Journal of Theoretical Biology 364 (2015) 113–120

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

On the hierarchy of trinucleotide n-circular codes and their corresponding amino acids

Elena Fimmel *, Lutz Strüngmann

Institute of Applied Mathematics, Faculty for Computer Sciences, Mannheim University of Applied Sciences, 68163 Mannheim, Germany

HIGHLIGHTS

- Five new hierarchically ordered classes of trinucleotide codes are introduced.
- An easy test-criterion for circularity of codes is developed.
- The maximal number of amino acids encoded by circular codes is investigated.
- The circularity of the RNY-primeval code and related codes is shown.

ARTICLE INFO

Article history:
Received 9 May 2014
Received in revised form
3 September 2014
Accepted 7 September 2014
Available online 16 September 2014

Keywords:
Comma-free code, Error-detecting codes

ABSTRACT

Circular codes are putative remnants of primeval comma-free codes and are potentially involved in detecting and maintaining the normal reading frame in protein coding sequences. In Michel and Pirillo (2013a) it was shown by computer algorithm that no maximal trinucleotide circular code can encode more than 18 different amino acids under the standard version of the genetic code. For comma-free codes the maximum is even less, namely 13 (Michel, 2014). The main purpose of this paper is to investigate these facts from a mathematical point of view and to show why the codes with the best-known error detecting properties are limited in the number of amino acids they can encode. We introduce five hierarchically ordered classes of trinucleotide codes including the well-known comma-free and circular codes and prove combinatorically that it is impossible to encode all amino acids using codes from four out of the five classes that have the strongest error detecting properties. However, it is possible to encode all 20 amino acids using codes from the largest class with the weakest properties. Additionally, we develop a handy criterion for circularity. As an application, it is shown that all codes from a special class of trinucleotide codes which includes the RNY-primeval code (Shepherd, 1986) are automatically circular. We also list which amino acids these codes encode.

1. Introduction

In 1953 Crick and Watson published their pioneering discovery of the right-handed double-helix structure of DNA. Their article 'A Structure for Deoxyribose Nucleic Acid' (Watson and Crick, 1953) turned biology upside down and since then many attempts have been made to model the assignment of codons to amino acids so as to explain the chemical and biological symmetries of the genetic code (see for example Gonzalez, 2008; Gonzalez et al., 2008, 2011; Jestin, 2006; Koch and Lehmann, 1997; Negadi, 2009; Rumer, 1969; Sciarrino, 2003). It was Gamow who first initiated in his letter to Crick and Watson (1953) a search for formal rules describing the genetic code (Nanjundiah, 2004).

It is well known that the genetic code is written with words of three letters called codons which are built on an alphabet of four letters, nucleotide bases Uracil (Thymine), Cytosin, Adenine, and Guanine, in short $U(T), C, A, G$. One of the first natural ideas is to find a proper classification of the elementary words of the genetic code, the codons, explaining the degeneracy in their assignment to amino acids. In the late 1950s biologists became aware of the so-called frame-shift problem: a sequence of codons can be translated only in a right frame into the right amino acids. For this problem Crick suggested a solution that was received enthusiastically. He assumed that the adaptor molecules might exist for only a subset of all possible codons, the remaining codons are the 'nonsense codons' (Hayes, 1998). The codes in which all out-of-frame triplets are nonsense and which can be meaningfully read in only one frame are called comma-free codes. However, there is still no

* Corresponding author.
E-mail addresses: e.fimmel@hs-mannheim.de (E. Fimmel), l.struengmann@hs-mannheim.de (L. Strüngmann).

http://dx.doi.org/10.1016/j.jtbi.2014.09.011
0022-5193/

experimental evidence for the existence of comma-free codes in nature. Golomb and collaborators showed that there are 408 maximal comma-free codes and gave a method to find them (Golomb et al., 1958, 1958). Later a less restrictive family of codes called circular codes was introduced and discovered (see, for instance, Arquès and Michel, 1996; Michel et al., 2008, 2008, 2012; Michel and Pirillo, 2010, 2013a, 2013b; Michel, 2012, 2014). The circular codes have the property that only some out-of-frame triplets are nonsense, which still suffices to recognize an out-of-frame shift. Thus, the comma-free and circular codes are error-detecting codes for the genetic information. In contrast to comma-free codes, circular codes were indeed found on large gene populations of eukaryotes and prokaryotes (Arquès and Michel, 1996). This has initiated intensive studies, mostly experimental, since it seems that circular codes are indeed potentially involved in maintaining the normal reading frame.

The paper is structured as follows. In Section 2 of the present paper we will discuss the basic definitions. In Section 3.3 we introduce five hierarchically ordered classes of trinucleotide codes. Among them comma-free codes build the smallest class with the most restrictive properties. The next class with less restrictive properties are the circular codes, followed by the trinucleotide 3-circular, the 2-circular and the 1-circular codes introduced in the present paper. Numerous examples will be given to show that all these classes are different. The results obtained in Section 3.3 together with the discussion from Section 4 prove that all these classes are ordered by proper inclusion.

In Section 4 we will give a useful criterion to test a given trinucleotide code for its circularity and afterwards discuss the question how many different amino acids can be coded by codes with properties introduced in Section 3.3. It is known that the maximal number of amino acids that can be encoded by a circular code under the standard version of the genetic code is 18 (Michel and Pirillo, 2013a). This result was obtained by an intensive computer algorithm listing all possible maximal trinucleotide circular codes (Herrmann et al., 2013). Among 12,964,440 possible maximal trinucleotide circular codes there are only 10 which code 18 amino acids (see Michel and Pirillo, 2013a). We will show in Section 4 of this paper that it is theoretically impossible under the standard version of the genetic code to encode all 20 amino acids having a comma-free, a trinucleotide circular or even a 2-circular code but it is possible to encode all 20 amino acids using trinucleotide 1-circular codes. It will also be shown in Section 4 that for some special class of trinucleotide codes which include the so-called RNY-primeval code (see Shepherd, 1986) the circularity is automatically given. Section 5 contains the conclusions.

2. Definitions

In this section we recall the basic notions and definitions from Michel and Pirillo (2013a) and reformulate them in a way which is suitable for our considerations. Let us denote the nucleotide bases alphabet as

$$\mathcal{B} := \{U(T), C, A, G\}$$

where U stands for Uracil, C stands for Cytosine, A stands for Adenine and G stands for Guanine. Given a codon $x = N_1N_2N_3 \in \mathcal{B}^3$ we let

$$\alpha_0(x) = id(x) = N_1N_2N_3, \quad \alpha_1(x) = N_2N_3N_1, \quad \text{and} \quad \alpha_2(x) = N_3N_1N_2$$

be the codons obtained from x by a shift of 0,1, and 2 positions, respectively. We will say that two codons $x_1, x_2 \in \mathcal{B}^3$ are *cyclically equivalent* if $x_1 \in \{\alpha_0(x_2), \alpha_1(x_2), \alpha_2(x_2)\}$. It is easy to see that the relation defined above is an equivalence relation on the set of codons $\mathcal{B}^3$, i.e. it is reflexive, symmetric, and transitive. For instance, *ATC* is equivalent to *TCA* and also to *CAT*.

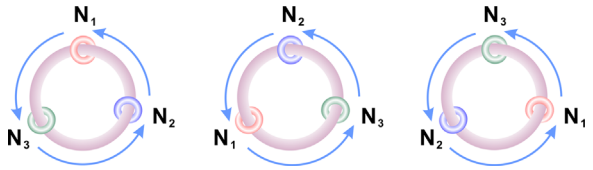

Fig. 1. The codons $N_1N_2N_3$, $N_2N_3N_1$, and $N_3N_1N_2$ are cyclically equivalent for any bases $N_1, N_2, N_3 \in \mathcal{B}$.

Table 1
The table of equivalence (conjugacy) classes of codons containing three elements.

$D_1 = \{AAC, ACA, CAA\}$	$D_2 = \{AAG, AGA, GAA\}$
$D_3 = \{AAT, ATA, TAA\}$	$D_4 = \{ACC, CCA, CAC\}$
$D_5 = \{ACG, CGA, GAC\}$	$D_6 = \{ACT, CTA, TAC\}$
$D_7 = \{AGC, GCA, CAG\}$	$D_8 = \{AGG, GGA, GAG\}$
$D_9 = \{AGT, GTA, TAG\}$	$D_{10} = \{ATC, TCA, CAT\}$
$D_{11} = \{ATG, TGA, GAT\}$	$D_{12} = \{ATT, TTA, TAT\}$
$D_{13} = \{CCG, CGC, GCC\}$	$D_{14} = \{CCT, CTC, TCC\}$
$D_{15} = \{CGG, GGC, GCG\}$	$D_{16} = \{CGT, GTC, TCG\}$
$D_{17} = \{CTG, TGC, GCT\}$	$D_{18} = \{CTT, TTC, TCT\}$
$D_{19} = \{GGT, GTG, TGG\}$	$D_{20} = \{GTT, TTG, TGT\}$

Geometrically the equivalence relation means that two equivalent codons cannot be distinguished when *read on a circle*, i.e. one codon can be obtained from the other by a cyclic permutation of the defining bases, a rotation of the circle (Fig. 1).

Let us consider the equivalence classes (conjugacy classes) of codons with respect to the cyclical equivalence. Clearly, for four codons *AAA*, *CCC*, *GGG*, *UUU* the corresponding equivalence classes contain only one element. The remaining 20 equivalence classes contain three elements each and, thus, are *complete classes* (compare for example Pearson, 2003).

Table 1 lists all the complete conjugacy classes (see also Michel et al., 2012) and the corresponding amino acids they code for.

It will be of interest to us in the sequel how the amino acids are distributed among the equivalence classes of codons and vice versa. Table 2 shows which amino acids are coded by the complete conjugacy classes. In particular, it can be seen from Table 2 that codons from the same equivalence class never code for the same amino acid. Hence each complete equivalence class of codons codes for three different amino acids or the stop signal.

Table 3 combines all the information on the codons with respect to their conjugacy class and the amino acids they code for. In particular it shows in which conjugacy classes a given amino acid participates. As mentioned above every complete equivalence class codes for three different amino acids (or the stop signal), hence the number of equivalence classes that code for a particular amino acid equals the degeneracy of this amino acid. This fact can already be interpreted as a kind of coding: a coding sequence of codons can be uniquely reconstructed by its corresponding sequence of amino acids and the sequence of equivalence class numbers.

Finally, we give a technical definition that will be helpful to determine the structure of certain codes in the next section.

Definition 2.1. Let $X \subseteq \mathcal{B}^3$. We will denote by π_i $(i = 1, 2, 3)$ the map

$$\pi_i : \mathcal{B}^3 \to \mathcal{B}$$

which assigns to each codon $x \in \mathcal{B}^3$ its ith coordinate (projection onto the ith coordinate). $\pi_i(X) \subseteq \mathcal{B}$ denotes the set of all ith coordinates of the elements of X.

Table 2
The table of amino acids encoded by the equivalence (conjugacy) classes of codons containing three elements, $A(D_i)$ denotes the set of amino acids encoded by codons from D_i.

$A(D_1)=\{asparagine, glutamine, threonine\}$	$A(D_2)=\{arginine, glutamate, lysine\}$
$A(D_3)=\{asparagine, isoleucine, stop\}$	$A(D_4)=\{histidine, proline, threonine\}$
$A(D_5)=\{arginine, aspartate, threonine\}$	$A(D_6)=\{leucine, threonine, tyrosine\}$
$A(D_7)=\{alanine, glutamine, serine\}$	$A(D_8)=\{glutamate, glycine, arginine\}$
$A(D_9)=\{serine, valine, stop\}$	$A(D_{10})=\{histidine, isoleucine, serine\}$
$A(D_{11})=\{aspartate, methionine, stop\}$	$A(D_{12})=\{isoleucine, leucine, tyrosine\}$
$A(D_{13})=\{alanine, arginine, proline\}$	$A(D_{14})=\{leucine, proline, serine\}$
$A(D_{15})=\{alanine, arginine, glycine\}$	$A(D_{16})=\{arginine, serine, valine\}$
$A(D_{17})=\{alanine, cysteine, leucine\}$	$A(D_{18})=\{leucine, phenylalanine, serine\}$
$A(D_{19})=\{glycine, tryptophan, valine\}$	$A(D_{20})=\{cysteine, leucine, valine\}$

Table 3
The standard nuclear correspondence between codons and amino acids. The index i on the top of the codons assigns the equivalence class D_i (compare Table 1).

Amino acid	5′→3′ Codon sequence
Alanine	GCT^{17}, GCC^{13}, GCA^{7}, GCG^{15}
Arginine	CGT^{16}, CGC^{13}, CGA^{5}, CGG^{15}, AGA^{2}, AGG^{8}
Asparagine	AAT^{3}, AAC^{1}
Aspartate	GAT^{11}, GAC^{5}
Cysteine	TGT^{20}, TGC^{17}
Glutamate	GAA^{2}, GAG^{8}
Glutamine	CAA^{1}, CAG^{7}
Glycine	GGT^{19}, GGC^{15}, GGA^{8}, GGG
Histidine	CAT^{10}, CAC^{4}
Isoleucine	ATT^{12}, ATC^{10}, ATA^{3}
Leucine	TTA^{12}, TTG^{20}, CTT^{18}, CTC^{14}, CTA^{6}, CTG^{17}
Lysine	AAA, AAG^{2}
Methionine	ATG^{11}
Phenylalanine	TTT, TTC^{18}
Proline	CCT^{14}, CCC, CCA^{4}, CCG^{13}
Serine	TCT^{18}, TCC^{14}, TCA^{10}, TCG^{16}, AGT^{9}, AGC^{7}
Threonine	ACT^{6}, ACC^{4}, ACA^{1}, ACG^{5}
Tryptophan	TGG^{19}
Tyrosine	TAT^{12}, TAC^{6}
Valine	GTT^{20}, GTC^{16}, GTA^{9}, GTG^{19}

Note that for any subset $X\subseteq\mathcal{B}^3$ the sets $\pi_1(X)$, $\pi_2(X)$ and $\pi_3(X)$ can contain at most four elements. For instance, $\pi_2(TGA)=G$, $\pi_1(CCG)=C$ and $\pi_3(TAG)=G$ as well as

$\pi_1(\{CCG, TGA, GAA, GTC, TCT\})=\{C, G, T\}$.

3. Trinucleotide codes

Based on and motivated by Michel et al. (2012) we will consider various classes of trinucleotide codes in this section. A *trinucleotide code* is a subset X of the set $\mathcal{B}^3$ of all codons, e.g. $\mathcal{B}^3$ itself is a trinucleotide code but also just a single codon $\{N_1N_2N_3\}$. It has been observed that in the coding sequences certain codons are used more often than others. This seems to indicate that nature has built in some error detecting and probably also correcting mechanism in the sense that subcodes of $\mathcal{B}^3$ that allow the detection of frame shifts are used more frequently. Here we discuss a hierarchy of such codes including the classical comma-free codes and circular codes.

3.1. Trinucleotide n-circular codes

We begin with the definition of trinucleotide n-circular codes for natural numbers $n\in N$.

Definition 3.1. Let $n\in N$ and $X\subseteq\mathcal{B}^3$ be a trinucleotide code. We say that X is a *trinucleotide n-circular code* if for any concatenation $x_1\cdots x_m$ of $m\leq n$ codons from X, there is only one partition into codons from X when read on a circle, i.e. the next letter after the last letter being the first letter.

For instance the sequence $ACTGTAAAC$ would read on a circle as the infinite sequence

$ACTGTAAACACTGTAAACACTGTAAAC\cdots\cdots$.

Before we give several examples we would like to state an easy lemma.

Lemma 3.2. *The following holds for $n\in N$:*

(a) *If X is a trinucleotide n-circular code, then X is also m-circular for all $m\leq n$.*
(b) *A trinucleotide n-circular code can contain at most one element from each complete equivalence class and cannot contain the codons AAA, CCC, GGG, TTT.*
(c) *A trinucleotide code $X\subseteq\mathcal{B}^3$ is1-circular if and only if X contains at most one codon from each complete conjugacy class and none of the codons AAA, CCC, GGG, TTT.*

The following examples show that the classes of n-circular codes build a proper hierarchy for $n=1,2,3,4$. It will be shown in Theorem 4.3 that for $n\geq 5$ the class of trinucleotide n-circular codes coincides with the class of trinucleotide 4-circular codes.

Example 3.3. Let us consider the following examples:

- The set of codons

  $X:=\{TGG, GTG\}$

  is not a trinucleotide 1-circular code since it contains two codons from the conjugacy class 19.
- The set of codons

  $X:=\{TGG, CTG, GGC, TGT\}$

  is a trinucleotide 1-circular but not a trinucleotide 2-circular code: It contains only codons from different conjugacy classes while for example the word $w=TGGCTG$ has two factorizations into words from X since GGC and TGT are also in the code.
- The set of codons

  $X:=\{ACG, GTA, CGT, CGG, TAC\}$

  is a trinucleotide 2-circular but not a trinucleotide 3-circular code.
  The word $w=ACGGTACGT$ has two factorizations on a circle

  $ACG|GTA|CGT$ and $CGG|TAC|GTA$.

  At the same time it is easy to see that the concatenation of any two codons from X has a unique factorization over X.

- The set of codons
  $X = \{CGT, ACG, TAC, GTA\}$
  is a trinucleotide 3-circular but not a trinucleotide 4-circular code.
  The word $w = CGTACGTACGTA$ has two factorizations on a circle
  $CGT|ACG|TAC|GTA$ and $GTA|CGT|ACG|TAC$.
  Thus X is not 4-circular. However, X is clearly 1-circular and at the same time the concatenation of any two different codons from X either do not contain codons from X at all in shift one or two, like in
  $CGTTAC$, $ACGGTA$, $TACCGT$, $GTAACG$
  or they contain subcodons (three adjacent letters from $\mathcal{B}$) from X in shift one or two but do not have a second factorization over X which can be easily checked for the remaining pairs
  $CGTACG$, $CGTGTA$, $ACGCGT$, $ACGTAC$, $TACACG$,
  $TACGTA$, $GTACGT$, $GTATAC$

  The concatenations of three different codons from X on a circle always contain a subcodon that has the same nucleotide basis in two successive positions. However, every codon from X contains a nucleotide basis at most one time. This shows that every concatenation of three codons from X has a unique decomposition over X on a circle.

The above examples prove that the notions of trinucleotide n-circularity differ for $n = 1, 2, 3, 4$. As mentioned above we will show in Section 4 (see Theorem 4.3) that trinucleotide 4-circularity coincides with the traditional notion of trinucleotide circularity and hence with n-circularity for all $n \geq 5$.

3.2. Trinucleotide circular codes

In this section we discuss the classical notion of trinucleotide circular codes (see for instance Michel et al., 2012).

Definition 3.4. Let $X \subseteq \mathcal{B}^3$. We call X a *trinucleotide circular code* if it is trinucleotide n-circular for all $n \in N$. Equivalently, this means that any word over the alphabet $\mathcal{B}$ written on a circle has at most one decomposition into words from X.

The following lemma shows that the definition of a trinucleotide circular code as given above is equivalent to Definition 5 from Michel et al. (2012). We will say that a subword $\mathbf{v}$ of the word $w \in \mathcal{B}^*$ is *at position* $(\mathbf{i}, \mathbf{j})$, $1 \leq \mathbf{i} \leq \mathbf{j}$ *in* $\mathbf{w}$ if v starts with the ith and stops with the jth letter of w.

Lemma 3.5. *Let $k \in \mathbb{N}$, $X \subseteq \mathcal{B}^3$, $w \in X^*$ a concatenation of k words of X written on a circle, v_j a subword of length 3 of the word w at position $(j, (j+2) \bmod 3k)$. Then X is a trinucleotide circular code if and only if the following property holds: there exist $j_1, j_2 \in \{1, \ldots, 3k\}$ so that*

- $v_{j_1} \in X$, $v_{j_1-1} \notin X$ *and*
- $v_{j_2} \in X$, $v_{j_2+1} \notin X$

Proof. Clearly, if the property does not hold then there are at least two factorizations of a word w on a circle. The converse direction is trivial. □

Obviously, a trinculeotide n-circular code can contain at most 20 codons since it must not contain more than one codon from each complete equivalence class and must exclude the codons AAA, CCC, GGG, and TTT. This inspires the following definition.

Definition 3.6. We will call a trinucleotide n-circular code X *maximal* if it contains exactly 20 codons.

Thus, there are at most 3^{20} potential different maximal trinucleotide circular codes. In reality (compare Michel et al., 2012) there is less than one per cent of this number, namely 12,964,440 maximal trinucleotide circular codes which can code for at most 18 amino acids under the standard version of the genetic code. The number of maximal trinucleotide n-circular codes for $n \leq 3$ is not known yet.

3.3. Comma-free codes

In this section we consider an even more restrictive class of codes than the trinucleotide circular codes, the so-called comma-free codes. Comma-free codes were introduced by Crick in the early 1950s as a possible solution to the frame-shift problem. His assumption was that the 20 amino acids coded as comma-free codes can also contain at most 20 codons. This conjecture turned out to be wrong but the comma-free codes still provide an interesting class of codes (compare for example Golomb et al., 1958; Michel et al., 2012).

Definition 3.7. A trinucleotide code $X \subseteq \mathcal{B}^3$ is called a *comma-free* code provided that given any two codons $x_1, x_2 \in X$ any sub-codon of the concatenation x_1x_2 except for x_1, x_2 themselves does not belong to X.

Note, that in the above definition it is intended that the set $\{AAA\}$ is not comma-free. We give some examples.

Example 3.8.

- The set $X = \{GGC, GCC\}$ is a trinucleotide comma-free code.
- The set of codons $Y = \{ATC, TCC, CAA\}$ on the contrary is not a comma-free code since the concatenation CAA and TCC contains as a substring $ATC \in Y$

  $CAATCC$.

As for trinucleotide circular codes any comma-free code can contain at most one codon from each complete equivalence class and must not contain the codons AAA, CCC, GGG, or TTT. We therefore say that a comma-free code is *maximal* if it contains exactly 20 codons. Thus, there are at most 3^{20} potential different maximal comma-free codes but it was shown in Golomb et al. (1958, 1958) that there are in fact only 408 maximal comma-free codes.

The following lemma shows that the definition given above is equivalent to Definition 4 from Michel et al. (2012).

Lemma 3.9. *Let $k \in \mathbb{N}, X \subseteq \mathcal{B}^3$, $w \in X^*$ a concatenation of k words of X. Then X is a comma-free code if and only if the following property holds.*

For all $1 \leq j \leq 3k-3, j \not\equiv 1 \bmod 3$ a subword v of length 3 of the word w at position $(j, j+2)$ is not *in X.*

Proof. It is obvious that the formulated property implies that a given code is comma-free. Let us assume that a subset $X \subseteq \mathcal{B}^3$ is a comma-free code according to the definition above and there is a subword $v \in X$ of a concatenation w of k words from X with $k > 2$ at position $(j, j+2)$ with $j \not\equiv 1 \bmod 3, j < 3k-4$. Without lost of generality let us assume that $j \equiv 2 \bmod 3$. Then a subword of w at position $(j-1, j+4)$ is a concatenation of two codons from X which contains as a subword v. That is in contradiction to the comma-free property of X. □

It is easy to see comparing Lemmas 3.9 and 3.5 that every trinucleotide comma-free code is automatically a trinucleotide circular code. The converse direction is not true.

Consider the following set of codons:

$X = \{ATC, TCC, CAA\}$.

X is not a comma-free code as shown above but X is a trinucleotide circular code (compare Corollary 4.6).

Thus, we have the following picture which visualizes the hierarchy of trinucleotide codes related to their error-detecting properties (Fig. 2).

4. Results

Throughout this section we assume the standard nuclear genetic code. It has been shown in Michel (2014) that a maximal comma-free code can encode at most 13 amino acids. Here is an example.

Example 4.1. The trinucleotide comma-free code

$X = \{AAC, AAG, ATA, CAC, GAC, CTA, CAG, GAG, GTA, ATC,$

$ATG, TTA, CCG, CTC, GCG, GTC, CTG, TTC, GTG, TTG\}$

encodes the 13 amino acids asparagine, lysine, isoleucine, histidine, aspartate, leucine, glutamine, glutamate, valine, methionine, proline, alanine, and phenylalanine.

The maximal number of amino acids that can be encoded by a circular code is 18 (Michel and Pirillo, 2013a). Here is an example.

Example 4.2. The trinucleotide circular code

$X = \{AAC, AAG, AAT, ACC, ATC, ATG, ATT, CAG, CCT, CGC,$

$GAC, GAG, GGC, GTA, TAC, TCG, TGC, TGG, TTC, TTG\}$

encodes the 18 amino acids tryptophan, glutamine, proline, cysteine, arginine, glycine, tyrosine, lysine, phenylalanine, asparagine, isoleucine, methionine, leucine, aspartate, glutamate, threonine, serine and valine.

In both cases the result was obtained using a computer programme (compare Fig. 1, Arquès and Michel, 1996). In this section we will give a short combinatoric proof that, apart from 1-circular codes, a single code cannot represent the whole set of 20 amino acids.

4.1. Characterizing trinucleotide codes

First of all we give a criterion which allows us to test a given code for its circularity in an easier way than considering all possible words (an infinite set!) over the alphabet $\mathcal{B}$ written on a circle. We would like to remark at this point that also the flower automaton allows us to decide if a trinucleotide code is circular or not without considering all possible words (see Fig. 1, Arquès and Michel, 1996). The theorem below shows in addition that the property of n-circularity for a given code X for all $n \geq 4$ coincides with its circularity.

Theorem 4.3. *Let $X \subseteq \mathcal{B}^3$ and put $n(X) = \max\{|\pi_1(X)|, |\pi_3(X)|\}$. Then X is trinucleotide circular if and only if it is trinucleotide $n(X)$-circular. In particular, X is trinucleotide circular if and only if it is trinucleotide 4-circular.*

Proof. See Appendix A. □

To illustrate the above theorem let us remark the following: while comma-free codes allow us to detect a frame-shift immediately, circular codes by definition just make sure that eventually a frame-shift will be discovered. However, Theorem 4.3 shows that in average a frame-shift will be detected after reading four codons. The reason is that letters in a coding sequence are soon repeated since the alphabet is finite. Moreover, the above theorem shows that circularity is much more connected to the first and third positions of codons rather than to the second position. Clearly, this is intuitive since positions one and third are the positions where codons are concatenated. Thus the theorem just reflects the fact that if the bases used at positions one and third are quite disjoint, e.g. when they come from different chemical classes like purine/pyrimidine or strong/weak, then a frame-shift will be easily detected. The next useful corollaries of Theorem 4.3 can be easily verified and explain this phenomenon.

Lemma 4.4. *Let $X \subseteq \mathcal{B}^3$ with $\pi_1(X) \cap \pi_3(X) = \emptyset$. Then X is trinucleotide circular.*

Proof. See Appendix B. □

Corollary 4.5. *The RNY-primeval code is trinucleotide circular.*

Proof. The RNY- primeval code consists of 8 amino acids $\{Gly, Thr, Asp, Ser, Val, Asn, Ile, Ala\}$ and the 16 codons

$X = \{GGT, GGC, ACT, ACC, AGC, AGT, GAC, GAT,$
$GTC, GTT, AAT, ATT, AAC, ATC, GCT, GCC\}$

with purine as first base and pyrimidine as third base. Thus we have $\pi_1(X) \cap \pi_3(X) = \emptyset$. According to Lemma 4.4 the primeval code is circular. □

Motivated by Lemma 4.4 we now look at the codes X having disjoint sets $\pi_1(X)$ and $\pi_3(X)$. All such codes satisfy Lemma 4.4 and hence are trinucleotide circular codes. We will consider for each partition of $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ into disjoint sets the unique maximal possible set of codons X_{max} with

$\pi_1(X_{max}) = \mathcal{B}_1, \ \pi_3(X_{max}) = \mathcal{B}_2$.

Clearly, for any code X with disjoint sets $\pi_1(X)$ and $\pi_3(X)$ there is a maximal code X_{max} with $\pi_1(X) \subseteq \pi_1(X_{max})$, $\pi_3(X) \subseteq \pi_3(X_{max})$ and $X \subseteq X_{max}$. Simply take X_{max} to be the set of all codons starting with a base from $\pi_1(X)$ and ending with a base from $\pi_3(X)$. The following table characterizes the amino acids that can be coded by trinucleotide circular codes X_{max} with disjoint sets $\pi_1(X_{max})$ and

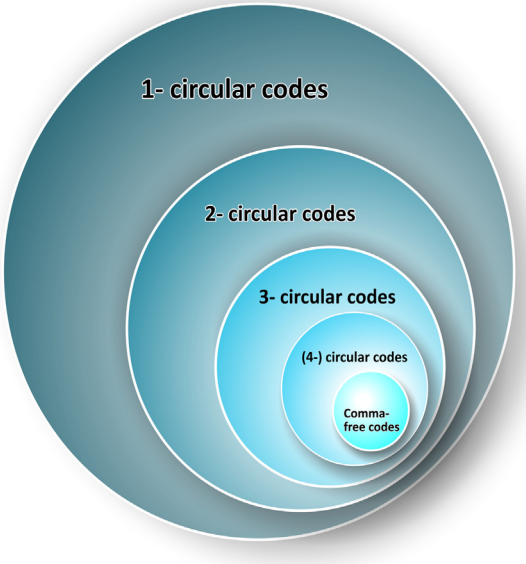

Fig. 2. Hierarchy of the classes of trinucleotide n-circular and comma-free codes.

$\pi_3(X_{max})$. As one can see the maximal number of amino acids that can be coded by such a code is 12 for example for $\mathcal{B}_1 = \{A, C, G\}$ and $\mathcal{B}_3 = \{U\}$. The row corresponding to the primeval code in the table below is indicated in gray.

$\pi_1(X)$	$\pi_3(X)$	Maximal set of coded amino acids
$\{A\}$	$\{C, G, T\}$	Arginine, asparagine, isoleucine, lysine, methionine, serine,and threonine
$\{C\}$	$\{A, G, T\}$	Arginine, glutamine, histidine, leucine, and proline
$\{G\}$	$\{A, C, T\}$	Alanine, aspartate, glutamate, glycine, and valine
$\{T\}$	$\{A, C, G\}$	Cysteine, leucine, phenylalanine, serine, tryptophan, and tyrosine
$\{A, C\}$	$\{G, T\}$	Arginine, asparagine, glutamine, histidine, isoleucin, leucine, lysine, methiodine, proline, serine, and threonine
$\{A, G\}$	$\{C, T\}$	Alanine, asparagine, aspartate, glycine, isoleucine, serine, threonine, and valine
$\{A, T\}$	$\{C, G\}$	Arginine, asparagine, cysteine, isoleucine, leucine, lysine, methionine, phenylalanine, serine, threonine, tryptophan, and tyrosine
$\{C, G\}$	$\{A, T\}$	Alanine, arginine, aspartate, glutamate, glutamine, glycine, histidine, leucine, proline, and valine
$\{C, T\}$	$\{A, G\}$	Arginine, glutamine, leucine, proline, serine, and tryptophan
$\{G, T\}$	$\{A, C\}$	Alanine, aspartate, cysteine, glutamate, glycine, leucine, phenylalanine, serine, tyrosine, and valine
$\{A, C, G\}$	$\{T\}$	Alanine, arginine, asparagine, aspartate, glycine, histidine, isoleucine, leucine, proline, serine, threonine, and valine
$\{A, C, T\}$	$\{G\}$	Arginine, glutamine, leucine, lysine, methiodine, proline, serine, threonine, and tryptophan
$\{A, G, T\}$	$\{C\}$	Alanine, asparagine, aspartate, cysteine, glycine, isoleucine, phenylalanine, serine, threonine, tyrosine, and valine
$\{C, G, T\}$	$\{A\}$	Alanine, arginine, glutamate, glutamine, glycine, leucine, proline, serine, and valine

The following corollary will be needed in the next section.

Corollary 4.6. *The sets*

1. $X = \{ATC, TCC, CAA\}$,
2. $X = \{TGG, ATG, TTC, AAG, GAG, GAC, GGC\}$,
3. $X = \{TGG, ATG, TTC, AAG, GAG, GAC, GGC, CAG, GTA, CGT\}$

are trinucleotide circular codes.

Proof. See Appendix C. □

4.2. Coding 20 amino acids

In this section we will be interested in finding comma-free, trinucleotide circular or at least trinucleotide 2-circular codes which code for a maximal number of amino acids. It is known from Michel and Pirillo (2013a) that there are maximal trinucleotide circular codes that code for 18 amino acids. Hence, passing to n-circular codes we are only interested in codes coding for more than 18 amino acids. We start with the trinucleotide 1-circular codes.

Lemma 4.7. *Assume that X is a trinucleotide 1-circular code that codes for all amino acids corresponding to the standard nuclear table. Then X must contain the set*

$X' = \{TGG, ATG, TTC, AAG, GAG, GAC, GGC\}$

which is a trinucleotide circular code itself.

Proof. Note that X must contain exactly one codon from each complete conjugacy class and exactly one codon for each amino acid. Now such a code must contain the codons TGG^{19}, ATG^{11}, TTC^{18} and AAG^{2} since these codons are the only coding codons (except 'non-acceptable' codons AAA and TTT) coding the amino acids tryptophan, methionine, phenylalanine and lysine. Furthermore such a code must contain the codon GAC^{8} since the second codon coding the amino acid glutamate is GAA^{2} belonging to the same equivalence class as AAG^{2}. For the same reason GAC^{5} (aspartate) and GGC^{15} (glycine) must be a member of the code. Finally, X' is a trinucleotide code by Corollary 4.6. □

The next example shows that there are trinucleotide 1-circular codes that code for all 20 amino acids. The two codes given below are the only trinucleotide 1-circular codes which code all 20 amino acids. The result can be obtained using Lemma 4.7 and Table 3 by straightforward calculations.

Example 4.8. The following codes are the only trinucleotide 1-circular coding for all 20 amino acids:

$X_1 = \{TGG, ATG, TTC, AAG, GAG, GAC, GGC, CAG, GTA,$
$CGT, TGT, AAC, GCC, CTG, TAT,$

$ACT, ATA, TAA, CCA, ACA, TCA\}$

$X_2 = \{TGG, ATG, TTC, AAG, GAG, GAC, GGC, CAG, GTA, CGT,$
$TGT, AAC, GCC, CTG, TAT,$

$ACT, ATA, TAA, CCA, CAT, TCC\}$

We now turn to comma-free codes. Note that the comma-freeness property is stronger than circularity, hence it is known from Michel and Pirillo (2013a) that there is no circular (and hence no comma-free) code that codes for more than 18 amino acids. However, here we give a strictly combinatorial proof rather than computational checking.

Theorem 4.9. *There exists no comma-free code coding all 20 amino acids with respect to the standard (eukaryote) genetic code (Table 2).*

Proof. According to Lemma 4.7 such a code must contain the codons ATG, GAG and TGG. However, the concatenation of ATG and GAG contains as a substring TGG

$ATGGAG$.

Thus, the codons TGG, ATG, GAG cannot be elements of any comma-free code at the same time. □

We now look at 2-circular codes.

Theorem 4.10. *There is no trinucleotide 2-circular code that codes all amino acids with respect to the standard (eukaryote) genetic code (Table 3).*

Proof. See Appendix D. □

Certainly, Theorem 4.10 poses a question for a corresponding result for 19 amino acids. Since a trinucleotide circular code is in particular 2-circular and, consequently, there is a 2-circular code which encodes 18 amino acids, the case of 19 amino acids remains the sole open case. The following result is obtained by computer

calculations of all possible 2-circular codes. We cannot yet present combinatorial proof of the result.

Theorem 4.11. *There is no trinucleotide 2-circular code that codes for 19 amino acids with respect to the standard (eukaryote) genetic code (Table 3).*

We conclude this section with the following observation. In the proof of Theorem 4.10 it turns out that a contradiction to the existence of a maximal trinucleotide 2-circular code coding for all amino acids is obtained after using the fact that it must contain the amino acids *arginine, cysteine, lysine, methionine, tryptophan, phenylalanine, glutamate, aspartate, glycine, glutamine, valine, asparagine, alanine, and leucine*. Moreover, the maximal trinucleotide circular code identified on a large gene population of eukaryotes and prokaryotes (Michel et al., 2012) does not code the amino acids *arginine, cysteine, histidine, lysine, methionine, proline, serine, and tryptophan*. Hence we pose the following challenge.

Problem 4.12. *Identify possible sets of amino acids which cannot be encoded by* (n)*-circular codes for* $n=2,3,4$.

5. Conclusions

In the present paper five hierarchically ordered classes of trinucleotide codes which include well-known comma-free and circular codes are introduced. Numerous examples and results obtained show that all these classes are different and can be ordered by inclusion. The most restrictive property considered is the comma-freeness, followed by the circularity, then by 3-circularity, then by 2-circularity and finally by 1-circularity. The hierarchy obtained makes it possible to test the circularity of a given trinucleotide code in an easier way than what is currently practised.

The main concern of the paper is to investigate the maximal number of different amino acids which can be coded taking codes from different classes of trinucleotide codes. In particular, it is shown in the present work that it is theoretically impossible to code all 20 amino acids using comma-free, circular or even less restrictive 2-circular codes but it is possible to code them with trinucleotide 1-circular codes. Additionally, it is shown that all codes from a special class of trinucleotide codes which includes the primeval RNY-code are automatically circular, and which amino acids they can code are listed.

Acknowledgments

The authors would like to thank Diego Gonzalez (Italian National Research Council and UNIBO, Bologna, Italy) and Simone Giannerini (UNIBO, Bologna, Italy) for introducing us to the subject and for many helpful and valuable discussions and advice. The authors are deeply grateful to Alberto Danielli (UNIBO, Bologna, Italy) for his patience in discussing with us the biological background and his many pieces of valuable advice and to Christian Michel (University of Strasbourg) for fruitful discussions. Last but not least the authors would like to thank the students of the University of Applied Sciences Mannheim Michail Polishuk and Dorothea Stahl for technical support.

Appendix A. Proof of Theorem 4.3

Proof. We will be using the following notation. If $x\in\mathcal{B}^3$ is a codon, then we write $x=x(1)x(2)x(3)$ with $x(i)\in\mathcal{B}$.

One implication is trivial, hence we assume that X is trinucleotide $n(X)$-circular. Assume that $w=x_1,\ldots,x_k$ is a word over $\mathcal{B}$ with $x_i\in X$ for all i such that k is minimal with respect to the condition that w has two different partitions into words from X on a circle. By assumption $k>n(X)$. We have to distinguish two cases.

Case I: The second partition of the word w comes from a shift of 1 base, i.e. $x_1(2)x_1(3)x_2\cdots x_kx_1(1)$ is an element of X^k. By definition of $n(X)$ there must exist $s<t\le k$ such that $x_s(1)=x_t(1)$ because $k>n(X)$. But then the word $x_s\cdots x_{t-1}$ has two different partitions into codons from X and has length smaller than w – a contradiction to the minimality of k.

Case II: The second partition of the word w comes from a shift of 2 bases, i.e. $x_1(3)x_2\cdots x_kx_1(1)x_1(2)$ is an element of X^k. By definition of $n(X)$ there must exist $s<t\le k$ such that $x_s(3)=x_t(3)$ because $k>n(X)$. If $t=k$ define $x_{t+1}=x_1$. But then the word $x_{t+1}\cdots x_kx_1\cdots x_s$ has two different partitions into codons from X and has length smaller than w – a contradiction to the minimality of k. □

Appendix B. Proof of Theorem 4.4

Proof. Let us assume that a concatenation of codons from X w has two different partitions into words from X on a circle.

Case I: The second partition of the word w comes from a shift of 1 base. Then the third bases of one decomposition become first bases of the second one. It is a contradiction since $\pi_1(X)\cap\pi_3(X)=\emptyset$.

Case II: The second partition of the word w comes from a shift of 2 bases. Then the first bases of one decomposition become third bases of the second one. Again it is a contradiction since $\pi_1(X)\cap\pi_3(X)=\emptyset$. □

Appendix C. Proof of Corollary 4.6

Proof.

1. By Theorem 4.3 we need to check that X is trinucleotide 3-circular. Note that $|\pi_1(X)|=3$ and $|\pi_3(X)|=2$. Therefore, we list all the combinations of two codons from X

   ATCATC, *ATCTCC*, *ATCCAA*, *TCCTCC*, *TCCCAA*,

   TCCATC, *CAACAA*, *CAATCC*, *CAAATC*

   The only potential candidates for a second partition are *ATC CAA* (shift by one)and *CAA TCC* (shift by two) because *TCC* and *ATC* are in X. However, *AAA* and *CCA* are not in X, hence X is trinucleotide *2*-circular.
   By the same argument any combination $x_1x_2x_3$ of three codons from X that allows a second partition into codons from$X$ on a circle must start with either *ATC CAA* (shift by one) or *CAA TCC* (shift by two). Here are the possible combinations

   ATCCAAATC, *ATCCAATCC*, *ATCCAACAA*, *CAATCCATC*,

   CAATCCTCC, *CAATCCCAA*

   However, in each case we obtain a contradiction since *AAA*, *AAT*, *AAC*, *CAT*, *CTC* and *CCA* are not in X.
2. As in the proof above it suffices to check that X is 3 -circular. For trinucleotide 2-circularity we learn that the only potential candidates to violate are *ATG GAG* and *ATG GAC* since *TGG* is in X. However, *AGA* and *ACA* are not in X, so X is trinucleotide 2-circular. Moreover, any combination of three codons from X that might have a second partition on a circle must start with *ATG GAG* or *ATG GAC*. Easy checking shows that one of them combined with a third codon from X has a second partition.
3. It suffices to check that X is trinucleotide 4-circular. The arguments are completely similar to those above. □

Appendix D. Proof of Theorem 4.10

Proof. By Lemma 4.7 $\{TGG, ATG, TTC, AAG, GAG, GAC, GGC\}$ is a subset of X. Note again that X must contain exactly one codon from each conjugacy class and exactly one codon for each amino acid.

Since X codes for Glutamine it must contain either CAA or CAG. Assume that it contains CAA, hence $AAC \notin X$. Then it contains AAT since X codes for Asparagine and AAT and AAC are the only codons for Asparagine. But now $TGGCAA$ has two different partitions on a circle since GGC and AAT are in X – contradiction. Thus CAG is in X. So

$$\{TGG, ATG, TTC, AAG, GAG, GAC, GGC, CAG\} \subseteq X$$

Now it follows that $AGT \notin X$ since otherwise $TGG\ CAG$ would contradict 2-circularity of X. Hence GTA is in X since it is the only codon left that has conjugacy class 9. Note that TAG codes for the Stop signal. Thus

$$\{TGG, ATG, TTC, AAG, GAG, GAC, GGC, CAG, GTA\} \subseteq X$$

Next, let us look at conjugacy class 16. There are only two codons left, namely TCG and CGT in that class. Note that GTC codes for valine as $GTA \in X$ does, so $GTC \notin X$. Assume that $TCG \in X$, then by coding for arginine we conclude $CGC \in X$ and hence by coding for alanine we have $GCT \in X$. But then $CGCTGG$ has two different partitions since GCT and GGC are in X – contradiction. So $CGT \in X$

$$\{TGG, ATG, TTC, AAG, GAG, GAC, GGC, CAG, GTA, CGT\} \subseteq X$$

Note that this is a trinucleotide circular code by Lemma 4.6. By trinucleotide 2-circularity applied to $ATGCGT \in X^2$ we now conclude that $TGC \notin X$ since $GTA \in X$. Hence, since X codes for cysteine, we have that $TGT \in X$. So

$$\{TGG, ATG, TTC, AAG, GAG, GAC, GGC, CAG, GTA, CGT, TGT\} \subseteq X$$

Now we have the situation that TGG and ATG are both contained in the code and the word $ATGTGT$ has two partitions into words from X on a circle since $TGT, GTA \in X$ is true. This contradicts the 2-circular property of the code X. □

References

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

Golomb, S.W., Gordon, B., Welch, L.R., 1958. Comma-free codes. Can. J. Math. 10.

Golomb, S.W., Delbruck, M., Welch, L.R., 1958. Construction and properties of comma-free codes. Biol. Medd. K. Danske Vidensk. Selsk. 23 (9).

Gonzalez, D.L., 2008. The mathematical structure of the genetic code. In: Barbieri (Ed.), Codes of Life: The Rules of Microevolution, Springer Science + Business Media B.V.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2008. Strong short-range correlations and dichotomic codon classes in coding DNA sequences. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 78(November (5 Pt 1)):051918. Epub November 19, 2008.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2011. Circular codes revisited: a statistical approach. J. Theor. Biol. 275 (1).

Hayes, B., 1998. The invention of the genetic code. Am. Sci. 86 (1), 8–14.

Herrmann, M., Michel, C.J., Zugmeyer, B., 2013. A necklace algorithm to determine the growth function of trinucleotide circular codes. J. Appl. Math. Bioinf. 3, 1–40.

Jestin, J.L., 2006. Degeneracy in the genetic code and its symmetries by base substitutions. C. R. Biol. 329, 168–171.

Koch, A.J., Lehmann, J., 1997. About a symmetry of the genetic code. J. Theor. Biol. 189, 171–174.

Michel, C.J., Pirillo, G., Pirillo, M.A., 2008. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theor. Comput. Sci. 401, 17–26.

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

Michel, C.J., Pirillo, G., 2013a. A permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes. J. Theor. Biol. 319, 116–121.

Michel, C.G., Pirillo, G., 2013b. Dinucleotide circular codes. ISRN Biomathematics (2013), Article ID 538631, 8.

Michel, C.J., Pirillo, G., Pirillo, M.A., 2012. A classification of 20-trinucleotide circular codes. Inf. Comput. 212, 55–63.

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

Michel, C.J., Pirillo, G., 2010. Identification of all trinucleotide circular codes. Comput. Biol. Chem. 34, 122–125.

Michel, C.J., 2014. A genetic scale of reading frame coding. J. Theor. Biol. 355, 83–94.

Nanjundiah, V., July 2004. George Gamow and the genetic code. In: Resonance.

Negadi, T., 2009. The genetic code degeneracy and the amino acids chemical composition are connected. NeuroQuantology 7 (March (1)), 181–187.

Pearson, J., 2003. Comma-free codes. In: Kishor S. Trivedi (Eds.), Computer Science Applications, 2nd ed.

Rumer, Y.B., 1969. Systematization of codons in the genetic code. Dokl. Akad. Nauk SSSR 187 (August (4)), 937–938.

Shepherd, J.C.W., 1986. Origins of life and molecular evolution of present-day genes. Chem. Scr. 268, 75–83.

Sciarrino, A., 2003. A mathematical model accounting for the organization in multiplets of the genetic code. BioSystems 69, 1–13.

Watson, J.D., Crick, F.H.C., 1953. A structure for deoxyribose nucleic acid. Nature 171, 737–738.