, see Table 1). However, we can already observe some statistical trends with the trinucleotides in the preferential frame.

Non-Maximal Circular Code X in Genes of Mitochondria

Surprisingly, in the genes of mitochondria $\mathbb{M}$, the four CP trinucleotide sets $T_9, T_7, T_8, T_3 \in X$ have occurrence probabilities $Pb(T, \mathbb{M})$ with the four highest ranks $Rk(T, \mathbb{M})$ among 30 (Table 2). The CP set $T_{28} \notin X$ with $Rk(T_{28}, \mathbb{M}) = 5$ explains that the two complementary trinucleotides $\{\mathrm{CAG}, \mathrm{CTG}\} \in X$ (T_5) do not occur preferentially in $\mathbb{M}$. The CP set $T_{25} \notin X$ with $Rk(T_{25}, \mathbb{M}) = 6$ determines that the two complementary trinucleotides $\{\mathrm{CTC}, \mathrm{GAG}\} \in X$ (T_6) do not occur preferentially in $\mathbb{M}$. The CP set $T_{24} \notin X$ with $Rk(T_{24}, \mathbb{M}) = 7$ implies that the two complementary trinucleotides $\{\mathrm{ATC}, \mathrm{GAT}\} \in X$ (T_4) do not occur preferentially in $\mathbb{M}$. The CP set $T_{17} \notin X$ with $Rk(T_{17}, \mathbb{M}) = 11$ explains that the two complementary trinucleotides $\{\mathrm{AAT}, \mathrm{ATT}\} \in X$ (T_2) do not occur preferentially in $\mathbb{M}$. Thus, a subset of X of 12 trinucleotides (a non-maximal C^3 self-complementary circular code) is identified in the genes of mitochondria $\mathbb{M}$:

$$X_{\mathbb{M}} = X \backslash Y_{\mathbb{M}} \text{ with } Y_{\mathbb{M}} = \{\mathrm{AAT}, \mathrm{ATC}, \mathrm{ATT}, \mathrm{CAG}, \mathrm{CTC}, \mathrm{CTG}, \mathrm{GAG}, \mathrm{GAT}\}. \quad (15)$$

This subset $X_{\mathbb{M}} = \{\mathrm{AAC}, \mathrm{ACC}, \mathrm{GAA}, \mathrm{GAC}, \mathrm{GCC}, \mathrm{GGC}, \mathrm{GGT}, \mathrm{GTA}, \mathrm{GTC}, \mathrm{GTT}, \mathrm{TAC}, \mathrm{TTC}\}$ is very close to the subset $X_{\tilde{\mathbb{M}}} = \{\mathrm{ACC}, \mathrm{ATC}, \mathrm{CTC}, \mathrm{GAA}, \mathrm{GAC}, \mathrm{GAT}, \mathrm{GCC}, \mathrm{GGC}, \mathrm{GGT}, \mathrm{GTA}, \mathrm{GTC}, \mathrm{GTT}, \mathrm{TTC}\}$ of X of 13 trinucleotides previously identified by inspection in mitochondrial genes [21], as $X_{\mathbb{M}} \bigcap X_{\tilde{\mathbb{M}}} = \{\mathrm{ACC}, \mathrm{GAA}, \mathrm{GAC}, \mathrm{GCC}, \mathrm{GGC}, \mathrm{GGT}, \mathrm{GTA}, \mathrm{GTC}, \mathrm{GTT}, \mathrm{TTC}\}$ has 10 trinucleotides in common.

Non-Maximal Circular Code X in Genes of Chloroplasts

In the genes of chloroplasts $\mathbb{C}$, the highest occurrences of CP trinucleotide sets again belong to the circular code X. The three CP trinucleotide sets $T_2, T_9, T_3 \in X$ have occurrence probabilities $Pb(T, \mathbb{C})$ with the three highest ranks $Rk(T, \mathbb{C})$ among 30 (Table 2). The CP set $T_{13} \notin X$ with $Rk(T_{13}, \mathbb{C}) = 4$ explains that the two complementary trinucleotides $\{\mathrm{GAC}, \mathrm{GTC}\} \in X$ (T_8) do not occur preferentially in $\mathbb{C}$. The CP set $T_{28} \notin X$ with $Rk(T_{28}, \mathbb{C}) = 5$ states that the two complementary trinucleotides $\{\mathrm{CAG}, \mathrm{CTG}\} \in X$ (T_5) do not occur preferentially in $\mathbb{C}$. The CP set $T_{14} \notin X$ with $Rk(T_{14}, \mathbb{C}) = 8$ implies that the two complementary trinucleotides $\{\mathrm{GTA}, \mathrm{TAC}\} \in X$ (T_{10}) do not occur preferentially in $\mathbb{C}$. The CP set $T_{18} \notin X$ with $Rk(T_{18}, \mathbb{C}) = 10$ explains that the two complementary trinucleotides $\{\mathrm{ATC}, \mathrm{GAT}\} \in X$ (T_4) do not occur preferentially in $\mathbb{C}$. The CP set $T_{25} \notin X$ with $Rk(T_{25}, \mathbb{C}) = 12$ implies that the two complementary trinucleotides $\{\mathrm{CTC}, \mathrm{GAG}\} \in X$ (T_6) do not occur preferentially in $\mathbb{C}$. Thus, a subset of X of 10 trinucleotides (a non-maximal C^3 self-complementary circular code) is identified in the genes of chloroplasts $\mathbb{C}$:

$$X_{\mathbb{C}} = X \backslash Y_{\mathbb{C}} \text{ with } Y_{\mathbb{C}} = \{\mathrm{ATC}, \mathrm{CAG}, \mathrm{CTC}, \mathrm{CTG}, \mathrm{GAC}, \mathrm{GAG}, \mathrm{GAT}, \mathrm{GTA}, \mathrm{GTC}, \mathrm{TAC}\}. \quad (16)$$

3.1.7. Circular Code X in Genes of Viruses

In the genes of viruses $\mathbb{V}$, the nine CP trinucleotide sets $T_1, \ldots, T_4, T_6, \ldots, T_{10} \in X$ (except T_5) have occurrence probabilities $Pb(T, \mathbb{V})$ with the nine highest ranks $Rk(T, \mathbb{V})$ among 30 (Table 2). The highest rank with $Pb(T_8, \mathbb{V}) = 7.2\%$ is again related to the complementary pair $\{\mathrm{GAC}, \mathrm{GTC}\} \in X$. The CP set $T_{15} \notin X$ with $Rk(T_{15}, \mathbb{V}) = 10$ explains that the two complementary trinucleotides $\{\mathrm{CAG}, \mathrm{CTG}\} \in X$

(T_5) do not occur preferentially in $\mathbb{V}$. Thus, a subset of X of 18 trinucleotides (a non-maximal C^3 self-complementary circular code) is identified in the genes of viruses:

$$X_{\mathbb{V}} = X \backslash Y_{\mathbb{V}} \text{ with } Y_{\mathbb{V}} = \{\text{CAG}, \text{CTG}\}. \quad (17)$$

The statistical method of viral genes at the gene population level [1] could not decide between the two codes $X_{18} = X\backslash\{\text{CAG}, \text{CTG}\}$ and $X_{16} = X\backslash\{\text{CAG}, \text{CTG}, \text{GTA}, \text{TAC}\}$. The statistical analysis at the gene level confirms the code $X_{\mathbb{V}} = X_{18}$ of 18 trinucleotides in the genes of viruses.

3.2. Circular Code X Found in DNA and RNA Genomes and in Double-Stranded and Single-Stranded Genomes

The self-complementary property of the circular code X has been related since 1996 to the complementary property of the DNA double helix. In order to deepen this idea, we searched with this statistical approach the circular code X in five important sub-classes of viral genes using either DNA genome or RNA genome, and either double-stranded genome or single-stranded genome, i.e., in the genes of double-stranded DNA viruses $\mathbb{V}_{\text{dsDNA}}$, double-stranded RNA viruses $\mathbb{V}_{\text{dsRNA}}$, single-stranded DNA viruses $\mathbb{V}_{\text{ssDNA}}$, single-stranded RNA viruses $\mathbb{V}_{\text{ssRNA}}$, and retro-transcribing viruses $\mathbb{V}_{\text{rt}}$.

In the genes of double-stranded DNA viruses $\mathbb{V}_{\text{dsDNA}}$, the 10 CP trinucleotide sets $T_1, \ldots, T_{10} \in X$ have occurrence probabilities $Pb(T, \mathbb{V}_{\text{dsDNA}})$ with the 10 highest ranks $Rk(T, \mathbb{V}_{\text{dsDNA}})$ among 30 (Table 2). Thus, the circular code X is found in $\mathbb{V}_{\text{dsDNA}}$:

$$X_{\mathbb{V}_{\text{dsDNA}}} = X. \quad (18)$$

In the genes of double-stranded RNA viruses $\mathbb{V}_{\text{dsRNA}}$, single-stranded RNA viruses $\mathbb{V}_{\text{ssRNA}}$, and retro-transcribing viruses $\mathbb{V}_{\text{rt}}$, respectively, the nine CP trinucleotide sets $T_1, \ldots, T_4, T_6, \ldots, T_{10} \in X$ (except T_5) have occurrence probabilities $Pb(T, \mathbb{V}_{\text{dsRNA}})$, $Pb(T, \mathbb{V}_{\text{ssRNA}})$, and $Pb(T, \mathbb{V}_{\text{rt}})$, respectively, with the nine highest ranks $Rk(T, \mathbb{V}_{\text{dsRNA}})$, $Rk(T, \mathbb{V}_{\text{ssRNA}})$, and $Rk(T, \mathbb{V}_{\text{rt}})$, respectively, among 30 (Table 2). Note that the ranks $Rk(T, \mathbb{V}_{\text{dsRNA}})$, $Rk(T, \mathbb{V}_{\text{ssRNA}})$, and $Rk(T, \mathbb{V}_{\text{rt}})$ for a given CP trinucleotide set are not identical (Table 2). Thus, by using the reasoning mentioned previously ($T_{15} \notin X$ with $Rk(T_{15}, \mathbb{V}) > Rk(T_5, \mathbb{V})$ for $\mathbb{V}$ in $\mathbb{V}_{\text{dsRNA}}$, $\mathbb{V}_{\text{ssRNA}}$, and $\mathbb{V}_{\text{rt}}$), a subset of X of 18 trinucleotides is observed in $\mathbb{V}_{\text{dsRNA}}$, $\mathbb{V}_{\text{ssRNA}}$, and $\mathbb{V}_{\text{rt}}$:

$$X_{\mathbb{V}_{\text{dsRNA}}} = X \backslash Y_{\mathbb{V}_{\text{dsRNA}}} \text{ with } Y_{\mathbb{V}_{\text{dsRNA}}} = \{\text{CAG}, \text{CTG}\}, \quad (19)$$

$$X_{\mathbb{V}_{\text{ssRNA}}} = X \backslash Y_{\mathbb{V}_{\text{ssRNA}}} \text{ with } Y_{\mathbb{V}_{\text{ssRNA}}} = \{\text{CAG}, \text{CTG}\}, \quad (20)$$

$$X_{\mathbb{V}_{\text{rt}}} = X \backslash Y_{\mathbb{V}_{\text{rt}}} \text{ with } Y_{\mathbb{V}_{\text{rt}}} = \{\text{CAG}, \text{CTG}\}. \quad (21)$$

In the genes of single-stranded DNA viruses $\mathbb{V}_{\text{ssDNA}}$, the eight CP trinucleotide sets $T_1, \ldots, T_4, T_6, \ldots, T_9 \in X$ (except T_5 and T_{10}) have occurrence probabilities $Pb(T, \mathbb{V}_{\text{ssDNA}})$ with the eight highest ranks $Rk(T, \mathbb{V}_{\text{ssDNA}})$ among 30 (Table 2). Thus, by using the reasoning as previously mentioned ($T_{15} \notin X$ with $Rk(T_{15}, \mathbb{V}_{\text{ssDNA}}) > Rk(T_5, \mathbb{V}_{\text{ssDNA}})$ and $T_{14} \notin X$ with $Rk(T_{14}, \mathbb{V}_{\text{ssDNA}}) > Rk(T_{10}, \mathbb{V}_{\text{ssDNA}})$), a subset of X of 16 trinucleotides is observed in $\mathbb{V}_{\text{ssDNA}}$:

$$X_{\mathbb{V}_{\text{ssDNA}}} = X \backslash Y_{\mathbb{V}_{\text{ssDNA}}} \text{ with } Y_{\mathbb{V}_{\text{ssDNA}}} = \{\text{CAG}, \text{CTG}, \text{GTA}, \text{TAC}\}. \quad (22)$$

All these results show that the circular code X is found almost perfectly in DNA genomes, RNA genomes, double-stranded genomes, and single-stranded genomes. The very few exceptions, either the two trinucleotides $\{\text{CAG}, \text{CTG}\}$ or the four trinucleotides $\{\text{CAG}, \text{CTG}, \text{GTA}, \text{TAC}\}$ for one case, are related to the CP set or the two CP sets having the lowest occurrence among the 10 CP sets $T_1, \ldots, T_{10} \in X$.

Table 2. Identification of the maximal C^3 self-complementary trinucleotide circular code X in gene kingdoms $\mathbb{K}$ of bacteria $\mathbb{B}$, archaea $\mathbb{A}$, plasmids $\mathbb{P}$, eukaryotes $\mathbb{E}$, chromosomes of eukaryotes $\mathbb{E}_{chr}$, mitochondria $\mathbb{M}$, chloroplasts $\mathbb{C}$, viruses $\mathbb{V}$, and its five taxonomic groups: double-stranded DNA viruses $\mathbb{V}_{dsDNA}$, double-stranded RNA viruses $\mathbb{V}_{dsRNA}$, single-stranded DNA viruses $\mathbb{V}_{ssDNA}$, single-stranded RNA viruses $\mathbb{V}_{ssRNA}$, and retro-transcribing viruses $\mathbb{V}_{rt}$ (Table 1). Occurrence probability $Pb(T,\mathbb{K})$ (%) of the 30 complementary and permutation (CP) trinucleotide sets $T = \{T^0, T^1, T^2\}$ with $T^0 = \{t, \mathcal{C}(t)\}$ in frame 0, $T^1 = \mathcal{P}(T^0) = \{\mathcal{P}(t), \mathcal{P}(\mathcal{C}(t))\}$ in frame 1, $T^2 = \mathcal{P}^2(T^0) = \{\mathcal{P}^2(t), \mathcal{P}^2(\mathcal{C}(t))\}$ in frame 2, in a gene kingdom $\mathbb{K}$ computed according to Equation (9) and its rank $Rk(T,\mathbb{K})$, the 1st rank being associated to the highest value of $Pb(T,\mathbb{K})$ and the 30th rank, to the lowest value of $Pb(T,\mathbb{K})$. The 20 trinucleotides of the C^3 self-complementary circular code X are in bold, the 20 trinucleotides of the circular code $X_1 = \mathcal{P}(X)$ are in italics, and the 20 trinucleotides of the circular code $X_2 = \mathcal{P}^2(X)$ are both in bold and italics. The first 10 CP sets $T_1, \ldots, T_{10}$ belong to the circular code X ($T^0 = \{t, \mathcal{C}(t)\} \in X$ with $T^1 = \mathcal{P}(T^0) \in \mathcal{P}(X) = X_1$ and $T^2 = \mathcal{P}^2(T^0) \in \mathcal{P}^2(X) = X_2$) and are in lexicographical order with respect to the trinucleotide $t \in X$ in bold, and the 20 remaining CP sets $T_{11}, \ldots, T_{30}$ are in lexicographical order with respect to the trinucleotide $t \in X_1$ in italics. The numbers in italics occurring with the CP sets $T_1, \ldots, T_{10}$ are associated with the two trinucleotides $T^0 = \{t, \mathcal{C}(t)\}$ of X which do not occur preferentially in the gene kingdom.

			$\mathbb{B}$		$\mathbb{A}$		$\mathbb{P}$		$\mathbb{E}$		$\mathbb{E}_{chr}$		$\mathbb{M}$		$\mathbb{C}$		$\mathbb{V}$		$\mathbb{V}_{dsDNA}$		$\mathbb{V}_{dsRNA}$		$\mathbb{V}_{ssDNA}$		$\mathbb{V}_{ssRNA}$		$\mathbb{V}_{rt}$	
T	***t***	$\mathcal{C}(t)$	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***
T_1	**AAC**	**GTT**	6.1	6	7.4	4	6.0	4	8.5	3	8.4	3	4.4	9	4.9	9	6.6	3	6.8	4	6.4	3	5.9	1	7.0	3	5.0	5
T_2	**AAT**	**ATT**	7.4	2	5.3	8	7.3	2	8.7	2	8.7	2	*3.4*	*14*	5.8	1	6.6	2	7.5	2	5.9	4	5.6	2	6.3	4	5.3	4
T_3	**ACC**	**GGT**	5.1	9	*3.9*	*12*	5.1	8	5.1	8	4.7	10	5.7	4	5.2	3	4.9	8	4.9	9	4.6	8	4.6	5	5.3	7	3.7	9
T_4	**ATC**	**GAT**	6.5	4	6.7	5	6.2	3	8.1	4	8.2	4	*3.6*	*13*	*4.6*	*14*	6.5	4	6.7	5	6.4	2	5.4	4	7.1	2	6.0	2
T_5	**CAG**	**CTG**	4.6	10	3.7	13	3.9	10	4.2	10	4.9	9	*0.7*	*30*	*0.0*	*30*	*2.9*	*15*	3.8	10	*2.4*	*18*	*2.8*	*19*	*1.5*	*24*	*2.9*	*18*
T_6	**CTC**	**GAG**	6.2	5	7.5	3	5.9	6	7.0	5	7.5	5	*2.6*	*18*	*0.4*	*27*	5.6	6	6.3	6	5.3	7	4.1	10	5.4	6	4.9	6
T_7	**GAA**	**TTC**	5.8	7	5.3	7	5.5	7	5.0	9	5.2	7	6.3	2	4.9	7	5.2	7	5.6	7	5.3	6	4.6	6	4.9	8	5.4	3
T_8	**GAC**	**GTC**	8.2	1	9.7	1	7.8	1	9.0	1	9.1	1	5.8	3	*0.6*	*26*	7.2	1	8.0	1	7.3	1	5.4	3	7.3	1	6.4	1
T_9	**GCC**	**GGC**	6.7	3	8.2	2	6.0	5	5.7	6	5.2	8	7.1	1	5.3	2	5.9	5	7.0	3	5.5	5	4.3	8	5.7	5	4.7	7
T_{10}	**GTA**	**TAC**	5.4	8	6.6	6	5.0	9	5.4	7	5.7	6	4.6	8	*4.8*	*11*	4.7	9	5.3	8	4.6	9	*4.0*	*11*	4.5	10	4.3	8
T_{11}	*AAG*	***CTT***	3.0	13	4.1	11	2.9	16	3.4	14	3.5	13	1.4	27	3.2	18	3.3	12	3.0	14	3.2	13	3.6	12	3.6	13	2.9	16
T_{12}	*ACA*	***TGT***	1.1	26	1.4	20	1.1	26	0.3	30	0.2	30	4.3	10	2.9	19	1.2	27	0.9	26	1.3	28	1.4	28	1.5	26	1.5	29
T_{13}	*ACG*	***CGT***	1.6	22	0.3	28	1.8	21	0.6	26	0.6	25	1.9	24	5.0	4	1.8	22	1.4	23	1.6	25	3.2	16	1.6	22	2.1	24
T_{14}	*ACT*	***AGT***	2.9	15	1.3	22	3.3	13	3.0	15	2.5	15	2.1	22	4.9	8	3.6	11	3.1	13	3.7	11	4.3	7	4.1	11	3.4	14
T_{15}	*AGC*	***GCT***	2.9	14	3.4	14	3.1	15	3.4	13	3.1	14	3.9	12	4.9	6	4.0	10	3.6	11	4.2	10	4.3	9	4.6	9	3.5	13
T_{16}	*AGG*	***CCT***	2.3	19	1.5	19	2.5	19	2.4	16	2.0	17	2.2	21	4.8	13	2.7	17	2.5	17	2.6	17	3.5	13	2.7	17	2.4	22
T_{17}	*ATA*	***TAT***	1.6	21	4.1	10	1.6	24	0.8	23	0.9	23	4.2	11	2.7	20	2.3	19	1.7	20	2.9	16	2.8	21	2.7	16	2.9	17
T_{18}	*ATG*	***CAT***	3.1	12	2.5	16	3.2	14	1.3	20	1.4	19	1.8	26	4.8	10	2.5	18	2.6	15	2.3	19	3.1	17	1.8	20	2.7	20
T_{19}	*CCA*	***TGG***	1.6	23	1.3	21	1.8	20	1.1	22	0.8	24	2.8	16	4.5	15	2.0	21	1.6	22	2.3	20	2.6	22	1.8	19	2.8	19
T_{20}	*CCG*	***CGG***	0.6	28	0.1	29	0.7	28	0.8	24	1.0	22	0.8	29	0.6	25	1.5	26	0.8	28	1.4	27	2.8	20	1.5	25	2.3	23

Table 2. *Cont.*

			$\mathbb{B}$		$\mathbb{A}$		$\mathbb{P}$		$\mathbb{E}$		$\mathbb{E}_{\mathbf{chr}}$		$\mathbb{M}$		$\mathbb{C}$		$\mathbb{V}$		$\mathbb{V}_{\mathbf{dsDNA}}$		$\mathbb{V}_{\mathbf{dsRNA}}$		$\mathbb{V}_{\mathbf{ssDNA}}$		$\mathbb{V}_{\mathbf{ssRNA}}$		$\mathbb{V}_{\mathbf{rt}}$	
T	***t***	$\mathcal{C}(t)$	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***	***Pb***	***Rk***
T_{21}	*GCG*	***CGC***	2.7	17	1.7	18	3.4	11	3.5	12	3.9	12	2.0	23	4.2	17	2.9	16	2.4	18	3.2	15	3.4	14	3.0	14	3.6	12
T_{22}	*GTG*	***CAC***	3.3	11	4.7	9	3.4	12	3.8	11	4.5	11	1.8	25	0.3	28	3.3	13	3.5	12	3.2	14	3.2	15	2.9	15	3.7	10
T_{23}	*TAG*	***CTA***	1.7	20	2.1	17	1.6	22	1.6	19	1.8	18	3.0	15	0.3	29	1.7	24	1.6	21	2.1	22	1.8	26	1.6	23	2.0	25
T_{24}	*TCA*	***TGA***	0.4	29	0.8	25	0.6	29	0.6	25	0.4	28	4.8	7	0.7	24	0.7	30	0.6	30	1.0	29	0.9	30	0.8	30	0.9	30
T_{25}	*TCC*	***GGA***	1.5	24	1.0	24	1.6	23	0.6	27	0.6	26	5.0	6	4.8	12	1.7	23	1.1	25	1.9	23	2.4	23	2.0	18	2.7	21
T_{26}	*TCG*	***CGA***	0.2	30	0.1	30	0.4	30	0.4	29	0.3	29	2.6	17	4.3	16	1.0	29	0.7	29	0.8	30	1.7	27	1.0	28	1.7	28
T_{27}	*TCT*	***AGA***	1.2	25	0.6	27	1.5	25	1.6	18	1.3	21	2.4	19	2.1	22	1.6	25	1.4	24	1.8	24	2.0	25	1.7	21	1.7	27
T_{28}	*TGC*	***GCA***	2.5	18	3.0	15	2.8	18	2.4	17	2.0	16	5.2	5	4.9	5	3.0	14	2.6	16	3.4	12	2.9	18	3.7	12	3.2	15
T_{29}	*TTA*	***TAA***	0.9	27	0.6	26	1.0	27	0.5	28	0.4	27	2.3	20	1.6	23	1.1	28	0.9	27	1.5	26	1.4	29	1.0	29	1.9	26
T_{30}	*TTG*	***CAA***	2.8	16	1.2	23	2.9	17	1.2	21	1.4	20	1.2	28	2.3	21	2.1	20	2.2	19	2.2	21	2.3	24	1.4	27	3.6	11

4. Conclusions

The "universal" occurrence in genes of a same set X of 20 trinucleotides, which has in addition the mathematical property to be a circular code, must be confirmed by several statistical approaches and various gene data analyses at different levels: kingdom, taxonomic group, genome, and gene. All the previous approaches have studied and identified the circular code X at the gene population level (kingdom, taxonomic group, and genome) [1,2,15–17,21]. The statistical approach at the gene level developed here, for the first time since 1996, analyses the preferential occurrence of trinucleotides among the three frames of each gene. This new methodology allows all genes, i.e., of large and small lengths, to be considered with the same weight. As a consequence, the concept of circular code, in particular the reading frame retrieval, is directly associated to each gene. Thus, X motifs from the circular code X at different locations in a gene may assist the ribosome to maintain and synchronize the reading frame. The number, the cardinality, and the length of X motifs in genes may be associated to the length, the function, and the ancestry of genes. This research work is currently under investigation.

At the gene level, the circular code X is strengthened in the genes of bacteria, eukaryotes, plasmids, and viruses, and is now also identified in the genes of archaea. In addition to eukaryotic genomes, it is also found in the genes of eukaryotic chromosomes. The genes of mitochondria and chloroplasts contain a subset of the circular code X. It should be stressed that some mitochondrial and chloroplast genes lack the stop codon and are excluded from this data acquisition. Such a statistical bias may prevent a proper detection of preferential frames for some trinucleotides in the genes of eukaryotic organelles. The circular code X is searched in the large class of $\begin{pmatrix} 30 \\ 10 \end{pmatrix} = 30,045,015$ C^3 self-complementary trinucleotide codes which contains in particular the 216 maximal C^3 self-complementary circular codes. Thus, for a basic order of magnitude, the probability to retrieve the same circular code X in four independent gene kingdoms (bacteria $\mathbb{B}$, plasmids $\mathbb{P}$, eukaryotes $\mathbb{E}$, double-stranded DNA viruses $\mathbb{V}_{\text{dsDNA}}$) is equal to $1/\begin{pmatrix} 30 \\ 10 \end{pmatrix}^4 \approx 10^{-30}$.

In the genes of the bacterial, eukaryotic, and plasmid kingdoms, 14 among the 47 studied gene taxonomic groups (about 30%) have variant X codes [1], i.e., trinucleotide codes which differ from X. Seven variant X codes are identified. However, all have at least 16 trinucleotides of X. Two variant X codes X_A (according to the notation in [1]) in cyanobacteria and plasmids of cyanobacteria, and X_D in birds, are self-complementary, without permuted trinucleotides, but are 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. Thus, two variant X codes X_A and X_D are not circular and do not belong to the set of the 216 maximal C^3 self-complementary circular codes [2] having the strong mathematical structure of the dihedral group [20]. The reason could be related to the gene data or to a biological property which remains to be identified. All these variant X codes in the genes are identified at the taxonomic group level. However, as the circular code X is now also identified at the gene level, variant X codes may also be associated with genes belonging to the same genome but with different protein coding functions. This interesting and open problem should be investigated in the future.

A probability measure of the reading frame retrieval (*RFR*) of each trinucleotide of X has been introduced in [22] and [23] (Section 2.2 and 1st row of Table 1). The *RFR* probability *PrRFR* of the circular code X, i.e., the average *RFR* probability of the 20 trinucleotides of X, is equal to $PrRFR(X) = 82.5\%$ (Result 5 in [22]; 1st row of Table 1 in [23]). This *RFR* measure can be applied to the non-maximal C^3 self-complementary circular codes, precisely to the excluded trinucleotides $Y_{\mathbb{A}} = \{\text{ACC}, \text{GGT}\}$ of archaea (Equation (11)), $Y_{\mathbb{M}} = \{\text{AAT}, \text{ATC}, \text{ATT}, \text{CAG}, \text{CTC}, \text{CTG}, \text{GAG}, \text{GAT}\}$ of mitochondria (Equation (15)), $Y_{\mathbb{C}} = \{\text{ATC}, \text{CAG}, \text{CTC}, \text{CTG}, \text{GAC}, \text{GAG}, \text{GAT}, \text{GTA}, \text{GTC}, \text{TAC}\}$ of chloroplasts (Equation (16)), $Y_{\mathbb{V}} = \{\text{CAG}, \text{CTG}\}$ of viruses (Equation (17)), and $Y_{\mathbb{V}_{\text{ssDNA}}} = \{\text{CAG}, \text{CTG}, \text{GTA}, \text{TAC}\}$ of single-stranded

DNA viruses (Equation (22)). The computation leads to $PrRFR(Y_{\mathbb{A}}) = 69.0\%$, $PrRFR(Y_{\mathbb{M}}) = 88.5\%$, $PrRFR(Y_{\mathbb{C}}) = 87.1\%$, $PrRFR(Y_{\mathbb{V}}) = 100.0\%$, and $PrRFR\left(Y_{\mathbb{V}_{\text{ssDNA}}}\right) = 85.7\%$. Archaeal genes miss two trinucleotides of X which have the lowest RFR values. In contrast, mitochondrial, chloroplast, and viral genes miss trinucleotides of X with high RFR values. Thus, the genes in reduced genomes are more flexible in translation, allowing overlap coding by frameshifting in agreement with [24] (and the cited references). However, it should be stressed that this result may vary with the increase of gene data of eukaryotic organelles in the future. The circular code X (20 trinucleotides) with the functions of reading frame retrieval and maintenance in regular RNA transcription, may also have, through its bijective transformation codes, the same functions in nucleotide exchanging RNA transcription in mitochondrial genes [23]. Indeed, as the mitochondrial gamma polymerase has bacterial origins (e.g., [25]), mitochondrial polymerization and its associated bijective transformations might use the circular code X. However at the translational level, the ribosome might follow the non-maximal C^3 self-complementary circular code $X_{\mathbb{M}}$ observed in mitochondrial genes (Equation (15)). A similar explanation could be applied to the chloroplast genes which have also bacterial origins (cyanobacteria).

By a study of viral genes, the circular code X is found in DNA genomes, RNA genomes, double-stranded genomes, and single-stranded genomes. Thus, the reading frame retrieval property of X could operate for translating DNA and RNA genes, in particular for the "primitive" RNA genes. The C^3 property of X could be involved for translating the two shifted frames in DNA and RNA genes, in particular for optimizing the genomes of small sizes. The complementarity property of X is naturally associated to the double-stranded DNA and RNA genomes. It could also be used to pair single-stranded DNA genomes between them and single-stranded RNA genomes between them. Thus, the C^3 and complementary properties of X could be involved for translating the three frames (reading frame and its two shifted frames) in one strand and the three frames in the complementary strand of DNA and RNA genes.

In summary, this new statistical approach at the gene level which is applied to massive gene data identifies the maximal C^3 self-complementary trinucleotide circular code X in the genes of bacteria, archaea, eukaryotes, plasmids, and viruses, which may be involved in translation coding [3].

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

Conflicts of Interest: The author declares no conflict of interest.

References

1. Michel, C.J. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. *J. Theor. Biol.* **2015**, *380*, 156–177. [CrossRef] [PubMed]
2. Arquès, D.G.; Michel, C.J. A complementary circular code in the protein coding genes. *J. Theor. Biol.* **1996**, *182*, 45–58. [CrossRef] [PubMed]
3. Michel, C.J. Circular code motifs in transfer and 16S ribosomal RNAs: A possible translation code in genes. *Comput. Biol. Chem.* **2012**, *37*, 24–37. [CrossRef] [PubMed]
4. El Soufi, K.; Michel, C.J. Circular code motifs in genomes of eukaryotes. *J. Theor. Biol.* **2016**, *408*, 198–212. [CrossRef] [PubMed]
5. Michel, C.J. Circular code motifs in transfer RNAs. *Comput. Biol. Chem.* **2013**, *45*, 17–29. [CrossRef] [PubMed]
6. El Soufi, K.; Michel, C.J. Circular code motifs in the ribosome decoding center. *Comput. Biol. Chem.* **2014**, *52*, 9–17. [CrossRef] [PubMed]
7. El Soufi, K.; Michel, C.J. Circular code motifs near the ribosome decoding center. *Comput. Biol. Chem.* **2015**, *59*, 158–176. [CrossRef] [PubMed]
8. El Soufi, K.; Michel, C.J. Unitary circular code motifs in genomes of eukaryotes. *Biosystems* **2017**, in press. [CrossRef] [PubMed]
9. Canapa, A.; Cerioni, P.N.; Barucca, M.; Olmo, E.; Caputo, V. A centromeric satellite DNA may be involved in heterochromatin compactness in gobiid fishes. *Chromosome Res.* **2002**, *10*, 297–304. [CrossRef] [PubMed]

10. Gemayel, R.; Vinces, M.D.; Legendre, M.; Verstrepen, K.J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **2010**, *44*, 445–477. [CrossRef] [PubMed]
11. Pirillo, G. A characterization for a set of trinucleotides to be a circular code. In *Determinism, Holism, and Complexity*; Pellegrini, C., Cerrai, P., Freguglia, P., Benci, V., Israel, G., Eds.; Kluwer Academic Publisher: New York, NY, USA, 2003.
12. Michel, C.J.; Pirillo, G. Identification of all trinucleotide circular codes. *J. Theor. Biol.* **2010**, *34*, 122–125. [CrossRef] [PubMed]
13. Fimmel, E.; Michel, C.J.; Strüngmann, L. *n*-Nucleotide circular codes in graph theory. *Philos. Trans. R. Soc. A* **2016**, *374*, 20150058. [CrossRef] [PubMed]
14. Bussoli, L.; Michel, C.J.; Pirillo, G. On conjugation partitions of sets of trinucleotides. *Appl. Math.* **2012**, *3*, 107–112. [CrossRef]
15. Frey, G.; Michel, C.J. Circular codes in archaeal genomes. *J. Theor. Biol.* **2003**, *223*, 413–431. [CrossRef]
16. Frey, G.; Michel, C.J. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. *Comput. Biol. Chem.* **2006**, *30*, 87–101. [CrossRef] [PubMed]
17. Arquès, D.G.; Michel, C.J. A code in the protein coding genes. *Biosystems* **1997**, *44*, 107–134. [CrossRef]
18. Gonzalez, D.L.; Giannerini, S.; Rosa, R. Circular codes revisited: a statistical approach. *J. Theor. Biol.* **2011**, *275*, 21–28. [CrossRef] [PubMed]
19. Michel, C.J.; Pirillo, G.; Pirillo, M.A. A relation between trinucleotide comma-free codes and trinucleotide circular codes. *Theor. Comput. Sci.* **2008**, *401*, 17–26. [CrossRef]
20. Fimmel, E.; Giannerini, S.; Gonzalez, D.L.; Strüngmann, L. Circular codes, symmetries and transformations. *J. Math. Biol.* **2015**, *70*, 1623–1644. [CrossRef] [PubMed]
21. Arquès, D.G.; Michel, C.J. A circular code in the protein coding genes of mitochondria. *J. Theor. Biol.* **1997**, *189*, 273–290. [CrossRef] [PubMed]
22. Ahmed, A.; Frey, G.; Michel, C.J. Essential molecular functions associated with circular code evolution. *J. Theor. Biol.* **2010**, *264*, 613–622. [CrossRef] [PubMed]
23. Michel, C.J.; Seligmann, H. Bijective transformation circular codes and nucleotide exchanging RNA transcription. *Biosystems* **2014**, *118*, 39–50. [CrossRef] [PubMed]
24. Seligmann, H. Chimeric mitochondrial peptides from contiguous regular and swinger RNA. *Comput. Struct. Biotechnol. J.* **2016**, *14*, 283–297. [CrossRef] [PubMed]
25. Wolf, Y.I.; Koonin, E.V. Origin of an animal mitochondrial DNA polymerase subunit via lineage-specific acquisition of a glycyl-tRNA synthetase from bacteria of the *Thermus-Deinococcus* group. *Trends Genet.* **2001**, *17*, 431–433. [CrossRef]