Journal of Theoretical Biology 389 (2016) 40–46

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Maximal dinucleotide and trinucleotide circular codes

Christian J. Michel [a,*], Marco Pellegrini [b], Giuseppe Pirillo [c,d]

[a] Theoretical Bioinformatics, ICube, University of Strasbourg, CNRS, 300 Boulevard Sébastien Brant, 67400 Illkirch, France
[b] Dipartimento di Matematica e Informatica "U.Dini", viale Morgagni 67/A, 50134 Firenze, Italy
[c] Consiglio Nazionale delle Ricerche, Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Unità di Firenze, Dipartimento di Matematica e Informatica "U.Dini", viale Morgagni 67/A, 50134 Firenze, Italy
[d] Université de Marne-la-Vallée, 5 boulevard Descartes, 77454 Marne-la-Vallée Cedex 2, France

HIGHLIGHTS

- Maximal dinucleotide circular codes.
- Maximal self-complementary dinucleotide circular codes.
- Maximal trinucleotide circular codes.
- Maximal self-complementary trinucleotide circular codes.

ARTICLE INFO

Article history:
Received 2 March 2015
Received in revised form
28 July 2015
Accepted 29 August 2015
Available online 14 September 2015

Keywords:
Maximal dinucleotide circular code
Maximal self-complementary dinucleotide circular code
Maximal trinucleotide circular code
Maximal self-complementary trinucleotide circular code

ABSTRACT

We determine here the number and the list of maximal dinucleotide and trinucleotide circular codes. We prove that there is no maximal dinucleotide circular code having strictly less than 6 elements (maximum size of dinucleotide circular codes). On the other hand, a computer calculus shows that there are maximal trinucleotide circular codes with less than 20 elements (maximum size of trinucleotide circular codes). More precisely, there are maximal trinucleotide circular codes with 14, 15, 16, 17, 18 and 19 elements and no maximal trinucleotide circular code having less than 14 elements. We give the same information for the maximal self-complementary dinucleotide and trinucleotide circular codes. The amino acid distribution of maximal trinucleotide circular codes is also determined.

1. Introduction

We extend here our combinatorial study of circular codes in genes, i.e. on the nucleotide alphabet $\mathcal{A}_4 = \{A, C, G, T\}$. A dinucleotide is a word of two letters (diletter) on $\mathcal{A}_4$. A trinucleotide is a word of three letters (triletter) on $\mathcal{A}_4$. The two sets of 16 dinucleotides and 64 trinucleotides are codes in the sense of language theory but not circular codes (Berstel and Perrin, 1985; Lassez, 1976). In order to have an intuitive meaning of these notions, codes are written on a straight line while circular codes are written on a circle, but, in both cases, unique decipherability is required.

Trinucleotide comma free codes, a very particular case of trinucleotide circular codes, have been studied for a long time, e.g. Crick et al. (1957), Golomb et al. (1958a), and Golomb et al. (1958b). After the discovery of a trinucleotide circular code in genes with strong mathematical properties (Arquès and Michel, 1996), circular codes are mathematical objects studied in combinatorics, theoretical computer science and theoretical biology. This theory underwent a rapid development, e.g. Koch and Lehmann (1997), Béal and Senellart (1998), Bassino (1999), Frey and Michel (2003, 2006), Pirillo (2003), Pirillo and Pirillo (2005), Lassez et al. (2007), Michel et al. (2008a, 2008b, 2012), Pirillo (2008), Michel and Pirillo (2010, 2011, 2013a, 2013b), Bussoli et al. (2011, 2012), Gonzalez et al. (2011), Fimmel et al. (2014, 2015), and Fimmel and Strüngmann (2015a, 2015b).

Trinucleotides are the fundamental words for genes, i.e. the DNA sequences coding the amino acids constituting the protein sequences. However, dinucleotides are also words with important

* Corresponding author.
E-mail addresses: c.michel@unistra.fr (C.J. Michel), pellegrin@math.unifi.it (M. Pellegrini), pirillo@math.unifi.it (G. Pirillo).

http://dx.doi.org/10.1016/j.jtbi.2015.08.029
0022-5193/

biological functions in genomes (Michel and Pirillo, 2013a; Fimmel et al., 2015; Fimmel and Strüngmann, 2015b). They are involved in some genome sites, e.g. the splice sites of eukaryotic introns are based on the dinucleotides *GT* and *AT* (Burset et al., 2000; Mount, 1982). They are also involved in some genome regions, e.g. the eukaryotic tandem repeats, i.e. concatenated words $(l_1l_2)^+$, $l_1, l_2 \in \mathcal{A}_4$, in particular the dinucleotides *CA* (Buerger et al., 2004), *TG* (Cuppens et al. 1998), etc.

We determine here the maximal circular codes for dinucleotides and trinucleotides. Their numbers and lists have never been studied on the 4-letter alphabet.

2. Preliminaries

The following definitions are classical for any finite set of words on any finite alphabet (Berstel and Perrin, 1985). Let $\mathcal{A}_4 = \{A, C, G, T\}$ be the genetic alphabet (nucleotides or letters) lexicographically ordered by $A < C < G < T$. The set of non-empty words (words resp.) on $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^+$ ($\mathcal{A}_4^*$ resp.). The set of the 4^n words of length n on $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^n = \{A^n, \ldots, T^n\}$. Dinucleotides and trinucleotides, i.e. words of length $n=2$ (diletters) and $n=3$ (triletters) on $\mathcal{A}_4$, are studied here. Thus, the set of the 16 words of length $n=2$ on $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^2 = \{AA, \ldots, TT\}$. The set of the 64 words of length $n=3$ on $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^3 = \{AAA, \ldots, TTT\}$. Let $x_1 \cdots x_n$ be the concatenation of the words x_i for $i = 1, \ldots, n$.

There is an important biological map involved in codes in genes on $\mathcal{A}_4$.

Definition 1. The nucleotide complementarity map $\mathcal{C} : \mathcal{A}_4 \to \mathcal{A}_4$ is defined by $\mathcal{C}(A) = T$, $\mathcal{C}(C) = G$, $\mathcal{C}(G) = C$ and $\mathcal{C}(T) = A$.

According to the property of the complementary and antiparallel double helix, the word complementarity map is defined as follows:

Definition 2. The word complementarity map $\mathcal{C} : \mathcal{A}_4^n \to \mathcal{A}_4^n$ is defined by $\mathcal{C}(uv) = \mathcal{C}(v)\mathcal{C}(u)$ for all $u \in \mathcal{A}_4, v \in \mathcal{A}_4^{n-1}$.

Definition 3. The word set complementarity map $\mathcal{C} : \mathcal{P}(\mathcal{A}_4^n) \to \mathcal{P}(\mathcal{A}_4^n)$ is defined by $\mathcal{C}(\mathcal{S}) = \{v | u, v \in \mathcal{A}_4^n, u \in \mathcal{S}, v = \mathcal{C}(u)\}$.

Example 1. On $\mathcal{A}_4^2$, we have $\mathcal{C}(\{AC, AG\}) = \{CT, GT\}$ and on $\mathcal{A}_4^3$, we have $\mathcal{C}(\{ACG, AGT\}) = \{ACT, CGT\}$.

Remark 1. The complementarity map $\mathcal{C}$ is involutional, i.e. for each word set $\mathcal{S}$, $\mathcal{C}(\mathcal{C}(\mathcal{S})) = \mathcal{S}$.

Definition 4 (*Code*). A word set $\mathcal{S} \subset \mathcal{A}_4^n$ is a code if, for each $x_1, \ldots, x_n, y_1, \ldots, y_m \in \mathcal{S}$, $n, m \geq 1$, the condition $x_1 \cdots x_n = y_1 \cdots y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \ldots, n$.

Codes are read on a straight line.

Remark 2. The set $\mathcal{A}_4^n$ is a code.

Remark 3. The non-empty subsets of $\mathcal{A}_4^n$ are codes.

Definition 5. The subsets of $\mathcal{A}_4^2$ and $\mathcal{A}_4^3$ are called dinucleotide and trinucleotide codes, respectively.

Definition 6 (*Circular code*). A code $\mathcal{S} \subset \mathcal{A}_4^n$ is circular if, for each $x_1, \ldots, x_n, y_1, \ldots, y_m \in \mathcal{S}$, $n, m \geq 1$, $r \in \mathcal{A}_4^*$, $s \in \mathcal{A}_4^+$, the conditions $sx_2 \cdots x_n r = y_1 \cdots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word) and $x_i = y_i$ for $i = 1, \ldots, n$.

Circular codes are read on a circle.

The following necessary and sufficient conditions for a set of dinucleotides (trinucleotides) to be a circular code are based on the concept of letter necklace.

We have the following definition and proposition for dinucleotides.

Definition 7 (*Necklace*). Let $l_1, l_2, l_3, \ldots, l_n, l_{n+1}$ be letters in $\mathcal{A}_4$. We say that the ordered sequence $l_1l_2l_3 \cdots l_nl_{n+1}$ is a $(n+1)$-necklace for a code $\mathcal{S} \subset \mathcal{A}_4^2$ if $l_1l_2, l_2l_3, \ldots, l_nl_{n+1} \in \mathcal{S}$, i.e. each dinucleotide l_il_{i+1} belongs to $\mathcal{S}$.

Proposition 1 (*Michel and Pirillo, 2013a*). *Let $\mathcal{S}$ be a dinucleotide code on $\mathcal{A}_4^2$. The following conditions are equivalent*:

(i) *$\mathcal{S}$ is a dinucleotide circular code*;
(ii) *$\mathcal{S}$ has no 5-necklace*.

We have the following definition and proposition for trinucleotides.

Definition 8. Letter Diletter Continued Necklaces (*LDCN*) for trinucleotides: Let $l_1, l_2, l_3, \ldots, l_n, l_{n+1}$ be letters in $\mathcal{A}_4$ and let $d_1, d_2, d_3, \ldots, d_n$ be diletters in $\mathcal{A}_4^2$. We say that the ordered sequence $l_1d_1l_2d_2 \cdots d_{n-1}l_nd_nl_{n+1}$ is an $(n+1)$*LDCN* for a code $\mathcal{S} \subset \mathcal{A}_4^3$ if $l_1d_1, l_2d_2, \ldots, l_nd_n \in \mathcal{S}$, i.e. each trinucleotide l_id_i is in $\mathcal{S}$, and if $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n, d_nl_{n+1} \in \mathcal{S}$, i.e. each trinucleotide d_il_{i+1} belongs to $\mathcal{S}$.

Proposition 2 (*Pirillo, 2003*). *Let $\mathcal{S}$ be a trinucleotide code on $\mathcal{A}_4^3$. The following conditions are equivalent*:

(i) *$\mathcal{S}$ is a trinucleotide circular code*;
(ii) *$\mathcal{S}$ has no 5-LDCN*.

Definition 9 (*Self-complementary circular code*). A circular code $\mathcal{S} \subset \mathcal{A}_4^n$ is self-complementary if, for each $x \in \mathcal{S}$, $\mathcal{C}(x) \in \mathcal{S}$, i.e. $\mathcal{C}(\mathcal{S}) = \mathcal{S}$.

Definition 10 (*Maximal circular code*). A circular code $\mathcal{S} \subset \mathcal{A}_4^n$ is maximal if for each $x \in \mathcal{A}_4^n, x \notin \mathcal{S}$, $\mathcal{S} \cup \{x\}$ is not a circular code.

Definition 11. A dinucleotide (trinucleotide resp.) circular code of size l, i.e. containing exactly l elements, is called a l-dinucleotide (l -trinucleotide resp.) circular code.

Remark 4. A 6-dinucleotide circular code is always maximal. A 20-trinucleotide circular code is always maximal.

Example 2. The following 14-trinucleotide circular code

$$\mathcal{S} = \{AAC, AAG, AAT, ACG, ATG, CGG, GCA, GCC, GCT, GGA, GTA, GTC, GTT, TGG\}$$

is maximal as it is not contained in a l-trinucleotide circular code of length $l = 15$. Thus, $\mathcal{S}$ is also not contained in a l-trinucleotide circular code of length $l = 16, \ldots, 20$. Indeed, if $\mathcal{S}$ would be contained in a l-trinucleotide circular code of length $l > 15$ and as any subset of a trinucleotide circular code is also a trinucleotide circular code, then a 15-trinucleotide circular code contains S and S is not maximal. Contradiction.

Example 3. The following 14-trinucleotide circular code

$$\{AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, AGT, ATC, ATG, ATT, CCG, CCT\}$$

is not maximal as it is contained in at least one 15-trinucleotide circular code, e.g. in $\{AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, AGT,$

$ATC, ATG, ATT, CCG, CCT, CGG\}$, $\{AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, AGT, ATC, ATG, ATT, CCG, CCT, GGC\}$, etc.

3. Results

3.1. Dinucleotide circular codes

3.1.1. Growth function of dinucleotide circular codes

The number of dinucleotide circular codes having one element is obviously equal to 12. The list contains all the dinucleotides except $\{AA\}$, $\{CC\}$, $\{GG\}$ and $\{TT\}$.

Proposition 3 (*Michel and Pirillo, 2013a*). *The number of 6-dinucleotide circular codes is equal to 24. Furthermore, any 6-dinucleotide circular code has the following form* $\mathcal{S} = \{ab, ac, ad, bc, bd, cd\}$ *where* $\{a, b, c, d\}$ *is a permutation of* $\{A, C, G, T\}$.

The list of the 24 maximal dinucleotide circular codes (Michel and Pirillo, 2013a) is recalled here:

$\{AC, AG, AT, CG, CT, GT\}$, $\{AC, AG, AT, CG, CT, TG\}$, $\{AC, AG, AT, CG, TC, TG\}$,

$\{AC, AG, AT, CT, GC, GT\}$, $\{AC, AG, AT, GC, GT, TC\}$, $\{AC, AG, AT, GC, TC, TG\}$,

$\{AC, AG, CG, TA, TC, TG\}$, $\{AC, AG, GC, TA, TC, TG\}$, $\{AC, AT, CT, GA, GC, GT\}$,

$\{AC, AT, GA, GC, GT, TC\}$, $\{AC, GA, GC, GT, TA, TC\}$, $\{AC, GA, GC, TA, TC, TG\}$,

$\{AG, AT, CA, CG, CT, GT\}$, $\{AG, AT, CA, CG, CT, TG\}$, $\{AG, CA, CG, CT, TA, TG\}$,

$\{AG, CA, CG, TA, TC, TG\}$, $\{AT, CA, CG, CT, GA, GT\}$, $\{AT, CA, CT, GA, GC, GT\}$,

$\{CA, CG, CT, GA, GT, TA\}$, $\{CA, CG, CT, GA, TA, TG\}$, $\{CA, CG, GA, TA, TC, TG\}$,

$\{CA, CT, GA, GC, GT, TA\}$, $\{CA, GA, GC, GT, TA, TC\}$, $\{CA, GA, GC, TA, TC, TG\}$.

More generally, the number of maximum dinucleotide circular codes over a given alphabet A_n with $n \geq 2$ is $n!$ (Pellegrini and Pirillo, 2014).

Proposition 4. *In any 6-dinucleotide circular code, there are two self-complementary dinucleotides, i.e. one of the following pairs* $\{AT, CG\}$, $\{AT, GC\}$, $\{TA, CG\}$ *and* $\{TA, GC\}$.

Proof. Let $\mathcal{S} \subset \mathcal{A}_4^n$ be a self-complementary 6-dinucleotide circular code. By Proposition 3, we have $\mathcal{S} = \{ab, ac, ad, bc, bd, cd\}$ for some permutation (a, b, c, d) of $\{A, C, G, T\}$.

(i) Suppose that $b = \mathcal{C}(a)$ and $d = \mathcal{C}(c)$. Then, $\mathcal{S} = \{a\mathcal{C}(a), ac, a\mathcal{C}(c), \mathcal{C}(a)c, \mathcal{C}(a)\mathcal{C}(c), c\mathcal{C}(c)\}$ contains the two self-complementary dinucleotides $a\mathcal{C}(a)$ and $c\mathcal{C}(c)$.
(ii) Suppose that $c = \mathcal{C}(a)$ and $d = \mathcal{C}(b)$. Then, $\mathcal{S} = \{ab, a\mathcal{C}(a), a\mathcal{C}(b), b\mathcal{C}(a), b\mathcal{C}(b), \mathcal{C}(a)\mathcal{C}(b)\}$ also contains the two self-complementary dinucleotides $a\mathcal{C}(a)$ and $b\mathcal{C}(b)$.
(iii) Suppose that $d = \mathcal{C}(a)$ and $c = \mathcal{C}(b)$. Then, $\mathcal{S} = \{ab, a\mathcal{C}(b), a\mathcal{C}(a), b\mathcal{C}(b), b\mathcal{C}(a), \mathcal{C}(b)\mathcal{C}(a)\}$ also contains the two self-complementary dinucleotides $a\mathcal{C}(a)$ and $b\mathcal{C}(b)$. □

Table 1
Growth function of l-dinucleotide circular codes for $l = 1,\ldots,6$. The 1st and 2nd rows give the size l and the occurrence number $\mathrm{Nb}(l)$ of l-dinucleotide circular codes, respectively.

l	1	2	3	4	5	6
$\mathrm{Nb}(l)$	12	60	152	186	108	24

Proposition 5. *The growth function of l-dinucleotide circular codes for* $l = 1, \ldots, 6$ *has a minimum number* $\mathrm{NbMin} = 12$ *at* $l = 1$ *and a maximum number* $\mathrm{NbMax} = 186$ *at* $l = 4$ (*Table 1*).

3.1.2. Growth function of maximal dinucleotide circular codes

Proposition 6. *There is no maximal l-dinucleotide circular code for* $l < 6$.

Proof. Let $P(\mathcal{S})$ be the set of the prefixes used in dinucleotides of $\mathcal{S} = \{ab, ac, ad, bc, bd, cd\}$, $S(\mathcal{S})$ the set of the suffixes used in dinucleotides of $\mathcal{S}$. Let $p = |P(\mathcal{S})|$, $s = |S(\mathcal{S})|$ and $q = |P(\mathcal{S}) \cup S(\mathcal{S})|$. We now consider the different cases according to the size l of dinucleotide circular codes:

- Case $l = 1$: In this case, $\mathcal{S}$ has obviously the form $\{ab\}$. Then it is non-maximal as it is strictly included in a maximal code.
- Case $l = 2$: In this case, we consider the three cases for $q = 2, 3, 4$:
  If $q = 2$ then $P(\mathcal{S}) = S(\mathcal{S})$ leading to $\mathcal{S}$ with the form $\{ab, ba\}$ that is impossible.
  If $q = 3$, we consider the three following subcases:
  - Case $p = 1, s = 2$: Let a the nucleotide used as prefix, then $\mathcal{S}$ has the form $\{ab, ac\}$. Then it is non-maximal as it is strictly included in a maximal code.
  - Case $p = 2, s = 1$: Let c the nucleotide used as suffix, then $\mathcal{S}$ has the form $\{ac, bc\}$, then it is non-maximal, as above.
  - Case $p = 2, s = 2$: $\mathcal{S}$ has the form $\{ab, bc\}$, then it is non-maximal.

  If $q = 4$ then $\mathcal{S}$ has the form $\{ab, cd\}$, then it is non-maximal.
- Case $l = 3$: In this case, we consider the two cases for $q = 3, 4$:
  If $q = 3$, we consider the three following subcases:
  - Case $p = 1$: $\mathcal{S}$ with the form $\{ab, ac, ad\}$ leads to $q = 4$ that is impossible.
  - Case $p = 2$: $\mathcal{S}$ has the form $\{ab, ac, bc\}$, then it is non-maximal. Note that this code would be maximal over an alphabet of cardinality 3.
  - Case $p = 3$: Then $P(\mathcal{S}) = S(\mathcal{S})$ leading to $\mathcal{S}$ with the form $\{ab, bc, ca\}$ that is impossible.

  If $q = 4$, we consider the six following subcases:
  - Case $p = 1, s = 3$: $\mathcal{S}$ has the form $\{ab, ac, ad\}$, then it is non-maximal.
  - Case $p = 2, s = 2$: If a, b are the dinucleotides used as prefixes, then $\mathcal{S}$ has the form $\{ac, ad, bc\}$, then it is non-maximal.
  - Case $p = 2, s = 3$: If a, b are the dinucleotides used as prefixes, then $\mathcal{S}$ has one of the following forms $\{ab, ac, bd\}$ or $\{ab, bc, bd\}$, then it is non-maximal.
  - Case $p = 3, s = 1$: $\mathcal{S}$ has the form $\{ad, bd, cd\}$, then it is non-maximal.
  - Case $p = 3, s = 2$: $\mathcal{S}$ has one of the following forms $\{ab, bd, cd\}$ or $\{ac, bc, cd\}$, then it is non-maximal.
  - Case $p = 3, s = 3$: $\mathcal{S}$ has the form $\{ab, bc, cd\}$, then it is non-maximal.
- Case $l = 4$: In this case, q must be equal to 4. We consider the four following subcases:
  - Case $p = 2, s = 2$: $\mathcal{S}$ has the form $\{ac, ad, bc, bd\}$, then it is non-maximal.
  - Case $p = 2, s = 3$: $\mathcal{S}$ has one of the following forms: $\{ab, ac, ad, bc\}$ or $\{ab, ac, bc, bd\}$, then it is non-maximal.
  - Case $p = 3, s = 2$: $\mathcal{S}$ has one of the following forms: $\{ad, bc, bd, cd\}$ or $\{ac, bc, bd, cd\}$; then it is non-maximal.
  - Case $p = 3, s = 3$: $\mathcal{S}$ has one of the following forms: $\{ab, ad, bc, cd\}$, $\{ab, ac, bd, cd\}$, $\{ab, bc, bd, cd\}$ or $\{ab, ac, bc, cd\}$, then it is non-maximal.
- Case $l = 5$: In this case, q must be equal to 4. We consider the three following subcases:

- Case $p=2, s=3$: $\mathcal{S}$ has the form $\{ab, ac, ad, bc, bd\}$, then it is non-maximal.
- Case $p=3, s=2$: $\mathcal{S}$ has the form $\{ac, ad, bc, bd, cd\}$, then it is non-maximal.
- Case $p=3, s=3$: $\mathcal{S}$ has one of the following forms: $\{ab, ac, bc, bd, cd\}$, $\{ab, ac, ad, bc, cd\}$, $\{ab, ad, bc, bd, cd\}$ or $\{ab, ac, ad, bd, cd\}$, then it is non-maximal.□

A computer calculus also retrieves this result of maximality of l-dinucleotide circular codes (not shown).

3.1.3. Growth function of self-complementary dinucleotide circular codes

The number of self-complementary dinucleotide circular codes having one element is obviously equal to 4. The list contains $\{AT\}$, $\{CG\}$, $\{GC\}$ and $\{TA\}$.

Proposition 7. *The number of self-complementary 6-dinucleotide circular codes is equal to 8. Furthermore, any self-complementary 6-dinucleotide circular code has the following form* $\mathcal{S}=\{ab, a\mathcal{C}(b), a\mathcal{C}(a), b\mathcal{C}(b), b\mathcal{C}(a), \mathcal{C}(b)\mathcal{C}(a)\}$, *where* (a,b,c,d) *is a permutation of* $\{A,C,G,T\}$.

Proof. Let $\mathcal{S}\subset\mathcal{A}_4^n$ be a self-complementary 6-dinucleotide circular code. By Proposition 3, we have $\mathcal{S}=\{ab, ac, ad, bc, bd, cd\}$ for some permutation (a,b,c,d) of $\{A,C,G,T\}$. We analyze all the possibilities with the complementarity map $\mathcal{C}$.

(i) Suppose that $b=\mathcal{C}(a)$ and $d=\mathcal{C}(c)$. Then, $\mathcal{S}=\{a\mathcal{C}(a), ac, a\mathcal{C}(c), \mathcal{C}(a)c, \mathcal{C}(a)\mathcal{C}(c), c\mathcal{C}(c)\}$. We are in contradiction with the self-complementarity of $\mathcal{S}$ as, for example, $\mathcal{C}(\mathcal{C}(a)c)=\mathcal{C}(c)a\notin\mathcal{S}$ (a is not suffix of any element of $\mathcal{S}$).

(ii) Suppose that $c=\mathcal{C}(a)$ and $d=\mathcal{C}(b)$. Then, $\mathcal{S}=\{ab, a\mathcal{C}(a), a\mathcal{C}(b), b\mathcal{C}(a), b\mathcal{C}(b), \mathcal{C}(a)\mathcal{C}(b)\}$. We are in contradiction with the self-complementarity of $\mathcal{S}$ as $\mathcal{C}(\mathcal{C}(a)\mathcal{C}(b))=ba\notin\mathcal{S}$ (a is not suffix of any element of $\mathcal{S}$).

(iii) Suppose that $d=\mathcal{C}(a)$ and $c=\mathcal{C}(b)$. Then, $\mathcal{S}=\{ab, a\mathcal{C}(b), a\mathcal{C}(a), b\mathcal{C}(b), b\mathcal{C}(a), \mathcal{C}(b)\mathcal{C}(a)\}$ which is clearly self-complementary.

The number of self-complementary 6-dinucleotide circular codes is 8 since we can first choose a in 4 different ways and then b in the 2 remaining different ways.□

Remark 5. In the code $\mathcal{S}=\{ab, a\mathcal{C}(b), a\mathcal{C}(a), b\mathcal{C}(b), b\mathcal{C}(a), \mathcal{C}(b)\mathcal{C}(a)\}$, there are two self-complementary dinucleotides $a\mathcal{C}(a)$ and $b\mathcal{C}(b)$, and two dinucleotides ab and $a\mathcal{C}(b)$ complementary to the two dinucleotides $\mathcal{C}(b)\mathcal{C}(a)$ and $b\mathcal{C}(a)$, respectively.

Proposition 8. *The number of 6-dinucleotide circular codes which are not self-complementary is equal to 16.*

Proof. Consequence of Propositions 3 and 7.□

As a dinucleotide can be self-complementary, a self-complementary dinucleotide circular code can contain an odd number of elements. In particular, it can contain 1, 3 or 5 dinucleotides.

Table 2
Growth function of self-complementary l-dinucleotide circular codes for $l=1,...,6$. The 1st and 2nd rows give the size l and the occurrence number $\mathrm{Nb}(l)$ of self-complementary l-dinucleotide circular codes, respectively.

l	1	2	3	4	5	6
$\mathrm{Nb}(l)$	4	8	16	16	12	8

Table 3
Growth function of maximal l-trinucleotide circular codes for $l=1,...,20$. The 1st and 2nd rows give the size l and the occurrence number $\mathrm{Nb}(l)$ of maximal l-trinucleotide circular codes, respectively.

l	1,…,13	14	15	16	17	18	19	20
$\mathrm{Nb}(l)$	0	192	1008	17,040	113,616	960,608	3,617,664	12,964,440

Proposition 9. *The growth function of self-complementary dinucleotide circular codes for* $l=1,\ldots,6$ *has a minimum number* $\mathrm{NbMin}=4$ *at* $l=1$ *and a maximum number* $\mathrm{NbMax}=16$ *at* $l=\{3,4\}$ (Table 2).

3.2. Trinucleotide circular codes

The growth function of trinucleotide circular codes was determined in Michel and Pirillo (2010).

3.2.1. Growth function of maximal trinucleotide circular codes

Proposition 10. *The growth function of maximal l-trinucleotide circular codes for* $l=1,\ldots,20$ *has a non-zero minimum number* $\mathrm{NbMin}=192$ *at* $l=14$ *and a maximum number* $\mathrm{NbMax}=12,964,440$ *at* $l=20$ (Table 3). *There is no maximal l-trinucleotide circular code for* $l<14$.

Table 4 gives an element of each class of maximum l-trinucleotide circular codes for l = 14,...,20.

The growth function of self-complementary trinucleotide circular codes was determined in Pirillo and Pirillo (2005).

3.2.2. Growth function of maximal self-complementary trinucleotide circular codes

There are two possible notions of maximal self-complementary trinucleotide circular codes.

Definition 12. If $\mathcal{S}\subset\mathcal{A}_4^3$ is a self-complementary trinucleotide circular code, it is maximal if for each $x\in\mathcal{A}_4^3$, $x\notin\mathcal{S}$, $\mathcal{S}\cup\{x\}$ is not a circular code.

Definition 13. If $\mathcal{S}\subset\mathcal{A}_4^3$ is a self-complementary trinucleotide circular code, it is dimaximal if for each $x\in\mathcal{A}_4^3$, $x\notin\mathcal{S}$, $\mathcal{S}\cup\{x,\mathcal{C}(x)\}$ is not a circular code.

Example 4. The following dimaximal self-complementary 16-trinucleotide circular code

$\{AAC, AAG, AAT, ACC, ACG, ATT, CCG, CGG, CGT, CTA, CTT, GGT, GTT, TAG, TCA, TGA\}$

is not maximal as it contains in the following 17-trinucleotide circular code:

$\{AAC, AAG, AAT, ACC, ACG, ATT, CCG, CGG, CGT, CTA, CTT, GCA, GGT, GTT, TAG, TCA, TGA\}$

with the additional trinucleotide GCA.

Note that Definition 12 implies Definition 13 but the converse is not true. Indeed, there are self-complementary trinucleotide circular codes $\mathcal{S}$ that are dimaximal but not maximal, i.e. there is $x\in\mathcal{A}_4^3$, $x\notin\mathcal{S}$, such that $\mathcal{S}\cup\{x\}$ is a circular code, although it cannot be self-complementary. By a computer search we see that such codes do exist (Proposition 11).

Proposition 11. *The growth functions of self-complementary trinucleotide circular codes that are maximal or dimaximal are given in Table 5.*

Tables 6 and 7 give the class of 32 maximal self-complementary 18-trinucleotide circular codes and the class of 16 dimaximal self-complementary 16-trinucleotide circular codes.

Table 4
An element of each class of maximal l-trinucleotide circular codes for $l = 14, \ldots, 20$. Note that the maximal 14-trinucleotide circular code was already given in Example 2.

l	Maximal l-trinucleotide circular code
14	{*AAC*, *AAG*, *AAT*, *ACG*, *ATG*, *CGG*, *GCA*, *GCC*, *GCT*, *GGA*, *GTA*, *GTC*, *GTT*, *TGG*}
15	{*AAC*, *AAG*, *AAT*, *ACC*, *AGT*, *CAG*, *CAT*, *CTA*, *GTT*, *TCC*, *TCG*, *TGA*, *TGC*, *TGG*, *TTA*}
16	{*AAC*, *AAG*, *AAT*, *ACC*, *ACG*, *ACT*, *CAG*, *CCT*, *CGC*, *GCT*, *GGA*, *GTA*, *TCA*, *TCT*, *TGA*, *TTA*}
17	{*AAC*, *AAG*, *AAT*, *ACC*, *ACG*, *ACT*, *AGC*, *AGG*, *ATG*, *CCT*, *CGG*, *CTG*, *GCC*, *GTA*, *TCA*, *TGG*, *TTA*}
18	{*AAC*, *AAG*, *AAT*, *ACC*, *ACG*, *ACT*, *AGC*, *AGG*, *AGT*, *ATC*, *ATG*, *CCT*, *CGT*, *GCT*, *GGT*, *TAT*, *TCT*, *TGT*}
19	{*AAC*, *AAG*, *AAT*, *ACC*, *ACG*, *ACT*, *AGC*, *AGG*, *AGT*, *ATC*, *ATG*, *ATT*, *CCG*, *CCT*, *CGT*, *GTT*, *TGC*, *TGG*, *TTC*}
20	{*AAC*, *AAG*, *AAT*, *ACC*, *ACG*, *ACT*, *AGC*, *AGG*, *AGT*, *ATC*, *ATG*, *ATT*, *CCG*, *CCT*, *CGG*, *CGT*, *CTG*, *CTT*, *GGT*, *GTT*}

Table 5
Growth functions of maximal or dimaximal self-complementary l-trinucleotide circular codes for $l = 1,...,20$. The 1st row gives the size l, the 2nd and 3rd rows give the occurrence numbers Nb1(l) and Nb2(l) of self-complementary l-trinucleotide circular codes that are maximal or dimaximal, respectively.

l	2,4,…,12,14	16	18	20
Nb1(l)	0	0	32	528
Nb2(l)	0	16	96	528

Table 6
The class of 32 maximal self-complementary 18-trinucleotide circular codes (see Table 5). This class of codes is also dimaximal.

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

The class of 96 dimaximal self-complementary 18-trinucleotide circular codes is the union of the class of 32 maximal self-complementary 18-trinucleotide circular codes (Table 6) and the class of 64 non-maximal dimaximal self-complementary 18-trinucleotide circular codes (not shown).

3.2.3. Amino acid coding of maximal trinucleotide circular codes

Maximal trinucleotide circular codes constitute an important class of circular codes for its variety representation. Indeed, a maximal trinucleotide circular code cannot be included in another circular code. The distribution of amino acid coding according to the universal genetic code in the 12,964,440 maximal 20-trinucleotide circular codes was determined in Michel (2014). We extend the amino acid distribution for the maximal l-trinucleotide circular codes for $l = 14,...,19$.

Table 8 shows that the (17,674,568-1) maximal l-trinucleotide circular codes for $l = 14,...,20$ code maximum numbers of amino acids strictly less than l, except for one case with the maximal 16-trinucleotide circular code

{*AAG*, *AAT*, *ACG*, *CAG*, *CAT*, *CCG*, *CTT*, *GAG*, *GAT*, *GCT*, *GGC*, *GTA*, *TAT*, *TCG*, *TGG*, *TGT*}

coding the 16 amino acids

{*Ala*, *Asn*, *Asp*, *Cys*, *Gln*, *Glu*, *Gly*, *His*, *Leu*, *Lys*, *Pro*, *Ser*, *Thr*, *Trp*, *Tyr*, *Val*}.

The distribution of amino acid coding according to the universal genetic code in the 528 maximal self-complementary

Table 7
The class of 16 dimaximal self-complementary 16-trinucleotide circular codes (see Table 5) .

{*AAC*, *AAG*, *AAT*, *ACC*, *ACG*, *ATT*, *CCG*, *CGG*, *CGT*, *CTA*, *CTT*, *GGT*, *GTT*, *TAG*, *TCA*, *TGA*}
{*AAC*, *AAG*, *AAT*, *ACC*, *ACG*, *ATT*, *CGT*, *CTA*, *CTT*, *GGA*, *GGT*, *GTT*, *TAG*, *TCA*, *TCC*, *TGA*}
{*AAC*, *AAG*, *AAT*, *AGC*, *AGG*, *ATT*, *CCA*, *CCT*, *CTT*, *GCT*, *GTA*, *GTT*, *TAC*, *TCA*, *TGA*, *TGG*}
{*AAC*, *AAG*, *AAT*, *AGC*, *AGG*, *ATT*, *CCT*, *CTT*, *GCC*, *GCT*, *GGC*, *GTA*, *GTT*, *TAC*, *TCA*, *TGA*}
{*AAC*, *AAG*, *ACC*, *CAG*, *CGA*, *CTG*, *CTT*, *GCC*, *GGA*, *GGC*, *GGT*, *GTA*, *GTT*, *TAC*, *TCC*, *TCG*}
{*AAC*, *AAG*, *AGG*, *CCA*, *CCG*, *CCT*, *CGG*, *CTA*, *CTT*, *GAC*, *GCA*, *GTC*, *GTT*, *TAG*, *TGC*, *TGG*}
{*AAC*, *ACC*, *CAG*, *CGA*, *CTG*, *GCC*, *GGA*, *GGC*, *GGT*, *GTA*, *GTT*, *TAA*, *TAC*, *TCC*, *TCG*, *TTA*}
{*AAG*, *AGG*, *CCA*, *CCG*, *CCT*, *CGG*, *CTA*, *CTT*, *GAC*, *GCA*, *GTC*, *TAA*, *TAG*, *TGC*, *TGG*, *TTA*}
{*AAT*, *ACC*, *ACG*, *ATC*, *ATT*, *CAG*, *CGT*, *CTG*, *GAA*, *GAT*, *GCC*, *GGA*, *GGC*, *GGT*, *TCC*, *TTC*}
{*AAT*, *AGC*, *AGG*, *ATG*, *ATT*, *CAA*, *CAT*, *CCA*, *CCG*, *CCT*, *CGG*, *GAC*, *GCT*, *GTC*, *TGG*, *TTG*}
{*ACC*, *ACG*, *ATC*, *CAA*, *CAG*, *CGT*, *CTG*, *GAA*, *GAT*, *GCC*, *GGA*, *GGC*, *GGT*, *TCC*, *TTC*, *TTG*}
{*ACC*, *ACT*, *AGT*, *ATG*, *CAA*, *CAT*, *CGA*, *GAA*, *GGA*, *GGT*, *TAA*, *TCC*, *TCG*, *TTA*, *TTC*, *TTG*}
{*ACT*, *AGG*, *AGT*, *ATC*, *CAA*, *CCA*, *CCT*, *GAA*, *GAT*, *GCA*, *TAA*, *TGC*, *TGG*, *TTA*, *TTC*, *TTG*}
{*ACT*, *AGT*, *ATC*, *CAA*, *CCA*, *GAA*, *GAT*, *GCA*, *GCC*, *GGC*, *TAA*, *TGC*, *TGG*, *TTA*, *TTC*, *TTG*}
{*ACT*, *AGT*, *ATG*, *CAA*, *CAT*, *CCG*, *CGA*, *CGG*, *GAA*, *GGA*, *TAA*, *TCC*, *TCG*, *TTA*, *TTC*, *TTG*}
{*AGC*, *AGG*, *ATG*, *CAA*, *CAT*, *CCA*, *CCG*, *CCT*, *CGG*, *GAA*, *GAC*, *GCT*, *GTC*, *TGG*, *TTC*, *TTG*}

Table 8
Amino acid distribution (mean, 25th and 75th percentiles, minimum and maximum values) of each maximal l-trinucleotide circular codes for $l = 14,...,20$.

l	Nb(l)	Mean	25th %ile	75th %ile	Min	Max
20	12,964,440	11.4	10	12	5	18
19	3,617,664	11.3	10	12	5	17
18	960,608	11.0	10	12	5	17
17	113,616	10.7	10	12	5	16
16	17,040	10.4	9	11	5	16
15	1008	10.1	9	11	6	14
14	192	9.7	9	11	6	12

Table 9
Amino acid distribution (mean, 25th and 75th percentiles, minimum and maximum values) of each maximal or dimaximal self-complementary l-trinucleotide circular codes for $l = 16,18,20$.

l	Nb(l)	Mean	25th %ile	75th %ile	Min	Max
20	Nb1(l)=Nb2(l)=528	11.3	10	12	6	15
18	Nb1(l)=32	11.1	10	12	9	15
18	Nb2(l)=96	11.2	10	12	8	15
16	Nb2(l)=16	11.1	10	12	9	13

20-trinucleotide circular codes was determined in Michel (2014). Table 9 gives the amino distribution for the maximal or dimaximal self-complementary l-trinucleotide circular codes for $l = 16, 18$.

4. Discussion

These results constitute a contribution to the combinatorial properties of circular codes, precisely to the number and the list of maximal dinucleotide and trinucleotide circular codes which have never been studied on a 4-letter alphabet. There is no maximal l-dinucleotide circular code for $l < 6$ and no maximal l-trinucleotide circular code for $l < 14$.

The concept of maximal circular codes will be important to understand evolution of circular codes in genes. In 1996, a statistical analysis of occurrence frequencies of the 64 trinucleotides $\{AAA, ..., TTT\}$ in the three frames of genes of both prokaryotes and eukaryotes showed that the trinucleotides are not uniformly distributed in these three frames (Arquès and Michel, 1996). By excluding the four periodic trinucleotides and by assigning each trinucleotide to a preferential frame (frame of its highest occurrence frequency), three subsets $X = X_0$, X_1 and X_2 of 20 trinucleotides are found in the frames 0 (reading frame), 1 (frame 0 shifted by one nucleotide) and 2 (frame 0 shifted by two nucleotides) in genes of both prokaryotes and eukaryotes. This set X contains the 20 following trinucleotides:

$$X = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}. \quad (4.1)$$

The trinucleotide set X presents several strong mathematical properties, particularly the fact that X is a maximal C^3 self-complementary trinucleotide circular code (Arquès and Michel, 1996). Very recently, in 2015, by quantifying the approach used in 1996 for identifying a preferential frame and by applying a massive statistical analysis of gene taxonomic groups, the circular code X is strengthened in genes of prokaryotes (7,851,762 genes, 2,481,566,882 trinucleotides) and eukaryotes (1,662,579 genes, 824,825,761 trinucleotides), and now also identified in genes of plasmids (237,486 genes, 68,244,356 trinucleotides) and viruses (184,344 genes, 45,688,798 trinucleotides) (Michel, 2015). Thus, the circular code X is the "universal" (average) circular code in genes. However, at the gene taxonomic group level, variant X codes are observed in a few cases of bacteria and eukaryotes (Michel, 2015)

$$X_A = \{AAC, AAT, ACC, ATC, ATT, CAG, CGC, CTC, CTG, GAA, GAC, GAG, GAT, GCG, GGT, GTA, GTC, GTT, TAC, TTC\} \quad (4.2)$$

in cyanobacteria,

$$X_B = \{AAC, AAT, ATC, ATT, CAC, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GTA, GTC, GTG, GTT, TAC, TTC\} \quad (4.3)$$

in deinococcus, mammals and kinetoplasts,

$$X_C = \{AAC, AAT, ACC, AGC, ATC, ATT, CTC, GAA, GAC, GAG, GAT, GCC, GCT, GGC, GGT, GTA, GTC, GTT, TAC, TTC\} \quad (4.4)$$

in elusimicrobia and apicomplexans,

$$X_D = \{AAC, AAT, ATC, ATT, CAC, CAG, CGC, CTC, CTG, GAA, GAC, GAG, GAT, GCG, GTA, GTC, GTG, GTT, TAC, TTC\} \quad (4.5)$$

in birds,

$$X_E = \{AAC, AAG, AAT, ATC, ATT, CAC, CAG, CTC, CTG, CTT, GAC, GAG, GAT, GCC, GGC, GTA, GTC, GTG, GTT, TAC\} \quad (4.6)$$

in fishes,

$$X_F = \{AAC, AAG, AAT, ACC, ATC, ATT, CAG, CTC, CTG, CTT, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC\} \quad (4.7)$$

in insects, and

$$X_G = \{AAC, AAT, ACC, ACT, AGT, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTC, GTT, TTC\} \quad (4.8)$$

in basidiomycetes. The codes X_A and X_D are self-complementary but non-circular. The codes X_B, X_C, X_E, X_F and X_G are maximal C^3 self-complementary circular (Michel, 2015). In genes of viruses, a subset of X is identified which may have either 18 or 16 trinucleotides, i.e. two non-maximal C^3 self-complementary circular codes (Michel, 2015). At the gene species level, the variation of the circular code X increases, e.g. a subset of X_B of 18 trinucleotides is observed in the human genes (Michel, 2015). So far, no evolutionary study of circular codes has been investigated. Several interesting questions remain to be analyzed. What are the successive evolutionary steps to explain the circular code X observed in genes (its 20 trinucleotides) from the expansion of circular codes of short lengths, i.e. containing a few trinucleotides ("primitive" circular codes) . What are the evolutionary processes to transform the circular code X to other circular codes, or more

generally a given circular code to another circular code, e.g. a variant X code to another variant X code. The concept of maximal circular codes, presented here for the first time, is an interesting combinatorial property for this purpose. Indeed, circular codes which are maximal cannot be extended in a circular code. Such maximal circular codes stop circular code evolution. Thus, expansion and variation of circular codes may avoid maximal circular codes.

Acknowledgment

The third author thanks the Dipartimento di matematica "U. Dini" for giving him their friendly hospitality and the Project Interomics of CNR for the financial support.

References

Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182 (1), 45–58.
Bassino, F., 1999. Generating functions of circular codes. Adv. Appl. Math. 22 (1), 1–24.
Béal, M.P., Senellart, J., 1998. On the bound of the synchronization delay of a local automaton. Theor. Comput. Sci. 205 (1-2), 297–306.
Berstel, J., Perrin, D., 1985. Theory of Codes. Academic Press, London, UK.
Buerger, H., Packeisen, J., Boecker, A., et al., 2004. Allelic length of a CA dinucleotide repeat in the egfr gene correlates with the frequency of amplifications of this sequence-first results of an inter-ethnic breast cancer study. J. Pathol. 203 (1), 545–550.
Burset, M., Seledtsov, I.A., Solovyev, V.V., 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 28 (21), 4364–4375.
Bussoli, L., Michel, C.J., Pirillo, G., 2011. On some forbidden configurations for self-complementary trinucleotide circular codes. J. Algebra Number Theory Acad. 2, 223–232.
Bussoli, L., Michel, C.J., Pirillo, G., 2012. On conjugation partitions of sets of trinucleotides. Appl. Math. 3, 107–112.
Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. In: Proceedings of the National Academy of Sciences, vol. 43, pp. 416–421.
Cuppens, H., Lin, W., Jaspers, M., et al., 1998. Polyvariant mutant cystic fibrosis transmembrane conductance regulator genes: the polymorphic (TG)m locus explains the partial penetrance of the T5 polymorphism as a disease mutation. J. Clin. Investig. 101 (2), 487–496.
Fimmel, E., Giannerini, S., Gonzalez, D., Strüngmann, L., 2014. Circular codes, symmetries and transformations. J. Math. Biol. http://10.1007/s00285-014-0806-7.
Fimmel, E., Giannerini, S., Gonzalez, D., Strüngmann, L., 2015. Dinucleotide circular codes and bijective transformations. J. Theor. Biol. in press.
Fimmel, E., Strüngmann, L., 2015a. On the hierarchy of trinucleotide n-circular codes and their corresponding amino acids. J. Theor. Biol. 364, 113–120.
Fimmel, E., Strüngmann, L., 2015b. Maximal dinucleotide comma-free codes, Submitted (Personal Communication).
Frey, G., Michel, C.J., 2003. Circular codes in archaeal genomes. J. Theor. Biol. 223 (4), 413–431.
Frey, G., Michel, C.J., 2006. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Comput. Biol. Chem. 30 (2), 87–101.
Golomb, S.W., Gordon, B., Welch, L.R., 1958a. Comma-free codes. Can. J. Math. 10, 202–209.
Golomb, S.W., Welch, L.R., Delbrück, M., 1958b. Construction and properties of comma-free codes. Biol. Medd. K. Danske Videnskabernes Selsk. 23 (9).
Gonzalez, D.L., Giannerini, S., Rosa, R., 2011. Circular codes revisited: a statistical approach. J. Theor. Biol. 275 (1), 21–28.
Koch, A.J., Lehmann, J., 1997. About a symmetry of the genetic code. J. Theor. Biol. 189 (2), 171–174.
Lassez, J.L., 1976. Circular codes and synchronization. Int. J. Comput. Inf. Sci. 5 (2), 201–208.
Lassez, J.L., Rossi, R.A., Bernal, A.E., 2007. Crick's hypothesis revisited: the existence of a universal coding frame. In: Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07), Niagara Falls, Canada, May, pp. 745–751.
Michel, C.J., 2014. A genetic scale of reading frame coding. J. Theor. Biol. 355, 83–94.
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., Pirillo, G., 2010. Identification of all trinucleotide circular codes. Comput. Biol. Chem. 34 (2), 122–125.
Michel, C.J., Pirillo, G., 2011. Strong trinucleotide circular codes. Int. J. Combin. 2011, 14 pp., Article ID 659567.
Michel, C.J., Pirillo, G., 2013a. Dinucleotide circular codes. ISRN Biomath. 8 pp., Article ID 538631.
Michel, C.J., Pirillo, G., 2013b. A permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes. J. Theor. Biol. 319, 116–121.
Michel, C.J., Pirillo, G., Pirillo, M.A., 2008a. Varieties of comma-free codes. Comput. Math. Appl. 55 (5), 989–996.
Michel, C.J., Pirillo, G., Pirillo, M.A., 2008b. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theor. Comput. Sci. 401 (1–3), 17–26.
Michel, C.J., Pirillo, G., Pirillo, M.A., 2012. A classification of 20-trinucleotide circular codes. Inf. Comput. 212, 55–63.
Mount, S.M., 1982. A catalogue of splice junction sequences. Nucleic Acids Res. 10 (2), 459–472.
Pellegrini, M., Pirillo, G., 2014. On the dinucleotide circular codes of maximum cardinality. Theor. Biol. Forum 107 (1–2), 89–95.
Pirillo, G., 2003. A characterization for a set of trinucleotides to be a circular code. In: Pellegrini, C., Cerrai, P., Freguglia, P., Benci, V., Israel, G. (Eds.), Determinism, Holism, and Complexity. Kluwer Academic Publisher, New York, NY, USA.
Pirillo, G., 2008. A hierarchy for circular codes. RAIRO-Theor. Inf. Appl. 42 (4), 717–728.
Pirillo, G., Pirillo, M.A., 2005. Growth function of self-complementary circular codes. Biol. Forum 98 (1), 97–110.