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.
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 \(\text{SGC}\), 61 trinucleotides code the 20 amino acids as there are three termination trinucleotides \(TAA, TAG\) and \(TGA\). The trinucleotide \(ATG\) coding the amino acid \(\text{Met} (M)\) is also the initiation trinucleotide (noted \(i\)) in the general case. Two amino acids are encoded by a single trinucleotide: \(\text{Met} (M)\) and \(\text{Trp} (W)\). Nine amino acids are encoded by two trinucleotides: \(\text{Asn} (N), \text{Asp} (D), \text{Cys} (C), \text{Gln} (Q), \text{Glu} (E), \text{His} (H), \text{Lys} (K), \text{Phe} (F)\) and \(\text{Tyr} (Y)\). One amino acid is encoded by three trinucleotides: \(\text{Ile} (I)\). Five amino acids are encoded by four trinucleotides: \(\text{Ala} (A), \text{Gly} (G), \text{Pro} (P), \text{Thr} (T)\) and \(\text{Val} (V)\). Three amino acids are encoded by six trinucleotides: \(\text{Arg} (R), \text{Leu} (L)\) and \(\text{Ser} (S)\). 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 \(S\) 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 \(\text{SGC}\)). There is no bijection, unfortunately (in a certain way), between a 20-trinucleotide circular code and a set \(S\). 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 \(\mathcal{A}_3^4 \setminus \{\text{AAA}, \text{CCC}, \text{GGG}, \text{TTT}\}\) [6]. Indeed, each circular code \(X\) can be associated with two other subsets \(X_1\) and \(X_2\) of \(\mathcal{A}_3^4 \setminus \{\text{AAA}, \text{CCC}, \text{GGG}, \text{TTT}\}\) simply by operating two circular permutations \(P\) of one letter and two letters on the trinucleotides of \(X\), respectively, i.e. \(X_1 = P(X)\) and \(X_2 = P^2(X)\). During this research work, we identify a trinucleotide circular code \(Y = \{\text{ACG}, \text{ACT}, \text{AGA}, \text{AGG}, \text{AGT}, \text{ATA}, \text{ATC}, \text{CAA}, \text{CAC}, \text{CAG}, \text{CCT}, \text{GCC}, \text{GCG}, \text{GCT}, \text{GGT}, \text{TCG}, \text{TCT}, \text{TGA}, \text{TGT}, \text{TTA}\}\) which has a permuted set \(P^2(Y) = \{\text{AAG}, \text{AAT}, \text{ACA}, \text{ATG}, \text{ATT}, \text{CAT}, \text{CCA}, \text{CGC}, \text{GAC}, \text{GAG}, \text{GCA}, \text{GGC}\}\).
GT C, TAC, TAG, TCC, TGC, TGG, TTC, TT G \} 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 [4]. Let \( A_4 = \{ A, C, G, T \} \) denote the genetic alphabet, lexicographically ordered with \( A < C < G < T \). We use the following notation:

- \( A_4^+ \) (respectively \( A_4^* \)) is the set of words (respectively non-empty words) over \( A_4 \).
- \( A_4^2 \) is the set of the 16 words of length two (diletters or dinucleotides) and
- \( A_4^3 \) 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 \( \mathcal{P} : A_4^3 \rightarrow A_4^3 \) 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 \( x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X, n, m \geq 1, p \in A_4^1, s \in A_4^+ \), 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 \( A_4^3 \setminus \{ AAA, CCC, GGG, TTT \} \), 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 \( A_4, d_1, d_2, \ldots, d_{n-1}, d_n, \ldots \) diletters in \( A_4^2 \) 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 \( X \subset A_4^3 \) if

\[
\begin{align*}
&l_1d_1, l_2d_2, \ldots, l_nd_n \in X \\
\text{and} \\
&d_1l_2, d_2l_3, \ldots, d_{n-1}l_n, d_nl_{n+1} \in X.
\end{align*}
\]

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:

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

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.

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:
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 and the trinucleotides TAA and TGA coding no amino acid are the termination codons Ter. The permuted set \(P^2(Y)\) 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

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

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

**Proposition 3.3.** The trinucleotide circular code

\[
Y = \{ACG, ACT, AGA, AGG, AGT, ATA, ATC, CAA, CAC, CAG, CCA, CCG, CGG, CGT, TCG, TCT, TGA, TGG, TTC, TTA\}
\]

has a permuted set

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

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


