ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

C. J. Michel, G. Pirillo

A PERMUTED SET OF A TRINUCLEOTIDE CIRCULAR CODE CODING THE 20 AMINO ACIDS IN VARIANT NUCLEAR CODES

R. 20, 2012

Christian J. Michel – Equipe de Bioinformatique Théorique BFO, LSIIT (UMR 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France. Email: `michel@dpt-info.u-strasbg.fr`.

G. Pirillo – Consiglio Nazionale delle Ricerche, Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti", Unità di Firenze, Dipartimento di Matematica "U.Dini", viale Morgagni 67/A, 50134 Firenze, Italia and Université de Marne-la-Vallée, 5 boulevard Descartes, 77454 Marne-la-Vallée Cedex 2, France. Email: `pirillo@math.unifi.it`.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

During our study of the combinatorial properties of trinucleotide circular codes, we identify a permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes. This circular code property allows a set of 20 trinucleotides to retrieve the reading frame in genes and one of its permuted set of 20 trinucleotides to code the 20 amino acids. This result is a contribution to the research field analysing the mathematical properties of genetic codes.

Key words: trinucleotide circular code; trinucleotide permuted set; genetic code; amino acid.

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 64 trinucleotides is a code in the sense of language theory, more precisely a uniform code, but not a circular code [4, 18]. 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. [7, 10, 11]. After the discovery of a circular code in genes with strong mathematical properties [1], circular codes are mathematical objects studied in combinatorics, theoretical computer science and theoretical biology. This theory underwent a rapid development e.g. [17, 3, 2, 35, 14, 8, 27, 28, 21, 32, 9, 19, 24, 25, 29, 15, 22, 30, 31, 23, 5, 12, 6, 26].

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 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. The variant genetic codes differ form the standard one by the number of trinucleotides coding the 20 amino acids or by a coding reassignment of trinucleotides. All genetic codes are surjective maps. There are $2^9 \times 3 \times 4^5 \times 6^3 = 339,738,624$ sets $\mathcal{S}$ of 20 trinucleotides coding the 20 amino acids, i.e. with a bijective map.

There are exactly $12,964,440$ circular codes $\mathcal{X}$ of 20 trinucleotides [1, 22]. 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 and a set $\mathcal{S}$. Ten 20-trinucleotide circular codes code 18 amino acids. The common 20-trinucleotide circular code of eukaryotes and prokaryotes [1] 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 $\mathcal{A}_4^3 \setminus \{AAA, CCC, GGG, TTT\}$ [6]. Indeed, each circular code X can be associated with two other subsets X_1 and X_2 of $\mathcal{A}_4^3 \setminus \{AAA, CCC, GGG, TTT\}$ 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 trinucleotide circular code $Y = \{ACG, ACT, AGA, AGG, AGT, ATA, ATC, CAA, CAC,$
$CAG, CCT, GCC, GCG, GCT, GGT, TCG, TCT, TGA, TGT, TTA\}$ which has a permuted set $\mathcal{P}^2(Y) = \{AAG, AAT, ACA, ATG, ATT, CAT, CCA, CGC, GAC, GAG, GCA, GGC,$

$GTC, TAC, TAG, TCC, TGC, TGG, TTC, TTG\}$ coding the 20 amino acids in the variant nuclear codes 6 and 15.

2. Definitions

The classical notions of langage theory can be found in [4]. 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. [4, 1].

Definition 2.1. *The circular permutation map $\mathcal{P} : \mathcal{A}_4^3 \to \mathcal{A}_4^3$ permutes circularly each trinucleotide $l_0l_1l_2$ as follows $\mathcal{P}(l_0l_1l_2) = l_1l_2l_0$.*

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 2.2. *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 $i = 1, \ldots, n$.*

Definition 2.3. *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 $sx_2 \cdots x_np = x'_1 \cdots x'_m$ and $x_1 = ps$ imply $n = m$, $p = \varepsilon$ (empty word) and $x_i = x'_i$ for $i = 1, \ldots, n$.*

Definition 2.4. *If X is a subset of $\mathcal{A}_4^3 \setminus \{AAA, CCC, GGG, TTT\}$, 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 2.5. *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 [28] 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 2.6. *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*

$$l_1d_1, l_2d_2, \ldots, l_nd_n \in X$$

and

$$d_1l_2, d_2l_3, \ldots, d_{n-1}l_n, d_nl_{n+1} \in X.$$

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

Proposition 2.7. *[28]. Let X be a trinucleotide code. The following conditions are equivalent:*
(i) X is a circular code;
(ii) X has no $5LDCN$.

The nuclear code of ciliatea (Oxytricha, Stylonychia, Paramecium, Tetrahymena; [13]), dasycladacean (Acetabularia, Batophora; [33, 34]), hexamita [16] (variant nuclear code 6 according to the GenBank convention, National Center for Biotechnology Information (NCBI), July 07 2010) is defined by Table 1:

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 1. Nuclear code of ciliate, dasycladacean and hexamita (variant nuclear code 6) showing the correspondence between the 64 trinucleotides $\{AAA, ..., TTT\}$ 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.

The two trinucleotides TAA coding Gln (Q) 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 [20]) (variant nuclear code 15 according to the GenBank convention, National Center for Biotechnology Information (NCBI), July 07 2010) is defined by Table 2:

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

Table 2. The nuclear code of ciliate (Blepharisma) (variant nuclear code 15) showing the correspondence between the 64 trinucleotides $\{AAA, ..., TTT\}$ 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.

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 3.1. *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 3.2. *The following set of trinucleotides*

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

is a circular code. More precisely, Y *is the* $11,056,585th$ *among* $12,964,440$ *circular codes (in the lexicographical order) and belongs to the classes* $C^{5LDN} = C^{5LDCN} = C^{5DLN}$ *[26].*

We give here a direct proof based on dinucleotides.
Proof. Y **is a circular code.** We use Proposition 1. By way of contradiction, suppose that Y admits a $5LDCN$ $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_5$. By Lemma 3.1, 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 $\{AC, AG, CC, GG, TC\}$. So, it is enough to prove that each choice $d_4 \in \{AC, AG, CC, GG, TC\}$ leads to a contradiction.

Claim 1. *$AC \neq d_4$ and $AC \neq d_3$.*

Proof. By way of contradiction, suppose $d_4 = AC$. We have $l_4 = C$ and, consequently, $d_3 \in \{AT, CA, GC\}$. But, as $\{AT, CA, GC\} \cap \{AC, AG, CC, GG, TC\} = \emptyset$, we are in contradiction with Lemma 3.1. So, $AC \neq d_4$. In the same way, we prove $AC \neq d_3$.

Claim 2. *$AG \neq d_4$ and $AG \neq d_3$.*

Proof. By way of contradiction, suppose $d_4 = AG$. We have $l_4 = C$ and, consequently, $d_3 \in \{AT, CA, GC\}$. But, as $\{AT, CA, GC\} \cap \{AC, AG, CC, GG, TC\} = \emptyset$, we are in contradiction with Lemma 3.1. So, $AG \neq d_4$. In the same way, we prove $AG \neq d_3$.

Claim 3. *$TC \neq d_4$ and $TC \neq d_3$.*

Proof. 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 $\{AT, CA, TG, TT\} \cap \{AC, AG, CC, GG, TC\} = \emptyset$), we are in contradiction with Lemma 1, and if $AG = d_3$ we are in contradiction with Claim 2. So, $TC \neq d_4$. In the same way, we prove $TC \neq d_3$.

Claim 4. *$CC \neq d_4$.*

Proof. 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 $\{CA, GC\} \cap \{AC, AG, CC, GG, TC\} = \emptyset$), we are in contradiction with Lemma 1; if $AC = d_3$ we are in contradiction with Claim 1; if $AG = d_3$ we are in contradiction with Claim 2; and if $TC = d_3$ we are in contradiction with Claim 3. So, $CC \neq d_4$.

Claim 5. *$GG \neq d_4$.*

Proof. 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 Claim 3, we are in contradiction with Lemma 1 and Claim 2. So, $GG \neq d_4$.
By Claims 1, 2, 3, 4 and 5, $d_4 \notin \{AC, AG, CC, GG, TC\}$. So, by Lemma 3.1, we are in contradiction. So, Y is a circular code. ∎

Proposition 3.3. *The trinucleotide circular code*

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

has a permuted set

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

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

Proof. $\mathcal{P}^2(Y)$ **is not a circular code.** Consider the subset $\{AAT, ATT, TAC, ACA\}$ 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 $\{AAT, ATT, TAC, ACA\}$ 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). ∎

Proposition 3 allows a set of 20 trinucleotides to retrieve the reading frame in genes and one of its permuted set of 20 trinucleotides to code the 20 amino acids. This circular code property involves two sets of 20 trinucleotides in the coding process of amino acids. The remaining trinucleotides allow an additional coding function which remains to be discovered. This result is a contribution to the identification of mathematical properties of genetic codes.

Acknowledgement. We thank Jacques Justin for his advices. The second author thanks the Dipartimento di matematica "U. Dini" for giving him a friendly hospitality.

References

[1] D.G. Arquès, C.J. Michel. A complementary circular code in the protein coding genes. *J. Theor. Biol.* **182**, 45-58 (1996).

[2] F. Bassino. Generating function of circular codes. *Adv. Appl. Math.* **22**, 1-24 (1999).

[3] M.-P. Béal, J. Senellart. On the bound of the synchronization delay of a local automaton. *Theoret. Comput. Sci.* **205**, 297-306 (1998).

[4] J. Berstel, D. Perrin. *Theory of Codes.* Academic Press, London, 1985.

[5] L. Bussoli, C.J. Michel, G. Pirillo. On some forbidden configurations for self-complementary trinucleotide circular codes. *J. Algebra Number Theory Academia* **2**, 223-232 (2011).

[6] L. Bussoli, C.J. Michel, G. Pirillo. On conjugation partitions of sets of trinucleotides. *Applied Math.* **3**, 107-112 (2012).

[7] F.H.C. Crick, J.S. Griffith, L.E. Orgel. Codes without commas. *Proc. Natl. Acad. Sci.* **43**, 416-421 (1957).

[8] G. Frey, C.J. Michel. Circular codes in archaeal genomes. *J. Theor. Biol.* **223**, 413-431 (2003).

[9] 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**, 87-101 (2006).

[10] S.W. Golomb, B. Gordon, L.R. Welch. Comma-free codes. *Canad. J. Math.* **10**, 202-209 (1958).

[11] S.W. Golomb, L.R. Welch, M. Delbrück. Construction and properties of comma-free codes. *Biol. Medd. Dan. Vid. Selsk.* **23** (1958).

[12] D.L. Gonzalez, S. Giannerini, R. Rosa. Circular codes revisited: a statistical approach. *J. Theor. Biol.* **275**, 21-28 (2011).

[13] D.C. Hoffman, R.C. Anderson, M.L. DuBois, D.M. Prescott. Macronuclear gene-sized molecules of hypotrichs. *Nucleic Acids Res.* **23**, 1279-1283 (1995).

[14] R. Jolivet, F. Rothen. Peculiar symmetry of DNA sequences and evidence suggesting its evolutionary origin in a primeval genetic code. Proceedings of the First European Workshop in Exo-/astro-biology. Eds.: P. Ehrenfreund, O. Angerer & B. Battrick. ESA SP-496, Noordwijk, 173-176 (2001).

[15] M.V. José, T. Govezensky, J.A. García, J.R. Bobadilla. 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 (2009).

[16] P.J. Keeling, W.F. Doolittle. A non-canonical genetic code in an early diverging eukaryotic lineage. *EMBO J.* **15**, 2285-2290 (1996).

[17] A.J. Koch, J. Lehman. About a symmetry of the genetic code. *J. Theor. Biol.* **189**, 171-174 (1997).

[18] J.-L. Lassez. Circular codes and synchronization. *Int. J. Comput. Syst. Sciences* **5**, 201-208 (1976).

[19] J.-L. Lassez, R.A. Rossi, A.E. Bernal. Crick's hypothesis revisited: the existence of a universal coding frame. *IEEE AINAW'07* (2007).

[20] A. Liang, K. Heckmann. Blepharisma uses UAA as a termination codon. *Naturwissenschaften* **80**, 225-226 (1993).

[21] E.E. May, M.A. Vouk, D.L. Bitzer, D.I. Rosnick. An error-correcting framework for genetic sequence analysis. *J. Franklin Inst.* **341**, 89-109 (2004).

[22] C.J. Michel, G. Pirillo. Identification of all trinucleotide circular codes. *Comput. Biol. Chem.* **34**, 122-125 (2010).

[23] C.J. Michel, G. Pirillo. Strong trinucleotide circular codes. *Int. J. Combinatorics* **2011 ID 659567**, 1-14 (2011).

[24] C.J. Michel, G. Pirillo, M.A. Pirillo. Varieties of comma-free codes. *Comput. Math. Appl.* **55**, 989-996 (2008).

[25] C.J. Michel, G. Pirillo, M.A. Pirillo. A relation between trinucleotide comma-free codes and trinucleotide circular codes. *Theoret. Comput. Sci.* **401**, 17-25 (2008).

[26] C.J. Michel, G. Pirillo, M.A. Pirillo. A classification of 20-trinucleotide circular codes. *Information and Computation* **212**, 55-63 (2012).

[27] C. Nikolaou, Y. Almirantis. Mutually symmetric and complementary triplets: difference in their use distinguish systematically between coding and non-coding genomic sequences. *J. Theor. Biol.* **223**, 477-487 (2003).

[28] G. Pirillo. A characterization for a set of trinucleotides to be a circular code. In *Determinism, Holism, and Complexity* (Edited by C. Pellegrini, P. Cerrai, P. Freguglia, V. Benci and G. Israel), Kluwer (2003).

[29] G. Pirillo. A hierarchy for circular codes. *RAIRO-Theor. Inf. Appl.* **42**, 717-728 (2008).

[30] G. Pirillo. Some remarks on prefix and suffix codes. *Pure Math. Appl.* **19**, 53-60 (2008).

[31] G. Pirillo. Non sharing border codes. *Adv. Appl. Math. Sci.* **3**, 215-223 (2010).

[32] G. Pirillo, M.A. Pirillo. Growth function of self-complementary circular codes. *Biology Forum* **98**, 97-110 (2005).

[33] S.U. Schneider, M.B. Leible, X.P. Yang. 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 (1989).

[34] S.U. Schneider, E.J. de Groot. Sequences of two rbcS cDNA clones of Batophora oerstedii: structural and evolutionary considerations. *Curr. Genet.* **20**, 173-175 (1991).

[35] N. Štambuk. On circular coding properties of gene and protein sequences. *Croatica Chemica Acta* **72**, 999-1008 (1999).