Available online at www.sciencedirect.com
ScienceDirect
Computers and Mathematics with Applications 55 (2008) 989–996
An International Journal
computers & mathematics
with applications
www.elsevier.com/locate/camwa

Varieties of comma-free codes

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

[a] *Equipe de Bioinformatique Théorique, LSIIT (UMR CNRS - ULP 7005), Université Louis Pasteur de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France*
[b] *Consiglio Nazionale delle Ricerche, Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Unità di Firenze, Dipartimento di Matematica "U.Dini", viale Morgagni 67/A, 50134 Firenze, Italy*
[c] *Université de Marne-la-Vallée, 5 boulevard Descartes, Champs sur Marne, 77454 Marne-la-Vallée Cedex 2, France*
[d] *Istituto Statale SS. Annunziata, Piazzale del Poggio Imperiale, Firenze, Italy*

Abstract

New varieties of comma-free codes CFC of length 3 on the 4-letter alphabet are defined and analysed: self-complementary comma-free codes (CCFC), C^3 comma-free codes (C^3CFC), C^3 self-complementary comma-free codes (C^3CCFC), self-complementary maximal comma-free codes (CMCFC), C^3 maximal comma-free codes (C^3MCFC) and C^3 self-complementary maximal comma-free codes (C^3CMCFC). New properties with words of length 3, 4, 5 and 6 in comma-free codes are used for the determination of growth functions in the studied code varieties.

Keywords: Comma-free code; Word; Letter; Occurrence number; Occurrence probability

1. Introduction

A code in genes has been proposed by Crick et al. [1] in order to explain how the reading of a series of nucleotides could code for the amino acids constituting the proteins. The two problems stressed were: why are there more trinucleotides than amino acids and how to choose the reading frame? Crick et al. [1] have then proposed that only 20 among 64 trinucleotides code for the 20 amino acids. Such a bijective code implies that the coding trinucleotides are found only in one frame. Such a particular code is called a comma-free code (CFC) or a code without commas. However, the determination of a set of 20 trinucleotides forming a comma-free code has several constraints:

(i) A trinucleotide with identical nucleotides must be excluded from such a code. Indeed, the concatenation of AAA with itself, for example, does not allow the reading (original) frame to be retrieved as there are three possible decompositions: $\ldots AAA, AAA, AAA, \ldots$, $\ldots A, AAA, AAA, AA \ldots$ and $\ldots AA, AAA, AAA, A \ldots$ (the commas showing the construction frame).

(ii) Two trinucleotides related to circular permutation, for example, AAC and ACA, must be also excluded from such a code. Indeed, the concatenation of AAC with itself, for example, also does not allow the reading frame to be retrieved as there are two possible decompositions: $\ldots AAC, AAC, AAC, \ldots$ and $\ldots A, ACA, ACA, AC \ldots$.

* Corresponding author. Tel.: +33 3 90 24 44 62.
E-mail addresses: michel@dpt-info.u-strasbg.fr (C.J. Michel), pirillo@math.unifi.it (G. Pirillo), map@conmet.it (M.A. Pirillo).

0898-1221/$ - see front matter
doi:10.1016/j.camwa.2006.12.091

Therefore, by excluding AAA, CCC, GGG and TTT and by gathering the 60 remaining trinucleotides in 20 classes of three trinucleotides such that, in each class, three trinucleotides are deduced from each other by circular permutations, e.g. AAC, ACA and CAA, a comma-free code has only one trinucleotide per class and therefore contains at most 20 trinucleotides. This trinucleotide number is identical to the amino acid one, thus leading to a comma-free code assigning one trinucleotide per amino acid without ambiguity. Some investigations have been proposed by Golomb et al. [2,3]. However, the determination of comma-free codes and their properties are unrealizable without computer as there are billions of potential codes. Furthermore, in the late fifties, the two discoveries that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes for phenylalanine [4] and that genes are placed in reading frames with a particular start trinucleotide, have led to the concept of comma-free code over the alphabet $\{A, C, G, T\}$ being given up. For several biological reasons, in particular the interaction between mRNA and tRNA, this concept is taken again over the purine/pyrimidine alphabet $\{R, Y\}$ (purine $= R = \{A, G\}$, pyrimidine $= Y = \{C, T\}$) with two comma-free codes for primitive genes: RRY [5] and RNY ($N = \{R, Y\}$) [6].

By analysing the trinucleotide occurrence frequencies in the three frames of genes, several circular codes, but no comma-free codes, have been identified in genes [7–10]. A circular code also allows the reading frames of genes to be retrieved but with weaker conditions compared to a comma-free code. It is a set of words over an alphabet such that any word written on a circle (the next letter after the last letter of the word being the first letter) has at most one decomposition into words of the circular code.

This paper studies comma-free codes of length three on the four-letter alphabet, i.e. comma-free codes associated with trinucleotides in the gene structure. New varieties of comma-free codes CFC are defined and analysed such as self-complementary comma-free codes (CCFC), C^3 comma-free codes (C^3CFC), C^3 self-complementary comma-free codes (C^3CCFC), maximal comma-free codes (MCFC), self-complementary maximal comma-free codes (CMCFC), C^3 maximal comma-free codes (C^3MCFC) and C^3 self-complementary maximal comma-free codes (C^3CMCFC). These varieties of comma-free codes could explain the origin of circular codes in genes.

2. Definitions

The definitions hereafter are useful in order to introduce the different varieties of comma-free codes.

2.1. Genetic sequences

The *letters* (or *nucleotides* or *bases*) of the genetic alphabet, denoted by β_4, are A, C, G and T.

The set of *nonempty sequences* (resp. *sequences*) on β_4 is denoted by β_4^+ (resp. β_4^*). The set of the 16 sequences of length two (or *diletters* or *dinucleotides*) is denoted by β_4^2. The set of the 64 sequences of length three (or *triletters* or *trinucleotides*) is denoted by β_4^3.

The *total order* on the alphabet $\beta_4 = \{A, C, G, T\}$ is $A < C < G < T$. Consequently, β_4^+ is *lexicographically ordered*: given two words $u, v \in \beta_4^+$, *u is smaller than v* in the *lexicographical order*, noted $u < v$, if and only if either u is a proper left factor of v or there exist $x, y \in \beta_4$, $x < y$, and $r, s, t \in \beta_4^*$ such that $u = rxs$ and $v = ryt$.

Let $w = w[0]w[1]w[2]\dots w[i]\dots w[j]\dots w[n]$ a word of length $n+1$ on β_4. Then, we say that the factor $w[i]\dots w[j]$ is in frame $f \in \{0, 1, 2\}$ if $i = f \bmod 3$.

2.2. Two important maps

(i) The *complementarity*

$$\mathcal{C} : \beta_4^+ \to \beta_4^+$$

is an involutional antiisomorphism of β_4^+ given by

$$\mathcal{C}(A) = T, \qquad \mathcal{C}(T) = A, \qquad \mathcal{C}(C) = G, \qquad \mathcal{C}(G) = C$$

and naturally

$$\mathcal{C}(uv) = \mathcal{C}(v)\mathcal{C}(u)$$

for any $u, v \in \beta_4^+$.

(ii) The (left) circular *permutation*

$$\mathcal{P} : \beta_4^3 \to \beta_4^3$$

which circularly permutes each triletter $l_1l_2l_3$ as follows

$$\mathcal{P}(l_1l_2l_3) = l_2l_3l_1.$$

2.3. Varieties of comma-free codes

Code. A subset $X \subset \beta_4^+$ is a code if, for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, the condition

$$x_1 \cdots x_n = x'_1 \cdots x'_m$$

implies $n = m$, and, for $i = 1, \ldots, n$,

$$x_i = x'_i.$$

Comma-free code (CFC). A code $X \subset \beta_4^3$ is comma-free if, for each $y \in X$ and $u, v \in \beta_4^*$ such that $uyv = x_1 \cdots x_n$ with $x_1, \ldots, x_n \in X$, $n \geq 1$, then $u, v \in X^*$.

Maximal comma-free code (MCFC). A CFC $X \subset \beta_4^3$ is maximal if, for each $y \in \beta_4^3$, $X \cup \{y\}$ is not a CFC.

Self-complementary comma-free code (CCFC). A CFC $X \subset \beta_4^3$ is self-complementary if, for each $y \in X$, $\mathcal{C}(y) \in X$.

C^3 comma-free code (C^3CFC). A CFC $X \subset \beta_4^3$ is C^3 if $\mathcal{P}(X)$ and $\mathcal{P}(\mathcal{P}(X))$ are also CFC.

The other notions of maximality of codes on β_4^3 in Results 1–8 are defined in a similar way.

2.4. Necklace concept

The concept of necklace has been introduced by Pirillo [11] for circular codes and has been used for studying self-complementary circular codes [12]. Here, it is applied for the comma-free codes with the concepts of Letter Diletter Necklace (LDN) and Diletter Letter Necklace (DLN).

Letter Diletter Necklaces (LDN). Let $l_1, l_2, \ldots, l_{n-1}, l_n$ be letters in β_4 and let $d_1, d_2, \ldots, d_{n-1}, d_n$ be diletters in β_4^2. We say that the ordered sequence $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n$ is a nLDN for a subset $X \subset \beta_4^3$ if $l_1d_1, l_2d_2, \ldots, l_nd_n \in X$ and $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n \in X$.

Diletter Letter Necklaces (DLN). Let $l_1, l_2, \ldots, l_{n-1}, l_n$ be letters in β_4 and let $d_1, d_2, \ldots, d_{n-1}, d_n$ be diletters in β_4^2. We say that the ordered sequence $d_1, l_1, d_2, l_2, \ldots, l_{n-1}, d_n, l_n$ is a nDLN for a subset $X \subset \beta_4^3$ if $d_1l_1, d_2l_2, \ldots, d_nl_n, \in X$ and $l_1d_2, l_2d_3, \ldots, l_{n-1}d_n \in X$.

3. Properties of comma-free codes

Proposition 1. *Let X be a subset of β_4^3. The following conditions are equivalent:*

(a) *X is a comma-free code;*
(b) *for each triletters t_i, t_j and t_k belonging to X, if t_k is a factor of t_it_j then t_k is in frame 0;*
(c) *for each tetraletter $l_1l_2l_3l_4$ such that $l_1l_2l_3$ and $l_2l_3l_4$ belong to X then no triletter of X has l_4 as a prefix and no triletter of X has l_1 as a suffix;*
(d) *for each pentaletter $l_1l_2l_3l_4l_5$ such that $l_1l_2l_3$ and $l_3l_4l_5$ belong to X then no triletter of X starts with l_4l_5 and no triletter of X ends with l_1l_2;*
(e) *for each hexaletter $l_1l_2l_3l_4l_5l_6$ such that $l_1l_2l_3$ and $l_4l_5l_6$ belong to X then the triletter $l_2l_3l_4$ does not belong to X and the triletter $l_3l_4l_5$ does not belong to X;*
(f) *X has no* 2LDN *and no* 2DLN.

Proof. (a) $\Rightarrow$ (b). Let X be a comma-free code. By way of contradiction, suppose that, for some triletters $t_i, t_j, t_k \in X$, the triletter t_k is a factor of t_it_j and in frame 1, the case of frame 2 being similar. Then, there is a letter $l_1 \in \beta_4$ and a diletter $l_5l_6 \in \beta_4^2$ such that $t_it_j = l_1t_kl_5l_6$. As l_1 is a letter and consequently not in X^*, we are in contradiction with the assumption that X is a comma-free code.

Table 1
Growth function of potential comma-free codes PCFC on β_4^3

l	1	2	3	4	5	6	7	8	9	10	11
Nb(l)	60	1710	30780	392445	3767472	28256040	169536240	826489170	3305956680	10909657044	29753610120

12	13	14	15	16	17	18	19	20
66945622770	123591918960	185387878440	222465454128	208561363245	147219785820	73609892910	23245229340	3486784401

The first (second resp.) row gives the length l (occurrence number Nb(l) resp.) of PCFC.

(b) $\Rightarrow$ (c). Suppose that X verifies condition (b). By way of contradiction, there exists a tetraletter $l_1l_2l_3l_4$ such that $l_1l_2l_3$ and $l_2l_3l_4$ belong to X and there exists a triletter of X having l_4 as a prefix or a triletter of X having l_1 as a suffix.

Case of a triletter of X having l_4 as a prefix. For some $l_5l_6 \in \beta_4^2$, the word $l_1l_2l_3l_4l_5l_6 \in X^2$. As $l_2l_3l_4$ is also in X and in frame 1 in $l_1l_2l_3l_4l_5l_6$ we are in contradiction with property (b) of X.

Case of a triletter of X having l_1 as a suffix. For some $l_5l_6 \in \beta_4^2$, the word $l_5l_6l_1l_2l_3l_4 \in X^2$. As $l_ll_2l_3$ is also in X and in frame 2 in $l_5l_6l_1l_2l_3l_4$ we are in contradiction with property (b) of X.

(c) $\Rightarrow$ (d). Suppose that X verifies condition (c). By way of contradiction, there is a pentaletter $l_1l_2l_3l_4l_5$ such that $l_1l_2l_3$ and $l_3l_4l_5$ belong to X and there is a triletter of X having l_4l_5 as a prefix, the case of a triletter of X having l_1l_2 as a suffix being similar. For some $l_6 \in \beta_4$, the word $l_1l_2l_3l_4l_5l_6 \in X^2$. Consider the tetraletter $l_3l_4l_5l_6$. The triletters $l_3l_4l_5$ and $l_4l_5l_6$ belong to X and the triletter $l_1l_2l_3$ also in X has l_3 as a suffix. So X does not verify condition (c). Contradiction.

(d) $\Rightarrow$ (e). Suppose that X verifies condition (d) and, by way of contradiction, X does not verify condition (e). Let $l_1l_2l_3l_4l_5l_6$ be a hexaletter such that $l_1l_2l_3$ and $l_4l_5l_6$ belong to X. There are 2 cases:

Case $l_2l_3l_4$ belong to X. Consider the pentaletter $l_1l_2l_3l_4l_5$. The triletters $l_1l_2l_3$ and $l_2l_3l_4$ belong to X. Moreover, the triletter $l_4l_5l_6$ also belonging to X, starts with l_4l_5. Contradiction.

Case $l_3l_4l_5$ belong to X. Consider the pentaletter $l_2l_3l_4l_5l_6$. The triletters $l_3l_4l_5$ and $l_4l_5l_6$ belong to X. Moreover, the triletter $l_1l_2l_3$ also belonging to X, ends with l_2l_3. Contradiction.

(e) $\Rightarrow$ (f). Suppose that X verifies condition (e) and, by way of contradiction, X does not verify condition (f).

Case X has a 2LDN, i.e. a sequence l_1, d_1, l_2, d_2 with $l_1, l_2 \in \beta_4$ and $d_1, d_2 \in \beta_4^2$. Consider the hexaletter $l_1d_1l_2d_2$. The triletters l_1d_1, l_2d_2 and also d_1l_2 belong to X. Consequently X does not verify condition (e). Contradiction.

Case X has a 2DLN, i.e. a sequence d_1, l_1, d_2, l_2 with $l_1, l_2 \in \beta_4$ and $d_1, d_2 \in \beta_4^2$. Consider the hexaletter $d_1l_1d_2l_2$. The triletters d_1l_1, d_2l_2 and also l_1d_2 belong to X. Consequently X does not verify condition (e). Contradiction.

(f) $\Rightarrow$ (a). Suppose that X has no 2LDN and no 2DLN and that X is not a comma-free code. There exist $x_1, \ldots, x_n$, $n \geq 1$, in X such that $uyv = x_1 \cdots x_n$ and either $u \notin X^*$ or $v \notin X^*$. As X is homogeneous (i.e. all its elements have the same length) we can suppose, without loss of generality, that $u \notin X^*$. Let k be the greatest integer such that $3k \leq |u|$. Then, y is a factor of $x_{k+1}x_{k+2}$. Let w be such that $x_1x_2 \ldots x_kw = u$.

There are two possible cases:

Case $w \in \beta_4$. Let $w = l_1$. Then, there exist $l_2 \in \beta_4$ and $d_1, d_2 \in \beta_4^2$ such that $x_{k+1} = l_1d_1$, $x_{k+2} = l_2d_2$ and $y = d_1l_2$. In this case, l_1, d_1, l_2, d_2 is a 2LDN for X.

Case $w \in \beta_4^2$. Let $w = d_3$. Then, there exist $d_4 \in \beta_4^2$ and $l_3, l_4 \in \beta_4$ such that $x_{k+1} = d_3l_3$, $x_{k+2} = d_4l_4$ and $y = l_3d_4$. In this case, d_3, l_3, d_4, l_4 is a 2DLN for X.

In both cases, we are in contradiction. □

4. Growth functions of varieties of comma-free codes

By developing algorithms based on the Proposition 1, the growth functions of varieties of comma-free codes (CFC) on β_4^3 are determined. The occurrence number Nb(l) and its probability Pr(l) of CFC of length l, by varying l between 1 and 20 (maximal length with words of length three on a four-letter alphabet), are given for each table. There are $\binom{20}{l} \times 3^l$ potential CFC (PCFC) of length $l \in \{1, 20\}$ (Table 1).

Result 1. Table 2a shows the growth function of comma-free codes CFC on β_4^3. The CFC of length one are the 60 words on $\beta_4^3 - \{AAA, CCC, GGG, TTT\}$. This function has a maximum with about 111 billions of CFC of

Table 2a
Growth function of comma-free codes CFC on β_4^3

l	1	2	3	4	5	6	7	8	9	10
Nb(l)	60	1656	25608	244008	1530060	6638340	20708460	47742654	82816632	109358220
Pr(l)	1	9.7×10^{-1}	8.3×10^{-1}	6.2×10^{-1}	4.1×10^{-1}	2.3×10^{-1}	1.2×10^{-1}	5.8×10^{-2}	2.5×10^{-2}	1.0×10^{-2}

11	12	13	14	15	16	17	18	19	20
110895036	87031844	53227980	25473732	9519912	2743080	591864	90420	8760	408
3.7×10^{-3}	1.3×10^{-3}	4.3×10^{-4}	1.4×10^{-4}	4.3×10^{-5}	1.3×10^{-5}	4.0×10^{-6}	1.2×10^{-6}	3.8×10^{-7}	1.2×10^{-7}

The first (second and third resp.) row gives the length l (occurrence number Nb(l) and probability Pr(l) resp.) of CFC.

Table 2b
The 28 codes invariant by letter permutation associated with the 408 comma-free codes CFC of length 20

{aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, ccd, dab, dac, dad, dbc, dbd, dcd}
{aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, ccd, dab, dac, dad, dbc, dbd, ddc}
{aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, cdb, cdc, dab, dac, dad, ddb, ddc}
{aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, cdc, cdd, dab, dac, dad, dbc, dbd}
{aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, cdd, dab, dac, dad, dbc, dbd, dcc}
{aab, aac, aad, bab, bac, bad, bbc, bda, bdb, bdc, cab, cac, cad, cbc, cda, cdb, cdc, dda, ddb, ddc}
{aab, aac, aad, bab, bac, bad, bbc, bda, bdb, bdc, cab, cac, cad, ccb, cda, cdb, cdc, dda, ddb, ddc}
{aab, aac, aad, bab, bac, bad, bbc, bdb, bdc, bdd, cab, cac, cad, cbc, cdb, cdc, cdd, dab, dac, dad}
{aab, aac, aad, bab, bac, bad, bbc, bdb, bdc, bdd, cab, cac, cad, ccb, cdb, cdc, cdd, dab, dac, dad}
{aab, aac, aad, bab, bac, bad, bca, bcb, bcd, bdb, bdd, cca, ccb, ccd, dab, dac, dad, dca, dcb, dcd}
{aab, aac, aad, bab, bac, bad, bca, bcb, bcd, bdd, cca, ccb, ccd, dab, dac, dad, dbb, dca, dcb, dcd}
{aab, aac, aad, bab, bac, bad, bcb, bcc, bcd, bdb, bdd, cab, cac, cad, dab, dac, dad, dcb, dcc, dcd}
{aab, aac, aad, bab, bac, bad, bcb, bcc, bcd, bdd, cab, cac, cad, dab, dac, dad, dbb, dcb, dcc, dcd}
{aab, aac, aad, bab, bac, bad, bcb, bcc, bdb, bdd, cab, cac, cad, cdb, cdd, dab, dac, dad, dcb, dcc}
{aab, aac, ada, adb, adc, add, bab, bac, bbc, bda, bdb, bdc, bdd, cab, cac, cbc, cda, cdb, cdc, cdd}
{aab, aac, ada, adb, adc, add, bab, bac, bbc, bda, bdb, bdc, bdd, cab, cac, ccb, cda, cdb, cdc, cdd}
{aab, aac, ada, adb, adc, add, bab, bac, bca, bcb, bda, bdb, bdc, bdd, cca, ccb, cda, cdb, cdc, cdd}
{aab, aac, ada, adb, adc, add, bab, bac, bcb, bcc, bda, bdb, bdc, bdd, cab, cac, cda, cdb, cdc, cdd}
{aab, aac, ada, adb, adc, add, bab, bac, bcc, bda, bdb, bdc, bdd, cab, cac, cbb, cda, cdb, cdc, cdd}
{aab, aca, acb, acc, acd, ada, adb, add, bab, bca, bcb, bcc, bcd, bda, bdb, bdd, dca, dcb, dcc, dcd}
{aab, aca, acb, acc, acd, ada, adb, add, bba, bca, bcb, bcc, bcd, bda, bdb, bdd, dca, dcb, dcc, dcd}
{aab, aca, acb, acc, ada, adb, add, bab, bca, bcb, bcc, bda, bdb, bdd, cda, cdb, cdd, dca, dcb, dcc}
{aab, aca, acb, acc, ada, adb, add, bba, bca, bcb, bcc, bda, bdb, bdd, cda, cdb, cdd, dca, dcb, dcc}
{aba, abb, abc, abd, aca, acc, acd, ada, add, cba, cbb, cbc, cbd, dba, dbb, dbc, dbd, dca, dcc, dcd}
{aba, abb, abc, abd, aca, acc, acd, add, cba, cbb, cbc, cbd, daa, dba, dbb, dbc, dbd, dca, dcc, dcd}
{aba, abb, abc, abd, aca, acc, ada, add, cba, cbb, cbc, cbd, cda, cdd, dba, dbb, dbc, dbd, dca, dcc}
{aba, abb, abc, aca, acc, ada, adc, add, bda, bdc, bdd, cba, cbb, cbc, cda, cdc, cdd, dba, dbb, dbc}
{aba, abb, abc, acc, ada, adc, add, bda, bdc, bdd, caa, cba, cbb, cbc, cda, cdc, cdd, dba, dbb, dbc}

length 11. The 408 CFC of length 20 have the lowest occurrence probability (1.2×10^{-7}) and can be presented by 28 codes invariant by letter permutation (Table 2b).

Result 2. Table 3a shows the growth function of self-complementary comma-free codes CCFC on β_4^3. It reaches a maximum with 642 CCFC of length eight. There is no CCFC of lengths 18 and 20. The four CCFC of length 16 have the lowest occurrence probability (1.9×10^{-11}) and can be presented by a unique code invariant by letter permutation and based on the complementarity map $a = A$, $b = C$, $c = T$ and $d = G$ (Table 3b).

Result 3. Table 4a shows the growth function of C^3 comma-free codes C^3CFC on β_4^3. It reaches a maximum with 854532 C^3CFC of length seven. There is no C^3CFC of lengths 17, 18, 19 and 20. The 18 C^3CFC of length 16 have the lowest occurrence probability (8.6×10^{-11}) and can be presented by three codes invariant by letter permutation (Table 4b).

Table 3a
Growth function of self-complementary comma-free codes CCFC on β_4^3

l	2	4	6	8	10	12	14	16	18	20
Nb(l)	28	210	556	642	396	152	36	4	0	0
Pr(l)	1.6×10^{-2}	5.4×10^{-4}	2.0×10^{-5}	7.8×10^{-7}	3.6×10^{-8}	2.3×10^{-9}	1.9×10^{-10}	1.9×10^{-11}	0	0

The first (second and third resp.) row gives the length l (occurrence number Nb(l) and probability Pr(l) resp.) of CCFC.

Table 3b
The unique code invariant by letter permutation associated with the four self-complementary comma-free codes CCFC of length 16 based on the complementarity map a $= A$, b $= C$, c $= T$ and d $= G$

{aab, aac, abb, abc, acb, acc, adb, adc, dab, dac, dbb, dbc, dcb, dcc, ddb, ddc}

Table 4a
Growth function of C^3 comma-free codes C^3CFC on β_4^3

l	1	2	3	4	5	6	7	8	9	10
Nb(l)	60	1548	18504	109824	353988	680616	854532	751842	493920	256800
Pr(l)	1	9.1×10^{-1}	6.0×10^{-1}	2.8×10^{-1}	9.4×10^{-2}	2.4×10^{-2}	5.0×10^{-3}	9.1×10^{-4}	1.5×10^{-4}	2.4×10^{-5}

11	12	13	14	15	16	17	18	19	20
109692	38604	10764	2196	288	18	0	0	0	0
3.7×10^{-6}	5.8×10^{-7}	8.7×10^{-8}	1.2×10^{-8}	1.3×10^{-9}	8.6×10^{-11}	0	0	0	0

The first (second and third resp.) row gives the length l (occurrence number Nb(l) and probability Pr(l) resp.) of C^3CFC.

Table 4b
The three codes invariant by letter permutation associated with the 18 C^3 comma-free codes C^3CFC of length 16

{aab, aac, abb, abc, acb, acc, adb, adc, dab, dac, dbb, dbc, dcb, dcc, ddb, ddc}
{aab, aac, adb, adc, bab, bac, bdb, bdc, cab, cac, cdb, cdc, dab, dac, ddb, ddc}
{aba, abb, abc, abd, aca, acb, acc, acd, dba, dbb, dbc, dbd, dca, dcb, dcc, dcd}

Table 5
Growth function of C^3 self-complementary comma-free codes C^3CCFC on β_4^3

l	2	4	6	8	10	12	14	16	18	20
Nb(l)	28	182	424	498	340	144	36	4	0	0
Pr(l)	1.6×10^{-2}	4.6×10^{-4}	1.5×10^{-5}	6.0×10^{-7}	3.1×10^{-8}	2.2×10^{-9}	1.9×10^{-10}	1.9×10^{-11}	0	0

The first (second and third resp.) row gives the length l (occurrence number Nb(l) and probability Pr(l) resp.) of C^3CCFC. The unique code invariant by letter permutation associated with the four C^3CCFC of length 16 is identical to the code associated with the four CCFC of length 16 (Table 3b).

Result 4. Table 5 shows the growth function of C^3 self-complementary comma-free codes C^3CCFC on β_4^3. It reaches a maximum with 498 C^3CCFC of length eight. There is no C^3CCFC of lengths 18 and 20. The four C^3CCFC of length 16 can be presented by the code invariant by letter permutation associated with the four CCFC of length 16 (Table 3b).

Result 5. Table 6 shows the growth function of maximal comma-free codes MCFC on β_4^3. It reaches a maximum with 10488 MCFC of length 16. There is no MCFC of lengths 1–8. There are several unexpected results. The number of MCFC of length 13 is less than the ones of MCFC of lengths 12 and 14. The 408 MCFC of length 20, obviously maximal and presented by the 28 codes invariant by letter permutation associated with the 408 CFC of length 20 (Table 2b), do not occur with the lowest probability which is observed with the 96 MCFC of length nine.

Table 6
Growth function of maximal comma-free codes MCFC on β_4^3

l	1	2	3	4	5	6	7	8	9	10
Nb(l)	0	0	0	0	0	0	0	0	96	1152
Pr(l)	0	0	0	0	0	0	0	0	2.9×10^{-8}	1.1×10^{-7}

11	12	13	14	15	16	17	18	19	20
4224	6708	4632	8040	8568	10488	4848	3072	960	408
1.4×10^{-7}	1.0×10^{-7}	3.7×10^{-8}	4.3×10^{-8}	3.9×10^{-8}	5.0×10^{-8}	3.3×10^{-8}	4.2×10^{-8}	4.1×10^{-8}	1.2×10^{-7}

The first (second and third resp.) row gives the length l (occurrence number Nb(l) and probability Pr(l) resp.) of MCFC. The 28 codes invariant by letter permutation associated with the 408 MCFC of length 20 are (obviously) identical with the 28 codes associated with the 408 CFC of length 20 (Table 2b).

Table 7
Growth function of self-complementary maximal comma-free codes CMCFC on β_4^3

l	2	4	6	8	10	12	14	16	18	20
Nb(l)	0	0	0	0	0	4	0	4	0	0
Pr(l)	0	0	0	0	0	6.0×10^{-11}	0	1.9×10^{-11}	0	0

The first (second and third resp.) row gives the length l (occurrence number Nb(l) and probability Pr(l) resp.) of CMCFC. The unique code invariant by letter permutation associated with the four CMCFC of length 16 is identical to the code associated with the four CCFC of length 16 (Table 3b).

Table 8a
Growth function of C^3 maximal comma-free codes C^3MCFC on β_4^3

l	1	2	3	4	5	6	7	8	9	10
Nb(l)	0	0	0	0	0	0	0	0	24	72
Pr(l)	0	0	0	0	0	0	0	0	7.3×10^{-9}	6.6×10^{-9}

11	12	13	14	15	16	17	18	19	20
192	124	72	24	0	6	0	0	0	0
6.5×10^{-9}	1.9×10^{-9}	5.8×10^{-10}	1.3×10^{-10}	0	2.9×10^{-11}	0	0	0	0

The first (second and third resp.) row gives the length l (occurrence number Nb(l) and probability Pr(l) resp.) of C^3MCFC.

Table 8b
The unique code invariant by letter permutation associated with the six C^3 maximal comma-free codes C^3MCFC of length 16

{aab, aac, abb, abc, acb, acc, adb, adc, dab, dac, dbb, dbc, dcb, dcc, ddb, ddc}

Result 6. Table 7 shows the growth function of self-complementary maximal comma-free codes CMCFC on β_4^3. There are only four CMCFC of length 12 and four CMCFC of length 16. Unexpectedly, there is no CMCFC of length 14. The four CMCFC of length 16 occur with the lowest probability 1.9×10^{-11}.

Result 7. Table 8a shows the growth function of C^3 maximal comma-free codes C^3MCFC on β_4^3. It reaches a maximum with 192 C^3MCFC of length 11. There is obviously no C^3MCFC of lengths 1–8 such as the MCFC (Table 6). There is obviously no C^3MCFC of lengths 17–20 such as the C^3CFC (Table 4a). Unexpectedly, there is also no C^3MCFC of length 15. The six C^3MCFC of length 16 have the lowest occurrence probability (2.9×10^{-11}) and can be presented by a unique code invariant by letter permutation (Table 8b).

Result 8. Table 9 shows the growth function of C^3 self-complementary maximal comma-free codes C^3CMCFC on β_4^3 which is identical to the one of CMCFC (Table 7).

Table 9
Growth function of C^3 self-complementary maximal comma-free codes C^3CMCFC on β_4^3

l	2	4	6	8	10	12	14	16	18	20
Nb(l)	0	0	0	0	0	4	0	4	0	0

The first (second resp.) row gives the length l (occurrence number Nb(l) resp.) of C^3CMCFC, the occurrence probabilities of C^3CMCFC being identical to Table 7. The unique code invariant by letter permutation associated with the four C^3CMCFC of length 16 is identical to the code associated with the four CCFC of length 16 (Table 3b).

5. Conclusion

New varieties of comma-free codes CFC of length three on the four-letter alphabet are defined and analysed: self-complementary comma-free codes CFC (CCFC), C^3 comma-free codes (C^3CFC), C^3 self-complementary comma-free codes (C^3CCFC), self-complementary maximal comma-free codes (CMCFC), C^3 maximal comma-free codes (C^3MCFC) and C^3 self-complementary maximal comma-free codes (C^3CMCFC). New properties with words of length three, four, five and six in comma-free codes are used for the determination of growth functions in these code varieties. New unexpected results are observed. In particular, the distributions of maximal comma-free codes MCFC, CMCFC and C^3MCFC present unexplained variations, and there are no self-complementary comma-free codes CCFC of length 20, in contrast with the circular codes of length 20 which can be self-complementary [7].

Acknowledgment

We thank T. Ludwig for verifying the correctness of a few computer results.

References

[1] F.H.C. Crick, J.S. Griffith, L.E. Orgel, Codes without commas, Proc. Natl. Acad. Sci. USA 43 (1957) 416–421.
[2] S.W. Golomb, B. Gordon, L.R. Welch, Comma-free codes, Canad. J. Math. 10 (1958) 202–209.
[3] S.W. Golomb, L.R. Welch, M. Delbrück, Construction and properties of comma-free codes, Biol. Medd. Dan. Vid. Selsk. 23 (1958).
[4] M.W. Nirenberg, J.H. Matthaei, The dependance of cell-free protein synthesis in E. Coli upon naturally occurring or synthetic polyribonucleotides, Proc. Natl. Acad. Sci. USA 47 (1961) 1588–1602.
[5] F.H.C. Crick, S. Brenner, A. Klug, G. Pieczenik, A speculation on the origin of protein synthesis, Origins of Life 7 (1976) 389–397.
[6] M. Eigen, P. Schuster, The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle, Naturwissenschaften 65 (1978) 341–369.
[7] D.G. Arquès, C.J. Michel, A complementary circular code in the protein coding genes, J. Theoret. Biol. 182 (1996) 45–58.
[8] D.G. Arquès, C.J. Michel, A circular code in the protein coding genes of mitochondria, J. Theoret. Biol. 189 (1997) 273–290.
[9] G. Frey, C.J. Michel, Circular codes in archaeal genomes, J. Theoret. Biol. 223 (2003) 413–431.
[10] G. Frey, C.J. Michel, Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes, J. Comput. Biol. Chem. 30 (2006) 87–101.
[11] 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, 2003.
[12] G. Pirillo, M.A. Pirillo, Growth function of self-complementary circular codes, Biology Forum 98 (2005) 97–110.