International Journal of Pure and Applied Mathematics
Volume 76 No. 3 2012, 431-437
ISSN: 1311-8080 (printed version)
url: http://www.ijpam.eu

ijpam.eu

A NECESSARY AND SUFFICIENT CONDITION FOR A SET OF SEQUENCES ON THE GENETIC ALPHABET TO BE A CIRCULAR CODE

Giuseppe Pirillo

Consiglio Nazionale delle Ricerche
Dipartimento di Matematica "U.Dini"
Università di Firenze
viale Morgagni 67/A, 50134, Firenze, ITALIA
and
Université de Marne-la-Vallée
5, Boulevard Descartes
Champs sur Marne
77454, Marne-la-Vallée Cedex 2, FRANCE

Abstract: The maximal circular code discovered by Arquès and Michel contains 20 trinucleotides, i.e. sequences of length three on the genetic alphabet. Necessary and sufficient conditions for a set of trinucleotides to be a circular code was stated and proved in a previous paper. Here we present an analogous result for sets of sequences of length four on the genetic alphabet. More precisely we prove that these sets are circular codes if and only if they have no unbalanced 5-necklaces and no balanced 9-necklaces.

*

Genetic Alphabets and Genetic Sequences. The *letters* (or *nucleotides* or *bases*) of the *genetic alphabet* β_4 are A, C, G, T.

The set of *non-empty sequences* (resp. *sequences*) on β is denoted by β^+ (resp. β^*). The set of the 16 sequences of lenght two (or *binucleotides* or *domi-*

Received: February 10, 2012

noes) is denoted by β_4^2. The set of the 64 sequences of lenght three (or *trinucleotides*) over β_4 is denoted by β_4^3. The set of the 256 sequences of lenght four over β_4 is denoted by β_4^4. For the theory of variable lenght codes we refer to [4], where one can find the following two definitions.

Code. *A subset (or language) X in β_4^+ is a* code *if, for each $n, m \geq 1$ and for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m$ in X, the condition*

$$x_1 \cdots x_n = x'_1 \cdots x'_m$$

implies $n = m$, and, for $i = 1, \ldots, n$,

$$x_i = x'_i.$$

Circular Code. *A language X in β_4^+ is a* circular code *if, for each $n, m \geq 1$ and for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m$ in X, $p \in \beta_4^*$ and $s \in \beta_4^+$, the conditions*

$$sx_2 \cdots x_n p = x'_1 \cdots x'_m$$

and

$$x_1 = ps$$

imply $n = m$, $p = \epsilon$, the empty sequence*, and, for $i = 1, \ldots, n$,*

$$x_i = x'_i.$$

The sets β_4^2, β_4^3 and β_4^4 are codes (more precisely, *uniform codes*, see [4]) but they are not circular codes.

Conjugate Sequences. *Two sequences u and v are conjugate if there exist two sequences u' and u'' such that $u = u'u''$ and $v = u''u'$.*

The following two propositions are well known [4].

Proposition 1. *A circular code cannot contain a sequence of the form u^n, where u is a non-empty sequence and $n \geq 2$.*

Proposition 2. *A circular code cannot contain two distinct conjugate sequences.*

The maximal circular code discovered by Arquès and Michel consists of the following 20 trinucleotides:

$AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC,$
$GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC.$

and has remarquable properties, see [1, 2, 5, 6]. Arquès and Michel discovered other 215 codes similar to above mentioned one, see [3].

Necklaces. In [7] we introduced the following

Definition. Let $l_1, l_2, l_3, l_4, \ldots, l_n, l_{n+1}$ be letters in β_4, and let $d_1, d_2, d_3, \ldots, d_n$ be dominoes in β_4^2. We say that the ordered sequence

$$l_1, d_1, l_2, d_2, l_3, d_3, \ldots, l_n, d_n, l_{n+1}$$

is a $(n+1)$-necklace for a subset $X \subset \beta_4^3$ if

$$l_1d_1, l_2d_2, l_3d_3, \ldots, l_nd_n$$

and

$$d_1l_2, d_2l_3, d_3l_4, \ldots, d_nl_{n+1} \in X.$$

Again in [7] we proved the following

Proposition 3. *Let X be a subset of β_4^3. The following conditions are equivalent:*

a) X is circular code;

b) X has no 5-necklace.

In this paper we are interested to subsets of β_4^4 which are circular codes. We prove a necessary and sufficient condition which can help in a investigation of these codes using computer. We do not pretend that our result is optimal, but only that it reduces the number of cases to examine to a reasonable one.

Definition. *Let $l_1, l_2, l_3, l_4, \ldots, l_n, l_{n+1}$ be letters in β_4, and let $t_1, t_2, t_3, \ldots, t_n$ be trinucleotides in β_4^3. We say that the ordered sequence*

$$l_1, t_1, l_2, t_2, l_3, t_3, \ldots, l_n, t_n, l_{n+1}$$

is a unbalanced $(n+1)$-necklace *for a subset $X \subset \beta_4^4$ if*

$$l_1t_1, l_2t_2, l_3t_3, \ldots, l_nt_n \in X$$

and

$$t_1l_2, t_2l_3, t_3l_4, \ldots, t_nl_{n+1} \in X.$$

Definition. We say that the ordered sequence of dominoes in β_4^2

$$d_1, d_2, d_3, \ldots, d_{2n}, d_{2n+1}$$

is a balanced $(n+1)$-necklace for a subset $X \subset \beta_4^4$ if

$$d_1d_2, d_3d_4, \ldots, d_{2n-1}d_{2n} \in X$$

and

$$d_2d_3, d_4d_5, \ldots, d_{2n}d_{2n+1} \in X.$$

Example 1. The sequence $C, ATC, G, GAT, A, GGC, C$ is a unbalanced 4-necklace for $X = \{CATC, ATCG, GGAT, GATA, AGGC, GGCC\}$. Note that X is not a circular code because

$$C(ATCG)(GATA)GGC = (CATC)(GGAT)(AGGC).$$

Example 2. The sequence $AC, TC, GG, AT, AG, GC, TT, AA, AC$ is a balanced 5-necklace for

$$Y = \{ACTC, GGAT, AGGC, TTAA, TCGG, ATAG, GCTT, AAAC\}.$$

Note that Y is not a circular code because

$$TC(GGAT)(AGGC)(TTAA)AC = (TCGG)(ATAG)(GCTT)(AAAC).$$

Example 3. The sequence $AC, TC, GG, AT, AG, GC, TT, AC, AA$ is a balanced 5-necklace for

$$Z = \{ACTC, GGAT, AGGC, TTAC, TCGG, ATAG, GCTT, ACAA\}.$$

Note that Z is not a circular code because

$$\begin{aligned} &TC(GGAT)(AGGC)(TTAC)(TCGG)(ATAG)(GCTT)AC \\ &\quad = (TCGG)(ATAG)(GCTT)(ACTC)(GGAT)(AGGC)(TTAC). \end{aligned}$$

Remark. In the proof of the following proposition the arguments of part **a)** $\rightarrow$ **b)** are similar to those of Examples 1, 2 and 3.

As for Proposition 3 concerning trinucleotides, see [7], the following one does not require any hypothesis of auto complementarity.

Proposition. *Let X be a subset of β_4^4. The following condition are equivalent:*

a) X is circular code;

b) X has no unbalanced 5-necklace and no balanced 9-necklace.

Proof. **a)** $\rightarrow$ **b).** Let X be a circular code. We have to prove: **i)** X has no unbalanced 5-necklaces and **ii)** X has no balanced 9-necklaces.

i) By way of contradiction suppose that

$$l_1, t_1, l_2, t_2, l_3, t_3, l_4, t_4, l_5$$

is a unbalanced 5-necklace for X. As β_4 contains four letters, for some $i, j \in \{1, 2, 3, 4, 5\}$, $i < j$, we have

$$l_i = l_j.$$

Put $x_i = l_i t_i, \ldots, x_{j-1} = l_{j-1} t_{j-1}$ and $x'_i = t_i l_{i+1}, \ldots, x'_{j-1} = t_{j-1} l_j = t_{j-1} l_i$.

If $j = i + 1$ then $l_i t_i$ and $t_i l_i$ are both in X and are conjugate. This contradicts the above Proposition 2.

So assume that $j > i + 1$. We have

$$t_i x_{i+1} \cdots x_{j-1} l_i = x'_i \cdots x'_{j-1}.$$

Since $x_i = l_i t_i$ and l_i, being a letter, is non-empty, X is not a circular code. This contradicts the definition of circular code.

ii) By way of contradiction suppose that

$$d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{14}, d_{15}, d_{16}, d_{17}$$

is a balanced 9-necklace for X.

As β_4 is a four letter alphabet, for some $i, j \in \{1, 2, \ldots, 17\}$, $i < j$, we have

$$d_i = d_j.$$

If $j = i+1$ then $d_i d_{i+1} = (d_i)^2$ is in X. This contradicts the above Proposition 1. If $j = i+2$ then $d_i d_{i+1}$ and $d_{i+1} d_i$ are both in X and are conjugate. This contradicts the above Proposition 2. So suppose that $j - i \neq 1$ and $j - i \neq 2$.

If $j - i$ is odd (and consequently $j \geq i + 3$), put $x_i = d_i d_{i+1}, \ldots, x_{j-1} = d_{j-1} d_j = d_{j-1} d_i$ and $x'_{i+1} = d_{i+1} d_{i+2}, \ldots, x'_{j-2} = d_{j-2} d_{j-1}$ and consider the following equality

$$d_{i+1} x_{i+2} \cdots x_{j-1} x'_{i+1} \cdots x'_{j-2} d_i = x'_{i+1} \cdots x'_{j-2} x_i \cdots x_{j-1}.$$

If $j - i$ is even (and consequently $j \geq i + 2$), put $x_i = d_i d_{i+1}, \ldots, x_{j-2} = d_{j-2} d_{j-1}$ and $x'_{i+1} = d_{i+1} d_{i+2}, \ldots, x'_{j-1} = d_{j-1} d_j = d_{j-1} d_i$ and consider the following equality

$$d_{i+1} x_{i+2} \cdots x_{j-2} d_i = x'_{i+1} \cdots x'_{j-1}.$$

In both case, since $x_i = d_i d_{i+1}$ and d_i, being a domino, is non-empty, X is not a circular code. This contradicts the definition of circular code.

b) → a) Let X be without unbalanced 5-necklace and without balanced 9-necklace and, by way of contradiction, suppose that X is not a circular code.

As all the elements of X have length four, there exists $n \geq 1$, $x_1, \ldots, x_n, x'_1, \ldots, x'_n$ in X, $p_1, s_1 \in \beta_4^+$, such that

$$s_1 x_2 \cdots x_n p_1 = x'_1 \cdots x'_n$$

and that

$$x_1 = p_1 s_1.$$

Moreover, there exists $p_2, \ldots, p_n \in \beta_4 \cup \beta_4^2 \cup \beta_4^3$ and $s_2, \ldots, s_n \in \beta_4 \cup \beta_4^2 \cup \beta_4^3$, such that

$$|p_1| = |p_2| = \cdots = |p_n|,$$
$$|s_1| = |s_2| = \cdots = |s_n|,$$
$$|p_1| + |s_1| = |p_2| + |s_2| = \cdots = |p_n| + |s_n| = 4,$$

$$p_2s_2 = x_2, \ldots, p_{n-1}s_{n-1} = x_{n-1}, p_ns_n = x_n,$$

and that

$$s_1p_2 = x'_1, s_2p_3 = x'_2, \ldots, s_{n-1}p_n = x'_{n-1}, s_np_1 = x'_n.$$

Since we can exchange the x_i's with the x'_i's, without loss of generality, we can consider only two cases: **i)** $|p_1| = 1$ and **ii)** $|p_1| = 2$. In case **i)** we have

$$|p_1| = |p_2| = \cdots = |p_n| = 1$$

and consequently

$$|s_1| = |s_2| = \cdots = |s_n| = 3$$

and in case **ii)** we have

$$|p_1| = |p_2| = \cdots = |p_n| = 2$$

and consequently

$$|s_1| = |s_2| = \cdots = |s_n| = 2$$

In analogy with the proof of Proposition 3, see [7], for each value of the positive integer n, we find in case **i)** a unbalanced 5-necklace and in case **ii)** a balanced 9-necklace. So, in any case we have a contradiction.

The previous result concerns subsets of β_4^4. Similar results hold for subsets of β_4^m, $m \geq 2$.

Acknowledgment

The author thanks the Dipartimento di matematica "U.Dini" for giving him a friendly hospitality.

References

[1] D.G. Arquès, C.J. Michel, A possible code in the genetic code, In: *STACS 95 - 12th Annual Symposium on Theoretical Aspects of Computer Science, Munich, Germany, March 2 - 4, 1995 Proceedings* (Ed-s: Ernst W. Mayr Claude Puech), Lecture Notes in Computer Science, Springer Verlag, Volume 900 (1995), 640-651.

[2] D.G. Arquès, C.J. Michel, A complementary circular code in the protein coding genes, *Journal of Theoretical Biology*, **182** (1996), 45-58.

[3] D.G. Arquès, C.J. Michel, *Personal communication.*

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

[5] G. Pirillo, *Maximal Circular Codes and Applications to Theoretical Biology*, Mathematical and computational biology (Aizu-Wakamatsu City, 1997) 187-190, Lectures Math. Life Sci., **26**, Amer. Math. Soc., Providence, RI (1999).

[6] G. Pirillo, Remarks on the Arquès-Michel code, *Biology Forum*, **94** (2001), 327-330.

[7] G. Pirillo, *A characterization for a set of trinucleotides to be a circular code*, To Appear.