Journal of Theoretical Biology 319 (2013) 116–121

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

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

A permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes

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

[a] Equipe de Bioinformatique Théorique, BFO, LSIIT (UMR 7005), Université 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, 77454 Marne-la-Vallée Cedex 2, France

HIGHLIGHTS

- Trinucleotide circular code with a property for coding the 20 amino acids.
- Two sets of 20 trinucleotides in the coding process of amino acids.
- Mathematical property in genetic codes.

ARTICLE INFO

Article history:
Received 27 February 2012
Received in revised form
19 November 2012
Accepted 21 November 2012
Available online 1 December 2012

Keywords:
Trinucleotide circular code
Trinucleotide permuted set
Genetic code
Amino acid

ABSTRACT

We identify here a combinatorial property between circular code and genetic code. A circular code of 20 trinucleotides which allows to retrieve the reading frame has a permuted set of 20 trinucleotides which is a code, but not circular, coding the 20 amino acids in variant nuclear codes. This result is a contribution to the research field analysing the mathematical properties of genetic codes.

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 $\{A,C,G,T\}$. The set of the 64 words of length 3 (trinucleotides or triletters) on $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^3 = \{AAA,AAC,\ldots,TTT\}$. The set $\mathcal{A}_4^3$ is a code in the sense of language theory, more precisely a uniform code, but not a circular code (Berstel and Perrin, 1985; Lassez, 1976). In order to have an intuitive meaning of these notions, codes are written on a straight line while circular codes are written on a circle, but, in both cases, unique decipherability is required.

Comma free codes, a very particular case of circular codes, have been studied for a long time, e.g. Crick et al. (1957) and Golomb et al. (1958a,b). After the discovery of a circular code in genes with strong mathematical properties (Arquès and Michel, 1996), circular codes are mathematical objects studied in combinatorics, theoretical computer science and theoretical biology. This theory underwent a rapid development e.g. Koch and Lehman (1997), Béal and Senellart (1998), Bassino (1999), Štambuk (1999), Jolivet and Rothen (2001), Frey and Michel (2003, 2006), Nikolaou and Almirantis (2003), Pirillo (2003, 2008a,b, 2010), May et al. (2004), Pirillo and Pirillo (2005), Lassez et al. (2007), Michel et al. (2008a,b, 2012), José et al. (2009), Michel and Pirillo (2010, 2011), Bussoli et al. (2011, 2012), Gonzalez et al. (2011).

A genetic code is a coding correspondence table between the 64 trinucleotides (words of three letters on the gene alphabet also called codons) and the 20 amino acids (words of one letter on the protein alphabet). There are several genetic codes, the standard genetic code and several variant genetic codes (www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=tgencodes#SG1, July 07, 2010). The variant genetic codes are divided into nuclear codes (codes 6–5, 10, 12, 15), mitochondrial codes (codes 2–5, 9, 13, 14, 16, 21–24) and the bacterial, archaeal and plant plastid code (code 11). Note that the numbering presents gaps as some codes have been deleted with the biological advances of knowledge. In the standard genetic code, called here

* Corresponding author. Tel.: +33 3 68 85 44 62.
E-mail addresses: michel@dpt-info.u-strasbg.fr (C.J. Michel), pirillo@math.unifi.it (G. Pirillo).

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.11.023

SGC, 61 trinucleotides code the 20 amino acids as there are three termination trinucleotides *TAA*, *TAG* and *TGA*. The trinucleotide *ATG* coding the amino acid *Met* (*M*) is also the initiation trinucleotide (noted *i*) in the general case. Two amino acids are encoded by a single trinucleotide: *Met* (*M*) and *Trp* (*W*). Nine amino acids are encoded by two trinucleotides: *Asn* (*N*), *Asp* (*D*), *Cys* (*C*), *Gln* (*Q*), *Glu* (*E*), *His* (*H*), *Lys* (*K*), *Phe* (*F*) and *Tyr* (*Y*). One amino acid is encoded by three trinucleotides: *Ile* (*I*). Five amino acids are encoded by four trinucleotides: *Ala* (*A*), *Gly* (*G*), *Pro* (*P*), *Thr* (*T*) and *Val* (*V*). Three amino acids are encoded by six trinucleotides: *Arg* (*R*), *Leu* (*L*) and *Ser* (*S*). No amino acid is encoded by five trinucleotides. There are $\mathrm{card}(\mathcal{S}) = 2^9 \times 3 \times 4^5 \times 6^3 = 339{,}738{,}624$ sets $\mathcal{S}$ of 20 trinucleotides coding the 20 amino acids, i.e. with a bijective map. The variant genetic codes differ from the standard one by the number of trinucleotides coding the 20 amino acids or by a coding reassignment of trinucleotides. In the nuclear code 6 (see Definition below), the number of sets of 20 trinucleotides coding the 20 amino acids is $2 \times \mathrm{card}(\mathcal{S})$ and in the nuclear code 15 (see Definition below), this number is $\frac{3}{2} \times \mathrm{card}(\mathcal{S})$. All genetic codes are surjective maps.

There are exactly 12,964,440 circular codes $\mathcal{X}$ of 20 trinucleotides (Arquès and Michel, 1996; Michel and Pirillo, 2010). None 20-trinucleotide circular code among these 12,964,440 ones codes 20 or 19 amino acids with *SGC*. There is no bijection, unfortunately (in a certain way), between a 20-trinucleotide circular code $\mathcal{X}$ and a set $\mathcal{S}$. Note that $\mathrm{card}(\mathcal{X})/\mathrm{card}(\mathcal{S}) \approx 3.8\%$. Ten 20-trinucleotide circular codes code 18 amino acids with *SGC*. The common 20-trinucleotide circular code of eukaryotes and prokaryotes (Arquès and Michel, 1996) only codes 12 amino acids, but it has exceptional properties, in particular the properties of C^3 and self-complementary (see also below).

Some combinatorial properties were recently identified with the conjugation partitions of sets of trinucleotides in $\mathcal{A}_4^3 \backslash \{AAA, CCC, GGG, TTT\}$ (Bussoli et al., 2012). Indeed, each circular code X can be associated with two other subsets X_1 and X_2 of $\mathcal{A}_4^3 \backslash \{AAA, CCC, GGG, TTT\}$ simply by operating two circular permutations $\mathcal{P}$ of one letter and two letters on the trinucleotides of X, respectively, i.e. $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$. During this research work, we identify a circular code $Y = \{ACG, ACT, AGA, AGG, AGT, ATA, ATC, CAA, CAC, CAG, CCT, GCC, GCG, GCT, GGT, TCG, TCT, TGA, TGT, TTA\}$ of 20 trinucleotides (maximal trinucleotide circular code) which has a permuted set $\mathcal{P}^2(Y) = \{AAG, AAT, ACA, ATG, ATT, CAT, CCA, CGC, GAC, GAG, GCA, GGC, GTC, TAC, TAG, TCC, TGC, TGG, TTC, TTG\}$ of 20 trinucleotides coding the 20 amino acids in the variant nuclear codes 6 and 15.

2. Definitions

The classical notions of language theory can be found in Berstel and Perrin (1985). Let $\mathcal{A}_4 = \{A, C, G, T\}$ denote the genetic alphabet, lexicographically ordered with $A < C < G < T$. We use the following notation:

- $\mathcal{A}_4^*$ (respectively $\mathcal{A}_4^+$) is the set of words (respectively non-empty words) over $\mathcal{A}_4$,
- $\mathcal{A}_4^2$ is the set of the 16 words of length two (diletters or dinucleotides) and
- $\mathcal{A}_4^3$ is the set of the 64 words of length three (triletters or trinucleotides).

We now recall the circular permutation map, the definitions of code and circular code, and the property of C^3 for a circular code, e.g. Berstel and Perrin (1985), Arquès and Michel (1996).

Definition 1. The circular permutation map $\mathcal{P} : \mathcal{A}_4^3 \to \mathcal{A}_4^3$ permutes circularly each trinucleotide $l_0 l_1 l_2$ as follows $\mathcal{P}(l_0 l_1 l_2) = l_1 l_2 l_0$.

The map $\mathcal{P}$ on words is naturally extended to a trinucleotide set X: its permuted trinucleotide set $\mathcal{P}(X)$ is obtained by applying the circular permutation map $\mathcal{P}$ to all the trinucleotides of X. We shortly write $\mathcal{P}^2(X)$ for $\mathcal{P}(\mathcal{P}(X))$.

Definition 2. A set X of words is a code if, for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, the condition $x_1 \cdots x_n = x'_1 \cdots x'_m$ implies $n = m$ and $x_i = x'_i$ for $i = 1, \ldots, n$.

In a code, the words are read on a straight line, i.e. the beginning and the end of a word are different. Fig. 1 gives an example of a word set X which is not a code.

Definition 3. A trinucleotide code X is circular if, for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, $p \in \mathcal{A}_4^*$, $s \in \mathcal{A}_4^+$, the conditions $s x_2 \cdots x_n p = x'_1 \cdots x'_m$ and $x_1 = ps$ imply $n = m$, $p = \varepsilon$ (empty word) and $x_i = x'_i$ for $i = 1, \ldots, n$.

In a circular code, the words are read on a circle, i.e. the beginning and the end of a word are not distinguished. Fig. 2 gives an example of a code X which is not circular.

Definition 4. If X is a subset of $\mathcal{A}_4^3 \backslash \{AAA, CCC, GGG, TTT\}$, we denote by X_1 the permuted trinucleotide set $\mathcal{P}(X)$ and by X_2 the permuted trinucleotide set $\mathcal{P}^2(X)$ and we call X_1 and X_2 the conjugated classes of X.

Definition 5. A trinucleotide circular code X is C^3 if X, X_1 and X_2 are circular codes.

The concept of necklace was introduced by Pirillo for circular codes (Pirillo, 2003) in order to characterize the circular codes for an efficient algorithm development. Let $l_1, l_2, \ldots, l_{n-1}, l_n, \ldots$ be letters in $\mathcal{A}_4$, $d_1, d_2, \ldots, d_{n-1}, d_n, \ldots$ diletters in $\mathcal{A}_4^2$ and $n \geq 2$ an integer.

Definition 6. Letter Diletter Continued Necklaces (*LDCN*): We say that the ordered sequence $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n, l_{n+1}$ is an $(n+1)$*LDCN* for a subset $X \subset \mathcal{A}_4^3$ if

$l_1 d_1, l_2 d_2, \ldots, l_n d_n \in X$

and

$d_1 l_2, d_2 l_3, \ldots, d_{n-1} l_n, d_n l_{n+1} \in X$.

Only a few trinucleotide sets are circular codes. We have the following proposition.

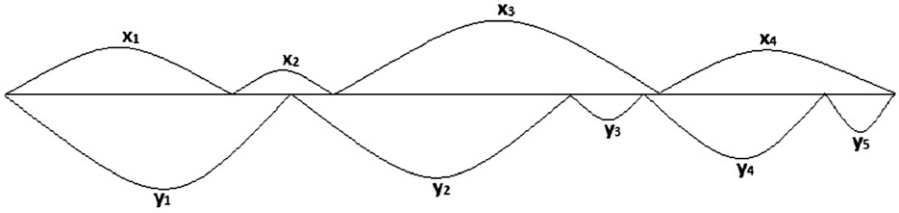

Fig. 1. The word set X is not a code as there is the relation $x_1 x_2 x_3 x_4 = y_1 y_2 y_3 y_4 y_5$.

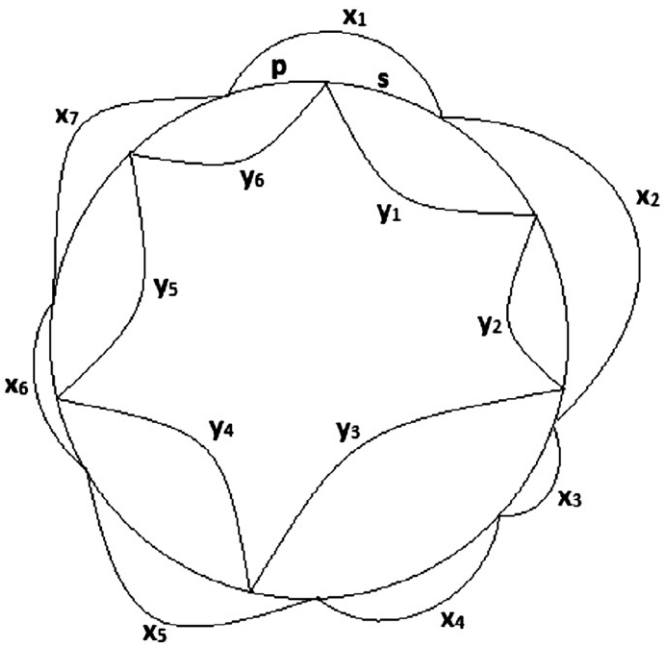

Fig. 2. The code X is not circular as there are the relations $sx_2x_3x_4x_5x_6x_7p = y_1y_2y_3y_4y_5y_6$ and $x_1 = ps$.

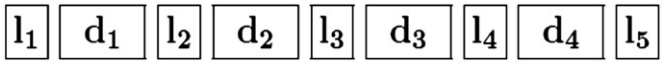

Fig. 3. The 5*LDCN* forbidden configuration for a circular code X: $l_1d_1, l_2d_2, l_3d_3, l_4d_4 \in X$ and $d_1l_2, d_2l_3, d_3l_4, d_4l_5 \in X$.

Proposition 1 (*Pirillo, 2003*). Let X be a trinucleotide code. The following conditions are equivalent: (i) X is a circular code; (ii) X has no 5*LDCN*.

Fig. 3 represents the 5*LDCN* forbidden configuration ($n+1=5$) for a circular code X.

The nuclear code of ciliate (Oxytricha, Stylonychia, Paramecium, Tetrahymena; Hoffman et al., 1995), dasycladacean (Acetabularia, Batophora; Schneider et al., 1989; Schneider and de Groot, 1991), hexamita (Keeling and Doolittle, 1996) (variant nuclear code 6 according to the GenBank convention, National Center for Biotechnology Information (NCBI), July 07 2010) is defined by Table 1.

The two trinucleotides *TAA* and *TAG* coding *Gln* (*Q*) in the variant nuclear code 6 are termination codons *Ter* in the standard code.

The nuclear code of ciliate (Blepharisma; Liang and Heckmann, 1993) (variant nuclear code 15 according to the GenBank convention, National Center for Biotechnology Information (NCBI), July 07 2010) is defined by Table 2.

The trinucleotide *TAG* coding *Gln* (*Q*) in the variant nuclear code 15 is a termination codon *Ter* in the standard code.

3. Results

In order to prove the following proposition, we need a very easy lemma.

Lemma 1. *If $l_1,d_1,l_2,d_2,l_3,d_3,l_4,d_4,l_5$ is a 5LDCN for a set of trinucleotides X then for each $i \in \{1,2,3,4\}$, the dinucleotide d_i must have at least one occurrence in prefix position in X and at least one occurrence in suffix position in X.*

Proof. Trivial.

Proposition 2. *The following set Y of 20 trinucleotides*

$Y = \{ACG,ACT,AGA,AGG,AGT,ATA,ATC,CAA,CAC,CAG,$
$CCT,GCC,GCG,GCT,GGT,TCG,TCT,TGA,TGT,TTA\}$

is a circular code (maximal). More precisely, Y is the 11,056,585*th among* 12,964,440 *maximal circular codes (in the lexicographical order) of* 20 *trinucleotides and belongs to the classes $C^{5LDN} = C^{5LDCN} = C^{5DLN}$* (Michel et al., 2012).

A proof based on dinucleotides is given in Appendix A.

Proposition 3. *The trinucleotide circular code Y (Proposition 2) has a permuted set $\mathcal{P}^2(Y)$ of* 20 *trinucleotides*

$\mathcal{P}^2(Y) = \{AAG,AAT,ACA,ATG,ATT,CAT,CCA,CGC,GAC,GAG,$
$GCA,GGC,GTC,TAC,TAG,TCC,TGC,TGG,TTC,TTG\}$

which is not circular and codes the 20 *amino acids in the variant nuclear codes* 6 *and* 15.

A proof is given in Appendix B.

4. Discussion

According to Proposition 3, a circular code of 20 trinucleotides (maximal) which allows to retrieve the reading frame has a permuted set of 20 trinucleotides which is a code, but not circular, coding the 20 amino acids in variant nuclear codes. This circular code property involves two sets of 20 trinucleotides in the coding process of amino acids. The four trinucleotides {*AAA*,*CCC*,*GGG*,*TTT*} may be involved, in contrast to reading frame retrieval, in frameshifting as shown particularly in Ahmed et al. (2007) and Seligmann (2012). The 20 remaining trinucleotides allow an additional coding function which remains to be discovered. They could have a redundant property of the previous functions, in particular in reading frame synchronization and error correction. Both the theory of circular code and a mathematical model of the genetic code based on a number theory with the information of dichotomic classes (nonlinear functions of the information contained in a dinucleotide) suggest these two properties (Frey and Michel, 2006; Giannerini et al., 2012). Our combinatorial result is a contribution to the identification of mathematical properties of genetic codes. In itself, this result is interesting as this combinatorial property is a representation of a biological reality. However, a main question is to elucidate whether the two sets Y and $\mathcal{P}^2(Y)$ have an independent biological function or a combined one. To date, the authors have no idea how such a code could operate in genes even if some biological arguments tend to promote a combined biological function. Correlated to this problem, which code, Y, $\mathcal{P}^2(Y)$ or a combination of both, could be reflected in the codon usage of coding sequences or in quadruplet codons (tetranucleotides)? While the triplet genetic code is near universally conserved, some mRNAs have additional or deleted bases with respect to a "canonical" gene. A $+1$ frameshift is a sequence with an additional base in the canonical gene which creates a quadruplet codon in the reading frame. This biological process called frameshifting decodes a quadruplet codon to restore the reading frame in the canonical gene (Atkins et al., 1991; Gesteland et al., 1992; Farabaugh, 1996). A study of the limits of codon and anticodon size shows that the contemporaneous translational machinery is the most efficient using three-base codons and seven-base anticodon tRNAs. However, four-base and even five-base codons can be processed by the ribosome with tRNAs containing up to at least 10 nucleotides in their anticodon loops (Anderson et al., 2002). Otherwise, experimental approaches

Table 1
Nuclear code of ciliate, dasycladacean and hexamita (variant nuclear code 6) showing the correspondence between the 64 trinucleotides $\{AAA,\ldots,TTT\}$ and the 20 amino acids given in the one-letter and the three-letter symbols. The trinucleotide *ATG* coding *Met* is also the initiator codon *i* and the trinucleotide *TGA* coding no amino acid is the termination codon *Ter*. The permuted set $\mathcal{P}^2(Y)$ of the trinucleotide circular code *Y* coding the 20 amino acids is in bold.

TTT	*F*	*Phe*	*TCT*	*S*	*Ser*	*TAT*	*Y*	*Tyr*	*TGT*	*C*	*Cys*
TTC	***F***	***Phe***	***TCC***	***S***	***Ser***	***TAC***	***Y***	***Tyr***	***TGC***	***C***	***Cys***
TTA	*L*	*Leu*	*TCA*	*S*	*Ser*	*TAA*	*Q*	*Gln*	*TGA*	*	*Ter*
TTG	***L***	***Leu***	*TCG*	*S*	*Ser*	***TAG***	***Q***	***Gln***	***TGG***	***W***	***Trp***
CTT	*L*	*Leu*	*CCT*	*P*	*Pro*	***CAT***	***H***	***His***	*CGT*	*R*	*Arg*
CTC	*L*	*Leu*	*CCC*	*P*	*Pro*	*CAC*	*H*	*His*	***CGC***	***R***	***Arg***
CTA	*L*	*Leu*	***CCA***	***P***	***Pro***	*CAA*	*Q*	*Gln*	*CGA*	*R*	*Arg*
CTG	*L*	*Leu*	*CCG*	*P*	*Pro*	*CAG*	*Q*	*Gln*	*CGG*	*R*	*Arg*
ATT	***I***	***Ile***	*ACT*	*T*	*Thr*	***AAT***	***N***	***Asn***	*AGT*	*S*	*Ser*
ATC	*I*	*Ile*	*ACC*	*T*	*Thr*	*AAC*	*N*	*Asn*	*AGC*	*S*	*Ser*
ATA	*I*	*Ile*	***ACA***	***T***	***Thr***	*AAA*	*K*	*Lys*	*AGA*	*R*	*Arg*
ATG	***M***	***Met*** *i*	*ACG*	*T*	*Thr*	***AAG***	***K***	***Lys***	*AGG*	*R*	*Arg*
GTT	*V*	*Val*	*GCT*	*A*	*Ala*	*GAT*	*D*	*Asp*	*GGT*	*G*	*Gly*
GTC	***V***	***Val***	*GCC*	*A*	*Ala*	***GAC***	***D***	***Asp***	***GGC***	***G***	***Gly***
GTA	*V*	*Val*	***GCA***	***A***	***Ala***	*GAA*	*E*	*Glu*	*GGA*	*G*	*Gly*
GTG	*V*	*Val*	*GCG*	*A*	*Ala*	***GAG***	***E***	***Glu***	*GGG*	*G*	*Gly*

Table 2
The nuclear code of ciliate (Blepharisma) (variant nuclear code 15) showing the correspondence between the 64 trinucleotides $\{AAA,\ldots,TTT\}$ and the 20 amino acids given in the one-letter and the three-letter symbols. The trinucleotide *ATG* coding *Met* is also the initiator codon *i* and the trinucleotides *TAA* and *TGA* coding no amino acid are the termination codons *Ter*. The permuted set $\mathcal{P}^2(Y)$ of the trinucleotide circular code *Y* coding the 20 amino acids is in bold.

TTT	*F*	*Phe*	*TCT*	*S*	*Ser*	*TAT*	*Y*	*Tyr*	*TGT*	*C*	*Cys*
TTC	***F***	***Phe***	***TCC***	***S***	***Ser***	***TAC***	***Y***	***Tyr***	***TGC***	***C***	***Cys***
TTA	*L*	*Leu*	*TCA*	*S*	*Ser*	*TAA*	*	*Ter*	*TGA*	*	*Ter*
TTG	***L***	***Leu***	*TCG*	*S*	*Ser*	***TAG***	***Q***	***Gln***	***TGG***	***W***	***Trp***
CTT	*L*	*Leu*	*CCT*	*P*	*Pro*	***CAT***	***H***	***His***	*CGT*	*R*	*Arg*
CTC	*L*	*Leu*	*CCC*	*P*	*Pro*	*CAC*	*H*	*His*	***CGC***	***R***	***Arg***
CTA	*L*	*Leu*	***CCA***	***P***	***Pro***	*CAA*	*Q*	*Gln*	*CGA*	*R*	*Arg*
CTG	*L*	*Leu*	*CCG*	*P*	*Pro*	*CAG*	*Q*	*Gln*	*CGG*	*R*	*Arg*
ATT	***I***	***Ile***	*ACT*	*T*	*Thr*	***AAT***	***N***	***Asn***	*AGT*	*S*	*Ser*
ATC	*I*	*Ile*	*ACC*	*T*	*Thr*	*AAC*	*N*	*Asn*	*AGC*	*S*	*Ser*
ATA	*I*	*Ile*	***ACA***	***T***	***Thr***	*AAA*	*K*	*Lys*	*AGA*	*R*	*Arg*
ATG	***M***	***Met*** *i*	*ACG*	*T*	*Thr*	***AAG***	***K***	***Lys***	*AGG*	*R*	*Arg*
GTT	*V*	*Val*	*GCT*	*A*	*Ala*	*GAT*	*D*	*Asp*	*GGT*	*G*	*Gly*
GTC	***V***	***Val***	*GCC*	*A*	*Ala*	***GAC***	***D***	***Asp***	***GGC***	***G***	***Gly***
GTA	*V*	*Val*	***GCA***	***A***	***Ala***	*GAA*	*E*	*Glu*	*GGA*	*G*	*Gly*
GTG	*V*	*Val*	*GCG*	*A*	*Ala*	***GAG***	***E***	***Glu***	*GGG*	*G*	*Gly*

allow the incorporation of unnatural amino acids in response to quadruplet codons using extended anticodon tRNAs and orthogonal ribosomes. The decoding of quadruplet codons by extended anticodon tRNAs uses in most cases a potential Watson–Crick or wobble base pair between the fourth base in the codon and the anticodon. Two main molecular mechanisms are consistent with the experimental data: the yardstick model proposes triplet or quadruplet interactions between the codon and anticodon in the A site of the ribosome with subsequent quadruplet translocation (Farabaugh, 1996) while the slippery model proposes a triplet interaction in the A site and a triplet translocation followed by a slip of the mRNA by one base (Qian et al., 1998). This experimental biological field allows the combinatorial biosynthesis of materials and therapeutics, and also investigations into whether life with additional genetically encoded polymers can evolve to perform functions that natural biological systems cannot. It has provoked great interest and underwent a rapid development, see for example the recent review "Reprogramming the genetic code: from triplet to quadruplet codes" (Wang, 2012). The two sets Y and $\mathcal{P}^2(Y)$ could be combined to form a quadruplet code associated to quadruplet codons. Quadruplet codons, i.e. words $l_1l_2l_3l_4$ of four letters $l_i \in \mathcal{A}_4$, can contain the two sets Y and $\mathcal{P}^2(Y)$. For example, if the trinucleotide $t = l_1l_2l_3 \in Y$ then $t^2 = l_1l_2l_3l_1l_2l_3 \in Y^2$ contains the quadruplet codon $l_3l_1l_2l_3$ where $l_3l_1l_2 \in \mathcal{P}^2(t)$. Note that other combinations of trinucleotides t and t' with $t,t' \in Y$ and $t \neq t'$, can generate quadruplet codons containing Y and $\mathcal{P}^2(Y)$. Interestingly, quadruplet codons are identified in mitochondrial genomes using various computational and statistical methods, in particular based on alignment, predicted peptide secondary structure, *GC* content, deamination gradient and circular code tests (Seligmann, 2012). This approach also shows that the mitochondrial quadruplet codons are codons expanded by a 4th silent site which are decoded by antisense tRNAs (from the complementary strand) with reduced anticodons in general, but sometimes with expanded anticodons. Thus, this result suggests that a codon prefix of a quadruplet codon is the coding process ($\mathbb{P}_1$) in quadruplet codons (Seligmann, 2012) while our combinatorial analysis may involve two codons belonging to two different codes (Y and $P^2(Y)$) in quadruplet codons (coding process $\mathbb{P}_2$). Both coding processes $\mathbb{P}_1$ and $\mathbb{P}_2$ in quadruplet codons are based on a codon information even if their concepts are different. The previous approach based on an integer number representation leads to a primitive mitochondrial genetic code also composed by quadruplet codons (Gonzalez et al., in press). However, this approach suggests that the quadruplet codons are generated by a concatenation of dinucleotides (coding process $\mathbb{P}_3$). In summary, all these coding processes $\mathbb{P}_1$, $\mathbb{P}_2$ and $\mathbb{P}_3$ in quadruplet codons constitute a fascinating and open biological problem

which has to be investigated by these different theories, i.e. computational and statistical methods, number theory and combinatorics.

5. Conclusion

We identify here a combinatorial property between circular code and genetic code. This theoretical result opens new research directions in combinatorics, particularly in the identification of tetranucleotide circular codes (necklace propositions, list, numbers, etc.) and their generation from trinucleotide circular codes. Such a combinatorial approach may give new insights in quadruplet codons, extended genetic codes and frameshifting.

Dedication

Giuseppe Pirillo dedicates this paper to Professor Renzo Pinzani for his 70th birthday.

Acknowledgments

We thank the reviewer and Jacques Justin for their advices. The second author thanks the Dipartimento di matematica "U. Dini" for giving him a friendly hospitality. The work of Giuseppe Pirillo is partially supported by Project Interomics of CNR.

Appendix A. Proof of Proposition 2

Proof. Y is a circular code. We use Proposition 1. By way of contradiction, suppose that Y admits a 5$LDCN$ $l_1,d_1,l_2,d_2,l_3,d_3,l_4,d_4,l_5$. By Lemma 1, for each $i \in \{1,2,3,4\}$, each d_i must appear as a prefix and as a suffix in Y. With Y, the set of dinucleotides with this property is $\mathcal{D} = \{AC,AG,CC,GG,TC\}$. So, it is enough to prove that each choice $d_4 \in \mathcal{D}$ leads to a contradiction.

Claim 1.

(i) $AC \neq d_4$ *and* $AC \neq d_3$.
(ii) $AG \neq d_4$ *and* $AG \neq d_3$.
(iii) $TC \neq d_4$ *and* $TC \neq d_3$.
(iv) $CC \neq d_4$.
(v) $GG \neq d_4$.

Proof. (i) By way of contradiction, suppose $d_4 = AC$. We have $l_4 = C$ and, consequently, $d_3 \in \{AT,CA,GC\}$. But, as $\{AT,CA,GC\} \cap \mathcal{D} = \emptyset$, we are in contradiction with Lemma 1. So, $AC \neq d_4$. In the same way, we prove $AC \neq d_3$. (ii) By way of contradiction, suppose $d_4 = AG$. We have $l_4 = C$ and, consequently, $d_3 \in \{AT,CA,GC\}$. But, as $\{AT,CA,GC\} \cap \mathcal{D} = \emptyset$, we are in contradiction with Lemma 1. So, $AG \neq d_4$. In the same way, we prove $AG \neq d_3$. (iii) By way of contradiction, suppose $d_4 = TC$. We have $l_4 = A$ and, consequently, $d_3 \in \{AG,AT,CA,TG,TT\}$. But, if $d_3 \in \{AT,CA,TG,TT\}$ (as $\{AT,CA,TG,TT\} \cap \mathcal{D} = \emptyset$), we are in contradiction with Lemma 1, and if $AG = d_3$, we are in contradiction with (ii). So, $TC \neq d_4$. In the same way, we prove $TC \neq d_3$. (iv) By way of contradiction, suppose $d_4 = CC$. We have $l_4 = G$ and, consequently, $d_3 \in \{AC,AG,CA,GC,TC\}$. But, if $d_3 \in \{CA,GC\}$ (as $\{CA,GC\} \cap \mathcal{D} = \emptyset$), we are in contradiction with Lemma 1; if $AC = d_3$, we are in contradiction with (i); if $AG = d_3$, we are in contradiction with (ii); and if $TC = d_3$, we are in contradiction with (iii). So, $CC \neq d_4$. (v) By way of contradiction, suppose $d_4 = GG$. We have $l_4 = A$ and, consequently, $d_3 \in \{AG,AT,CA,TG,TT\}$. As with (iii), we are in contradiction with Lemma 1 and (ii). So, $GG \neq d_4$.

By (i), (ii), (iii), (iv) and (v), $d_4 \notin \mathcal{D}$. So, by Lemma 1, we are in contradiction. So, Y is a circular code.

Appendix B. Proof of Proposition 3

Proof. $\mathcal{P}^2(Y)$ is not a circular code. Consider the subset $\mathcal{S} = \{AAT,ATT,TAC,ACA\}$ of $\mathcal{P}^2(Y)$. Note that it admits the necklace A,AT,T,AC,A and consequently cannot be a circular code. A fortiori, $\mathcal{P}^2(Y)$ containing $\mathcal{S}$ is also not a circular code.

$\mathcal{P}^2(Y)$ codes the 20 amino acids in the variant nuclear codes 6 and 15. Obvious by inspection (Tables 1 and 2).

References

Ahmed, A., Frey, G., Michel, C.J., 2007. Frameshift signals in genes associated with the circular code. In Silico Biol. 7, 155–168.
Anderson, J.C., Magliery, T.J., Schultz, P.G., 2002. Exploring the limits of codon and anticodon size. Chem. Biol. 9, 237–244.
Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182, 45–58.
Atkins, J.F., Weiss, R.B., Thompson, S., Gesteland, R.F., 1991. Towards a genetic dissection of the basis of triplet decoding, and its natural subversion: programmed reading frame shifts and hops. Annu. Rev. Genet. 25, 201–228.
Bassino, F., 1999. Generating function of circular codes. Adv. Appl. Math. 22, 1–24.
Béal, M.-P., Senellart, J., 1998. On the bound of the synchronization delay of a local automaton. Theor. Comput. Sci. 205, 297–306.
Berstel, J., Perrin, D., 1985. Theory of Codes. Academic Press, London.
Bussoli, L., Michel, C.J., Pirillo, G., 2011. On some forbidden configurations for self-complementary trinucleotide circular codes. J. Algebra Number Theory Acad. 2, 223–232.
Bussoli, L., Michel, C.J., Pirillo, G., 2012. On conjugation partitions of sets of trinucleotides. Appl. Math. 3, 107–112.
Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Natl. Acad. Sci. 43, 416–421.
Farabaugh, P.J., 1996. Programmed translational frameshifting. Annu. Rev. Genet. 30, 507–528.
Frey, G., Michel, C.J., 2003. Circular codes in archaeal genomes. J. Theor. Biol. 223, 413–431.
Frey, G., Michel, C.J., 2006. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Comput. Biol. Chem. 30, 87–101.
Gesteland, R.F., Weiss, R.B., Atkins, J.F., 1992. Recoding: reprogrammed genetic decoding. Science 257, 1640–1641.
Giannerini, S., Gonzalez, D.L., Rosa, R., 2012. DNA, dichotomic classes and frame synchronization: a quasi-crystal framework. Philos. Trans. R. Soc. A 370, 2987–3006.
Golomb, S.W., Gordon, B., Welch, L.R., 1958a. Comma-free codes. Can. J. Math. 10, 202–209.
Golomb, S.W., Welch, L.R., Delbrück, M., 1958b. Construction and properties of comma-free codes. Biol. Medd. Dan. Vidensk. Selsk. 23.
Gonzalez, D.L., Giannerini, S., Rosa, R., 2011. Circular codes revisited: a statistical approach. J. Theor. Biol. 275, 21–28.
Gonzalez, D.L., Giannerini, S., Rosa, R. On the origin of the mitochondrial genetic code: towards a unified mathematical framework for the management of genetic information. Nat. Prec. http://dx.doi.org/10.1038/npre.2012.7136.1, in press.
Hoffman, D.C., Anderson, R.C., DuBois, M.L., Prescott, D.M., 1995. Macronuclear gene-sized molecules of hypotrichs. Nucleic Acids Res. 23, 1279–1283.
Jolivet, R., Rothen, F., Peculiar symmetry of DNA sequences and evidence suggesting its evolutionary origin in a primeval genetic code. In: Ehrenfreund, P., Angerer, O., Battrick, B. (Eds.), Proceedings of the First European Workshop in Exo-/astro-biology. ESA SP-496, Noordwijk, 2001, pp. 173–176 .
José, M.V., Govezensky, T., García, J.A., Bobadilla, J.R., 2009. On the evolution of the standard genetic code: vestiges of critical scale invariance from the RNA world in current prokaryote genomes. PLoS ONE 4 (2), e4340.
Keeling, P.J., Doolittle, W.F., 1996. A non-canonical genetic code in an early diverging eukaryotic lineage. EMBO J. 15, 2285–2290.
Koch, A.J., Lehman, J., 1997. About a symmetry of the genetic code. J. Theor. Biol. 189, 171–174.
Lassez, J.-L., 1976. Circular codes and synchronization. Int. J. Comput. Syst. Sci. 5, 201–208.
Lassez, J.-L., Rossi, R.A., Bernal, A.E., 2007. Crick's hypothesis revisited: the existence of a universal coding frame, IEEE AINAW'07.
Liang, A., Heckmann, K., 1993. Blepharisma uses UAA as a termination codon. Naturwissenschaften 80, 225–226.
May, E.E., Vouk, M.A., Bitzer, D.L., Rosnick, D.I., 2004. An error-correcting framework for genetic sequence analysis. J. Franklin Inst. 341, 89–109.

Michel, C.J., Pirillo, G., 2010. Identification of all trinucleotide circular codes. Comput. Biol. Chem. 34, 122–125.
Michel, C.J., Pirillo, G., 2011. Strong trinucleotide circular codes. Int. J. Combinatorics 2011, 1–14, ID 659567.
Michel, C.J., Pirillo, G., Pirillo, M.A., 2008a. Varieties of comma-free codes. Comput. Math. Appl. 55, 989–996.
Michel, C.J., Pirillo, G., Pirillo, M.A., 2008b. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theor. Comput. Sci. 401, 17–25.
Michel, C.J., Pirillo, G., Pirillo, M.A., 2012. A classification of 20-trinucleotide circular codes. Inf. Comput. 212, 55–63.
Nikolaou, C., Almirantis, Y., 2003. Mutually symmetric and complementary triplets: difference in their use distinguish systematically between coding and non-coding genomic sequences. J. Theor. Biol. 223, 477–487.
Pirillo, G., 2003. A characterization for a set of trinucleotides to be a circular code. In: Pellegrini, C., Cerrai, P., Freguglia, P., Benci, V., Israel, G. (Eds.), Determinism, Holism, and Complexity. Kluwer.
Pirillo, G., 2008a. A hierarchy for circular codes. RAIRO-Theor . Inf. Appl. 42, 717–728.
Pirillo, G., 2008b. Some remarks on prefix and suffix codes. Pure Math. Appl. 19, 53–60.
Pirillo, G., 2010. Non sharing border codes. Adv. Appl. Math. Sci. 3, 215–223.
Pirillo, G., Pirillo, M.A., 2005. Growth function of self-complementary circular codes. Biol. Forum 98, 97–110.
Qian, Q., Li, J.N., Zhao, H., Hagervall, T.G., Farabaugh, P.J., Bjork, G.R., 1998. A new model for phenotypic suppression of frameshift mutations by mutant tRNAs. Mol. Cell 1, 471–482.
Schneider, S.U., Leible, M.B., Yang, X.P., 1989. Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of Acetabularia and the occurrence of unusual codon usage. Mol. Gen. Genet. 218, 445–452.
Schneider, S.U., de Groot, E.J., 1991. Sequences of two rbcS cDNA clones of *Batophora oerstedii*: structural and evolutionary considerations. Curr. Genet. 20, 173–175.
Seligmann, H., 2012. Putative mitochondrial polypeptides coded by expanded quadruplet codons, decoded by antisense tRNAs with unusual anticodons. Biosystems 110, 84–106.
Štambuk, N., 1999. On circular coding properties of gene and protein sequences. Croat. Chem. Acta 72, 999–1008.
Wang, K., Schmied, W.H., Chin, J.W., 2012. Reprogramming the genetic code: from triplet to quadruplet codes. Angew. Chem. Int. Ed. 51, 2288–2297.