On Conjugation Partitions of Sets of Trinucleotides

Lorenzo Bussoli¹, Christian J. Michel², Giuseppe Pirillo¹,³

¹Dipartimento di Matematica U.Dini, Firenze, Italia
²Equipe de Bioinformatique Théorique, Université de Strasbourg, Strasbourg, France
³Université de Marne-la-Vallée, Marne-la-Vallée, France

Email: {bussoli, pirillo}@math.unifi.it, michel@dpt-info.u-strasbg.fr

Received October 2, 2011; revised December 5, 2011; accepted December 13, 2011

ABSTRACT

We prove that a trinucleotide circular code is self-complementary if and only if its two conjugated classes are complement of each other. Using only this proposition, we prove that if a circular code is self-complementary then either both its two conjugated classes are circular codes or none is a circular code.

Keywords: Trinucleotide; Conjugated Trinucleotides; Code; Circular Code; Self-Complementary Circular Code; Complementary Circular 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 64 trinucleotides is a code in the sense of language theory, more precisely a uniform code, but not a circular code [1,2]. 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. [3-5]. After the discovery of a circular code in genes with important properties [6], circular codes are mathematical objects studied in combinatorics, theoretical computer science and theoretical biology, e.g. [7-23].

There are 528 self-complementary circular codes of 20 trinucleotides [6,24,25] and, as proved here, they are naturally partitioned into two quite symmetric classes.

Let \( T = \{AAA, CCC, GGG, TTT\} \) be the four trinucleotides with identical nucleotides. In this paper, we study some particular partitions of \( A_4^1 \setminus T \). Indeed, each circular code \( X_0 \) can be associated with two other subsets \( X_1 \) and \( X_2 \) of \( A_4^1 \setminus T \) simply by operating two circular permutations of one letter and two letters on the trinucleotides of \( X_0 \). Then, we prove our main result, i.e. a circular code is self-complementary if and only if the remaining two classes are complement of each other. Furthermore, we also show that a subset of \( A_4^1 \setminus T \) is a circular code if and only if the set consisting of all its complements is a circular code.

As a consequence of these results, we also prove that if a circular code is self-complementary then either both its two conjugated classes are circular codes or none is a circular code.

In Section 2, we give the necessary definitions and a characterization for a set of trinucleotides to be a circular code. In Section 3, we give the results, mainly expressed by Proposition 7 and Proposition 8.

2. Definitions

The classical notions of alphabet, empty word, length, factor, proper factor, prefix, proper prefix, suffix, proper suffix, lexicographical order, etc. are those of [1]. 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 \( A_4 \);
- \( A_4^1 \) is the set of the 16 words of length 2 (diletters or dinucleotides);
- \( A_4^3 \) is the set of the 64 words of length 3 (triletters or trinucleotides).

We now recall two important genetic maps, the definitions of code and circular code, and the property of \( C^3 \)-self-complementarity for a circular code, in particular [1,6,17,24,25].

Definition 1. The complementarity map \( C : A_4^+ \to A_4^+ \) is defined by \( C(A) = T \), \( C(T) = A \), \( C(C) = G \) and \( C(G) = C \), and by \( C(uv) = C(v)C(u) \) for all \( u, v \in A_4^+ \), e.g., \( C(AAC) = GTT \).

The map \( C \) on words is naturally extended to a word
set $X$: its complementary trinucleotide set $C(X)$ is obtained by applying the complementarity map $C$ to all the trinucleotides of $X$.

**Definition 2.** The circular permutation map $P: A_3^3 \rightarrow A_3^3$ permutes circularly each trinucleotide $l_1l_2l_3$ as follows $P(l_1l_2l_3) = l_1l_2l_3$.

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

**Definition 3.** 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 4.** A trinucleotide code $X$ is circular if, for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, $p \in A_3^3$, $s \in A_3^3$, the conditions $x_1 \cdots x_n = x'_1 \cdots x'_m$ and $x_i = ps$ imply $n = m$, $p = \varepsilon$ (empty word) and $x_i = x'_i$ for $i = 1, \ldots, n$.

**Definition 5.** A trinucleotide code $X$ is self-complementary if, for each $x \in X$, $C(x) \in X$.

**Definition 6.** If $X_o$ is a subset of $A_3^3 \setminus \{T\}$, we denote by $X_1$ the permuted trinucleotide set $\{P(x)\}$ and by $X_2$ the permuted trinucleotide set $\{P^2(x)\}$ and we call $X_1$ and $X_2$ the conjugated classes of $X_o$.

**Definition 7.** A trinucleotide circular code $X_o$ is $C^3$-self-complementary if $X_o$, $X_1$, and $X_2$ are circular codes satisfying the following properties: $X_o = C(X_o)$ (self-complementary), $C(X_1) = X_2$ and $C(X_2) = X_1$.

We have proved that there are exactly 528 self-complementary trinucleotide circular codes having 20 elements [6,24,25].

The concept of necklace was introduced by Pirillo [17] in order to characterize the circular codes for an efficient algorithm development. Let $l_1, l_2, \ldots, l_{n+1}, l_{n+2}, \ldots$ be letters in $A_3$, $d_1, d_2, \ldots, d_{n+1}, l_{n+1}, l_{n+2}, \ldots$ diletters in $A_3^2$ and $n \geq 2$ an integer.

**Definition 8.** Letter Diletter Continued Necklace (LDCN): We say that the ordered sequence $l_1, d_1, l_2, d_2, \ldots, d_{n+1}, l_{n+1}$ is an $(n+1)$-LDCN for a subset $X \subset A_3^3$ if

$$l_1l_2l_3, \ldots, l_nl_{n+1} \in X$$

The following conditions are equivalent:

1. $X$ is a circular code;
2. $X$ has no 5LDCN.

**Proposition 1.** [17] Let $X$ be a trinucleotide code. The following conditions are equivalent:

3. Results

**Proposition 2.** If $X_0$ is a trinucleotide circular code having 20 elements and $X_1$ and $X_2$ are its two conjugated classes then $X_0$, $X_1$ and $X_2$ constitute a partition of $A_3^3 \setminus \{T\}$.

**Proof.** It is enough to prove that $X_0 \cap X_1 = X_0 \cap X_2 = X_2 \cap X_0 = \emptyset$. Suppose that the trinucleotide $l_1l_2l_3$ belongs both to the classes $X_0$ and $X_1$. Then $l_1l_2l_3$ and $l_3l_1l_2$ are both in class $X_2$. As no two conjugated trinucleotides can belong to a circular code, we are in contradiction. Suppose that the trinucleotide $l_1l_2l_3$ belongs both to the classes $X_1$ and $X_2$. Then $l_1l_2l_3$ and $l_3l_1l_2$ are both in class $X_0$. As no two conjugated trinucleotides can belong to a circular code, we are in contradiction. So, $X_0 \cap X_1 = X_0 \cap X_2 = X_1 \cap X_2 = \emptyset$.\hfill $\Box$

**Proposition 3.** The class of self-complementary circular codes $X_o$ with both $X_1$ and $X_2$ in the class of circular codes is non-empty.

**Proof.** Consider, for example, the following set $X_0$ of 20 trinucleotides

$$X_0 = \{AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, AGT, ATC, ATT, CCT, CGT, CTT, GAT, GCC, GCT, GGC, GGT, GTT\}.$$

It is enough to prove that $X_0$ is a self-complementary circular code and that its two conjugated classes $X_1$ and $X_2$ are also circular codes.

$X_o$ is a self-complementary circular code. $X_o$ is self-complementary. Obvious by inspection.

$X_0$ is a circular code. We use Proposition 1 [17]. By way of contradiction, suppose that $X_0$ admits a 5LDCN. As $l_2$ can be $A$, $C$, $G$ or $T$, it is enough to prove that each choice leads to a contradiction.

1. If $l_2 = A$, then there is no possible $d_1$ as $A$ is not a suffix of any trinucleotide of $X_0$, contradiction.
2. If $l_2 = C$, there are three possible $d_2$:
   - if $d_2 = CT$ (a) or $d_2 = GT$ (b) then $l_2 = T$ (c) but there is no possible $d_3$ as $T$ is not a prefix of any trinucleotide of $X_0$, contradiction,
   - if $d_2 = TT$ (d), there is a contradiction as no trinucleotide of $X_0$ has a prefix $TT$.
3. If $l_2 = G$, there are six possible $d_2$:
   - if $d_2 = CT$ or $d_2 = GT$, contradiction (a) and (b),
   - if $d_2 = CC$ then $l_2 = T$, contradiction (c),
   - if $d_2 = GC$ or $d_2 = AT$ then $l_2 = C$ or $l_2 = T$.
We are in contradiction. Hence, (c), contradiction, and if \( d_3 = TT \), similarly to (d), contradiction.

- If \( l_3 = T \), contradiction (c).
- If \( d_3 = TT \), contradiction (d).

4) If \( l_3 = T \), similarly to (c), contradiction.

As, for each letter, we cannot complete the assumed \( 5\text{LDNC} \) for \( X_0 \), we are in contradiction. Hence, \( X_0 \) is a circular code.

\( X_1 = \mathcal{P}^1(X_0) \) is a circular code. We have to prove that

\[
\begin{align*}
X_1 &= \{ ACA, AGA, ATA, ATG, CCA, CCG, \\
& \quad \quad \quad CGA, CTA, CTC, CTG, GCA, GCG, GGA, \\
& \quad \quad \quad GTA, GTG, TCA, TTA, TTC, TTG \}
\end{align*}
\]

is a circular code. By way of contradiction, assume that \( X_1 \) admits a \( 5\text{LDNC} \).

1) If \( l_2 = A \), there are four possible \( d_2 \): \( CA, GA, TA \) and \( TG \), but no possible \( l_2 \), contradiction.

2) If \( \ell_2 = C \), there are three possible \( d_2 \): \( CT, GT \) and \( TT \), but no possible \( l_2 \), contradiction.

3) If \( l_2 = G \), there are six possible \( d_2 \): \( AT, CC \) and \( GC \), and the cases \( CT, GT \) and \( TT \) already seen, but no possible \( l_2 \), contradiction.

4) If \( l_2 = T \), there is no possible \( d_2 \), contradiction. Hence, \( X_2 \) is also a circular code.

\( X_2 = \mathcal{P}^2(X_0) \) is a circular code. Finally, we have to prove that

\[
\begin{align*}
X_2 &= \{ CAA, CAC, CAG, CAT, CGA, CGG, \\
& \quad \quad \quad GAA, GAC, GAG, TAA, TAC, TAG, TAT, \\
& \quad \quad \quad TCC, TCG, TCT, TGA, TGG, TTG \}
\end{align*}
\]

is a circular code. By way of contradiction, assume that \( X_2 \) admits a \( 5\text{LDNC} \).

1) If \( l_2 = A \), there is no possible \( d_2 \), contradiction.

2) If \( l_2 = C \), there are six possible \( d_2 \): \( AA, AC, AG, AT, GC \) and \( GG \), but no possible \( l_2 \), contradiction.

3) If \( l_2 = G \), there are three possible \( d_2 \): \( AA \), \( AC \) and \( AG \) which are cases already seen, contradiction.

4) If \( l_2 = T \), there are four possible \( d_2 \): \( CA, TA, TC \) and \( TG \), but no possible \( l_2 \), contradiction.

Hence, as \( X_0 \) and \( X_1 \), \( X_2 \) is also a circular code.

\[ \square \]

**Proposition 4.** The class of self-complementary circular codes \( X_0 \) having 20 elements with neither \( X_1 \) nor \( X_2 \) in the class of circular codes is non-empty.

**Proof.** Consider, for example, the following set \( X_0 \) of 20 trinucleotides

\[ X_0 = \{ AAC, AAG, AAT, ACC, ACG, ACT, \\
AGC, AGT, ATC, ATT, CGT, CTG, GAT, \\
GCC, GCT, GGA, GGC, GGT, GTT, TCC \} \]

It is enough to prove that \( X_0 \) is a self-complementary circular code and that neither its conjugated class \( X_1 \) nor its conjugated class \( X_2 \) are circular codes. \( X_0 \) is a self-complementary circular code. \( X_0 \) is self-complementary. Obvious by inspection.

\( X_0 \) is a circular code. We use Proposition 1 [17]. By way of contradiction, assume that \( X_0 \) admits a \( 5\text{LDNC} \).

1) If \( l_2 = A \) then there is one possible \( d_2 = GG \) but no possible \( l_2 \), contradiction.

2) If \( l_2 = C \), there are two possible \( d_2 \):

- if \( d_2 = GT \) then \( l_3 = T \) (a) and \( d_3 = CC \) (b) but there is no possible \( l_3 \), contradiction,

- if \( d_3 = TT \) (c) then there is no possible \( l_3 \), contradiction.

3) If \( l_2 = G \) we have seven possible \( d_2 \):

- if \( d_2 = GT \) then \( l_3 = C \) or \( l_3 = T \), contradiction.

4) If \( l_2 = T \), similarly to (a), contradiction.

Hence, \( X_0 \) is a circular code.

\( X_1 = \mathcal{P}^1(X_0) \) is not a circular code. We have

\[
X_1 = \{ ACA, AGA, ATA, ATG, CCA, CCG, \\
CTT, CGA, CTA, CTG, GCA, GCG, GTA, GTG, TCA, TTA, TTC, TTG \}
\]

We use a technique developed in [23]. Observe that \( X_1 \) contains \( \{ AGA, CCT, GAG, TTC \} \). So,

\[
\begin{align*}
& (l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4) \\
\end{align*}
\]

is a \( 5\text{LDNC} \) for this 4-element subset of \( X_1 \) and, a fortiori, for \( X_1 \) itself which, consequently, is not a circular code.

\( X_2 = \mathcal{P}^2(X_0) \) is not a circular code. We have

\[
X_2 = \{ AGG, CAA, CAC, CAG, CAT, CGC, \\
CGG, CTC, GAA, GAC, TAA, TAC, TAG, \\
TAT, TCG, TCT, TGA, TGC, TTG, TTG \}
\]
We again use a technique developed in [23]. Remark that \( X_2 \) contains \( \{GA, A, C, AT, GC, CT\} \). So, 
\[
(l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_5) \\
\]
is a SLDCN for this 4-element subset of \( X_2 \) and, a fortiori, for \( X_2 \) itself which, consequently, is not a circular code. □

We need the propositions hereafter and, in particular the following one which states a general property of the involutional antiisomorphisms such as the complementary map \( C \).

**Proposition 5.** A subset \( X \) of \( A_4^6 \setminus T \) is a circular code if and only if \( C(X) \) is a circular code.

**Proof.** Suppose, first, that \( X \) is not a circular code and that \( C(X) \) is a circular code. So \( X \) has a SLDCN. This means that there are 13 nucleotides, say 
\[
b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, b_{10}, b_{11}, b_{12}, b_{13}
\]
such that the trinucleotides 
\[
b_2 b_3 b_4, b_2 b_5 b_6, b_2 b_7 b_8, b_2 b_9 b_{10}, b_2 b_{11} b_{12}, b_2 b_{13} \in X
\]
and 
\[
b_1 b_2 b_3, b_1 b_4 b_5, b_1 b_6 b_7, b_1 b_8 b_9, b_1 b_{10} b_{11}, b_1 b_{12} b_{13} \in X.
\]
Now, consider the sequence 
\[
C(b_1), C(b_2), C(b_3), C(b_4), C(b_5), C(b_6), C(b_7), C(b_8), C(b_9), C(b_{10}), C(b_{11}), C(b_{12}), C(b_{13})
\]
All the following trinucleotides belong to \( C(X) \):
\[
C(b_2)C(b_3)b_4, C(b_2)C(b_4)b_5, C(b_2)C(b_5)b_6, C(b_2)C(b_6)b_7, C(b_2)C(b_7)b_8, C(b_2)C(b_8)b_9, C(b_2)C(b_9)b_{10}, C(b_2)C(b_{10})b_{11}, C(b_2)C(b_{11})b_{12}, C(b_2)C(b_{12})b_{13}
\]
and 
\[
C(b_1)C(b_2)b_3, C(b_1)C(b_3)b_4, C(b_1)C(b_4)b_5, C(b_1)C(b_5)b_6, C(b_1)C(b_6)b_7, C(b_1)C(b_7)b_8, C(b_1)C(b_8)b_9, C(b_1)C(b_9)b_{10}, C(b_1)C(b_{10})b_{11}, C(b_1)C(b_{11})b_{12}, C(b_1)C(b_{12})b_{13}
\]
as they are the complement of trinucleotides in \( X \). So, \( C(X) \) admits a SLDCN and it cannot be a circular code. Contradiction.

The case \( X \) is a circular code and \( C(X) \) is not a circular code is similar. □

**Proposition 6.** Let \( S \) be a self-complementary subset of \( A_4^6 \setminus T \). If \( S \) is partitioned into three classes such that two of them are the complement of each other then necessarily the third one is self-complementary.

**Proof.** Let \( X_1, Y \) and \( Z \) be the three classes of an arbitrary partition of \( S \) and suppose that \( Y \) and \( Z \) are complementary, i.e. \( Y \) and \( Z \) satisfy \( C(Y) = Z \). Let \( t \) be a trinucleotide of \( X \). We claim that \( C(t) \notin Y \). Indeed, in the opposite case, \( Z \) should not be the complement of \( Y \) because \( t \in X \). We also claim that \( C(t) \notin Z \). Indeed, in the opposite case, \( Y \) should not be the complement of \( Z \) because \( t \in X \). It remains the case \( C(t) \in X \). So, \( X \) is self-complementary. □

**Remark 1.** Clearly, if \( X, Y \) and \( Z \) constitute an arbitrary partition of \( A_4^6 \setminus T \) then the self-complementarity of \( X \) is not enough to ensure that \( Y \) and \( Z \) are complementary of each other. This remark is again true if, in addition, \( X \) is a self-complementary circular code having 20 elements. Indeed in this case, it is easy to make a partition \( A_4^6 \setminus (X \cup T) \) in two classes \( Y \) and \( Z \) that are not complementary of each other. Any case, if we consider the partition of \( A_4^6 \setminus T \) in the three classes given by a self-complementary trinucleotide circular code \( X_0 \) having 20 elements and by its two conjugated classes \( X_1 \) and \( X_2 \) then the necessary and sufficient condition holds (Proposition 7 below).

**Proposition 7.** A trinucleotide circular code \( X_0 \) having 20 elements is self-complementary if and only if \( X_1 \) and \( X_2 \) are complements.

**Proof if part.** It is a trivial consequence of Proposition 6.

**Only if part.** Suppose that \( X_0 \) is self-complementary and consider the partition \( X_0 = X_1, X_2 \) of \( A_4^6 \setminus T \). Suppose that the trinucleotide, say \( l_2l_1l_3 \), belongs to \( X_0 \). Then, also 
\[
C(l_1)C(l_2)C(l_3) \in X_0.
\]
We have 
\[
l_2l_1l_3, C(l_2)C(l_3)C(l_1) \in X_1
\]
and 
\[
l_3l_1l_2, C(l_3)C(l_1)C(l_2) \in X_2.
\]
As \( l_2l_1l_3 \) is a generic trinucleotide of \( X_0 \) and as \( l_2l_1l_3 \) is the complement of \( C(l_1)C(l_2)C(l_3) \) and 
\[
C(l_2)C(l_3)C(l_1) \in X_1
\]
then \( X_1 \) is the complement of \( X_2 \). □

As a consequence, we have the following proposition.

**Proposition 8.** If a trinucleotide circular code \( X_0 \) having 20 elements is self-complementary then either

1) \( X_1 \) and \( X_2 \) are both circular codes or

2) \( X_1 \) and \( X_2 \) are not circular codes (both have a necklace).

**Proof.** We have four possibilities:

1) \( X_1 \) is a circular code and \( X_2 \) is a circular code;

2) \( X_1 \) is a circular code and \( X_2 \) is not a circular code;

3) \( X_1 \) is not a circular code and \( X_2 \) is a circular code;

4) \( X_1 \) is not a circular code and \( X_2 \) is not a circular code.

Now, by applying Propositions 3 and 4, we have that...
the first and the last possibilities can be effectively realized.

Suppose that, by way of contradiction, the second possibility is realized. So, \( X_1 \) is a circular code. By Proposition 7, we have \( C(X_1) = X_2 \). So, by Proposition 5, \( X_2 \) must also be a circular code. Contradiction.

Suppose that, by way of contradiction, the third possibility is realized. So, \( X_2 \) is a circular code. By Proposition 7, we have \( C(X_2) = X_1 \). So, by Proposition 5, \( X_1 \) must also be a circular code. Contradiction.

So, only the first and the last possibilities can occur. \( \Box \)

Hence, our proposition holds.

**Proposition 9.** The 528 self-complementary circular codes having 20 elements are partitioned into two classes: one class contains codes with the two permuted sets \( X_1 \) and \( X_2 \) which are both circular codes while the other class contains codes with the two permuted sets \( X_1 \) and \( X_2 \) which both are not circular codes.

**Proof.** It is enough to apply Proposition 8 to each of the 528 trinucleotide circular codes having 20 elements. \( \Box \)

4. Acknowledgements

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

**REFERENCES**


[22] G. Pirillo, “Non Sharing Border Codes,” Advances in
