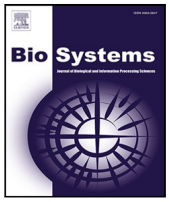
BioSystems 243 (2024) 105263

Contents lists available at ScienceDirect

BioSystems

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



Circular cut codes in genetic information

Elena Fimmel [a], Christian J. Michel [b,*], Lutz Strüngmann [a]

[a] Institute of Mathematical Biology, Faculty for Computer Sciences, Mannheim University of Applied Sciences, 68163 Mannheim, Germany
[b] Theoretical bioinformatics, ICube, University of Strasbourg, C.N.R.S., 300 Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Keywords:
Circular mixed codes
Dinucleotide circular codes
Trinucleotide circular codes
Graph properties

ABSTRACT

In this work we present an analysis of the dinucleotide occurrences in the three codon sites 1–2, 2–3 and 1–3, based on a computation of the codon usage of three large sets of bacterial, archaeal and eukaryotic genes using the same method that identified a maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria and eukaryotes in 1996 (Arquès and Michel, 1996). Surprisingly, two dinucleotide circular codes are identified in the codon sites 1–2 and 2–3. Furthermore, these two codes are shifted versions of each other. Moreover, the dinucleotide code in the codon site 1–3 is circular, self-complementary and contained in the projection of X onto the 1st and 3rd bases, i.e. by cutting the middle base in each codon of X. We prove several results showing that the circularity and the self-complementarity of trinucleotide codes is induced by the circularity and the self-complementarity of its dinucleotide cut codes. Finally, we present several evolutionary approaches for an emergence of trinucleotide codes from dinucleotide codes.

1. Introduction

The discovery almost 30 years ago of the universal distribution in nature amongst all species of the trinucleotide circular code X (Arquès and Michel, 1996) has raised many questions about its function in the modern genetic code, as well as the mechanism of its formation during the evolutionary process. Many theories of the genetic code assumed an increase of the coding number of amino acids during evolution and simultaneously a detection and correction of errors during the translation process (e.g. Koonin and Novozhilov, 2009; Dila et al., 2019). For example, the so-called $2-1-3$ theory (Taylor and Coates, 1989; Massey, 2006; Dragovich and Dragovich, 2010; Massey, 2016) claims that evolution went from encoding amino acids by single nucleotides, then by dinucleotides and finally by trinucleotides. In this context, we investigate if this nucleotide evolution can be retrieved in today's genes by analysing statistically the codon usage of three large data sets of bacterial, archaeal and eukaryotic genes.

The statistical analysis carried out is placed within the recent concept of circular mixed sets of words over an arbitrary finite alphabet (Fimmel et al., 2023). It offers an explanation of the frameshift robustness of the genetic code. More precisely, in this work we introduce the new concept of dinucleotide cut codes of a trinucleotide code, i.e. the dinucleotide codes created by cutting out the same position in each trinucleotide. The question is whether and under what conditions the circularity of the dinucleotide cut codes of a trinucleotide code follows its circularity. The same question is asked with regard to its self-complementarity, another important property of the genetic code. In both cases, a number of necessary and sufficient conditions are proven, supported by numerous examples. The identified relationships of the corresponding properties of the trinucleotide codes and their dinucleotide cut codes allow to better understand the concept of circularity and also the transition from encoding amino acids by words shorter than three letters.

2. Definitions and examples

In this section we recall the basic definitions and results that will be needed in the sequel. Let $\mathcal{B} = \{A, C, G, T\}$ be the genetic alphabet where the *nucleotides* A stands for Adenine, C for Cytosine, G for Guanine and T for Thymine. Triletter words $x = N_1N_2N_3 \in \mathcal{B}^3$ 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). Recall that for such a codon $x = N_1N_2N_3 \in \mathcal{B}^3$, its *complement* is indicated by $c(x) = c(N_1)c(N_2)c(N_3)$ where $c(A) = T$, $c(T) = A$, $c(C) = G$ and $c(G) = C$, and its reversed codon by $\bar{x} = N_3N_2N_1$. A trinucleotide code $Y \subseteq \mathcal{B}^3$ is called *self-complementary* if for each codon $x \in Y$ its *anticodon* $\overline{c(x)}$ is also in Y.

As mentioned in the introduction, a special subset X of 20 codons was statistically identified in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Arquès and Michel, 1996; Michel, 2015, 2017).

* Corresponding author.
E-mail addresses: e.fimmel@hs-mannheim.de (E. Fimmel), c.michel@unistra.fr (C.J. Michel), l.struengmann@hs-mannheim.de (L. Strüngmann).

https://doi.org/10.1016/j.biosystems.2024.105263
Received 22 May 2024; Received in revised form 30 June 2024; Accepted 30 June 2024
Available online 4 July 2024

This code X is self-complementary and is able to retrieve the correct reading frame during the translational process - a property that is called *circularity* which extends the Crick's classical notion of comma-freeness. Note that in our setting, any subset of $\mathcal{B}^3$ is a code while in general a code is a set of words Y (perhaps of different lengths) over some alphabet Σ such that any concatenation of words from Y can be read in a unique way.

Definition 1 (*Theorem 2.6 and Theorem 2.11 in Fimmel et al., 2016, and Definition 2.7 in Fimmel et al., 2017*). A code $Y \subseteq \mathcal{B}^n$ for some natural number n is

1. *comma-free* if every concatenation c_1c_2 of 2 words from Y does not contain as a substring any word from Y but c_1 and c_2 themselves.
2. *circular* if for any finite concatenation $c_1 \cdots c_m$ of elements from Y, where $m \in \mathbf{N}$, there is only one partition into elements from Y when read on a circle. Any such partition is a *circular decomposition* of $c_1 \cdots c_m$.

Circular codes (also over any finite alphabet) have been studied intensively over the last 30 years. One of the most fruitful approaches to the theory of circular codes was by graph theory introduced in Fimmel et al. (2016) by the authors. Recall from graph theory (see Clark and Holton, 1991) that for an associated graph $\mathcal{G}(X)$ of an l-letter code X, a *path* between two vertices v and w from $V(X)$ is a connection $[v, v_1][v_1, v_2] \ldots [v_{k-1}, v_k][v_k, w]$ of vertices from $V(X)$ that links v to w. A *cycle* in $\mathcal{G}(X)$ is a circle in $\mathcal{G}(X)$, that is an oriented closed path in $\mathcal{G}(X)$ that visits each vertex exactly once, except for the first vertex (since it is also the last vertex). The *length* of a directed cycle in $\mathcal{G}(X)$ is the number of edges (or vertices) of the cycle. A graph is called *acyclic* if it does not contain any cycle.

Proposition 1 (*Fimmel et al., 2016*). *Let $X \subseteq \mathcal{B}^n$ for some natural number n. Then,*

(i) X is circular if and only if its associated graph is acyclic.
(ii) X is comma-free if and only if the maximal length of a path in $\mathcal{G}(X)$ is 2.

Another approach to the theory of circular codes was given by group theory (Fimmel et al., 2014). In fact, group operations turned out to be very helpful for studying circular codes. Let $S_3 = \{\alpha : \{1,2,3\} \to \{1,2,3\} \mid \alpha \text{ is bijective}\}$, where α is a permutation of positions of nucleotides in a codon x, and consider $\mathcal{A}_3 := \{\alpha_0, \alpha_1, \alpha_2\} \subseteq S_3$, the subgroup of cyclical permutations of $(S_3, \circ)$. We will say that X is a C^3-*code* if $X = \alpha_0(X)$, as well as its circular 1-shift $\alpha_1(X)$ and its circular 2-shift $\alpha_2(X)$ are circular codes (Arquès and Michel, 1996; Fimmel et al., 2014). Clearly, in any circular code Y there is at most one element from each complete conjugacy class, and none of the *periodic* codons $\{AAA, TTT, CCC, GGG\}$.

A trinucleotide circular code $Y \subseteq \mathcal{B}^3$ is considered *maximal* when its size, denoted by $|X|$, is equal to 20 (Arquès and Michel, 1996) and a dinucleotide circular code $Y \subseteq \mathcal{B}^2$ is maximal when its size is 6 (Michel and Pirillo, 2013; Michel et al., 2016).

It turned out that the circular code X found in genes

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

is a maximal self-complementary C^3-code and belongs to the class of the 216 maximal self-complementary trinucleotide C^3-codes over the genetic alphabet $\mathcal{B}$ (Arquès and Michel, 1996) whose list is given in Michel et al. (2008a).

3. Identification of dinucleotide circular codes in codons of bacterial, archaeal and eukaryotic genes

The trinucleotide circular code X (1) has been identified in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Arquès and Michel, 1996; Michel, 2015, 2017). However, no observation of a maximal self-complementary dinucleotide circular code has been identified in genes. For this purpose, after computing the codon usage of 3 large sets of bacterial, archaeal and eukaryotic genes, we have analysed their dinucleotide occurrences in the 3 codon sites: 1–2, 2–3 and 1–3.

3.1. Data

An interesting codon statistics database (CSD) has recently been developed by the Alvarez-Ponce group (Subramanian et al., 2022). It provides the codon usage for all the species with reference or representative genomes in RefSeq. 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.

3.2. Codon usage of bacterial, archaeal and eukaryotic genes

In Table 1, the codon usage 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 3.1). 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.

3.3. Dinucleotides in codon sites of bacterial, archaeal and eukaryotic genes

In Table 2, the frequency f of the 16 dinucleotides in the 3 codon sites of $34,020,997$ 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, are computed from the codon usage Table 1: $f_{1,2}$ in the codon site 1–2, $f_{2,3}$ in the codon site 2–3 and $f_{1,3}$ in the codon site 1–3.

In order to identify dinucleotide circular codes, we are applying the same approach that has identified the trinucleotide circular code

Table 1
Codon usage (frequency in %) 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 3.1).

	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

Table 2
Frequency f (%) of the 16 dinucleotides in the 3 codon sites 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), computed from the codon usage in Table 1: $f_{1,2}$ in the codon site 1–2, $f_{2,3}$ in the codon site 2–3 and $f_{1,3}$ in the codon site 1–3. The highest dinucleotide frequencies between $f_{1,2}$ and $f_{2,3}$ are in bold. The highest dinucleotide frequencies $f_{1,3}$ between a dinucleotide and its permuted dinucleotide are in italics, the periodic dinucleotides being not considered. The 2 highest dinucleotide frequencies among 144 that does not comply with the rule are in bold and italics: $f_{1,3} = 7.46\%$ with AG in eukaryota and $f_{2,3} = 6.45\%$ with TT in eukaryota.

	Bacteria			Archaea			Eukaryota		
	$f_{1,2}$	$f_{2,3}$	$f_{1,3}$	$f_{1,2}$	$f_{2,3}$	$f_{1,3}$	$f_{1,2}$	$f_{2,3}$	$f_{1,3}$
AA	**7.80**	6.59	4.65	**6.68**	5.48	4.61	**9.78**	7.23	6.45
AC	5.76	**8.05**	*8.40*	5.94	**10.79**	*8.33*	5.54	**7.60**	*7.50*
AG	2.89	**7.45**	6.04	3.39	**8.81**	6.89	5.39	**9.60**	***7.46***
AT	**7.63**	6.02	*5.00*	**7.54**	4.49	*3.72*	**7.07**	7.00	*6.37*
CA	**5.42**	3.63	2.70	**4.23**	3.68	2.45	**6.89**	6.32	4.70
CC	4.83	**9.52**	7.15	4.45	**8.08**	8.41	5.66	**6.95**	5.71
CG	5.13	**8.11**	9.81	4.53	**9.15**	7.45	3.24	**3.94**	7.40
CT	**7.55**	3.40	3.26	**7.67**	2.86	2.56	**6.93**	6.60	4.89
GA	**11.80**	2.51	*6.78*	**15.38**	3.11	*6.89*	**11.86**	3.82	7.32
GC	**10.28**	7.70	*14.04*	**9.44**	6.88	*16.44*	**7.13**	5.71	*8.30*
GG	**7.98**	4.76	10.53	**8.03**	5.18	12.15	**6.37**	4.43	8.35
GT	**7.57**	3.32	*6.29*	**8.40**	2.89	*5.77*	**6.21**	4.34	*7.60*
TA	3.09	**3.56**	2.15	3.27	**3.40**	1.73	2.91	**3.55**	2.45
TC	3.79	**9.77**	*5.45*	3.95	**13.64**	*6.20*	5.48	**7.46**	*6.21*
TG	2.29	**9.88**	3.82	2.10	**6.80**	3.46	3.31	**8.99**	3.75
TT	**6.18**	5.73	3.93	**5.00**	4.76	2.93	6.24	***6.45***	5.53

X in genes of bacteria and eukaryotes in 1996 (Arquès and Michel, 1996). Indeed, each trinucleotide was classified in the frame among 3 (reading frame and 2 the shifted frames 1 and 2) according to its highest frequency (see Tables 1(a) and 1(b) in Arquès and Michel, 1996). A similar situation occurs with the codon sites 1–2 and 2–3. Indeed, the codon site 2–3 is a shift of 1 nucleotide from the codon site 1–2. Thus, we assign the dinucleotides of highest frequency in the codon sites 1–2 or 2–3 by comparing their frequencies $f_{1,2}$ and $f_{2,3}$ (in bold in Table 2). For example in bacteria, the dinucleotide AC has $f_{1,2} = 5.76\%$ and $f_{2,3} = 8.05\%$, and thus AC is assigned to the codon site 2–3, etc. The periodic dinucleotides $\{AA, CC, GG, TT\}$ are not considered for a search of dinucleotide circular codes. Very surprisingly, two dinucleotide codes are identified:

$$D_{1,2} = \{AT, CA, CT, GA, GC, GT\}$$

in the codon site 1–2 and

$$D_{2,3} = \{AC, AG, CG, TA, TC, TG\}$$

in the codon site 2–3. Very interestingly, the dinucleotide codes $D_{1,2}$ and $D_{2,3}$ are maximal circular with 2 additional properties:

$$D_{2,3} = \alpha_1(D_{1,2})$$

and if the self-complementary dinucleotides are not considered, i.e. $\{AT, CG, GC, TA\}$, $D'_{1,2} = D_{1,2} \setminus \{AT, GC\}$ and $D'_{2,3} = D_{2,3} \setminus \{CG, TA\}$, then

$$D'_{1,2} = \overleftarrow{c(D'_{2,3})}.$$

Note that the symbol $A \setminus B$ or $A - B$ is the relative complement, i.e. the elements that belong to the set A and not to the set B.

Importantly, the same maximal circular codes $D_{1,2}$ and $D_{2,3}$ are observed in genes of bacteria, archaea as well as eukaryotes, without any exception in the three kingdoms (see Table 2). These 2 additional properties of dinucleotide circular codes $D_{1,2}$ and $D_{2,3}$ are close to the 2 properties of the trinucleotide circular code X (1) where $X_2 = \alpha_1(X_1)$ and $X_1 = \overleftarrow{c(X_2)}$ with $X_1 = \alpha_1(X)$ and $X_2 = \alpha_2(X)$. Periodic dinucleotides and periodic trinucleotides cannot belong to a circular code. Self-complementary trinucleotides do not exist. Thus, the self-complementary dinucleotides $\{AT, CG, GC, TA\}$ may also be a special code in the dinucleotide word.

It is very important to stress that the 3 following properties of the dinucleotides $D_{1,2}$ and $D_{2,3}$ are totally unexpected: (i) each have 6 dinucleotides; (ii) they are related by permutation and reverse $D_{2,3} = \alpha_1(D_{1,2}) = \overleftarrow{D_{1,2}}$; and in addition they are circular. This was already the case with the trinucleotide code X (1) where X, X_1 and X_2 each have 20 trinucleotides, they are related by permutation $X_1 = \alpha_1(X)$ and $X_2 = \alpha_2(X)$ and in addition they are circular. This remarkable property of X was studied in Michel (2020). We also show here that the 3 properties of the dinucleotides $D_{1,2}$ and $D_{2,3}$ are unforeseen. A random

genetic code is generated (see Table 3 in Appendix A) according to a uniform distribution of the 64 codon frequencies. Then, the frequencies $f_{1,2}$ and $f_{2,3}$ of the 16 dinucleotides in the codon sites 1–2 and 2–3, respectively, are computed in the same way as previously from this random codon usage (see Table 4 in Appendix A). The two random dinucleotide codes obtained (without the periodic dinucleotides) are:

$$R_{1,2} = \{AT, GT, TA, TC\}$$

in the codon site 1–2 and

$$R_{2,3} = \{AC, AG, CA, CG, CT, GA, GC, TG\}$$

in the codon site 2–3. The random dinucleotide codes $R_{1,2}$ and $R_{2,3}$ are not the same size, they are not related by permutation and they are not circular.

At this point, the self-complementary property observed with the trinucleotide circular code X (1) is missing with the dinucleotide codes. The dinucleotide code $D_{1,3}$ in the codon site 1–3 cannot be related to a frameshift as $D_{1,2}$ and $D_{2,3}$. This is why we were interested in finding the self-complementary property in $D_{1,3}$ that is related to the complementary and anti-parallel structure of DNA. A dinucleotide in one DNA strand cannot pair with its permuted dinucleotide in the complementary DNA strand. Thus, we assign the dinucleotides of highest frequency in the codon site 1–3 by comparing the frequencies $f_{1,3}$ of a dinucleotide and its permuted dinucleotide (in italics in Table 2). For example in bacteria, the dinucleotide AC has $f_{1,3} = 8.40\%$ and its permuted dinucleotide CA has $f_{1,3} = 2.70\%$, and thus AC is assigned to codon site 1–3, etc. Therefore, the dinucleotide code is identified:

$$D_{1,3} = \{AC, AT, GA, GC, GT, TC\}.$$

Very interestingly, the dinucleotide code $D_{1,3}$ is maximal circular with the additional property of self-complementary

$$D_{1,3} = \overline{c(D_{1,3})}.$$

Importantly, the same maximal circular code $D_{1,3}$ is observed in genes of bacteria, archaea as well as eukaryotes, with only one minor exception in eukaryota where the frequency $f_{1,3} = 7.32\%$ of GA is very slightly lower than $f_{1,3} = 7.46\%$ of AG (in bold and italics in Table 2).

In summary, the statistical analysis of dinucleotide occurrences in the three codon sites of three large sets of genes of bacteria, archaea and eukaryota, retrieves a maximal self-complementary dinucleotide circular code in the codon site 1–3 and two maximal dinucleotide circular codes related by permutation and partially by complementarity in the codon sites 1–2 and 2–3. As we will see in the next section, the maximal self-complementary dinucleotide circular code $D_{1,3}$ found in the codon site 1–3 is even related to the circular code X identified in genes of bacteria, archaea, eukaryotes, plasmids and viruses (Arquès and Michel, 1996; Michel, 2015, 2017).

4. Circularity and self-complementarity from cut codes

In the previous section, we have statistically identified three maximal circular codes of dinucleotides $D_{1,2}$, $D_{2,3}$ and $D_{1,3}$ that occur preferentially in the codon sites 1–2, 2–3 and 1–3, respectively. The code $D_{1,3}$ is even self-complementary and more surprisingly it is even related to the code X (1) in the following sense. Indeed in Fimmel et al. (2023), we have proved that dinucleotides can be added to the maximal self-complementary trinucleotide circular code X (1) still preserving circularity, i.e. the union is a *circular set* (see Fimmel et al., 2023, for definition and properties of circular sets). We recall this theorem:

Theorem 1 (*Theorem 7 from Fimmel et al., 2023*)**.**

The mixed set $X \cup D$ (not a code) of maximal size 26 is circular where X (1) is the maximal (20 trinucleotides) self-complementary trinucleotide circular code identified in genes (recalled for the reader's convenience)

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

and D is the maximal (6 dinucleotides) self-complementary dinucleotide circular code

$$D = \{AC, AT, GA, GC, GT, TC\}. \quad (2)$$

The circular code X codes for the following 12 amino acids (three and one letter notation)

$$\{Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr, Val\} = \{A, N, D, Q, E, G, I, L, F, T, Y, V\}. \quad (3)$$

It is readily seen that the code D in Theorem 1 matches exactly the code $D_{1,3}$ statistically found in Section 3. Hence, Theorem 1 and our statistical analysis inspire the question if the code $D_{1,3}$ or even the three dinucleotide circular codes $D_{1,2}$, $D_{2,3}$ and $D = D_{1,3}$ have any relation with the trinucleotide circular code X (1) and if some properties of X can be deduced from properties of $D_{1,2}$, $D_{2,3}$ and $D = D_{1,3}$. We therefore calculate the so-called dinucleotide cuts of X (1) in order to see their relation with $D_{1,2}$, $D_{2,3}$ and $D_{1,3}$.

We begin with the definition of the (i, j)-cuts of a code Y in the usual way using the projections on coordinates:

Definition 2. Let $Y \subseteq \mathcal{B}^3$ with $\mathcal{B} = \{A, C, G, T\}$. Then

(1) $Y^{1,2} = \pi_{1,2}(Y) = \{N_1N_2 : N_1N_2N_3 \in Y \text{ for some } N_3 \in \mathcal{B}\}$.
(2) $Y^{2,3} = \pi_{2,3}(Y) = \{N_2N_3 : N_1N_2N_3 \in Y \text{ for some } N_1 \in \mathcal{B}\}$.
(3) $Y^{1,3} = \pi_{1,3}(Y) = \{N_1N_3 : N_1N_2N_3 \in Y \text{ for some } N_2 \in \mathcal{B}\}$.

As an example and as a motivation for the sequel, we give all the relevant cuts for the trinucleotide circular code X (1) (recalled for the reader's convenience)

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

We have

$$X^{1,2} = \{AA, AC, AT, CA, CT, GA, GC, GG, GT, TA, TT\}$$

$$X^{2,3} = \{AA, AC, AG, AT, CC, GC, GT, TA, TC, TG, TT\}$$

$$X^{1,3} = \{AC, AT, CC, CG, GA, GC, GG, GT, TC\}.$$

The cut codes $X^{1,2}$, $X^{2,3}$ and $X^{1,3}$ are not circular since they contain periodic dinucleotides. However, if these periodic dinucleotides are removed, then we are very close to circularity. In fact, all the above codes contain (several) maximal dinucleotide circular codes and even more interestingly

$$D_{1,2} \subseteq X^{1,2} \setminus \{AA, CC, GG, TT\} = \{AC, AT, CA, CT, GA, GC, GT, TA\}$$

$$D_{2,3} \subseteq X^{2,3} \setminus \{AA, CC, GG, TT\} = \{AC, AG, AT, GC, GT, TA, TC, TG\}$$

$$D_{1,3} \subseteq X^{1,3} \setminus \{AA, CC, GG, TT\} = \{AC, AT, CG, GA, GC, GT, TC\}.$$

$D_{1,3}$ is almost a subset of $X^{1,3} \setminus \{AA, CC, GG, TT\}$ (except for the dinucleotide CG). Moreover, the dinucleotide $D_{1,3}$ was recently studied where it was shown that the set of 216 maximal self-complementary C^3-codes can be arranged as a *spiderweb* using a special subgroup L of the group of permutations of the genetic alphabet $\{A, C, G, T\}$ and a relation between two such codes if their intersection is of maximal size 18 (Fimmel and Strüngmann, 2023).

All these facts motivate the next section where we prove several conditions on the dinucleotide cuts of a trinucleotide code Y to be sufficient for showing its circularity.

4.1. Circularity of trinucleotide cut codes

The following useful observation was first made in Golomb et al. (1958) and used substantially in Fayazi et al. (2021). For the convenience of the reader, we include a proof.

Lemma 1. *Let $Y \subseteq \mathcal{B}^3$ be a code such that $Y^{1,2} \cap Y^{2,3} = \emptyset$. Then Y is comma-free.*

Proof. Let $N_1N_2N_3, M_1M_2M_3 \in Y$. Then by definition, we have

$N_1N_2, M_1M_2 \in Y^{1,2}$

$N_2N_3, M_2M_3 \in Y^{2,3}$.

If $N_2N_3M_1 \in Y$ then $N_2N_3 \in Y^{1,2} \cap Y^{2,3} = \emptyset$ — a contradiction. Similarly, $N_3M_1M_2 \in Y$ leads to a contradiction. Hence Y is comma-free. □

With Lemma 1, comma-freeness can be extended in less obvious cases. For example, the two dinucleotide codes $Y^{1,2} = D_{1,2}$ and $Y^{2,3} = D_{2,3}$ identified in the codon sites 1–2 and 2–3, respectively, of bacterial, archaeal and eukaryotic genes (see Section 3) can give a comma-free code according to the following corollary.

Corollary 1. *Let $Y \subseteq \mathcal{B}^3$ be a code such that $Y^{1,2} = \{AT, CA, CT, GA, GC, GT\}$ and $Y^{2,3} = \{AC, AG, CG, TA, TC, TG\}$. Then Y is comma-free.*

Proof. Obvious from Lemma 1. □

The addition of periodic dinucleotides $\{AA, CC, GG, TT\}$ to the above dinucleotide codes does not change the statement as long as the intersection of the cuts remains empty. This property is used in the following example which gives a comma-free subcode of X (1).

Example 1. Let $Z = \{AAC, ATC, CAG, CTC, CTG, GAC, GAG, GCC, GTA, GTC, TTC\}$. Then $Z \subseteq X$, $|\ Z\ | = 11$ and Z is comma-free. This follows from Corollary 1 since we have $Z^{1,2} = \{AA, AT, CA, CT, GA, GC, GT, TT\}$ and $Z^{2,3} = \{AC, AG, CC, TA, TC, TG\}$ and hence $Z^{1,2} \cap Z^{2,3} = \emptyset$.

The largest subset of the circular code X which is comma-free has size 12 and there are 20 such subsets (result not shown). The largest comma-free subset among the 216 maximal C^3 self-complementary trinucleotide circular codes has size 16 (see below Observation 1).

Observation 1. *The maximal C^3 self-complementary trinucleotide circular code $Y = \{AAC, AAG, AAT, ATC, ATT, CAC, CAG, CTC, CTG, CTT, GAC, GAG, GAT, GCC, GGC, GTA, GTC, GTG, GTT, TAC\}$ has a subset $Z = \{AAC, AAG, AAT, CAC, CAG, CTC, CTG, CTT, GAC, GAG, GAT, GCC, GTA, GTC, GTG, GTT\}$ of cardinality 16 which is comma-free. Indeed, $Z^{1,2} = \{AA, CA, CT, GA, GC, GT\}$ and $Z^{2,3} = \{AC, AG, AT, CC, TA, TC, TG, TT\}$ and hence $Z^{1,2} \cap Z^{2,3} = \emptyset$.*

Obviously, the 12,964,440 maximal trinucleotide circular codes contain 408 comma-free codes (see the growth function in Table 2a in Michel et al., 2008b).

We now want to find conditions on the cuts $Y^{1,2}$, $Y^{2,3}$ and $Y^{1,3}$ which imply circularity of the code Y. The main idea is to ensure that under these conditions a sequence of codewords over Y that can be read in two ways on the circle induces a sequence of codewords over $Y^{1,3}$ that can be read in two ways on the circle. Thus, circularity of $Y^{1,3}$ would then imply circularity of Y itself. We need a further definition.

Definition 3. Let $D, E \subseteq \mathcal{B}^2$ be two dinucleotide codes. Then

$$D \odot E = \{N_2M_1 \mid N_1N_2 \in D \text{ and } M_1M_2 \in E \text{ for some } N_1, M_2 \in \mathcal{B}\}.$$

Note that the symbol $D \odot E$ gives the dinucleotide obtained by concatenating the 2 dinucleotides D and E and by removing the first and last nucleotides of the obtained tetranucleotide $D{\cdot}E$, i.e. the middle dinucleotide of the tetranucleotide $D \cdot E$.

Here is the first handy criterion for circularity.

Theorem 2. *Let $Y \subseteq \mathcal{B}^3$ be a code and assume that*

(1) $Y^{1,3}$ is circular.
(2) $Y^{1,2} \cap (Y^{2,3} \odot (Y^{1,2} \cap Y^{2,3})) \subseteq Y^{1,3}$.

Then Y is a circular code.

Proof. See Appendix B. □

We use a simple Example 2 to demonstrate how Theorem 2 is applied.

Example 2. Consider $Y = \{ACG, TAC, CAG, CAA\}$. We have

$Y^{1,2} = \{AC, TA, CA\}$,

$Y^{2,3} = \{CG, AC, AG, AA\}$

$Y^{1,3} = \{AG, TC, CG, CA\}$ which is circular

and hence

$Y^{1,2} \cap Y^{2,3} = \{AC\}$,

$Y^{1,2} \cap (Y^{2,3} \odot (Y^{1,2} \cap Y^{2,3})) = \{CA\} \subseteq Y^{1,3}$.

It follows that Y is circular according to Theorem 2.

The circular code X (1) does not verify Theorem 2. Indeed, $X^{1,3} = \{AC, AT, CC, CG, GA, GC, GG, GT, TC\}$ is not circular. Then, we have searched the larger subsets of X verifying Theorem 2. No subset of X verifying Theorem 2 can be identified in the 20 subsets of size 19, the 190 subsets of size 18, the 1,140 subsets of size 17, the 4,845 subsets of size 16, the 15,504 subsets of size 15, the 38,760 subsets of size 14 and the 77,520 subsets of size 13. Two subsets V and W of X verify Theorem 2 in the 125,970 subsets of size 12 (see Appendix C for Examples 7 and 8).

We now apply Theorem 2 in the 216 maximal C^3 self-complementary trinucleotide circular codes in order to identify a subset with the largest size being 17 (see below Example 3).

Example 3. The maximal C^3 self-complementary trinucleotide circular code $Y = \{AAC, AAG, AAT, ACC, ACG, ACT, AGG, AGT, ATG, ATT, CAG, CAT, CCG, CCT, CGG, CGT, CTG, CTT, GGT, GTT\}$ has a subset $Z = \{AAG, AAT, ACG, ACT, AGG, AGT, ATG, ATT, CAG, CAT, CCG, CCT, CGG, CGT, CTG, CTT, GTT\}$, $Z \subseteq Y$, $|\ Z\ | = 17$ which is circular (of maximal path length 3) and satisfies the condition of Theorem 2.

Indeed, we have

$Z^{1,2} = \{AA, AC, AG, AT, CA, CC, CG, CT, GT\}$

$Z^{2,3} = \{AG, AT, CG, CT, GG, GT, TG, TT\}$

$Z^{1,3} = \{AG, AT, CG, CT, GT\}$ which is circular

and hence

$Z^{1,2} \cap Z^{2,3} = \{AG, AT, CG, CT, GT\}$

$Z^{2,3} \odot (Z^{1,2} \cap Z^{2,3}) = \{GA, GC, GG, TA, TC, TG\}$

$Z^{1,2} \cap (Z^{2,3} \odot (Z^{1,2} \cap Z^{2,3})) = \emptyset \subseteq Z^{1,3}$.

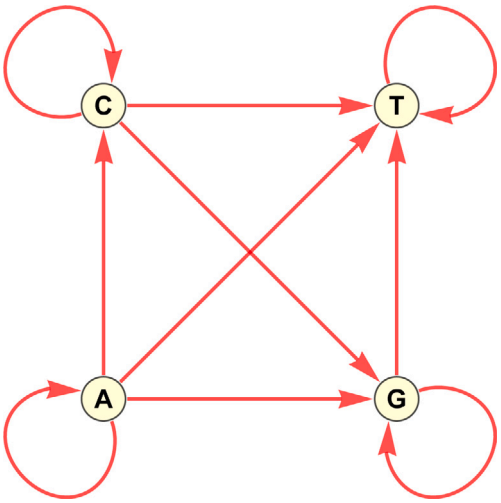


Fig. 1. A generalized dinucleotide circular code $D = \{AC, AG, AT, CG, CT, GT, AA, CC, GG, TT\}$.

Finally, we apply Theorem 2 in the 12,964,440 maximal trinucleotide circular codes in order to identify a subset with size greater than 17. Surprisingly, no set of size 18 or more is identified. Similarly, no set of size 18 or more is identified in the 960,608 circular codes of size 18 that cannot be embedded in a larger circular code (also called *maximal* in Michel et al., 2016) as well as in the 3,617,664 *maximal* circular codes of size 19 (see Table 3 in Michel et al., 2016). In summary, the largest sets have size 17 and already exist in the 216 maximal C^3 self-complementary trinucleotide circular codes.

As noted above, the addition of periodic dinucleotides to dinucleotide circular codes is not critical in relevant situations. Let us describe the situation formally.

Definition 4. Let $E \subseteq \mathcal{B}^2$ be a code. Then E is called generalized circular if the only cycles in the associated graph $\mathcal{G}(E)$ are loops.

From the definition of the graph $\mathcal{G}(Y)$ associated with a code Y and the fact that the associated graph of a circular code is acyclic (Fimmel et al., 2016), a generalized dinucleotide circular code $E' \subseteq \mathcal{B}^2$ can be represented as the union $E' = E \cup Y$ of a dinucleotide circular code $E \subseteq \mathcal{B}^2$ and a subset $Y \subseteq \{AA, CC, GG, TT\}$. Fig. 1 gives an example of a generalized circular code.

As a consequence, we have an easy but important observation.

Lemma 2. *Let $D \subseteq \mathcal{B}^2$ be a generalized circular code. Then a sequence of dinucleotides $d_1 d_2 \cdots d_n$ from D has a second decomposition over D if and only if $d_1 = d_2 = \cdots = d_n = NN$ for some $N \in \mathcal{B}$.*

Inspired by the above reasoning, we now state our second main criterion for circularity. Recall that a code $Y \subseteq \mathcal{B}^3$ is called *1-circular* (Fimmel et al., 2020) if for any codon $N_1N_2N_3 \in Y$ then the two cyclically shifted codons $N_2N_3N_1$ and $N_3N_1N_2$ are not in Y.

Theorem 3. *Let $Y \subseteq \mathcal{B}^3$ be a code and assume that*

(1) Y is 1-circular.
(2) $Y^{1,3} \cup \{AA, CC, GG, TT\}$ is generalized circular.
(3) $Y^{1,2} \cap (Y^{2,3} \odot (Y^{1,2} \cap Y^{2,3})) \subseteq Y^{1,3} \cup \{AA, CC, GG, TT\}$.

Then Y is a circular code.

Proof. See Appendix D. □

The main difference between the conditions of Theorem 3 and Theorem 2 is the condition (1). The following Example 4 shows that it must not be omitted in this case.

Example 4. Consider $Y = \{AAA\}$. We have

$$Y^{1,2} = Y^{2,3} = Y^{1,3} = \{AA\}.$$

The conditions (2) and (3) of Theorem 3 are obviously fulfilled, while condition (1) is not fulfilled and the code Y is not circular.

We have two subsets of X (1) of size 16 that satisfy the conditions of Theorem 3 (see Appendix E for Examples 9 and 10).

In fact, any of the 216 maximal self-complementary C^3-codes contains a subset of size at least 14 that satisfies the conditions of Theorem 3 (by computer calculations). Some of these codes even have such a subset of size 18 (see Appendix E for Example 11).

Some maximal trinucleotide circular codes among 12,964,440 verify Theorem 3.

Observation 2. *Let $Y = \{AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, AGT, ATC, ATG, ATT, CCG, CCT, CGG, CGT, CTG, CTT, GGT, GTT\}$. Then Y is a maximal ($|\ Y\ | = 20$) trinucleotide circular code verifying Theorem 3.*

Indeed, we have

$$Y^{1,2} = \{AA, AC, AG, AT, CC, CG, CT, GG, GT\}$$

$$Y^{2,3} = \{AC, AG, AT, CC, CG, CT, GC, GG, GT, TC, TG, TT\}$$

$$Y^{1,3} = \{AC, AG, AT, CG, CT, GT\}$$

and hence

$$Y^{1,2} \cap Y^{2,3} = \{AC, AG, AT, CC, CG, CT, GG, GT\}$$

$$Y^{23} \odot (Y^{1,2} \cap Y^{2,3}) = \{CA, CC, CG, GA, GC, GG, TA, TC, TG\}$$

$$Y^{1,2} \cap (Y^{23} \odot (Y^{1,2} \cap Y^{2,3})) = \{CC, CG, GG\}.$$

The latter is contained in $Y^{1,3} \cup \{CC, GG\}$ which is a generalized circular code.

In the next and final section, we investigate if self-complementarity can also be deduced from conditions on the cut codes.

4.2. Self-complementarity of trinucleotide cut codes

This section deals with the relationship between the self-complementarity of a trinucleotide code and its dinucleotide cut codes. First, we formulate the (almost obvious) necessary condition of self-complementarity, which was already observed in Fimmel et al. (2018) but without explicitly mentioning the dinucleotide cut codes.

Lemma 3. *Let $Y \subseteq \mathcal{B}^3$ be self-complementary. Then the following statements are true:*

(1) $\overleftarrow{c(Y^{1,2})} = Y^{2,3}$.
(2) $c(Y^{1,3}) = Y^{1,3}$, i.e. $Y^{1,3}$ is self-complementary.

Proof. Let $N_1N_2N_3 \in Y$. Because of the self-complementarity of Y, we have

$$c(N_3)c(N_2)c(N_1) \in Y.$$

It means that

$$N_1N_2,\ c(N_3)c(N_2) \in Y^{1,2}, \quad N_2N_3,\ c(N_2)c(N_1) \in Y^{2,3} \quad \text{and}$$

$$N_1N_3,\ c(N_3)c(N_1) \in Y^{1,3}.$$

Our statement follows immediately from the following observation

$$\overleftarrow{c(N_1N_2)} = c(N_2)c(N_1) \in Y^{2,3}, \quad \overleftarrow{c(N_2N_3)} = c(N_3)c(N_2) \in Y^{1,2} \quad \text{and}$$

$$\overleftarrow{c(N_1N_3)} = c(N_3)c(N_1) \in Y^{1,3}. \quad \square$$

The conditions from Lemma 3 are necessary but not sufficient as the following Example 5 shows.

Example 5. $Y = \{ACG, AGT, CCT, CGG\}$ is not self-complementary (e.g. $\overleftarrow{c(ACG)} = CGT \notin Y$) but for

$Y^{1,2} = \{AC, AG, CC, CG\}$, $Y^{2,3} = \{CG, CT, GG, GT\}$,

$Y^{1,3} = \{AG, AT, CG, CT\}$

the following relations from Lemma 3 take place:

$\overleftarrow{c(Y^{1,2})} = Y^{2,3}$ and $\overleftarrow{c(Y^{1,3})} = Y^{1,3}$.

Let us now formulate a sufficient condition for the self-complementarity of a trinucleotide code from its dinucleotide cuts.

Lemma 4. *Let $Y \subseteq \mathcal{B}^3$ such that*

(1) $\overleftarrow{c(Y^{1,2})} = Y^{2,3}$.
(2) for all $N_1N_2 \in Y^{1,2}, N_2N_3 \in Y^{2,3}$ it follows $N_1N_2N_3 \in Y$.

Then Y is self-complementary.

Proof. Let $N_1N_2N_3 \in Y$. We have to prove that $c(N_3)c(N_2)c(N_1) \in Y$. We have

$N_1N_2 \in Y^{1,2}$ and $N_2N_3 \in Y^{2,3}$.

Using Condition (1) we obtain

$c(N_3)c(N_2) \in Y^{1,2}$ and $c(N_2)c(N_1) \in Y^{2,3}$.

Then it follows from Condition (2) that

$c(N_3)c(N_2)c(N_1) \in Y$. □

The conditions from Lemma 4 are sufficient but not necessary as the following Example 6 shows.

Example 6. $Y = \{ACT, AGT, CGC, GCG\}$ is self-complementary, and

$Y^{1,2} = \{AC, AG, CG, GC\}$ and $Y^{2,3} = \{GC, GT, CG, CT\}$

are complementary to each other but Condition (2) of Lemma 4 is not fulfilled since

$AC \in Y^{1,2}$, $CG \in Y^{2,3}$ but $ACG \notin Y$.

The circular code X (1) does not verify Lemma 4. As in the previous section, we have searched the larger subsets of X verifying Lemma 4. Two subsets V and W of X (1) verify Lemma 4 in the 125,970 subsets of size 12 (see Appendix F).

Some maximal C^3 self-complementary trinucleotide circular codes among 216 verify Lemma 4.

Observation 3. *Let $Y = \{AAC, AAG, AAT, ACC, ACT, AGC, AGT, ATC, ATT, CTC, CTT, GAC, GAG, GAT, GCC, GCT, GGC, GGT, GTC, GTT\}$ with $|Y| = 20$ and Y is self-complementary.*

This follows by the above result since $Y^{1,2} = \{AA, AC, AG, AT, CT, GA, GC, GG, GT\}$ and $Y^{2,3} = \{AC, AG, AT, CC, CT, GC, GT, TC, TT\}$ and hence the conditions of Lemma 4 are satisfied.

5. Emergence of trinucleotide codes from dinucleotide cut codes

An exciting question is why the amino acids are coded by triplets of nucleotides, i.e. by codons? Taylor and Coates (1989) proposed the so-called $2-1-3$ hypothesis in the 1980s, which was later developed by Massey (2006, 2016). It states that the evolution of the genetic code began with the 2nd (middle) base in today's codons, then the 1st base was added and finally the 3rd base. This evolutionary process has made it possible to encode more amino acids on the one hand, and a minimization of translation errors on the other hand. In the following sections, we will present which trinucleotide codes could have been generated, taking into account the statistical results of the article.

5.1. Trinucleotide circular codes from a concatenation of a dinucleotide circular code and a nucleotide

The circular code $D_{1,2} = \{AT, CA, CT, GA, GC, GT\}$ that occurs preferentially in the codon site 1–2 (N_1N_2) of bacteria, archaea and eukaryota, is concatenated with the 4 nucleotides N leading to the trinucleotide code

$$Y = \{ATA, ATC, ATG, ATT, CAA, CAC, CAG, CAT, CTA, CTC, CTG, CTT, GAA, GAC, GAG, GAT, GCA, GCC, GCG, GCT, GTA, GTC, GTG, GTT\}$$

of size 24 with the codons of the form N_1N_2N. It codes 9 amino acids $\{Ala, Asp, Gln, Glu, His, Ile, Leu, Met, Val\}$. The code Y is obviously not circular as there are 4 pairs of permuted codons $\{ATC, CAT\}$, $\{ATG, GAT\}$, $\{CAG, GCA\}$ and $\{CTG, GCT\}$. Interestingly, 13 subsets of Y of size 20 among 10,626 are circular codes, and one is even comma-free (of maximal path length 2, thus comma-free; Definition 1)

$$\{ATA, ATC, ATG, ATT, CAA, CAC, CAG, CTA, CTC, CTG, CTT, GAA, GAC, GAG, GCC, GCG, GTA, GTC, GTG, GTT\}.$$

5.2. Trinucleotide circular codes from a concatenation of a nucleotide and a dinucleotide circular code

The 4 nucleotides N are concatenated with the circular code $D_{2,3} = \{AC, AG, CG, TA, TC, TG\}$ that occurs preferentially in the codon site 2–3 (N_2N_3) of bacteria, archaea and eukaryota, leading to the trinucleotide code

$$Z = \{AAC, AAG, ACG, ATA, ATC, ATG, CAC, CAG, CCG, CTA, CTC, CTG, GAC, GAG, GCG, GTA, GTC, GTG, TAC, TAG, TCG, TTA, TTC, TTG\}$$

of size 24 with the codons of the form NN_2N_3. It codes 16 amino acids

$$\{Ala, Asn, Asp, Gln, Glu, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Tyr, Val\}.$$

The code Z is obviously not circular as there are 4 pairs of permuted codons $\{ACG, GAC\}$, $\{CTA, TAC\}$, $\{GTA, TAG\}$ and $\{GTC, TCG\}$. Interestingly, 13 subsets of Z of size 20 among 10,626 are circular codes, and one is also comma-free

$$\{AAC, AAG, ATA, ATC, ATG, CAC, CAG, CCG, CTA, CTC, CTG, GAC, GAG, GCG, GTA, GTC, GTG, TTA, TTC, TTG\}.$$

There are 14 codons common to Y and Z, and the code $Y \cap Z$ is even comma-free

$$Y \cap Z = \{ATA, ATC, ATG, CAC, CAG, CTA, CTC, CTG, GAC, GAG, GCG, GTA, GTC, GTG\}.$$

5.3. Trinucleotide circular codes from a concatenation of dinucleotide circular codes

We now turn to the code $D_{1,2}$ and study the code Y of the 1st trinucleotides $N_1N_2M_1$ obtained by a concatenation of 2 dinucleotides $N_1N_2 \cdot M_1M_2$ in the circular code

$$D_{1,2} = \{AT, CA, CT, GA, GC, GT\}$$

that occurs preferentially in the codon site 1–2 of genes. Thus, $D_{1,2}^2$ is the code of 36 tetranucleotides, i.e. $D_{1,2}^2 = \{ATAT, ATCA, ATCT, \ldots, GTGA, GTGC, GTGT\}$. Note that the 1st trinucleotides $N_1N_2M_1$ has 2 nucleotides correctly positioned while the 2nd trinucleotides $N_2M_1M_2$ has no nucleotide correctly positioned.

The trinucleotide code

$$Y = \{ATA, ATC, ATG, CAA, CAC, CAG, CTA, CTC, CTG, GAA, GAC, GAG, GCA, GCC, GCG, GTA, GTC, GTG\}$$

of size 18 obtained from $D_{1,2}^2$ is not circular as there are 2 permuted trinucleotides CAG and GCA. It codes 9 amino acids $\{Ala, Asp, Gln, Glu, His, Ile, Leu, Met, Val\}$. Very interestingly, if the trinucleotide GCA is

removed from Y, then $Y \setminus \{GCA\}$ is a circular code, and even a comma-free code. On the other hand, if the trinucleotide CAG is removed from Y, then $Y \setminus \{CAG\}$ is a circular code (of maximal path length equal to 4).

We apply the same analysis with the code Z of the 2nd trinucleotides $N_3M_2M_3$ (2 nucleotides correctly positioned) obtained by a concatenation of 2 dinucleotides $N_2N_3 \cdot M_2M_3$ in the circular code

$$D_{2,3} = \{AC, AG, CG, TA, TC, TG\}$$

that occurs preferentially in the codon site 2–3 of genes. Note that the 1st trinucleotides $N_2N_3M_2$ has no nucleotide correctly positioned.

The trinucleotide code

$$Z = \{AAC, AAG, ACG, ATA, ATC, ATG, CAC, CAG, CCG, CTA, CTC, CTG, GAC, GAG, GCG, GTA, GTC, GTG\}$$

of size 18 obtained from $D_{2,3}^2$ is not circular as there are 2 permuted trinucleotides ACG and GAC. It codes 13 amino acids $\{Ala, Asn, Asp, Gln, Glu, His, Ile, Leu, Lys, Met, Pro, Thr, Val\}$. Very interestingly, if the trinucleotide ACG is removed from Z, then $Z \setminus \{ACG\}$ is a circular code, and even a comma-free code. On the other hand, if the trinucleotide GAC is removed from Z, then $Z \setminus \{GAC\}$ is a circular code (of maximal path length equal to 4).

6. Conclusion

In the present work, motivated by evolutionary hypotheses of the genetic code, a statistical evaluation of the occurrence of dinucleotides in the codon usage of large sets of bacteria, archaea and eukaryota genes, was carried out for the first time. In Section 3, the statistical data are evaluated using the same method that was used in 1996 to identify the universal circular code X (1) in genes, i.e. by comparing the frequencies in different frames (Arquès and Michel, 1996). As a result, two maximal dinucleotide circular codes $D_{1,2}$ and $D_{2,3}$ are identified in the sites 1–2 and 2–3, respectively, of today's codons. Interestingly, these two dinucleotide circular codes are related by permutation as the shifted trinucleotide circular codes X_1 and X_2 of the universal trinucleotide circular code X. The properties of these two dinucleotides codes, i.e. maximal (size 6), circular and permutation, are totally unexpected. Indeed, by generating a random genetic code and by applying the same statistical approach, the random dinucleotide codes in the codon sites 1–2 and 2–3 are not the same size, they are not related by permutation and they are not circular. In addition, a maximal circular and self-complementary dinucleotide code $D_{1,3}$ is identified by assigning the dinucleotides with the highest frequency in the codon site 1–3 by comparing the frequencies of a dinucleotide and its permuted dinucleotide. The reason in this case for using a different procedure is due to the fact that the codon site 1–3 cannot be assigned to a frameshift. The same three dinucleotide circular codes $D_{1,2}$, $D_{2,3}$ and $D_{1,3}$ are identified in the genes of bacteria, archaea and eukaryota, i.e. they appear to be universal in genes (regardless of the species).

The obtained statistical results raise several problems in coding theory (Section 4) and several questions in the evolution of the genetic code (Section 5).

In Section 4, we investigate into whether properties of trinucleotide codes such as circularity and self-complementarity can be determined using their dinucleotide cut codes, i.e. a specific position in each trinucleotide of a trinucleotide code is "cut out". Some sufficient and necessary conditions for the circularity or self-complementarity of a trinucleotide code are proven. The mathematical theory is supported by numerous examples.

In Section 5, some evolutionary scenarios of codes are presented making the transition between the world of dinucleotides and the world of trinucleotides, in particular how and which trinucleotide codes could be created from the universal maximal dinucleotide codes that are newly identified.

Table 3
Codon usage (frequency in %) of a random genetic code generated according to a uniform distribution of the 64 codon frequencies.

AAA	0.71	CAA	2.43	GAA	1.51	TAA	2.58
AAC	1.75	CAC	1.06	GAC	2.40	TAC	2.79
AAG	2.46	CAG	0.10	GAG	2.38	TAG	2.45
AAT	1.53	CAT	1.36	GAT	0.20	TAT	1.52
ACA	1.97	CCA	0.13	GCA	2.47	TCA	1.59
ACC	2.40	CCC	0.72	GCC	0.72	TCC	0.83
ACG	1.23	CCG	1.65	GCG	0.10	TCG	2.73
ACT	0.55	CCT	2.47	GCT	0.87	TCT	0.60
AGA	2.87	CGA	0.76	GGA	1.87	TGA	1.05
AGC	1.94	CGC	0.35	GGC	1.92	TGC	2.19
AGG	0.59	CGG	1.08	GGG	2.77	TGG	1.98
AGT	0.27	CGT	2.04	GGT	0.22	TGT	2.33
ATA	2.42	CTA	0.49	GTA	2.77	TTA	1.06
ATC	2.85	CTC	1.37	GTC	1.02	TTC	0.32
ATG	2.34	CTG	0.25	GTG	2.36	TTG	2.71
ATT	2.06	CTT	1.63	GTT	2.52	TTT	1.32

CRediT authorship contribution statement

Elena Fimmel: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Christian J. Michel:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Lutz Strüngmann:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The first and third authors would like to acknowledge the contribution of the COST Action CA21169, supported by COST (European Cooperation in Science and Technology).

Appendix A. Random genetic code

See Tables 3 and 4.

Appendix B. Proof of Theorem 2

Proof. Assume that Y is not circular. Then, there is a sequence of elements

$$(+) \qquad N_1^1N_2^1N_3^1 \cdot N_1^2N_2^2N_3^2 \cdots N_1^sN_2^sN_3^s \in Y^s$$

that has a second decomposition over Y on the circle. Assume first that this second decomposition is given by a shift of 2 positions. Hence

$$(++) \qquad N_3^1N_1^2N_2^2 \cdot N_3^2N_1^3N_2^3 \cdots N_3^sN_1^1N_2^1 \in Y^s.$$

We now conclude several properties with the dinucleotides. From (+), we deduce that $N_1^1N_3^1, N_1^2N_3^2, \cdots, N_1^sN_3^s \in Y^{1,3}$ and hence

$$(+++) \qquad N_1^1N_3^1 \cdot N_1^2N_3^2 \cdots N_1^sN_3^s$$

is a sequence of dinucleotides from $Y^{1,3}$. The idea now is to show that this sequence has a second decomposition on the circle which would contradict the circularity of $Y^{1,3}$. We claim that $N_3^iN_1^{i+1} \in Y^{1,3}$ and $N_3^sN_1^1 \in Y^{1,3}$ for all $i < s$.

Table 4
Frequency f (%) of the 16 dinucleotides in the 2 codon sites of a random genetic code, computed from the codon usage in Table 3: $f_{1,2}$ in the codon site 1–2 and $f_{2,3}$ in the codon site 2–3. The highest dinucleotide frequencies between $f_{1,2}$ and $f_{2,3}$ are in bold.

	$f_{1,2}$	$f_{2,3}$
AA	6.45	**7.23**
AC	6.15	**8.01**
AG	5.67	**7.39**
AT	**9.67**	4.62
CA	4.96	**6.17**
CC	**4.97**	4.67
CG	4.23	**5.70**
CT	3.73	**4.48**
GA	6.49	**6.55**
GC	4.15	**6.41**
GG	**6.79**	6.42
GT	**8.67**	4.87
TA	**9.35**	6.73
TC	**5.75**	5.56
TG	7.56	**7.67**
TT	5.41	**7.54**

(1) By (++), we have $N_3^i N_1^{i+1} \in Y^{1,2}$.
(2) By (+), we have $N_2^i N_3^i \in Y^{2,3}$ and $N_1^{i+1} N_2^{i+1} \in Y^{1,2}$. Moreover, by (++) we have $N_1^{i+1} N_2^{i+1} \in Y^{2,3}$ and thus $N_3^i N_1^{i+1} \in Y^{2,3} \odot (Y^{1,2} \cap Y^{2,3})$.

Together, we conclude that

$$N_3^i N_1^{i+1} \in Y^{1,2} \cap (Y^{2,3} \odot (Y^{1,2} \cap Y^{2,3})) \subseteq Y^{1,3}$$

by hypothesis for all $i < s$.

Similarly, it follows that $N_3^s N_1^1 \in Y^{1,3}$ for all $i < s$ which contradicts circularity of $Y^{1,3}$.

Assume now that the second decomposition of the sequence from (+) is given by a shift of 1 position. Hence

$$(+++) \qquad N_2^1 N_3^1 N_1^2 \cdot N_2^2 N_3^2 N_1^3 \cdots N_2^s N_3^s N_1^1 \in Y^s.$$

However, this means that the sequence from (+ + ++) has a second decomposition on the circle that is given by a shift of 2 positions since a shift 2 of (+ + ++) gives the original sequence from (+) — a contradiction. □

Appendix C. Examples verifying Theorem 2

Example 7. The code $V = \{AAC, AAT, ACC, ATC, ATT, GAC, GAT, GCC, GGC, GGT, GTC, GTT\}$, $V \subseteq X$, $|V| = 12$ is circular (of maximal path length 2, thus comma-free; Definition 1).

We have

$$V^{1,2} = \{AA, AC, AT, GA, GC, GG, GT\}$$

$$V^{2,3} = \{AC, AT, CC, GC, GT, TC, TT\}$$

$$V^{1,3} = \{AC, AT, GC, GT\} \text{ which is circular}$$

and hence

$$V^{1,2} \cap V^{2,3} = \{AC, AT, GC, GT\}$$

$$V^{2,3} \odot (V^{1,2} \cap V^{2,3}) = \{CA, CG, TA, TG\}$$

$$V^{1,2} \cap (V^{2,3} \odot (V^{1,2} \cap V^{2,3})) = \emptyset \subseteq V^{1,3}.$$

Example 8. The code $W = \{ACC, ATC, ATT, GAA, GAC, GAT, GCC, GGC, GGT, GTA, GTC, GTT\}$, $W \subseteq X$, $|W| = 12$ is circular (of maximal path length 3).

We have

$$W^{1,2} = \{AC, AT, GA, GC, GG, GT\}$$

$$W^{2,3} = \{AA, AC, AT, CC, GC, GT, TA, TC, TT\}$$

$$W^{1,3} = \{AC, AT, GA, GC, GT\} \text{ which is circular}$$

and hence

$$W^{1,2} \cap W^{2,3} = \{AC, AT, GC, GT\}$$

$$W^{2,3} \odot (W^{1,2} \cap W^{2,3}) = \{AA, AG, CA, CG, TA, TG\}$$

$$W^{1,2} \cap (W^{2,3} \odot (W^{1,2} \cap W^{2,3})) = \emptyset \subseteq W^{1,3}.$$

Appendix D. Proof of Theorem 3

Proof. We argue as in the proof of Theorem 2. Assume that Y is not circular and there is a sequence of elements

$$(+) \qquad N_1^1 N_2^1 N_3^1 \cdot N_1^2 N_2^2 N_3^2 \cdots N_1^s N_2^s N_3^s \in Y^s$$

that has a second decomposition over Y on the circle. Without loss of generality, it comes from a shift of 2 positions. As in the proof of Theorem 3, the sequence of dinucleotides over $Y^{1,3} \cup \{AA, CC, GG, TT\}$

$$(+++) \qquad N_1^1 N_3^1 \cdot N_1^2 N_3^2 \cdots N_1^s N_3^s$$

has a second decomposition over $Y^{1,3} \cup \{AA, CC, GG, TT\}$. However, $Y^{1,3}$ is generalized circular, hence the sequence (+ + +) must be of the form $NNNNNN\cdots$ for some $N \in \mathcal{B}$. This implies that the original sequence (+) is of the form

$$NM_1NNM_2N \cdots NM_sN$$

for some $M_i \in \mathcal{B}$ $(1 \le i \le s)$. However, a second decomposition of such a sequence would imply that a shifted codon M_iNN or NNM_i is in Y for some $NM_iN \in Y$, in contradiction with the 1-circularity of Y. This finishes the proof. □

Appendix E. Examples verifying Theorem 3

Example 9. Let $V = \{AAC, AAT, ACC, ATC, ATT, CTC, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT\}$. Then $V \subseteq X$, $|V| = 16$ and V is circular.

We have

$$V^{1,2} = \{AA, AC, AT, CT, GA, GC, GG, GT\}$$

$$V^{2,3} = \{AA, AC, AG, AT, CC, GC, GT, TA, TC, TT\}$$

$$V^{1,3} = \{AC, AT, CC, GA, GC, GG, GT\}$$

and hence

$$V^{1,2} \cap V^{2,3} = \{AA, AC, AT, GC, GT\}$$

$$V^{23} \odot (V^{1,2} \cap V^{2,3}) = \{AA, AG, CA, CG, GA, GG, TA, TG\}$$

$$V^{1,2} \cap (V^{23} \odot (V^{1,2} \cap V^{2,3})) = \{AA, GA, GG\}.$$

The latter is contained in $V^{1,3} \cup \{AA\}$ which is a generalized circular code. Theorem 3 now yields the result.

Example 10. Let $W = \{AAC, AAT, ACC, ATC, ATT, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TTC\}$. Then $W \subseteq X$, $|W| = 16$ and W is circular according to Theorem 3 (proof similar).

Example 11. Let $Y = \{AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, AGT, ATC, ATT, CCT, CGT, CTT, GAT, GCC, GCT, GGC, GGT, GTT\}$. Then Y is a maximal self-complementary C^3-code that has a subcode

$Z = \{AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, AGT, ATC, ATT, CCT, CTT, GCC, GCT, GGC, GGT, GTT\}$ of size 18 satisfying the conditions of Theorem 3.

We have

$$Z^{1,2} = \{AA, AC, AG, AT, CC, CT, GC, GG, GT\}$$

$$Z^{2,3} = \{AC, AG, AT, CC, CG, CT, GC, GG, GT, TC, TT\}$$

$$Z^{1,3} = \{AC, AG, AT, CT, GC, GT\}$$

and hence

$$Z^{1,2} \cap Z^{2,3} = \{AC, AG, AT, CC, CT, GC, GG, GT\}$$

$$Z^{23} \odot (Z^{1,2} \cap Z^{2,3}) = \{CA, CC, CG, GA, GC, GG, TA, TC, TG\}$$

$$Z^{1,2} \cap (Z^{23} \odot (Z^{1,2} \cap Z^{2,3})) = \{CC, GC, GG\}.$$

The latter is contained in $Z^{1,3} \cup \{CC, GG\}$ which is a generalized circular code.

We can now extend the above Example 8 from Appendix C as follows if we add the codon GAG to the code W:

Example 12. Let $Z = \{ACC, ATC, ATT, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT\}$. Then $Z \subseteq X$, $| Z | = 13$ and Z is circular (but not comma-free).

We have

$$Z^{1,2} = \{AC, AT, GA, GC, GG, GT\}$$

$$Z^{2,3} = \{AA, AC, AG, AT, CC, GC, GT, TA, TC, TT\}$$

$$Z^{1,3} = \{AC, AT, GA, GC, GG, GT\}$$ which is generalized circular

and hence

$$Z^{1,2} \cap Z^{2,3} = \{AC, AT, GC, GT\}$$

$$Z^{2,3} \odot (Z^{1,2} \cap Z^{2,3}) = \{AA, CA, GA, GG, TA, AG, CG, TG\}$$

$$Z^{1,2} \cap (Z^{2,3} \odot (Z^{1,2} \cap Z^{2,3})) = \{GA, GG\} \subseteq Z^{1,3}.$$

As in the proof of Theorem 2, a sequence over Z that has two decompositions induces that one sequence over $Z^{1,3}$ has also two decompositions. However, $Z^{1,3}$ is generalized circular, so this sequence must be of the form $GGGGGGGGG\cdots$ and hence the original sequence must be $GAG{\cdot}GAG\cdots GAG$, that does not have a second decomposition over Z.

Appendix F. Examples verifying Lemma 4

Example 13. Let $V = \{AAC, AAT, ACC, ATC, ATT, GAC, GAT, GCC, GGC, GGT, GTC, GTT\}$. Then $V \subseteq X$, $| V | = 12$ and V is self-complementary.

This follows by the above result since $V^{1,2} = \{AA, AC, AT, GA, GC, GG, GT\}$ and $V^{2,3} = \{AC, AT, CC, GC, GT, TC, TT\}$ and hence the conditions of Lemma 4 are satisfied.

Example 14. Let $W = \{ACC, ATC, CTC, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTC, TTC\}$. Then $W \subseteq X$, $| W | = 12$ and W is self-complementary.

This follows by the above result since $W^{1,2} = \{AC, AT, CT, GA, GC, GG, GT, TT\}$ and $W^{2,3} = \{AA, AC, AG, AT, CC, GC, GT, TC\}$ and hence the conditions of Lemma 4 are satisfied.

References

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

Clark, J., Holton, D.A., 1991. A First Look At Graph Theory. World Scientific, New Jersey.

Dila, G., Ripp, R., Mayer, C., Poch, O., Michel, C.J., Thompson, J.D., 2019. Circular code motifs in the ribosome: a missing link in the evolution of translation? RNA 25, 1714–1730.

Dragovich, B., Dragovich, A., 2010. p-Adic modelling of the genome and the genetic code. Comput. J. 53 (4), 432–442.

Fayazi, F., Fimmel, E., Strüngmann, L., 2021. Equivalence classes of circular codes induced by permutation groups. Theory Biosci. 140 (1), 107–121.

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., Pirot, F., Sereni, J.-S., Starman, M., Strüngmann, L., 2020. The relation between k-circularity and circularity of codes. Bull. Math. Biol. 82, 105, 1–34.

Fimmel, E., Michel, C.J., Starman, M., Strüngmann, L., 2018. Self-complementary circular codes in coding theory. Theory Biosci. 137 (1), 51–65.

Fimmel, E., Michel, C.J., Strüngmann, L., 2016. n-Nucleotide circular codes in graph theory. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences A 374, 20150058, 1–19.

Fimmel, E., Michel, C.J., Strüngmann, L., 2017. Strong comma-free codes in genetic information. Bull. Math. Biol. 79 (8), 1796–1819.

Fimmel, E., Michel, C.J., Strüngmann, L., 2023. Circular mixed sets. Biosystems 229, 104906, 1–11.

Fimmel, E., Strüngmann, L., 2023. The spiderweb of error-detecting codes in the genetic information. BioSystems 233, 105009, 1–14.

Golomb, S.W., Delbruck, M., Welch, L.R., 1958. Construction and properties of comma-free codes. Biol. Meddelelser Kongelige Danske Videnskabernes Selskab 23, 1–34.

Koonin, E.V., Novozhilov, A.S., 2009. Origin and evolution of the genetic code: the universal enigma. Life 61 (2), 99–111.

Massey, S.E., 2006. A sequential 2-1-3 model of genetic code evolution that explains codon constraints. J. Mol. Evol. 62, 809–810.

Massey, S.E., 2016. The neutral emergence of error minimized genetic codes superior to the standard genetic code. J. Theoret. Biol. 408, 237–242.

Michel, C.J., 2015. The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. J. Theor. 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., Pellegrini, M., Pirillo, G., 2016. Maximal dinucleotide and trinucleotide circular codes. J. Theoret. Biol. 389, 40–46.

Michel, C.J., Pirillo, G., 2013. Dinucleotide circular codes. ISRN Biomath. 2013, 1–8, Article ID 538631.

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

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

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.

Taylor, F., Coates, D., 1989. The code within the codons. Biosystems 22, 177–187.

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