Mathematical Biosciences 294 (2017) 120–129

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Diletter circular codes over finite alphabets

Elena Fimmel[1,a], Christian J. Michel[*,b], Lutz Strüngmann[1,a]

[a] Institute of Mathematical Biology, Faculty for Computer Sciences, University of Applied Sciences 68163 Mannheim, Germany
[b] Theoretical Bioinformatics, ICube, University of Strasbourg, CNRS, 300 Boulevard Sébastien Brant 67400 Illkirch, France

ARTICLE INFO

Keywords:
Diletter circular code
Finite alphabet
Enumerative combinatorics

ABSTRACT

The graph approach of circular codes recently developed (Fimmel et al., 2016) allows here a detailed study of diletter circular codes over finite alphabets. A new class of circular codes is identified, strong comma-free codes. New theorems are proved with the diletter circular codes of maximal length in relation to (i) a characterisation of their graphs as acyclic tournaments; (ii) their explicit description; and (iii) the non-existence of other maximal diletter circular codes. The maximal lengths of paths in the graphs of the comma-free and strong comma-free codes are determined. Furthermore, for the first time, diletter circular codes are enumerated over finite alphabets. Biological consequences of dinucleotide circular codes are analysed with respect to their embedding in the trinucleotide circular code X identified in genes and to the periodicity modulo 2 observed in introns. An evolutionary hypothesis of circular codes is also proposed according to their combinatorial properties.

1. Introduction

Circular codes have been involved in the genetic code. About 60 years ago, before the discovery of the genetic code, a class of trinucleotide codes, called comma-free codes, was proposed by Crick et al. [8] for explaining how the reading (original, correct) of a sequence of trinucleotides could code amino acids. In particular, how the reading frame can be retrieved and maintained. The four nucleotides $\{A, C, G, T\}$ as well as the 16 dinucleotides $\{AA,\ldots,TT\}$ are simple codes which are not appropriate for coding 20 amino acids. However, trinucleotides induce a redundancy in their coding. Thus, [8] conjectured that only 20 trinucleotides among the 64 possible trinucleotides $\{AAA,\ldots,TTT\}$ code the 20 amino acids. Such a bijective code implies that the coding trinucleotides are found only in one frame - the comma-freeness property. The determination of a set of 20 trinucleotides forming a comma-free code has several necessary conditions:

(i) A periodic trinucleotide from the set $\{AAA,CCC,GGG,TTT\}$ must be excluded from such a code. Indeed, the concatenation of AAA with itself, for instance, does not allow the reading frame to be retrieved as there are three possible decompositions: $\cdots AAA,AAA,AAA\cdots$ (original frame), $\cdots A,AAA,AAA,AA\cdots$ and $\cdots AA,AAA,AAA,A\cdots$, the commas showing the adopted decomposition.

(ii) Two non-periodic permuted trinucleotides, i.e. two trinucleotides related by a circular permutation, e.g. ACG and CGA, must also be excluded from such a code. Indeed, the concatenation of ACG with itself, for instance, does not allow the reading frame to be retrieved as there are two possible decompositions: $\cdots ACG,ACG,ACG\cdots$ (original frame) and $\cdots A,CGA,CGA,CG\cdots$

Therefore, by excluding the four periodic trinucleotides and by gathering the 60 remaining trinucleotides in 20 classes of three trinucleotides such that, in each class, the three trinucleotides are deduced from each other by a circular permutation, e.g. ACG, CGA and GAC, we see that a comma-free code can contain only one trinucleotide from each class and thus has at most 20 trinucleotides. This trinucleotide number is identical to the amino acid number, thus leading to a code assigning one trinucleotide per amino acid without ambiguity. A few combinatorial results on trinucleotide comma-free codes were obtained by Golomb et al. [15,16]. However, no trinucleotide comma-free code was identified in genes statistically. Furthermore, at the beginning of the 1960s, the discovery that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes phenylalanine, led to the abandonment of the concept of comma-freeness over the alphabet $\{A,C,G,T\}$.

The circular code theory initiated in 1996 proposes that genes are constituted of two trinucleotide codes: the classical genetic code with 61 trinucleotides for coding the 20 amino acids (except the three stop codons $\{TAA,TAG,TGA\}$) and a circular code based on 20 trinucleotides for retrieving, maintaining and synchronizing the reading frame. It relies on two main results: the identification of a maximal (20 words) C^3 self-complementary trinucleotide circular code X in genes of bacteria (15,735,053 genes, 5,222,267,667 trinucleotides), archaea (282,802 genes, 81,460,549 trinucleotides), eukaryotes (4,356,391 genes,

* 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).
[1] The first and third authors would like to thank the Karl Völker Foundation for their support.

http://dx.doi.org/10.1016/j.mbs.2017.10.001
Received 27 May 2017; Received in revised form 26 August 2017; Accepted 8 October 2017
Available online 10 October 2017

2,406,844,838 trinucleotides), plasmids (575,760 genes, 159,169,387 trinucleotides) and viruses (299,401 genes, 66,677,580 trinucleotides) [4,21,22] and the finding of X circular code motifs in tRNAs and rRNAs, in particular in the ribosome decoding center [9,20]. The universally conserved nucleotides A1492 and A1493 and the conserved nucleotide G530 in the ribosome decoding center are included in X circular code motifs. This circular code X contains the 20 following trinucleotides

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

A circular code, e.g. a trinucleotide circular code or a trinucleotide comma-free code, has the fundamental property to always retrieve the reading frame in any position of any sequence generated with the circular code. In particular, initiation and stop trinucleotides as well as any frame signals are not necessary to define the reading frame. Indeed, a window of a few nucleotides, whose nucleotide length depends on the class of circular codes, positioned anywhere in a sequence generated with the circular code always retrieves the reading frame (see, e.g., the example given in [12]). At most 13 consecutive nucleotides in a sequence of trinucleotides from the circular code X observed in genes are enough to retrieve the reading frame. In a trinucleotide comma-free code, this window is equal to three nucleotides.

The major combinatorial differences between the trinucleotide circular code X, explaining in particular its use in genes, and any trinucleotide comma-free code rely on the following results found in [25]. The maximal length of self-complementary trinucleotide comma-free codes is 16 trinucleotides and there exist four such codes (Tables 3a,b in [25]). The maximal length of C^3 trinucleotide comma-free codes is 16 trinucleotides and there exist 18 such codes (Tables 4a,b in [25]). The maximal length of C^3 self-complementary trinucleotide comma-free codes is 16 trinucleotides and there exist four such codes which coincide with the self-complementary trinucleotide comma-free codes (Table 5 in [25]). Thus, there is no trinucleotide comma-free code with the C^3 property with 20, 19, 18 and 17 trinucleotides and there is no trinucleotide comma-free code with the self-complementary property or the C^3 self-complementary property with 20 and 18 trinucleotides. The maximal length of a trinucleotide comma-free code having the C^3 property for retrieving the reading frame and its two shifted frames and the self-complementary property for DNA pairing which allow together to retrieve the three frames in the direct DNA strand and the three frames in the complementary DNA strand, is 16 trinucleotides. Thus, $(20-16)\cdot 3 = 12$ trinucleotides must be discarded compared to $20\cdot 3 = 60$ trinucleotides of the circular code X in genes. There is no chemical, evolutionary, statistical and combinatorial reason to avoid a particular set of 12 trinucleotides from our point of view.

Trinucleotides are the fundamental words for genes. Dinucleotides are also words with important biological functions in genomes as they are involved in some sites, e.g. the splice sites of eukaryotic introns, and in some regions, e.g. the tandem repeats (references in [12]). Dinucleotide circular codes have been studied according to three approaches, by the combinatorics theory [24], the group theory [11] and the graph approach [12]. The combinatorial results for dinucleotide circular codes are extended here to diletter circular codes over finite alphabets. For the first time, circular codes are enumerated. Biological consequences of dinucleotide circular codes are analysed with respect to their embedding in the trinucleotide circular code X identified in genes and to the periodicity modulo 2 observed in introns.

2. Graph theory of circular codes

Throughout this section, let $A_n = \{a_1,\ldots,a_n\}$, $n \in \mathbb{N}$, be a finite alphabet containing n letters. A very important subcase is $n = 4$ since it is biologically relevant with the alphabet being the set of nucleotide bases $\mathscr{B} = \{A, C, G, T\}$ of DNA sequences. Here, A stands for *Adenine, C* stands for *Cytosine, G* stands for *Guanine*, and T stands for *Thymine*. In the sequel, we will consider diletter circular codes over the finite alphabet A_n.

In [12], a directed graph $\mathscr{G}(X)$ was constructed for any k-nucleotide code X, i.e. a set of words of length $k \in \mathbb{N}$ over the 4-letter alphabet $\mathscr{B}$. The main idea of this graph approach was to deduce properties of the code X from its associated graph $\mathscr{G}(X)$. In this paper, the graph approach developed in [12] is extended to diletter codes over any finite alphabet A_n. Recall that a *diletter code* over the alphabet A_n is a subset $X \subseteq A_n^2$ of the set of words over A_n of length 2. Moreover, recall that a *graph* $G = (V, E)$ consists of a set of *vertices* V and *edges* E which is a subset of the set of pairs V^2 of vertices. The graph G is called *directed* if each edge $[v_1, v_2] \in E$ has an *orientation*, i.e. the order of the vertices v_1, v_2 is considered, i.e. $[v_1, v_2] \neq [v_2, v_1]$.

Definition 2.1. Given a diletter code $X \subseteq A_n^2$, the associated graph $\mathscr{G}(X)$ has nodes labeled by the letters of the alphabet A_n which appear in first or second position of diletters of the code X. Two nodes N_i, $N_j \in A_n$ of $\mathscr{G}(X)$ are connected by a directed edge $[N_i, N_j]$ if and only if the diletter N_iN_j belongs to the code X.

Fig. 1 gives an illustration of Definition 2.1.

We recall the notions of circularity and comma-freeness for diletter circular codes. Furthermore, we introduce here the new notion of strong comma-freeness. Indeed, diletter strong comma-free codes are studied for the first time here. They represent a subclass of the class of circular codes with interesting properties.

Definition 2.2. Let $X \subseteq A_n^2$ be a diletter code. We say that X is

(1) a *diletter circular code* if for any concatenation $c_1\cdots c_m$ of diletters from X there is only one partition into diletters from X when read on a circle;
(2) a *diletter comma-free code* if for any two diletters N_1N_2 and $N_1'N_2'$ from X the diletters N_2N_1' and $N_2'N_1$ do not belong to X;
(3) a *diletter strong comma-free code* if for any diletter N_1N_2 from X the diletter N_2N_3 does not belong to X for any $N_3 \in A_n$.

Furthermore, we say that a diletter circular code X is *maximal* if there is no diletter circular code $X' \subseteq A_n^2$ with $X \subseteq X'$ and $X \neq X'$, i.e. X cannot be extended to a larger diletter circular code. We also say that a diletter circular code X is of *maximal length* if there is no diletter circular code X' with $|X'| > |X|$ where $|X|$ denotes the cardinality of the code X, also called in the following *length of the code X*. Note that in [11] maximality was used in the sense of maximal length.

We recall the graph properties of circular codes identified in [12] which will be used for analyzing diletter circular codes. The associated (oriented) graph of circular codes is *simple*, meaning that it does not contain loops, i.e. edges between a node and itself (self-loop), and does not have multiple edges with the same orientation between two nodes. Note that the orientation for the multiple edges does play a role as for a simple oriented graph $\mathscr{G}$ the edges $[x, y] \in E(\mathscr{G})$ and $[y, x] \in E(\mathscr{G})$ imply that there is a cycle (circle) of length 2. However, this graph structure is also excluded for circular codes. Indeed, from the circularity of a diletter code X, each diletter $N_1N_2 \in X$ implies that the reversed diletter N_2N_1 is not a member of X since otherwise the word $N_1N_2N_1N_2$ would have two decompositions on the circle $N_1N_2\,N_1N_2$ and $N_2N_1\,N_2N_1$.

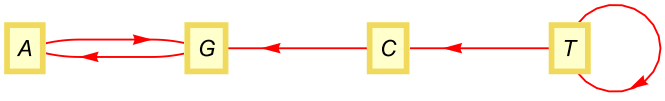


Fig. 1. Graph $\mathscr{G}(X)$ representing the dinucleotide code $X = \{AG,CG,GA,TC,TT\}$.

Thus, the underlying (not oriented) graph of a directed graph corresponding to circular codes is already simple.

Recall from graph theory [7] that a *cycle* in $\mathscr{G}$ is an oriented closed path in $\mathscr{G}$ and a graph is called *acyclic* if it does not contain cycles, i.e. oriented closed paths. A *Hamiltonian path* in a directed graph is a directed path that visits each vertex exactly once.

The following theorem is a generalization of Theorem 2.6 in [12] for the case of the alphabet of size 4 and the proof carries over verbatim.

Theorem 2.3. *Given a code $X \subseteq A_n^k$, $n, k \in \mathbb{N}$, the following statements are equivalent:*

(1) *X is circular.*
(2) *$\mathscr{G}(X)$ is acyclic.*

Remark 2.4. Theorem 2.3 shows, in particular, that any acyclic simple directed graph is the associated graph of a circular code.

A special class of directed graphs, called *tournaments*, is involved in diletter codes [12]. Recall from graph theory [7] that a graph $\mathscr{G}$ is a *tournament* if it is obtained by assigning an orientation to each edge of a complete (and hence simple) graph. In particular, it has $\frac{|V|(|V|-1)}{2}$ edges. The following lemma follows immediately from Theorem 2.3 and gives us a useful tool for studying diletter circular codes.

Lemma 2.5. *Given a diletter code $X \subseteq A_n^2$, the following statements are equivalent:*

(1) *X is a diletter circular code of maximal length.*
(2) *$\mathscr{G}(X)$ is an acyclic tournament on n vertices.*

Proof. Let X be a diletter circular code of maximal length. Obviously, its corresponding graph $\mathscr{G}(X)$ is embedded in some tournament and, thus, X cannot contain more diletters than the number of edges in a tournament on n vertices. Since for any $n \in \mathbb{N}$ there is an acyclic tournament (compare Theorem 2.7 below) and its corresponding diletter code is circular, the statement follows. The opposite direction is obvious. □

In the next section, the graph representation of circular codes will prove some enumerative combinatorial theorems about circular codes in a very elegant way according to the theory of tournaments. For a vertex v in a graph $G = (V, E)$, the number $d^+(v) = |\{[v, w] \in E: w \in V\}|$ is called the *out-degree* of $v \in V$. Analogously, the number $d^-(v)$ is defined as the *in-degree* of $v \in V$ $d^-(v) = |\{[w, v] \in E: w \in V\}|$.

Thus, we count how many edges start (or end) in each vertex.

Example 2.6. For instance in Fig. 1, the out-degree sequence is $d^+(A) = 1$, $d^+(C) = 1$, $d^+(G) = 1$, $d^+(T) = 2$ and the in-degree sequence is $d^-(A) = 1$, $d^-(C) = 1$, $d^-(G) = 2$, $d^-(T) = 1$.

The following property of the acyclicity of tournaments can be found, for instance, in [7]:

Theorem 2.7. *The following statements are equivalent for a tournament $T = (V, E)$ on n vertices:*

- *T is acyclic.*
- *The set of out-degrees of T (also called score sequence) is $\{0, 1, 2, ..., (n-1)\}$.*
- *T has exactly one Hamiltonian path.*

The graph representation allows an explicit description of diletter circular codes of maximal length.

Theorem 2.8. *Given a code $X \subseteq A_n^2$, the following statements are equivalent:*

(1) *X is a diletter circular code of maximal length.*
(2) *X has the form $X = \{N_iN_j | i, j = 1, ..., n, i < j, N_i, N_j \in A_n, N_i \neq N_j\}$.*

Proof. According to Lemma 2.5, the first condition is equivalent to the condition that $\mathscr{G}(X)$ is an acyclic tournament on n vertices. According to Theorem 2.7, the score sequence of $\mathscr{G}(X)$ is $\{0, 1, 2, ..., (n-1)\}$ which is obviously equivalent to the statement that X has the form $X = \{N_iN_j | i, j = 1, ..., n, i < j, N_i, N_j \in A_n, N_i \neq N_j\}$. Indeed, $N_1 \to N_2 \to ... \to N_n$ is obviously the unique Hamiltonian path in $\mathscr{G}(X)$ (compare Theorem 2.7 above). □

The strong comma-free codes newly studied have the following properties. Any strong comma-free code X is automatically comma-free. Furthermore, the oriented paths of comma-free codes (for a proof see Theorem 2.11 in [12]) and strong comma-free codes have the following properties (obvious extension of proof of Theorem 2.11 in [12]):

Theorem 2.9. *The following assertions hold:*

(I) *Given a code $X \subseteq A_n^k$, the following statements are equivalent:*
  (1) *The maximal length of a path in $\mathscr{G}(X)$ is 2.*
  (2) *The code X is comma-free.*
(II) *Given a code $X \subseteq A_n^k$ the following statements are equivalent:*
  (1) *The maximal length of a path in $\mathscr{G}(X)$ is 1.*
  (2) *The code X is strong comma-free.*

3. Enumerative combinatorics of diletter circular codes

Theorem 3.1 provides the first important result proving that the classes of maximal diletter circular codes and diletter circular codes of maximal length coincide. Recall that the result was already obtained in Proposition 6 in [23] for the particular case of dinucleotide circular codes, i.e. for $n = 4$. For trinucleotide circular codes, the analogous result is not true (Proposition 10 in [23]).

Theorem 3.1. *A maximal diletter circular code over A_n is of maximal length.*

Proof. Let $X \subseteq A_n^2$ be a maximal diletter circular code over the alphabet A_n. Then, the representing graph $\mathscr{G}(X)$ of X is acyclic and, thus, embedded in a tournament on n vertices. Assume that X has not the maximal length. Thus, there are two nodes $N_1, N_2 \in A_n$ which are not connected by an edge in $\mathscr{G}(X)$ and any orientation of the missing edge $[N_1, N_2]$ or $[N_2, N_1]$ leads to a cycle in the extended graph. Assume that we get a cycle $N_1N_2x_1...x_lN_1$ when adding the edge $[N_1, N_2]$ and a cycle $N_2N_1y_1...y_rN_2$ when adding the edge $[N_2, N_1]$. Then, $N_2x_1...x_lN_1y_1...y_rN_2$ is a cycle in $\mathscr{G}(X)$ which contradicts the circularity of X. □

Remark 3.2. In graph theory, the statement 3.1 means that every simple acyclic graph with n vertices can be embedded into some acyclic tournament on n vertices.

We give two formulas for the maximal size of diletter circular codes and diletter strong comma-free codes over the alphabet A_n.

Proposition 3.3. *Let A_n be an alphabet with n letters ($n \in \mathbb{N}$).*

(1) *Any diletter circular code over A_n can contain at most*

$$\frac{n(n-1)}{2}$$

diletters.
(2) *Any diletter strong comma-free code over A_n can contain at most*

$$\begin{cases} \frac{n^2}{4}, & \text{if } n \text{ is even} \\ \frac{(n-1)(n+1)}{4}, & \text{if } n \text{ is odd} \end{cases} \tag{3.1}$$

diletters.

Moreover, the upper bounds given above are sharp in the sense that there are diletter circular codes and diletter strong comma-free codes of the stated sizes.

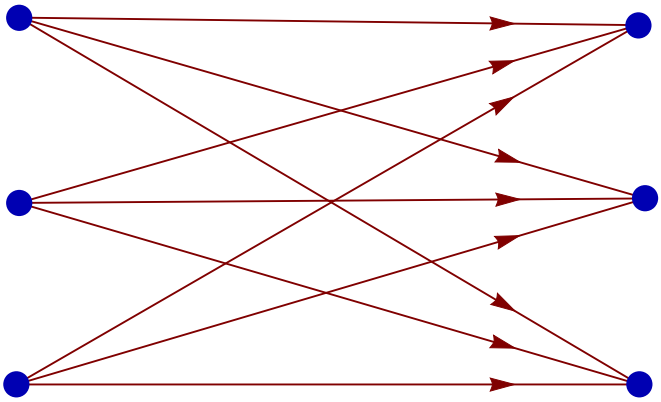


Fig. 2. Maximal diletter strong comma-free codes for $n = 6$.

Proof. The proof can be found in Appendix A □

At this point, we would like to mention that no formula for the maximal size of diletter comma-free codes has been identified so far. However, we will explain below some relation between diletter comma-free codes and diletter strong-comma-free codes. Nevertheless, we have the following remark that sheds light on the structure of diletter strong comma-free codes.

Remark 3.4. The proof of Proposition 3.3 also shows the structure of a maximal diletter strong comma-free code. In fact, there are two cases for the alphabet A_n. If n is even then the alphabet A_n is partitioned into two disjoint sets T^+ and T^- of equal size $\frac{n}{2}$ of vertices of the associated graph. Now, a vertex in T^- has to be connected with a vertex in T^+ in order to obtain a maximal diletter strong comma-free code. Fig. 2 gives an example for the case $n = 6$.

If n is odd, the alphabet A_n is also partitioned into two sets T^- and T^+ but one is of size $\frac{n+1}{2}$ and one is of size $\frac{n-1}{2}$. Fig. 3 shows the two possible cases for $n = 5$.

The relation between maximal diletter comma-free codes and maximal strong comma-free codes is more complicated as, for example, between maximal diletter circular codes and maximal comma-free codes. Firstly, not every maximal diletter comma-free code contains a maximal strong comma-free code. Fig. 4 shows this observation for $n = 4$. Indeed, it is impossible to remove only one edge from the maximal diletter comma-free code to obtain a maximal diletter strong comma-free code since there are two vertices with in-degree and out-degree not equal to zero.

Secondly, not every maximal diletter strong comma-free code can be extended to a maximal diletter comma-free code. In fact, no maximal diletter strong comma-free code on six vertices can be extended to a maximal diletter comma-free code on six vertices (compare Fig. 4 and Tables 2 and 3). Indeed, to pass from a maximal diletter strong comma-free code which has nine edges to a maximal diletter comma-free code which has 12 edges, three edges have to be added. By the structure of maximal diletter strong comma-free codes, these edges have to be chosen among the edges of a maximal Hamiltonian path. However, at most two edges can be added without obtaining a path of length 3.

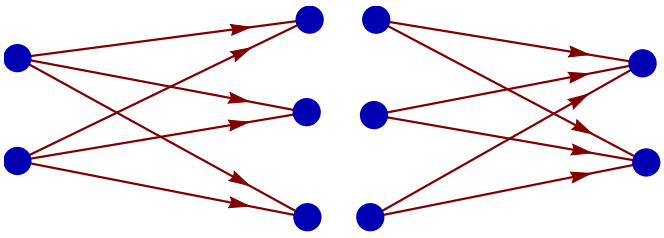


Fig. 3. Two possible maximal diletter strong comma-free codes for $n = 5$.

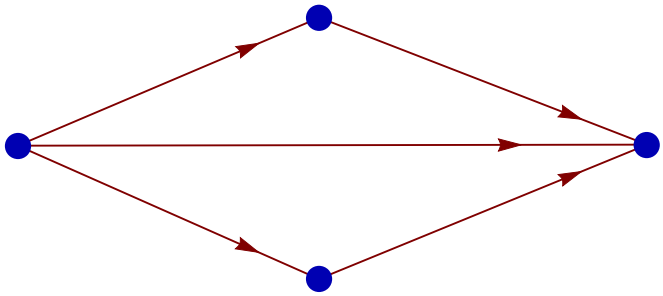


Fig. 4. A maximal diletter comma-free code not containing a maximal diletter strong comma-free code.

Now, we study the growth functions of diletter circular codes over the alphabet A_n in dependence of n, i.e. the numbers of different diletter circular codes of lengths $l \in \mathbb{N}$ for $1 \leq l \leq \frac{n(n-1)}{2}$ for a given $n \in \mathbb{N}$. The mathematical proofs and the computer calculation of the numbers for $n = 1, \ldots, 7$ are in agreement.

Theorem 3.5. *Let*[2] *A_n be an alphabet with n letters ($n \in \mathbb{N}$).*

(1) *The number of different diletter circular as well as comma-free as well as strong comma-free codes of length $l = 1$ over A_n is*

$$n(n-1).$$

(2) *The number of different diletter circular as well as comma-free codes of length $l = 2$ over A_n is*

$$\frac{2n!}{(n-3)!} + \frac{n!}{2(n-4)!}, \quad n \geq 3.$$

The number of different diletter strong comma-free codes of length $l = 2$ over A_n is

$$\frac{n!}{(n-3)!} + \frac{n!}{2(n-4)!}, \quad n \geq 3.$$

(3) *The number of different diletter circular codes of length $l = 3$ over A_n is*

$$\frac{n!}{(n-3)!} + \frac{16n!}{3(n-4)!} + \frac{2n!}{(n-5)!} + \frac{n!}{6(n-6)!}, \quad n \geq 3.$$

The number of different diletter comma-free codes of length $l = 3$ over A_n is

$$\frac{n!}{(n-3)!} + \frac{13n!}{3(n-4)!} + \frac{2n!}{(n-5)!} + \frac{n!}{6(n-6)!}, \quad n \geq 3.$$

The number of different diletter strong comma-free codes of length $l = 3$ over A_n is

$$\frac{4n!}{3(n-4)!} + \frac{n!}{(n-5)!} + \frac{n!}{6(n-6)!}, \quad n \geq 4.$$

(4) *The number of different diletter circular codes of length $l = 4$ over A_n is*

$$\frac{31n!}{4(n-4)!} + \frac{53n!}{3(n-5)!} + \frac{22n!}{3(n-6)!} + \frac{n!}{(n-7)!} + \frac{n!}{24(n-8)!}, \quad n \geq 4.$$

The number of different diletter comma-free codes of length $l = 4$ over A_n is

$$\frac{19n!}{4(n-4)!} + \frac{35n!}{3(n-5)!} + \frac{19n!}{3(n-6)!} + \frac{n!}{(n-7)!} + \frac{n!}{24(n-8)!}, \quad n \geq 4.$$

The number of different diletter strong comma-free codes of length $l = 4$

[2] The formulas below make sense for $l > n$ if we define $\frac{n!}{(n-l)!} = 0$ since $\frac{n!}{(n-l)!} = n(n-1)\ldots(n-n)\ldots(n-(l-1)) = 0$ if $l > n$.

over A_n is

$$\frac{n!}{4(n-4)!}+\frac{25n!}{12(n-5)!}+\frac{11n!}{6(n-6)!}+\frac{n!}{2(n-7)!}+\frac{n!}{24(n-8)!},\quad n\geq 4.$$

Proof. The proof can be found in Appendix B. □

Now, we study the growth function of circular codes from their maximal length. Recall that the particular result for the number of maximal diletter circular codes (i.e. $n!$ in Theorem 3.6 (1) (a)) was already obtained by Pellegrini and Pirillo [26] by a combinatorial approach. Our approach allows to prove the same result in a much simpler way.

Theorem 3.6. *Let A_n be an alphabet with n letters ($n \in \mathbb{N}$).*

(1) *(a) The number of different maximal diletter circular codes (of maximal length $l=\frac{n(n-1)}{2}$) over A_n is*

$n!$.

(b) The number of different diletter circular codes of length $l=\frac{n(n-1)}{2}-1$ over A_n is

$$\frac{n!(n-1)^2}{2}.$$

(2) *The number of different maximal diletter strong comma-free codes over A_n is*

$$\begin{cases}\binom{n}{\frac{n}{2}}=\frac{n!}{\left(\frac{n}{2}!\right)^2}, & \text{if } n \text{ is even}\\ 2\cdot\binom{n}{\frac{n+1}{2}}=\frac{2n!}{\left(\frac{n-1}{2}\right)!\left(\frac{n+1}{2}\right)!}, & \text{if } n \text{ is odd.}\end{cases}$$

Proof. The proof can be found in Appendix C. □

4. Some biological consequences of dinucleotide circular codes

Dinucleotide circular codes constitute an important particular case of diletter circular codes in biology, i.e. with an alphabet A_n where $n = 4$.

4.1. Maximal dinucleotide strong comma-free codes

The list of the 24 (Table 1) maximal (six dinucleotides) dinucleotide circular codes is given in [24]. The list of the 36 (Table 2) maximal (five dinucleotides) dinucleotide comma-free codes is given in [13]. As a new class of circular codes is identified here, we give the list of the six (Table 3) maximal (four dinucleotides) dinucleotide strong comma-free codes in Table 4. A maximal dinucleotide strong comma-free code has the form $X = \{N_1N_2, N_1N_3, N_4N_2, N_4N_3\}$ where $N_i \in \mathscr{B}$, $i = 1, 2, 3, 4$ and $N_i \neq N_j$, $i \neq j$.

4.2. Dinucleotide circular codes embedded in the circular code X in genes

New properties of the circular code X observed in genes are identified here. Let $M_{1,2}$ ($M_{1,3}$ and $M_{2,3}$, respectively) be the dinucleotide multisets associated with the three trinucleotide sites l_1l_2 (l_1l_3 and l_2l_3, respectively) of the maximal C^3 self-complementary trinucleotide circular code X. Then, by inspection of Table 5

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

$M_{1,3}$ = {AC,AC,AC,AT,AT,CC,CG,CG,GA,GA,GC,GC,GC,GC,GG,GT,GT,GT,TC,TC},

Table 1
Growth functions of diletter circular codes for an alphabet A_n with $n = 2, \ldots, 7$.

Code cardinality	A_2	A_3	A_4	A_5	A_6	A_7
1	2	6	12	20	30	42
2		12	60	180	420	840
3		6	152	940	3600	10,570
4			186	3050	20,790	93,030
5			108	6180	83,952	601,944
6			24	7960	240,480	2,934,568
7				6540	496,680	10,931,760
8				3330	750,810	31,528,980
9				960	838,130	71,331,470
10				120	691,020	128,047,374
11					416,160	183,963,066
12					178,230	212,642,360
13					51,480	198,066,960
14					9000	148,296,330
15					720	88,602,780
16						41,683,950
17						15,107,400
18						4,071,480
19						768,600
20						90,720
21						5040

Table 2
Growth functions of diletter comma-free codes for an alphabet A_n with $n = 2, \ldots, 7$.

Code cardinality	A_2	A_3	A_4	A_5	A_6	A_7
1	2	6	12	20	30	42
2		12	60	180	420	840
3		6	128	820	3240	9730
4			114	1970	14,670	70,350
5			36	2580	40,692	331,884
6				1840	71,760	1,059,688
7				660	82,440	2,365,440
8				90	62,070	3,787,560
9					30,110	4,427,150
10					8940	3,810,954
11					1440	2,414,706
12					90	1,113,140
13						363,300
14						79,590
15						10,500
16						630

Table 3
Growth functions of diletter strong comma-free codes for an alphabet A_n with $n = 2, \ldots, 7$.

Code cardinality	A_2	A_3	A_4	A_5	A_6	A_7
1	2	6	12	20	30	42
2		6	36	120	300	630
3			32	280	1320	4480
4			6	280	2910	17,220
5				120	3492	39,144
6				20	2400	56,294
7					960	53,760
8					210	35,070
9					20	15,680
10						4662
11						840
12						70

$M_{2,3}$ = {AA,AC,AC,AC,AG,AG,AT,AT,CC,CC,GC,GT,TA,TC,TC,TC,TC,TG,TT,TT}.

As the circular code X is self-complementary, i.e. $X = \overleftarrow{c(X)}$,[3] then

[3] Recall (compare, for instance, [13,24]) that, in the dinucleotide case, the mapping $\overleftarrow{c}$ is defined as $\overleftarrow{c(N_1N_2)} = c(N_2)c(N_1)$ where $N_1, N_2 \in \mathscr{B}$ and c: $A, C, G, T \to T, G, C, A$ is the complementing transformation of the nucleotide bases.

Table 4
The six maximal dinucleotide strong comma-free codes.

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

Table 5
Dinucleotides (type and number) in the three trinucleotide sites l_1l_2, l_1l_3 and l_2l_3 of the maximal C^3 self-complementary trinucleotide circular code X observed in genes.

Dinucleotide	l_1l_2	l_1l_3	l_2l_3
AA	2	0	1
AC	1	3	3
AG	0	0	2
AT	2	2	2
CA	1	0	0
CC	0	1	2
CG	0	2	0
CT	2	0	0
GA	4	2	0
GC	1	4	1
GG	2	1	0
GT	3	3	1
TA	1	0	1
TC	0	2	4
TG	0	0	1
TT	1	0	2

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

In order to identify dinucleotide circular codes embedded in the circular code X, the periodic dinucleotides $\{AA,CC,GG,TT\}$ and the permuted dinucleotides of low occurrence are excluded in the dinucleotide multisets $M_{1,2}$, $M_{1,3}$ and $M_{2,3}$ leading to the dinucleotide sets $D_{1,2}$, $D_{1,3}$ and $D_{2,3}$, respectively.

For the multiset $M_{1,3}$, by excluding the two periodic dinucleotides $\{CC,GG\}$ having the lowest occurrence (1 in Table 5) and the permuted dinucleotide CG with an occurrence number equal to 2 less than 4 with GC, the following dinucleotide set $D_{1,3}$ is identified

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

The set $D_{1,3}$ is a maximal self-complementary ($D_{1,3} = \overleftarrow{c(D_{1,3})}$) dinucleotide circular code (see Theorem 2.8; Proposition 29 and Table 1 in [11,24]).

For the multiset $M_{1,2}$, by excluding the three periodic dinucleotides $\{AA,GG,TT\}$ and the permuted dinucleotide TA with an occurrence number equal to 1 less than 2 with AT, two dinucleotide sets $D_{1,2}$ and $D'_{1,2}$ can be identified

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

and

$$D'_{1,2} = \{\text{AT,CA,CT,GA,GC,GT}\}.$$

The sets $D_{1,2}$ and $D'_{1,2}$ are also maximal dinucleotide circular codes.

For the multiset $M_{2,3}$, by applying complementarity, two dinucleotide sets $D_{2,3}$ and $D'_{2,3}$ are identified

$$D_{2,3} = \overleftarrow{c(D_{1,2})} = \{\text{AC,AG,AT,GC,GT,TC}\}$$

and

$$D'_{2,3} = \overleftarrow{c(D'_{1,2})} = \{\text{AC,AG,AT,GC,TC,TG}\}.$$

Consequently, the sets $D_{2,3}$ and $D'_{2,3}$ are also maximal dinucleotide circular codes. Note that $D_{1,2}$ and $D'_{1,2}$ are not related by permutation with $D_{2,3}$ and $D'_{2,3}$, respectively.

The two self-complementary dinucleotides AT and GC belong to the five identified circular codes $D_{1,3}$, $D_{1,2}$, $D'_{1,2}$, $D_{2,3}$ and $D'_{2,3}$ while the two other self-complementary dinucleotides TA and CG do not belong to them, a property also in agreement with their occurrence in Table 5. Thus, genes which are read in the $5' - 3'$ direction may have selected preferentially self-complementary dinucleotides of the type RY (purine $R = \{A,G\}$ and pyrimidine $Y = \{C,T\}$) in these codes.

4.3. A biological hypothesis: the periodicity modulo 2 in introns modelled by dinucleotide circular codes

Several statistical studies based on, for instance, the correlation function, the power spectrum, etc., have identified a nucleotide periodicity modulo 3 in genes. This signal is associated with the particular structure of genes constituted of words of the same length equal to 3 nucleotides, i.e. a series of trinucleotides. This classical result has been analysed in the past by several authors, in particular at the sequence level by Shepherd [27] and at the population level by Fickett [10] and Michel [19] Fig. 1). In 1986, it was shown that introns, in contrast to exons, have no nucleotide periodicity modulo 3 (Fig. 2 in [19], with a statistical analysis of 90 introns). One year after, with the increase in sequence data, a nucleotide periodicity modulo 2 was identified in introns (eukaryotes, viruses, mitochondria) by two different statistical methods [3,18].

Over the last 20 years, the nucleotide periodicity moduo 3, important as it is associated with genes, has been modelled, in particular by using the trinucleotide circular code X identified in genes. Several classes of models have been developed, by computer simulation [1], analytically with constant substitution rates [2] and numerically with chaotic substitution rates [5]. All these biomathematical models which explain perfectly the nucleotide periodicity modulo 3 in genes are based on two processes: a construction process of genes based on an independent mixing of the 20 trinucleotides (classically with equiprobability 1/20) of the circular code X followed by an evolutionary process with mutation (random substitutions, insertions and deletions).

Similarly to the modeling of the periodicity modulo 3 with the trinucleotide circular code X, the periodicity modulo 2 may be investigated by using dinucleotide circular codes, in particular with the class of the 24 maximal dinucleotide circular codes.

4.4. An evolutionary hypothesis of circular codes

Dinucleotide circular codes allow to retrieve the reading (correct) frame modulo 2 with at most 5 nucleotides. Trinucleotide circular codes allow to retrieve the reading frame modulo 3 with at most 13 nucleotides. The properties of these two classes of circular codes related to the reading frame retrieval, the cardinality range and the numbers of codes allow us to propose an evolutionary hypothesis of circular codes (Fig. 5).

According to this hypothesis, the evolution of circular codes is based on an increase in combinatorial flexibility, from the dinucleotide circular codes to the trinucleotide circular codes which contain the greatest number of codes and, in addition, the longest nucleotide window of reading frame retrieval.

Pure and long repeated dinucleotides and trinucleotides are very common in genomes (non-coding regions) of eukaryotes (e.g. [6,14,17]). We give here a few examples of pure and long repeated dinucleotides observed in genomes: $(AC)^{2250}$, $(AG)^{3824}$ and $(CT)^{3616}$ in *C. sinensis*, $(AT)^{5627}$ in *M. truncatula*, $(CG)^{216}$ in *C. sativus* and $(GT)^{1340}$ in *B. vulgaris*, and pure and long repeated trinucleotides: $(AAC)^{2885}$, $(GAA)^{3512}$, $(GTC)^{605}$ and $(TAC)^{576}$ in *S. pennellii*, $(AAT)^{4317}$, $(ATT)^{6425}$ and $(GAT)^{692}$ in *C. sinensis*, $(ACC)^{121}$, $(GCC)^{185}$ and $(GGC)^{70}$ in *O. brachyantha*, $(ATC)^{715}$ in *C. sativa*, $(CAG)^{1185}$, $(CTC)^{355}$, $(CTG)^{241}$ and $(GAG)^{274}$ in *F. albicollis*, $(GAC)^{48}$ in *B. terrestris*, $(GGT)^{210}$ in *H. sapiens*, $(GTA)^{642}$ in *Z. mays*, $(GTT)^{1413}$ and $(TTC)^{1421}$ in *C. arietinum*, etc. (from

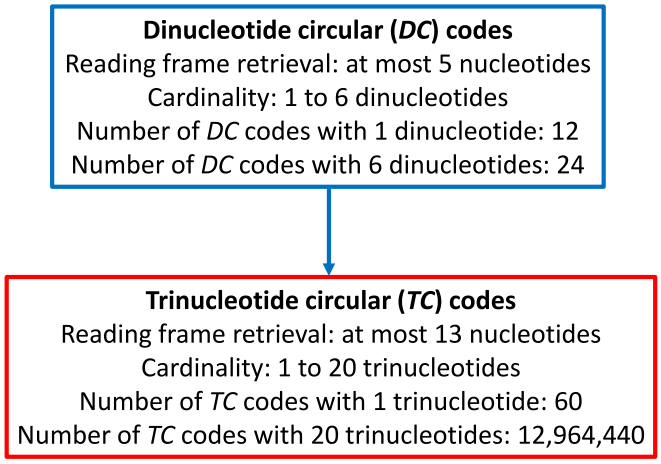


Fig. 5. An evolutionary hypothesis of circular codes.

Table 2 in [29]).

In the protein coding regions, 61 trinucleotides (without the 3 stop trinucleotides) are used. The code associated to these 61 trinucleotides is obviously not circular, i.e. the genetic code is not circular. However in genes, there is a preferential occurrence of a set of 20 trinucleotides which is a circular code.

Non-coding regions that represent about 90% (for an order of magnitude) of an eukaryotic genome contain different DNA structures along the chromosome: pseudogenes, RNA-coding genes, tandem repeats: minisatellites and microsatellites, retrotransposons: long terminal repeats (LTR), non-long terminal repeats (Non-LTR), long interspersed elements (LINE) and short interspersed elements (SINE), DNA transposons, etc. Surprisingly, some local structures in a genome can be associated to a circular code. For example, a repeated dinucleotide is generated by a dinucleotide circular code. Thus, a repeated dinucleotide has the property of retrieval of a modulo 2 frame. The dinucleotide repeats are highly enriched in enhancers which are genomic elements involved in gene expression [28]. However, how enhancer sequences relate to enhancer activity is unknown and remains an important open question in today's biology. We propose that the property of circular code with the repeated dinucleotides may be involved in an enhancer function.

From a theoretical perspective, which has been ignored so far, these pure and long repeated dinucleotides and trinucleotides are in fact motifs generated from dinucleotide and trinucleotide circular codes respectively. The expansion of dinucleotide circular codes to trinucleotide circular codes is a combinatorial problem which is open and difficult.

5. Conclusion

The graph approach of circular codes recently developed [12] allows here a detailed study of diletter circular codes over finite alphabets. A new class of circular codes is identified, strong comma-free codes (Definition 2.2). New theorems are proved with the diletter circular codes of maximal length in relation to (i) a characterization of their graphs as acyclic tournaments (Lemma 2.5); (ii) their explicit description (Theorem 2.8); and (iii) the non-existence of other maximal diletter circular codes (Theorem 3.1). The maximal lengths of paths in the graphs of comma-free and strong comma-free codes are determined (Theorem 2.9). Diletter circular codes are enumerated over finite alphabets (Proposition 3.3, Theorems 3.5 and 3.6). A biological hypothesis is proposed, based on using dinucleotide circular codes, to model and explain the periodicity modulo 2 observed in introns.

Appendix A. Proof of Proposition 3.3

Proof. We treat the cases separately.

(1) A maximal diletter circular code can be represented as an acyclic tournament on n vertices (see Lemma 2.5). Such a tournament contains $\frac{n(n-1)}{2}$ edges.

(2) Let us denote by $l_{max}(n)$ the maximal possible number of edges in a graph of a diletter strong comma-free code over A_n. According to Theorem 3.1, such a graph is a subgraph of an acyclic tournament on n vertices. Obviously, for each vertex v in such a graph either its out-degree $d^+(v) = 0$ or its in-degree $d^-(v) = 0$ due to the strong comma-freeness and the maximality of the code. We will show that for all $n \geq 2$ the following equality holds:

$$l_{max}(n) = l_{max}(n-2) + (n-2) + 1 \quad (*). \tag{A.1}$$

Let G be an acyclic tournament on n vertices with the unique Hamiltonian path $N_1 \to N_2 \to \ldots \to N_n$ with $N_i \in A_n$, $N_i \neq N_j$, $i \neq j$. Remove N_1 and N_n together with all corresponding edges from G that are connected to N_1 or N_n. We obtain an acyclic tournament on the $n-2$ vertices $N_2, \ldots, N_{n-1}$. Let G^* be a subgraph of $G \setminus \{N_1, N_n\}$ of a maximal diletter strong comma-free code on $N_2, \ldots, N_{n-1}$ with $l_{max}(n-2)$ edges. In order to extend this graph to a graph of a maximal diletter strong comma-free code over A_n, every vertex $v \in \{N_2, \ldots, N_{n-1}\}$ have to be connected with N_1 ($N_1 \to v$) if its out-degree $d^+(v) = 0$ or with N_n ($v \to N_n$) if its in-degree $d^-(v) = 0$. Additionally, the edge $N_1 \to N_n$ can always be added since no additional directed paths of length 2 can occur this way. Thus, the following inequality

$$l_{max}(n) \geq l_{max}(n-2) + (n-2) + 1$$

holds true. Let us consider now a subgraph of G of a maximal diletter strong comma-free code and remove the vertices N_1 and N_n and its corresponding edges. Thus, exactly $n-1$ edges are removed since no edge can have a connection to N_1 as well as to N_n due to the strong comma-freeness of the code. Thus, the following inequality

$$l_{max}(n) \leq l_{max}(n-2) + (n-2) + 1$$

is true and it proves Equality A.1. Finally, we prove Formula 3.1 per induction using (*) and the obvious facts

$$l_{max}(2) = 1,\ l_{max}(3) = 2.$$

□

Appendix B. Proof of Theorem 3.5

Proof.

(1) We have $\frac{n(n-1)}{2}$ possibilities to choose two vertices from n and two possibilities to orient the edge.

(2) First of all, let us consider the number of possibilities to draw a graph with two edges on n vertices without taking into account the orientation of the edges. We have the two possible cases represented by the following figures.

Fig. 6(a) shows that there are $\frac{n(n-1)}{2}$ possibilities to choose two outer vertices from n and then, $(n-2)$ possibilities to choose the vertices in the middle. Alltogether, there are $\frac{n!}{2(n-3)!}$ possibilities to draw the graph without orientation. (i) For the circular or comma-free codes, the edges can be oriented arbitrarily since no oriented cycle or a path of length more than 2 can be construct. So, there are $\frac{2n!}{(n-3)!}$ different oriented graphs since we have four possibilities to orient the edges. (ii) For the diletter strong comma-free codes, the edges can be oriented only according two possibilities as the oriented paths of length 2 are forbidden. So, there are $\frac{n!}{(n-3)!}$ different oriented graphs.

Fig. 6(b) shows that there are $2\cdot\frac{n(n-1)(n-2)(n-3)}{2\cdot 2}$ possibilities to construct such a non-oriented graph. For the circular, comma-free and strong comma-free codes, there are four possibilities to orient two egdes. So, there are $\frac{n!}{2(n-4)!}$ different oriented graphs.

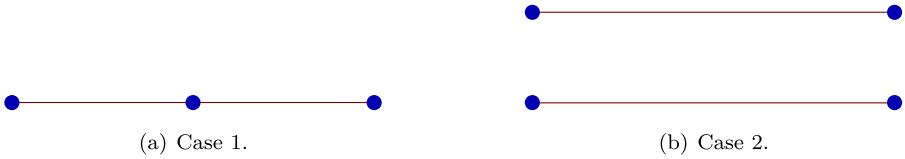
(3) Three edges of a simple graph can connect to at least three or at most six vertices.

Fig. 7(a) shows that there are $\frac{n!}{(n-3)!3!}$ possibilities to construct such a non-oriented graph on n vertices. (i) For the circular or comma-free codes, there are six possible edge orientations, two edge orientations leading to an oriented cycle being excluded. So, there are $\frac{n!}{(n-3)!}$ different oriented graphs. (ii) For the strong comma-free codes, no edge orientation is possible.

Fig. 7(b) shows that there are $\frac{2n!(n-2)!}{(n-2)!(n-4)!2!2!}=\frac{n!}{2(n-4)!}$ possibilities to construct such a non-oriented graph on n vertices. (i) For the circular codes, there are eight possible edge orientations leading to $\frac{4n!}{(n-4)!}$ different oriented graphs. (ii) For the comma-free codes, there are six possible edge orientations, two orientations leading to a directed path of length 3 being excluded, leading to $\frac{3n!}{(n-4)!}$ different oriented graphs. (iii) For the strong comma-free codes, there are two possible edge orientations leading to $\frac{n!}{(n-4)!}$ different oriented graphs.

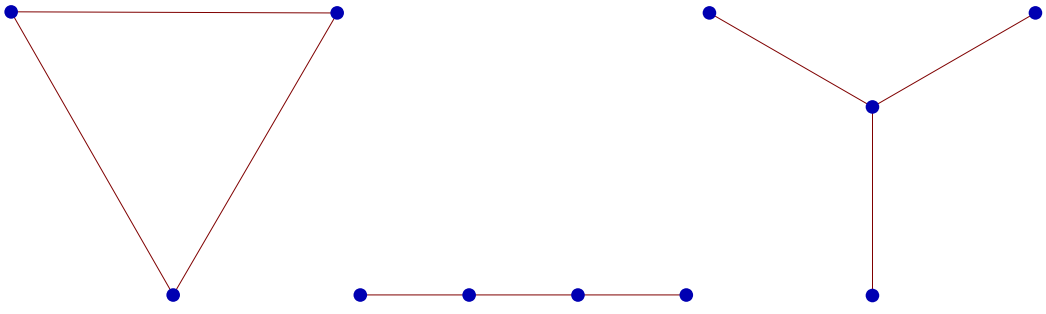
Fig. 7(c) shows that there are $\frac{n(n-1)!}{(n-4)!3!}=\frac{n!}{6(n-4)!}$ possibilities to construct such a non-oriented graph on n vertices. (i) For the circular or comma-free codes, there are eight possible edge orientations leading to $\frac{4n!}{3(n-4)!}$ different oriented graphs. (ii) For the strong comma-free codes, there are two possible edge orientations (three edges oriented to the common vertex or from it) leading to $\frac{n!}{3(n-4)!}$ different oriented graphs.

Fig. 7(d) shows that there are $\frac{n!(n-1)(n-3)!}{(n-2)!2!(n-5)!2!}=\frac{n!}{4(n-5)!}$ possibilities to construct such a non-oriented graph on n vertices. (i) For the circular or comma-free codes, there are eight possible edge orientations leading to $\frac{2n!}{(n-5)!}$ different oriented graphs. (ii) For the strong comma-free codes, there are four possible edge orientations, the sole edge being oriented arbitrary and the path of length 2 in two ways to avoid a directed path of



(a) Case 1. (b) Case 2.

Fig. 6. A graph with two edges.



(a) Case 1. (b) Case 2. (c) Case 3.

(d) Case 4. (e) Case 5.

Fig. 7. A graph with three edges.

length 2, leading to $\frac{n!}{(n-5)!}$ different oriented graphs.

Fig. 7(e) shows that there are $\frac{n!(n-2)!(n-4)!}{(n-2)!2!(n-4)!2!(n-6)!2!3!} = \frac{n!}{48(n-6)!}$ possibilities to construct such a non-oriented graph on n vertices. For the circular, comma-free and strong comma-free codes, there are eight possible edge orientations leading to $\frac{n!}{6(n-6)!}$ different oriented graphs. Adding all the numbers for each type of code, we obtain the formulas above.

(4) The formulas can be proven analoguously to the claim (3). □

Appendix C. Proof of Theorem 3.6

Proof. We treat the cases separately.

(1)

(a) According to Lemma 2.5 and Theorem 3.1, the graph $\mathscr{G}(X)$ of a maximal diletter circular code X over an alphabet A_n with n elements is an acyclic tournament on n vertices. By Theorem 2.7, $\mathscr{G}(X)$ has a unique Hamiltonian path. Since there are n elements there are exactly $n!$ many ways to choose such an Hamiltonian path. Thus, there are at most $n!$ maximal diletter circular codes over A_n. Since all results used above are equivalences there are exactly $n!$ such codes.

(b) According to Theorem 3.1, a diletter circular code of length $\frac{n(n-1)}{2} - 1$ can be embedded into a diletter circular code of maximal length $\frac{n(n-1)}{2}$. Thus, a diletter circular code of length $\frac{n(n-1)}{2} - 1$ corresponds to an acyclic graph with $\frac{n(n-1)}{2} - 1$ edges which can be obtained by removing an edge from an acyclic tournament on n vertices. Consequently, there are $\frac{n(n-1)}{2}$ different diletter circular codes of length $\frac{n(n-1)}{2} - 1$ from one maximal diletter circular code. The question now is when the same diletter circular code of length $\frac{n(n-1)}{2} - 1$ can be obtained from two different maximal diletter circular codes? Each acyclic tournament has a unique Hamiltonian path e.g. $N_1 \to N_2 \to \ldots \to N_n$ and the remaining edges can be oriented only in one way such that there is no cycle ($N_i \to N_j$, $i < j$). So, if an edge which does not belong to the Hamiltonian path is removed then the diletter circular code of length $\frac{n(n-1)}{2} - 1$ can be obtained only from one maximal diletter circular code. If an edge is removed from the Hamiltonian path then the diletter circular code can be obtained from two different maximal diletter circular codes since an edge from the Hamiltonian path can be oriented in another way by keeping an acyclic tournament. Thus, there are

$$n!\left(\frac{n(n-1)}{2} - (n-1) + \frac{n-1}{2}\right) = \frac{n!(n-1)^2}{2}$$

different diletter circular codes of length $\frac{n(n-1)}{2} - 1$ since there are $n!$ different maximal diletter circular codes according to the claim above.

(2) **n is even**: Then, the number of vertices with in-degree equal to zero is equal to the number of vertices with out-degree equal to zero. The order within each class of vertices (out-degree equal to zero or in-degree equal to zero) leads to the same strong comma-free code. So, the number of combinations to choose $\frac{n}{2}$ vertices from n is $\frac{n!}{\left(\frac{n}{2}!\right)^2}$.

n is odd: Then, the number of vertices with in-degree equal to zero is one more or less as the number of vertices with the out-degree equal to zero, i.e. the numbers are $\frac{n-1}{2}$ and $\frac{n+1}{2}$. Thus, $\frac{n-1}{2}$ vertices from n must be chosen without considering the order of the vertices and then, the set representing the vertices with in-degree equal to zero leads to a factor 2. Thus, the number of different maximal diletter strong comma-free codes is $2\cdot\binom{n}{\frac{n+1}{2}}$. □

References

[1] D.G. Arquès, J.-P. Fallot, C.J. Michel, An evolutionary model of a complementary circular code, J. Theor. Biol. 185 (1997) 241–253.

[2] D.G. Arquès, J.-P. Fallot, C.J. Michel, An evolutionary analytical model of a complementary circular code simulating the protein coding genes, the 5′ and 3′ regions, Bull. Math. Biol. 60 (1998) 163–194.

[3] D.G. Arquès, C.J. Michel, Periodicities in introns, Nucleic Acids Res. 15 (1987) 7581–7592.

[4] D.G. Arquès, C.J. Michel, A complementary circular code in the protein coding genes, J. Theor. Biol. 182 (1996) 45–58.

[5] J.M. Bahi, C.J. Michel, A stochastic model of gene evolution with chaotic mutations, J. Theor. Biol. 255 (2008) 53–63.

[6] A. Canapa, P.N. Cerioni, M. Barucca, E. Olmo, V. Caputo, A centromeric satellite DNA may be involved in heterochromatin compactness in gobiid fishes, Chromosome Res. 10 (2002) 297–304.

[7] J. Clark, D.A. Holton, A First Look at Graph Theory, World Scientific, Singapore, 1991.

[8] F. Crick, J.S. Griffith, L.E. Orgel, Codes without commas, Proc. Natl. Acad. Sci. U.S.A. 43 (1957) 416–421.

[9] K. El Soufi, C.J. Michel, Circular code motifs in the ribosome decoding center, Comput. Biol. Chem. 52 (2014) 9–17.

[10] J.W. Fickett, Recognition of protein coding regions in DNA sequences, Nucleic Acids Res. 10 (1982) 5303–5318.

[11] E. Fimmel, S. Giannerini, D. Gonzalez, L. Strüngmann, Dinucleotide circular codes and bijective transformations, J. Theor. Biol. 386 (2015) 159–165.

[12] E. Fimmel, C.J. Michel, L. Strüngmann, *n*–Nucleotide circular codes in graph theory, Philosoph. Trans. R. Soc. A 374 (2016) 20150058.

[13] E. Fimmel, L. Strüngmann, Maximal dinucleotide comma-free codes, J. Theor. Biol. 389 (2016) 206–213.

[14] R. Gemayel, M.D. Vinces, M. Legendre, K.J. Verstrepen, Variable tandem repeats accelerate evolution of coding and regulatory sequences, Annu. Rev. Genet. 44 (2010) 445–477.

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

[16] S.W. Golomb, B. Gordon, L.R. Welch, Comma-free codes, Can. J. Math. 10 (1958) 202–209.

[17] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (2004) 435–445.

[18] A.K. Konopka, G.W. Smythers, DISTAN - a program which detects significant distances between short oligonucleotides, Bioinformatics 3 (1987) 193–201.

[19] C.J. Michel, New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation, J. Theor. Biol. 120 (1986) 223–236.

[20] C.J. Michel, Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes, Comput. Biol. Chem. 37 (2012) 24–37.

[21] C.J. Michel, The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses, J. Theor. Biol. 380 (2015) 156–177.

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

[23] C.J. Michel, M. Pellegrini, G. Pirillo, Maximal dinucleotide and trinucleotide circular codes, J. Theor. Biol. 389 (2016) 40–46.

[24] C.J. Michel, G. Pirillo, Dinucleotide circular codes, ISRN Biomathematics, (2013), p. 538631.

[25] C.J. Michel, G. Pirillo, M.A. Pirillo, Varieties of comma free codes, Comput. Math. Appl. 55 (2008) 989–996.

[26] M. Pellegrini, G. Pirillo, On the dinucleotide circular codes of maximal cardinality, Theor. Biol. Forum 107 (2014) 89–95.
[27] J.C.W. Shepherd, Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification, Proc. Natl. Acad. Sci. U.S.A. 78 (1981), pp. 1596–1600.
[28] J.O. Yáñez-Cuna, C.D. Arnold, G. Stampfel, Ł.M. Boryń, D. Gerlach, M. Rath, A. Stark, Dissection of thousands of cell type–specific enhancers identifies dinucleotide repeat motifs as general enhancer features, Genome Res. 24 (2014) 1147–1156.
[29] K. El Soufi, C.J. Michel, Unitary circular code motifs in genomes of eukaryotes, Biosystems 153 (2017) 45–62.