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 $w = w[0]w[1]w[2] \ldots w[i] \ldots w[j] \ldots w[n]$ a word of length $n + 1$ on β_4. Then, we say that the factor $w[i] \ldots w[j]$ 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 $u, v \in \beta_4^+$.