Information and Computation 212 (2012) 55–63

Contents lists available at SciVerse ScienceDirect

Information and Computation

www.elsevier.com/locate/yinco

A classification of 20-trinucleotide circular codes

Christian J. Michel [a], Giuseppe Pirillo [b,c,*], Mario A. Pirillo [d]

[a] Equipe de Bioinformatique Théorique, BFO, LSIIT (UMR 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France
[b] CNR, IASI, Unità di Firenze, Dipartimento di Matematica "U. Dini", viale Morgagni 67/A, 50134 Firenze, Italy
[c] Université de Marne-la-Vallée, 5 boulevard Descartes, 77454 Marne-la-Vallée Cedex 2, France
[d] Istituto Statale SS. Annunziata, Piazzale del Poggio Imperiale, 50134 Firenze, Italy

ARTICLE INFO

Article history:
Received 20 April 2010
Revised 1 December 2011
Available online 28 December 2011

Keywords:
Code
Circular
Trinucleotide
Hierarchy
Classes of codes
Biological computation
Computational biology
Computational complexity

ABSTRACT

Trinucleotide comma-free codes and trinucleotide circular codes are two important classes of codes in code theory and theoretical biology. A trinucleotide circular code containing exactly 20 elements is called here a 20-trinucleotide circular code. In this paper, solving a combinatorial problem of hard computational complexity, we extend and improve our results of C.J. Michel, G. Pirillo, and M.A. Pirillo (2008) [14] concerning the small class of 528 self-complementary 20-trinucleotide circular codes, to the complete class of the 20-trinucleotide circular codes which contains 12,964,440 elements. A surprising relation with the symmetric group Σ_4 appears but it remains unexplained so far.

1. Introduction

We continue our study of the combinatorial properties of trinucleotide circular codes. A trinucleotide is a word of three letters (triletter) on the genetic alphabet $\{A, C, G, T\}$. For 50 years, codes, comma-free codes and circular codes have been mathematical objects studied in theoretical biology, mainly to understand the structure and the origin of the genetic code as well as the reading frame (construction) of genes, e.g. [5–7]. In order to have an intuitive meaning of these notions, codes are written on a straight line while comma-free codes and circular codes are written on a circle, but in both cases, unique decipherability is required.

The genetic code based on 64 trinucleotides is a code in the sense of language theory, more precisely a uniform code [4], but not a circular code [10] (see Remark 2 below). Before the discovery of the genetic code, Crick et al. [5] proposed a maximal comma-free code of 20 trinucleotides for coding the 20 amino acids. In 1996, a maximal circular code X_0 of 20 trinucleotides was identified statistically on a large gene population of eukaryotes and also on a large gene population of prokaryotes [1]:

$$\begin{aligned} X_0 = \{&AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC\\ &GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}. \end{aligned}$$

This code X_0 has remarkable properties. For example, X_0 is self-complementary: 10 trinucleotides are complementary to the 10 other trinucleotides, e.g. AAC is complementary to GTT, AAT to ATT, etc. The two sets of 20 trinucleotides, called X_1

* Corresponding author at: CNR, IASI, Unità di Firenze, Dipartimento di Matematica "U. Dini", viale Morgagni 67/A, 50134 Firenze, Italy.
E-mail addresses: michel@dpt-info.u-strasbg.fr (C.J. Michel), pirillo@math.unifi.it (G. Pirillo), map@conmet.it (M.A. Pirillo).

0890-5401/$ – see front matter
doi:10.1016/j.ic.2011.12.001

and X_2, obtained by a simple shift operation of X_0, one and two letters respectively, are also maximal circular codes [1]. This surprising result, still mysterious, was discussed in research works in mathematics/computer science and theoretical biology, e.g. [9,3,2,18,8,15,12,11,17]. Therefore, the mathematical study of trinucleotide circular codes is particularly important in theoretical biology as well as in code theory.

In this paper, a trinucleotide circular code containing exactly 20 elements is called a 20-trinucleotide circular code.

Recently, we described varieties of 20-trinucleotide comma-free codes [13]. Then, we proposed a hierarchy relation based on chains of inclusions between comma-free codes and circular codes. More precisely, all the trinucleotide codes in this hierarchy are circular, the strongest ones being comma-free [14]. In particular, we studied the case of the small class of the self-complementary 20-trinucleotide circular codes of cardinality 528. Here, we generalize our hierarchy relation to the case of the entire class of the 20-trinucleotide circular codes of cardinality $12,964,440$. Moreover, we identify some interesting equalities (Proposition 8).

In other words, solving a combinatorial problem of hard computational complexity, we extend and improve here our particular results of [14] to the class of all (maximal) 20-trinucleotide circular codes. Finally, we point out that Proposition 9 allows a computational calculus in order to determine the numbers of all (maximal) 20-trinucleotide circular codes in the different classes of the identified mathematical hierarchy.

2. Preliminaries

We refer the reader to [4] for the classical notions of an alphabet, empty word, length, factor, proper factor, prefix, proper prefix, suffix, proper suffix. Let $\mathcal{A}$ denote a finite alphabet and let $\mathcal{A}^*$ denote the set of all words over $\mathcal{A}$. Given a subset X of $\mathcal{A}^*$, X^n is the set of the words over $\mathcal{A}$ which are the product of n words from X, i.e. $X^n = \{x_1 x_2 \cdots x_n \mid x_i \in X\}$.

There is a correspondence between the genetic and language-theoretic concepts. The letters (or nucleotides or bases) define the genetic alphabet $\mathcal{A}_4 = \{A, C, G, T\}$. The set of non-empty words (resp. words) over $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^+$ (resp. $\mathcal{A}_4^*$). The set of the 16 words of length 2 (or dinucleotides or diletters) is denoted by $\mathcal{A}_4^2$. The set of the 64 words of length 3 (or trinucleotides or triletters) is denoted by $\mathcal{A}_4^3$. The total order over the alphabet $\mathcal{A}_4$ is $A < C < G < T$. Consequently, $\mathcal{A}_4^+$ is lexicographically ordered: given two words $u, v \in \mathcal{A}_4^+$, u is smaller than v in lexicographical order, written $u < v$, if and only if either u is a proper prefix of v or there exist $x, y \in \mathcal{A}_4$, $x < y$, and $r, s, t \in \mathcal{A}_4^*$ so that $u = rxs$ and $v = ryt$.

2.1. Two genetic maps

Definition 1. The complementary map $\mathcal{C}$: $\mathcal{A}_4^+ \to \mathcal{A}_4^+$ is defined by $\mathcal{C}(A) = T$, $\mathcal{C}(T) = A$, $\mathcal{C}(C) = G$ and $\mathcal{C}(G) = C$ and by $\mathcal{C}(uv) = \mathcal{C}(v)\mathcal{C}(u)$ for all $u, v \in \mathcal{A}_4^+$. For example, $\mathcal{C}(AAC) = GTT$. This map $\mathcal{C}$ is associated to the property of the complementary and antiparallel (one DNA strand chemically oriented in a $5' - 3'$ direction and the other DNA strand, in the opposite $3' - 5'$ direction) double helix. This map on words is naturally extended to word sets: a complementary trinucleotide set is obtained by applying the complementary map $\mathcal{C}$ to all its trinucleotides.

Moreover, the map $\mathcal{C}$ is involutional, i.e. for each trinucleotide set X, $X = \mathcal{C}(\mathcal{C}(X))$. More precisely, the map $\mathcal{C}$ is an involutional antiisomorphism.

Definition 2. The circular permutation map $\mathcal{P}$: $\mathcal{A}_4^3 \to \mathcal{A}_4^3$ permutes circularly each trinucleotide $l_1 l_2 l_3$ as follows $\mathcal{P}(l_1 l_2 l_3) = l_2 l_3 l_1$. For example, $\mathcal{P}(AAC) = ACA$. The kth iterate of $\mathcal{P}$ is denoted $\mathcal{P}^k$. This map on words is also naturally extended to word sets: a permuted trinucleotide set is obtained by applying the circular permutation map $\mathcal{P}$ to all its trinucleotides.

Remark 1. Two trinucleotides u and v are *conjugate* if there exist two words s and t such that $u = st$ and $v = ts$. Therefore, if u and v satisfy $\mathcal{P}^k(u) = v$ for some k, then u and v are conjugate.

2.2. Codes, trinucleotide comma-free codes and trinucleotide circular codes

Definition 3. Code: A set X of words is a code if, for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geqslant 1$, the condition $x_1 \cdots x_n = x'_1 \cdots x'_m$ implies $n = m$ and $x_i = x'_i$ for $i = 1, \ldots, n$.

The set $\mathcal{A}_4^3$ itself is a code. More precisely, it is a *uniform code* [4]. Consequently, any non-empty subset of $\mathcal{A}_4^3$ is a code called a *trinucleotide code* in this paper.

Definition 4. Trinucleotide comma-free code: A trinucleotide code X is comma-free if, for each $y \in X$ and $u, v \in \mathcal{A}_4^*$ such that $uyv = x_1 \cdots x_n$ with $x_1, \ldots, x_n \in X$, $n \geqslant 1$, it holds that $u, v \in X^*$.

Several varieties of trinucleotide comma-free codes were described in [13].

Definition 5. Trinucleotide circular code: A trinucleotide code X is circular if, for each $x_1,\dots,x_n,x'_1,\dots,x'_m \in X$, $n,m \geqslant 1$, $p \in \mathcal{A}_4^*$, $s \in \mathcal{A}_4^+$, the conditions $sx_2\cdots x_n p = x'_1\cdots x'_m$ and $x_1 = ps$ imply $n = m$, $p = \varepsilon$ (empty word) and $x_i = x'_i$ for $i = 1,\dots,n$.

Remark 2. $\mathcal{A}_4^3$ is obviously not a circular code and even less a comma-free code (see also Propositions 1 and 2 below).

Definition 6. Self-complementary code: A trinucleotide code X is self-complementary if, for each $x \in X$, $\mathcal{C}(x) \in X$.

Definition 7. C^3 self-complementary code: A trinucleotide code X is C^3 self-complementary if X, $\mathcal{P}(X)$ and $\mathcal{P}^2(X)$ are circular codes satisfying the following properties: $X = \mathcal{C}(X)$ (self-complementary) and $\mathcal{C}(\mathcal{P}(X)) = \mathcal{P}^2(X)$.

Definition 8. Maximal code: A trinucleotide circular code $X \in \mathcal{A}_4^3$ is maximal if for each $x \in \mathcal{A}_4^3$, $x \notin X$, $X \cup \{x\}$ is not a trinucleotide circular code.

The following lemma is very well known and is used several times in the paper.

Lemma 1. *For any letter α,β,γ and for any trinucleotide circular code X, then $\alpha\alpha\alpha \notin X$ and the set $\{\alpha\beta\gamma,\beta\gamma\alpha,\gamma\alpha\beta\} \cap X$ contains at most one element and exactly one when X has* 20 *elements.*

Remark 3. The conjugation class of the trinucleotide AAA has only one element: AAA itself. Obviously, this property is also true for the trinucleotides CCC, GGG, TTT. Otherwise, each other trinucleotide belongs to a conjugation class having exactly three trinucleotides. Consequently, the non-periodic trinucleotides, i.e. $\mathcal{A}_4^3 \setminus \{AAA, CCC, GGG, TTT\}$, are partitioned into exactly 20 classes. Finally, any trinucleotide circular code X with 20 words is maximal.

The set X_0 of 20 trinucleotides identified in the gene populations of both eukaryotes and prokaryotes is a maximal C^3 self-complementary circular code [1].

2.3. Necklaces

We recall the following definitions and some previous results. We denote by $l_1,l_2,\dots,l_{n-1},l_n,\dots$ the letters in $\mathcal{A}_4$, by $d_1,d_2,\dots,d_{n-1},d_n,\dots$ the diletters in $\mathcal{A}_4^2$, and by n an integer satisfying $n \geqslant 2$.

Definition 9. Letter Diletter Necklaces (LDN): We say that the ordered sequence $l_1,d_1,l_2,d_2,\dots,d_{n-1},l_n,d_n$ is an $nLDN$ for a subset $X \subset \mathcal{A}_4^3$ if $l_1d_1,l_2d_2,\dots,l_nd_n \in X$ and $d_1l_2,d_2l_3,\dots,d_{n-1}l_n \in X$.

Definition 10. Letter Diletter Continued Necklaces ($LDCN$): We say that the ordered sequence $l_1,d_1,l_2,d_2,\dots,d_{n-1},l_n,d_n,l_{n+1}$ is an $(n+1)LDCN$ for a subset $X \subset \mathcal{A}_4^3$ if $l_1d_1,l_2d_2,\dots,l_nd_n \in X$ and $d_1l_2,d_2l_3,\dots,d_{n-1}l_n,d_nl_{n+1} \in X$.

Definition 11. Diletter Letter Necklaces (DLN): We say that the ordered sequence $d_1,l_1,d_2,l_2,\dots,l_{n-1},d_n,l_n$ is an $nDLN$ for a subset $X \subset \mathcal{A}_4^3$ if $d_1l_1,d_2l_2,\dots,d_nl_n \in X$ and $l_1d_2,l_2d_3,\dots,l_{n-1}d_n \in X$.

Definition 12. Diletter Letter Continued Necklaces ($DLCN$): We say that the ordered sequence $d_1,l_1,d_2,l_2,\dots,l_{n-1},d_n,l_n,d_{n+1}$ is an $(n+1)DLCN$ for a subset $X \subset \mathcal{A}_4^3$ if $d_1l_1,d_2l_2,\dots,d_nl_n \in X$ and $l_1d_2,l_2d_3,\dots,l_{n-1}d_n,l_nd_{n+1} \in X$.

Proposition 1. *(See [16].) Let X be a trinucleotide code. The following conditions are equivalent*:

(i) *X is a circular code.*
(ii) *X has no $5LDCN$.*

Proposition 2. *(See [13].) Let X be a trinucleotide code. The following conditions are equivalent*:

(i) *X is a comma-free code.*
(ii) *X has no $2LDN$ and no $2DLN$.*

Definition 13. Let X be a trinucleotide code. For any integer $n \in \{2,3,4,5\}$, we say that X belongs to the class C^{nLDN} if X has no $nLDN$ and that X belongs to the class C^{nDLN} if X has no $nDLN$. Similarly, for any integer $n \in \{3,4,5\}$, we say that X belongs to the class C^{nLDCN} if X has no $nLDCN$ and that X belongs to the class C^{nDLCN} if X has no $nDLCN$.

Notation 1. For any integer $n \in \{2,3,4,5\}$, $I^n = C^{nLDN} \cap C^{nDLN}$ and $U^n = C^{nLDN} \cup C^{nDLN}$. Similarly, for any integer $n \in \{3,4,5\}$, $I^n C = C^{nLDCN} \cap C^{nDLCN}$ and $U^n C = C^{nLDCN} \cup C^{nDLCN}$.

Proposition 3. *(See [14].) The following chains of inclusions hold*:

(i) $C^{2LDN} \subset C^{3LDCN} \subset C^{3LDN} \subset C^{4LDCN} \subset C^{4LDN} \subset C^{5LDCN} \subset C^{5LDN}$.
(ii) $C^{2DLN} \subset C^{3DLCN} \subset C^{3DLN} \subset C^{4DLCN} \subset C^{4DLN} \subset C^{5DLCN} \subset C^{5DLN}$.
(iii) $C^{2LDN} \subset C^{3DLCN} \subset C^{3LDN} \subset C^{4DLCN} \subset C^{4LDN} \subset C^{5DLCN} \subset C^{5LDN}$.
(iv) $C^{2DLN} \subset C^{3LDCN} \subset C^{3DLN} \subset C^{4LDCN} \subset C^{4DLN} \subset C^{5LDCN} \subset C^{5DLN}$.
(v) $I^2 \subset I^3 C \subset I^3 \subset I^4 C \subset I^4 \subset I^5 C \subset I^5$.
(vi) $U^2 \subset U^3 C \subset U^3 \subset U^4 C \subset U^4 \subset U^5 C \subset U^5$.

Proposition 4. *(See [14].)* $C^{5LDN} = C^{5LDCN} = C^{5DLN}$.

Remark 4. By Propositions 1 and 4, $C^{5LDN} = C^{5LDCN} = C^{5DLN}$ is the class of circular codes. Therefore, all the chains of inclusions of Proposition 3 end with the class of circular codes. By Proposition 2, the chain of inclusions of Proposition 3(v) begins with I^2 which is the class of comma-free codes.

3. Mathematical results

Notation 2. Let X be a trinucleotide code. The mirror code of X, denoted by $\widetilde{X}$, is the set of the mirror images of the trinucleotides of X. Note that the mirror map is an involution.

Proposition 5. *Let X be a trinucleotide code. X is a circular code if and only if $\widetilde{X}$ is a circular code.*

Proof. By way of contradiction, suppose that X is a circular code and $\widetilde{X}$ is not a circular code. Then, there exists a $5LDCN$, i.e. $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_5$, for $\widetilde{X}$. Consequently, $l_5, \widetilde{d_4}, l_4, \widetilde{d_3}, l_3, \widetilde{d_2}, l_2, \widetilde{d_1}, l_1$ is a $5LDCN$ for X and, by Proposition 1, X is not a circular code. Contradiction. The other implication is proved by replacing in the proof X with $\widetilde{X}$, and conversely, and by using the fact that the mirror map is an involution. □

Proposition 6. *Let X be a trinucleotide code. For any integer $n \in \{2,3,4,5\}$, $X \in C^{nLDN}$ if and only if $\widetilde{X} \in C^{nDLN}$.*

Proof. We first prove the implication $X \in C^{2LDN} \Rightarrow \widetilde{X} \in C^{2DLN}$. Suppose that $X \in C^{2LDN}$ and, by way of contradiction, that $\widetilde{X} \notin C^{2DLN}$. Then, there exists a $2DLN$, i.e. d_1, l_1, d_2, l_2, for $\widetilde{X}$. Consequently, $l_2, \widetilde{d_2}, l_1, \widetilde{d_1}$ is a $2LDN$ for X. Contradiction. The implication $\widetilde{X} \in C^{2DLN} \Rightarrow X \in C^{2LDN}$ is proved in a similar way. The proofs of the equivalences for $n \in \{3,4,5\}$ use, as in the previous proposition, the fact that the mirror map is an involution. □

Definition 14. A trinucleotide circular code containing exactly l elements is called an l-trinucleotide circular code.

Remark 5. A 20-trinucleotide circular code is

- maximal (in the sense that it cannot be contained in a trinucleotide circular code with more words);
- maximum (in the sense that no trinucleotide circular code can contain more than 20 elements).

Proposition 7. *For 20-trinucleotide circular codes and for any integer $n \in \{2,3,4,5\}$, $|C^{nLDN}| = |C^{nDLN}|$.*

Proof. We first prove the equality $|C^{2LDN}| = |C^{2DLN}|$. Consider two codes X and Y, $X \neq Y$, in $(C^{2LDN} - C^{2DLN})$. By Proposition 6, $\widetilde{X}$ and $\widetilde{Y}$ are circular codes in $(C^{2DLN} - C^{2LDN})$ and $\widetilde{X} \neq \widetilde{Y}$. So, there is an injective map from $(C^{2LDN} - C^{2DLN})$ into $(C^{2DLN} - C^{2LDN})$. In a similar way, we prove that there is also an injective map from $(C^{2DLN} - C^{2LDN})$ into $(C^{2LDN} - C^{2DLN})$. Then, there is a bijection between $(C^{2LDN} - C^{2DLN})$ and $(C^{2DLN} - C^{2LDN})$, hence $|(C^{2LDN} - C^{2DLN})| = |(C^{2LDN} - C^{2DLN})|$. Consequently, $|C^{2LDN}| = |(C^{2LDN} - C^{2DLN})| + |I^2| = |(C^{2DLN} - C^{2LDN})| + |I^2| = |C^{2DLN}|$. The proofs of the equalities for $n \in \{3,4,5\}$ are similar. □

The main result of this article is the following one.

Proposition 8. *For 20-trinucleotide circular codes, the following chain of inclusions and equalities hold*:

$$I^2 \subset U^2 = I^3 C \subset U^3 C = I^3 \subset U^3 = I^4 C \subset U^4 C = I^4 \subset U^4 = I^5 C \subset U^5 C = I^5 = U^5.$$

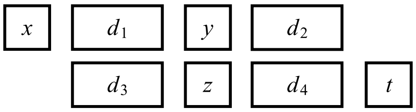

Fig. 1. Necklaces used in proof of $I^3C \subset U^2$.

Proof. The inclusions are trivial. We have only to prove the equalities. We begin with $U^2 = I^3C$ which is the most difficult to prove.

Proof of $U^2 \subset I^3C$. If X is a 20-trinucleotide circular code in U^2 then either X is in C^{2LDN} or X is in C^{2DLN}. Suppose that X is in C^{2LDN}. By Proposition 3(i), we have $C^{2LDN} \subset C^{3LDCN}$ and by Proposition 3(iii), we have $C^{2LDN} \subset C^{3DLCN}$. So, X is in $C^{3LDCN} \cap C^{3DLCN} = I^3C$. On the other hand, suppose that X is in C^{2DLN}. By Proposition 3(ii), we have $C^{2DLN} \subset C^{3DLCN}$ and by Proposition 3(iv), we have $C^{2DLN} \subset C^{3LDCN}$. So, X is in $C^{3DLCN} \cap C^{3LDCN} = I^3C$. Hence, in both cases X is in I^3C and the inclusion $U^2 \subset I^3C$ holds.

Proof of $I^3C \subset U^2$. By way of contradiction, suppose that a 20-trinucleotide circular code X is in I^3C but is not in U^2. Then, for some letters $x, y, z, t \in \mathcal{A}$ and for some diletters $d_1, d_2, d_3, d_4 \in \mathcal{A}^2$ we have $xd_1, d_1y, yd_2 \in X$ and $d_3z, zd_4, d_4t \in X$ (Fig. 1).

Claim 1. $\{x, y, z, t\} = \{A, C, G, T\}$.

Proof of Claim 1. Note that $x \neq y$. Otherwise, xd_1 and d_1x (which are conjugate) should both be in X, contradiction according to Lemma 1.

Note also that $z \neq t$. Otherwise, zd_4 and d_4z (which are conjugate) should both be in X, contradiction according to Lemma 1.

Finally, note that $\{x, y\} \cap \{z, t\} = \emptyset$. Otherwise,

- if $x = z$ then $d_3z, zd_1, d_1y, yd_2 \in X$, hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$, in contradiction with $X \in I^3C$;
- if $x = t$ then $d_3z, zd_4, d_4t, td_1, d_1y, yd_2 \in X$ (hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$), in contradiction with $X \in I^3C$;
- if $y = z$ then $xd_1, d_1y, yd_4, d_4t \in X$ (hence $X \notin C^{3LDCN}$ and so $X \notin I^3C$), in contradiction with $X \in I^3C$;
- if $y = t$ then $d_3z, zd_4, d_4t, td_2 \in X$ (hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$), in contradiction with $X \in I^3C$.

Claim 2. $xzt \in X$.

Proof of Claim 2. As X is a 20-trinucleotide circular code, it must contain at least an element in the conjugacy class of xzt, according to Lemma 1. If $ztx \in X$ then $ztx, xd_1, d_1y, yd_2 \in X$ hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$, in contradiction with $X \in I^3C$, and if $txz \in X$ then $d_3z, zd_4, d_4t, txz \in X$ (hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$), in contradiction with $X \in I^3C$. So, the unique element of X in the conjugacy class of xzt is xzt.

Claim 3. $xxz \in X$.

Proof of Claim 3. As X is a 20-trinucleotide circular code, it must contain at least an element in the conjugacy class of xxz, according to Lemma 1. If $xzx \in X$ then $xzx, xd_1, d_1y, yd_2 \in X$ hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$, in contradiction with $X \in I^3C$, and if $zxx \in X$ then $zxx, xd_1, d_1y, yd_2 \in X$ (hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$), in contradiction with $X \in I^3C$. So, the unique element of X in the conjugacy class of xxz is xxz.

Claim 4. $zyx \notin X$.

Proof of Claim 4. By way of contradiction, suppose that zyx is in X. We have $zyx, xd_1, d_1y, yd_2 \in X$ hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$, in contradiction with $X \in I^3C$.

Now, we consider the elements in the conjugacy class of zzx and we show that none of them can be in X.

Claim 5. $zzx \notin X$.

Proof of Claim 5. In the opposite case, $zzx, xd_1, d_1y, yd_2 \in X$ hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$, in contradiction with $X \in I^3C$.

Claim 6. $zxz \notin X$.

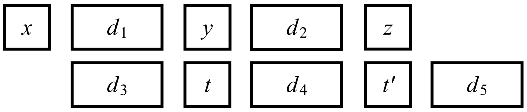

Fig. 2. Necklaces used in proof of $I^3 \subset U^3C$.

Proof of Claim 6. By way of contradiction, suppose that zxz is in X and note that, as X is a 20-trinucleotide circular code, exactly one element of the conjugacy class of zyx can be in X, according to Lemma 1. By Claim 4, i.e. $zyx \notin X$, we have to consider only two cases:

- $yxz \in X$. By Claim 2, i.e. $xzt \in X$, we have $xd_1, d_1y, yxz, xzt \in X$ hence $X \notin C^{3LDCN}$ and so $X \notin I^3C$, in contradiction with $X \in I^3C$;
- $xzy \in X$. We have $d_3z, zxz, xzy, yd_2 \in X$ (hence $X \notin C^{3DLCN}$ and so $X \notin I^3C$), in contradiction with $X \in I^3C$.

So, zxz cannot be in X.

Claim 7. $xzz \notin X$.

Proof of Claim 7. By Claim 3, i.e. $xxz \in X$, we have $xxz, xzz, zd_4, d_4t \in X$ hence $X \notin C^{3LDCN}$ and so $X \notin I^3C$, in contradiction with $X \in I^3C$. So, xzz cannot be in X.

By Claims 5, 6 and 7, the conjugacy class $\{zzx, zxz, xzz\}$ has no element in X, in contradiction with the maximality of X according to Lemma 1.

The inclusion $I^3C \subset U^2$ holds leading to the equality $U^2 = I^3C$.

The other equalities in the proposition are less difficult to prove than the equality $U^2 = I^3C$ as the Pigeon hole Principle can be used. For example, let us to prove the equality $U^3C = I^3$. We first prove the inclusion $U^3C \subset I^3$ and then the inclusion $I^3 \subset U^3C$.

Proof of $\boldsymbol{U^3C \subset I^3}$. If X is a 20-trinucleotide circular code in U^3C then either X is in C^{3LDCN} or X is in C^{3DLCN}. Suppose that X is in C^{3LDCN}. By Proposition 3(i), we have $C^{3LDCN} \subset C^{3LDN}$ and by Proposition 3(iv), we have $C^{3LDCN} \subset C^{3DLN}$. So, X is in $C^{3LDN} \cap C^{3DLN} = I^3$. On the other hand, suppose that X is in C^{3DLCN}. By Proposition 3(ii), we have $C^{3DLCN} \subset C^{3DLN}$ and by Proposition 3(iii), we have $C^{3DLCN} \subset C^{3LDN}$. So, X is in $C^{3DLN} \cap C^{3LDN} = I^3$. Hence, in both cases X is in I^3 and the inclusion $U^3C \subset I^3$ holds.

Proof of $\boldsymbol{I^3 \subset U^3C}$. By way of contradiction, suppose that a 20-trinucleotide circular code X is in I^3 but is not in U^3C. So, for some letters $x, y, z, t, t' \in \mathcal{A}$ and for some diletters $d_1, d_2, d_3, d_4, d_5 \in \mathcal{A}^2$ we have $xd_1, d_1y, yd_2, d_2z \in X$ and $d_3t, td_4, d_4t', t'd_5 \in X$ (Fig. 2).

As $\mathcal{A}$ contains four letters, we have, by the Pigeon hole Principle, at least two identical letters in $\{x, y, z, t, t'\}$.

If the equality holds in $\{x, y, z\}$ then we have $x = y$ or $x = z$ or $y = z$. If $x = y$ then xd_1 and d_1x (which are conjugate) should both be in X, in contradiction with $X \in I^3$. If $y = z$ then yd_2 and d_2y (which are conjugate) should both be in X, in contradiction with $X \in I^3$. If $x = z$ then $xd_1, d_1y, yd_2, d_2x, xd_1 \in X$ hence $X \notin C^{3LDN}$ and so $X \notin I^3$, in contradiction with $X \in I^3$.

If the equality holds in $\{t, t'\}$ then td_4, d_4t (which are conjugate) should both be in X, in contradiction with $X \in I^3$.

Finally, if $\{x, y, z\} \cap \{t, t'\}$ is non-empty then one of the following equalities holds: $t = x$, $t = y$, $t = z$, $t' = x$, $t' = y$ and $t' = z$. Now:

- if $t = x$ then $d_3x, xd_1, d_1y, yd_2, d_2z \in X$ (hence $X \notin C^{3DLN}$ and so $X \notin I^3$), in contradiction with $X \in I^3$;
- if $t = y$ then $xd_1, d_1y, yd_4, d_4t', t'd_5 \in X$ (hence $X \notin C^{3LDN}$ and so $X \notin I^3$), in contradiction with $X \in I^3$;
- if $t = z$ then $xd_1, d_1y, yd_2, d_2z, zd_4, d_4t', t'd_5 \in X$ (hence $X \notin C^{3LDN}$ and so $X \notin I^3$), in contradiction with $X \in I^3$;
- if $t' = x$ then $d_3t, td_4, d_4t', t'd_1, d_1y, yd_2, d_2z \in X$ (hence $X \notin C^{3DLN}$ and so $X \notin I^3$), in contradiction with $X \in I^3$;
- if $t' = y$ then $d_3t, td_4, d_4t', t'd_2, d_2z \in X$ (hence $X \notin C^{3DLN}$ and so $X \notin I^3$), in contradiction with $X \in I^3$;
- if $t' = z$ then $xd_1, d_1y, yd_2, d_2z, zd_5 \in X$ (hence $X \notin C^{3LDN}$ and so $X \notin I^3$), in contradiction with $X \in I^3$.

So, $\{x, y, z\} \cap \{t, t'\}$ is empty. Hence, there are no identical letters in $\{x, y, z, t, t'\}$, in contradiction with the fact that $\mathcal{A}$ has exactly four letters. Therefore, the inclusion $I^3 \subset U^3C$ holds leading to the equality $U^3C = I^3$.

The other equalities are proved in a similar way. □

For a fast computing of the number of 20-trinucleotide circular codes in the different classes C^{nLDN}, C^{nDLN}, I^n and U^n with $n \in \{2, 3, 4, 5\}$, and C^{nLDCN}, C^{nDLCN}, I^nC and U^nC with $n \in \{3, 4, 5\}$, the following definition of a closed necklace is now introduced.

Table 1
Partition of $\mathcal{A}_4^3 \setminus \{AAA, CCC, GGG, TTT\}$ into the 20 conjugacy classes.

$\mathcal{D}_1 = \{AAC, ACA, CAA\}$	$\mathcal{D}_2 = \{AAG, AGA, GAA\}$
$\mathcal{D}_3 = \{AAT, ATA, TAA\}$	$\mathcal{D}_4 = \{ACC, CCA, CAC\}$
$\mathcal{D}_5 = \{ACG, CGA, GAC\}$	$\mathcal{D}_6 = \{ACT, CTA, TAC\}$
$\mathcal{D}_7 = \{AGC, GCA, CAG\}$	$\mathcal{D}_8 = \{AGG, GGA, GAG\}$
$\mathcal{D}_9 = \{AGT, GTA, TAG\}$	$\mathcal{D}_{10} = \{ATC, TCA, CAT\}$
$\mathcal{D}_{11} = \{ATG, TGA, GAT\}$	$\mathcal{D}_{12} = \{ATT, TTA, TAT\}$
$\mathcal{D}_{13} = \{CCG, CGC, GCC\}$	$\mathcal{D}_{14} = \{CCT, CTC, TCC\}$
$\mathcal{D}_{15} = \{CGG, GGC, GCG\}$	$\mathcal{D}_{16} = \{CGT, GTC, TCG\}$
$\mathcal{D}_{17} = \{CTG, TGC, GCT\}$	$\mathcal{D}_{18} = \{CTT, TTC, TCT\}$
$\mathcal{D}_{19} = \{GGT, GTG, TGG\}$	$\mathcal{D}_{20} = \{GTT, TTG, TGT\}$

Definition 15. Letter Diletter Continued Closed Necklaces ($LDCCN$): We say that the ordered sequence $l_1, d_1, l_2, d_2, \dots, d_{n-1}, l_n, d_n, l_{n+1}$ is an $(n+1)LDCCN$ for a subset $X \subset \mathcal{A}_4^3$ if $l_1d_1, l_2d_2, \dots, l_nd_n \in X$ and $d_1l_2, d_2l_3, \dots, d_{n-1}l_n, d_nl_{n+1} \in X$ and $l_1 = l_{n+1}$.

Notation 3. An $(n+1)LDCCN$ $l_1, d_1, l_2, d_2, \dots, d_{n-1}, l_n, d_n, l_{n+1}$ is denoted it by $[l_1, d_1, l_2, d_2, \dots, d_{n-1}, l_n, d_n]$. Accordingly:

a $2LDCCN$, i.e. $[l_1, d_1]$, has the form l_1, d_1, l_1;
a $3LDCCN$, i.e. $[l_1, d_1, l_2, d_2]$, has the form l_1, d_1, l_2, d_2, l_1;
a $4LDCCN$, i.e. $[l_1, d_1, l_2, d_2, l_3, d_3]$, has the form $l_1, d_1, l_2, d_2, l_3, d_3, l_1$;
a $5LDCCN$, i.e. $[l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4]$, has the form $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_1$.

Remark 6. An $(n+1)LDCCN$ is an $(n+1)LDCN$ (Definition 10) in which the first and the last letters are identical.

The following proposition gives a relation between a trinucleotide circular code and the closed necklace $LDCCN$.

Proposition 9. *Let X be a trinucleotide circular code. The following conditions are equivalent*:

(i) *X is a trinucleotide circular code.*
(ii) *X has no $nLDCCN$ for any integer $n \in \{2, 3, 4, 5\}$.*

Proof. (i) $\Rightarrow$ (ii). By way of contradiction, suppose that X has some $nLDCCN$ for some integer $n \in \{2, 3, 4, 5\}$.
If it is a $2LDCCN$ then $l_1, d_1, l_1, d_1, l_1, d_1, l_1, d_1, l_1$ is a $5LDCN$ for X.
If it is a $3LDCCN$ then $l_1, d_1, l_2, d_2, l_1, d_1, l_2, d_2, l_1$ is a $5LDCN$ for X.
If it is a $4LDCCN$ then $l_1, d_1, l_2, d_2, l_3, d_3, l_1, d_1, l_2$ is a $5LDCN$ for X.
If it is a $5LDCCN$ then $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_1$ is a $5LDCN$ for X.
In each of these four cases, by Proposition 1, X is not a trinucleotide circular code. Contradiction.
(ii) $\Rightarrow$ (i). By way of contradiction, suppose that X is not a trinucleotide circular code. By Proposition 1, X has a $5LDCN$, say $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_5$. As $\mathcal{A}_4$ has four letters, then $l_i = l_j$ for some i, j, $1 \leqslant i \leqslant j \leqslant 5$.
If $j - i = 4$ then $l_1 = l_5$ and $[l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4]$ is a $5LDCCN$ for X.
If $j - i = 3$ then $[l_i, d_i, l_{i+1}, d_{i+1}, l_{i+2}, d_{i+2}]$ is a $4LDCCN$ for X.
If $j - i = 2$ then $[l_i, d_i, l_{i+1}, d_{i+1}]$ is a $3LDCCN$ for X.
If $j - i = 1$ then $[l_i, d_i]$ is a $2LDCCN$ for X.
In each of these four cases, by Proposition 1, there is a contradiction with (ii). $\square$

4. Computer results

4.1. Number of 20-trinucleotide circular codes

We consider the following partition of $\mathcal{A}_4^3 \setminus \{AAA, CCC, GGG, TTT\}$ into the 20 conjugacy classes (Table 1).

Let the length l of a word set $\mathcal{S}_l$, $1 \leqslant l \leqslant 20$, be the number of its words. In order to determine the number of 20-trinucleotide circular codes ($l = 20$ words), we have developed an algorithm that constructs trinucleotide sets $\mathcal{S}_l$ of increasing length l such that one and only one trinucleotide is chosen in each class $\mathcal{D}_l$ between the three possible ones. All sets $\mathcal{S}_1$ are circular codes (60 codes of length $l = 1$). Each set $\mathcal{S}_l$ is tested according to Proposition 9 verifying that it has no closed necklace $nLDCCN$ for any integer $n \in \{2, 3, 4, 5\}$. If a set $\mathcal{S}_l$ has no $nLDCCN$, then it is increased by a trinucleotide chosen in the next (in lexicographical order) conjugacy class $\mathcal{D}_{l+1}$. Indeed, if a set $\mathcal{S}_l$ is not a circular code then any set $\mathcal{S}_{l'}$, $1 \leqslant l < l' \leqslant 20$, containing $\mathcal{S}_l$ is also not a circular code. This algorithm ends with sets $\mathcal{S}_l$ of $l = 20$ trinucleotides.

The obtained number of 20-trinucleotide circular codes is 12, 964, 440.

Table 2
Mathematical hierarchy of 20-trinucleotide circular codes.

C^{2LDN}	C^{3LDCN}	C^{3LDN}	C^{4LDCN}	C^{4LDN}	C^{5LDCN}	C^{5LDN}
α_1	α_4	α_7	α_9	α_{12}	α_{14}	α_{14}
C^{2DLN}	C^{3DLCN}	C^{3DLN}	C^{4DLCN}	C^{4DLN}	C^{5DLCN}	C^{5DLN}
α_1	α_5	α_7	α_{10}	α_{12}	α_{13}	α_{14}
I^2	I^3C	I^3	I^4C	I^4	I^5C	I^5
α_2	α_3	α_6	α_8	α_{11}	α_{13}	α_{14}
U^2	U^3C	U^3	U^4C	U^4	U^5C	U^5
α_3	α_6	α_8	α_{11}	α_{13}	α_{14}	α_{14}

Table 3
Computational hierarchy of 20-trinucleotide circular codes.

C^{2LDN}	C^{3LDCN}	C^{3LDN}	C^{4LDCN}	C^{4LDN}	C^{5LDCN}	C^{5LDN}
1,584	294,912	423,552	5,088,264	5,528,688	12,964,440	12,964,440
C^{2DLN}	C^{3DLCN}	C^{3DLN}	C^{4DLCN}	C^{4DLN}	C^{5DLCN}	C^{5DLN}
1,584	4,920	423,552	578,496	5,528,688	5,940,648	12,964,440
I^2	I^3C	I^3	I^4C	I^4	I^5C	I^5
408	2,760	297,072	550,032	5,116,728	5,940,648	12,964,440
U^2	U^3C	U^3	U^4C	U^4	U^5C	U^5
2,760	297,072	550,032	5,116,728	5,940,648	12,964,440	12,964,440

4.2. *Mathematical and computational hierarchies of* 20-*trinucleotide circular codes*

According to Proposition 8, the number α_i of 20-trinucleotide circular codes in the different classes C^{nLDN}, C^{nDLN}, I^n and U^n with $n \in \{2,3,4,5\}$, and C^{nLDCN}, C^{nDLCN}, I^nC and U^nC with $n \in \{3,4,5\}$ must follow the hierarchy given in Table 2.

The computational hierarchy of 20-trinucleotide circular codes is given in Table 3 and agrees perfectly with the mathematical hierarchy.

The numbers of 20-trinucleotide circular codes in the classes from C^{2LDN} to C^{5LDN}, and from C^{2DLN} to C^{5DLN} are non-decreasing. The classes C^{2LDN} and C^{2DLN} are the first ones which are non-empty. Note that no self-complementary 20-trinucleotide circular codes are in these two classes C^{2LDN} and C^{2DLN} [14]. According to Proposition 4, the classes C^{5LDN}, C^{5LDCN} and C^{5DLN} contain all the 12,964,440 circular codes.

The numbers presented in Table 3 and the others symmetric relations identified (see, for example, Proposition 7) suggest us that the symmetric group Σ_4 can be involved in these problems. So far, its role is not very clear for the authors of this paper. A suitable mathematical formulation based on this symmetric group Σ_4 could simplify the definitions and the proofs of our results.

Acknowledgments

We thank Amy Glen and Jacques Justin for their advices. The second author thanks the Dipartimento di matematica "U. Dini" for giving him a friendly hospitality.

References

[1] D.G. Arquès, C.J. Michel, A complementary circular code in the protein coding genes, J. Theoret. Biol. 182 (1996) 45–58.
[2] F. Bassino, Generating function of circular codes, Adv. in Appl. Math. 22 (1999) 1–24.
[3] M.-P. Béal, J. Senellart, On the bound of the synchronization delay of a local automaton, Theoret. Comput. Sci. 205 (1998) 297–306.
[4] J. Berstel, D. Perrin, Theory of Codes, Pure Appl. Math., vol. 117, Academic Press, London, UK, 1985.
[5] F.H.C. Crick, J.S. Griffith, L.E. Orgel, Codes without commas, Proc. Natl. Acad. Sci. USA 43 (1957) 416–421.
[6] S.W. Golomb, B. Gordon, L.R. Welch, Comma-free codes, Canad. J. Math. 10 (1958) 202–209.
[7] S.W. Golomb, L.R. Welch, M. Delbrück, Construction and properties of comma-free codes, Biologiske Meddel Danske Vidensk Selsk 23 (1958) 1–34.
[8] R. Jolivet, F. Rothen, Peculiar symmetry of DNA sequences and evidence suggesting its evolutionary origin in a primeval genetic code, in: P. Ehrenfreund, O. Angerer, B. Battrick (Eds.), First European Workshop Exo-/Astro-Biology, vol. 496, Noordwijk, The Netherlands, May 2001, pp. 173–176.
[9] A.J. Koch, J. Lehman, About a symmetry of the genetic code, J. Theoret. Biol. 189 (1997) 171–174.
[10] J.-L. Lassez, Circular codes and synchronization, Int. J. Comput. Inform. Sci. 5 (1976) 201–208.
[11] J.-L. Lassez, R.A. Rossi, A.E. Bernal, 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/Symposia, vol. 2, AINAW'07, 2007, pp. 745–751.
[12] E.E. May, M.A. Vouk, D.L. Bitzer, D.I. Rosnick, An error-correcting framework for genetic sequence analysis, J. Franklin Inst. 341 (2004) 89–109.
[13] C.J. Michel, G. Pirillo, M.A. Pirillo, Varieties of comma-free codes, Comput. Math. Appl. 55 (2008) 989–996.
[14] C.J. Michel, G. Pirillo, M.A. Pirillo, A relation between trinucleotide comma-free codes and trinucleotide circular codes, Theoret. Comput. Sci. 401 (2008) 17–26.

[15] C. Nikolaou, Y. Almirantis, Mutually symmetric and complementary triplets: difference in their use distinguish systematically between coding and non-coding genomic sequences, J. Theoret. Biol. 223 (2003) 477–487.
[16] G. Pirillo, A characterization for a set of trinucleotides to be a circular code, in: C. Pellegrini, P. Cerrai, P. Freguglia, V. Benci, G. Israel (Eds.), Determinism, Holism, and Complexity, Kluwer, Boston, MA, 2003.
[17] G. Pirillo, A hierarchy for circular codes, Theor. Inform. Appl. 42 (2008) 717–728.
[18] N. Štambuk, On circular coding properties of gene and protein sequences, Croatica Chemica Acta 72 (1999) 999–1008.