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

C. J. Michel, G. Pirillo

STRONG TRINUCLEOTIDE CIRCULAR CODES

R. 11-05 2011

C. J. Michel – Equipe de Bioinformatique Théorique, FDBT, LSIIT (UMR UdS-CNRS 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France
`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, Italy and Université de Marne-la-Vallée, 5 boulevard Descartes, 77454 Marne-la-Vallée Cedex 2, France `pirillo@math.unifi.it`..

ISSN: 1128–3378

Collana dei Rapporti
Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti"
Consiglio Nazionale delle Ricerche

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

Recently, we identified a hierarchy relation between trinucleotide comma-free codes and trinucleotide circular codes (Proposition 3 in [18]). Here, we extend our hierarchy with two new classes of codes, called DLD and LDL codes, which are stronger than the comma-free codes. We also prove that no circular code with 20 trinucleotides is a DLD code and that a circular code with 20 trinucleotides is comma-free if and only if it is a LDL code. Finally, we point out the possible role of the symmetric group Σ_4 in the mathematical study of trinucleotide circular codes.

Key words: Circular code, strong circular code, comma-free code, trinucleotide.

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 (Remark 1 and [4, 14]). 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. Circular codes only belong to some subsets of the 64 trinucleotide set while comma-free codes are even more constrained subsets.

In the past 50 years, comma-free codes and circular codes have been studied in theoretical biology, mainly to understand the structure and the origin of the genetic code as well as the reading frame (construction) of genes, e.g. [5, 8, 9]. Before the discovery of the genetic code, Crick *et al.* [5] proposed a (maximal) comma-free code of 20 trinucleotides for coding the 20 amino acids. In 1996, a (maximal) circular code X_0 of 20 trinucleotides was identified statistically on two large and different gene populations, eukaryotes and prokaryotes [1]. During the last years, circular codes are mathematical objects studied in discrete mathematics, theoretical computer science and theoretical biology, e.g. [12, 3, 2, 24, 10, 6, 19, 16, 7, 15, 20, 17, 21, 11, 22, 23]. In particular, in theory of codes, there are some unexpected common notions between variable length circular codes and trinucleotide circular codes [20, 21, 22, 23].

Recently, we proposed a hierarchy relation between the trinucleotide comma-free codes and the trinucleotide circular codes (Proposition 3 in [18]). More precisely, all the trinucleotide codes in this hierarchy are circular, the strongest ones being comma-free. In this paper, we identify two new classes of trinucleotide circular codes which are stronger than the comma-free codes.

We introduce here the following new notions. A set X of trinucleotides has the property DLD if for any trinucleotides $t, t' \in X$, no letter occurs both as a proper suffix of t and a proper prefix of t'. A set X of trinucleotides has the property LDL if for any trinucleotides $t, t' \in X$, no diletter occurs both as a proper suffix of t and a proper prefix of t'. These sets DLD and LDL are not only trinucleotides circular codes but they are also stronger than the comma-free codes (Propositions 3.3 and 3.4, and Remarks 5 and 6). We also prove that no circular code with 20 trinucleotides is a DLD code (Proposition 3.7) and that a circular code with 20 trinucleotides is comma-free if and only if it is a LDL code (Proposition 3.8).

Therefore, our previous hierarchy (Proposition 3 in [18] recalled in Proposition 4 below) is extended with these new DLD and LDL classes of strong trinucleotides circular codes (Proposition 4.1).

Finally, a curious relation with the symmetric group Σ_4 appears again. The tables given here and the other symmetric relations identified previously (e.g. Proposition 6 in [18]) suggest us that the symmetric group Σ_4 can play an important role in the mathematical study of these trinucleotide circular codes. However, we have no formal mathematical explanation so far.

2. Preliminaries

Let $\mathcal{A}$ denote a finite alphabet, $\mathcal{A}^*$ the free monoid over $\mathcal{A}$ and $\mathcal{A}^+$ the free semigroup over $\mathcal{A}$. The elements of $\mathcal{A}^*$ are words and the empty word, denoted by ε, is the identity of $\mathcal{A}^*$. Given a subset X of $\mathcal{A}^*$, X^n is the set of the words over $\mathcal{A}$ which are the products of n words from X, i.e. $X^n = \{x_1 x_2 \cdots x_n \mid x_i \in X\}$. If X is a (finite) set, then $|X|$ denotes its cardinality and if u is a word, then $|u|$ denotes its length. A word u is a factor of a word v if there exist two words u' and u'' such that $v = u'uu''$. When $u' = \varepsilon$ (resp. $u'' = \varepsilon$), u is a prefix (resp. suffix) of v. A proper factor (resp. proper prefix, proper suffix) u of v is a factor (resp. prefix, suffix) u of v such that $|u| < |v|$.

There is a correspondence between the genetic and language-theoretic concepts. The letters (or nucleotides or bases) define the genetic alphabet $\mathcal{A}_4 = \{A, C, G, T\}$. The set of non-empty words (resp. words) over $\mathcal{A}_4$ is denoted by $\mathcal{A}_4^+$ (resp. $\mathcal{A}_4^*$). The set of the 16 words of length 2 (or dinucleotides or diletters) is denoted by $\mathcal{A}_4^2$. The set of the 64 words of length 3 (or trinucleotides or triletters) is denoted by $\mathcal{A}_4^3$. The total order over the alphabet $\mathcal{A}_4$ is $A < C < G < T$. Consequently, $\mathcal{A}_4^+$ is lexicographically ordered: given two words $u, v \in \mathcal{A}_4^+$, u is smaller than v in lexicographical order, written $u < v$, if and only if either u is a proper prefix of v or there exist $x, y \in \mathcal{A}_4$, $x < y$, and $r, s, t \in \mathcal{A}_4^*$ such that $u = rxs$ and $v = ryt$.

Definition 2.1. *Code: A subset X of $\mathcal{A}^+$ is a* code *over $\mathcal{A}$ 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 $i = 1, \ldots, n$.*

For any k-letter alphabet, $k \geq 1$, and for any word length l, $l \geq 1$, $\mathcal{A}_k^l$ is a code. In particular, $\mathcal{A}_4^3$ is a code. More precisely, it is a uniform code [4]. Consequently, any non-empty subset of $\mathcal{A}_4^3$ is a code, called *trinucleotide code* in this paper.

Definition 2.2. *Trinucleotide comma-free code: A trinucleotide code $X \subset \mathcal{A}_4^3$ is* comma-free *if for each $y \in X$ and $u, v \in \mathcal{A}_4^*$ such that $uyv = x_1 \cdots x_n$ with $x_1, \ldots, x_n \in X$, $n \geq 1$, it results that $u, v \in X^*$.*

Several varieties of trinucleotide comma-free codes were described in [17].

Definition 2.3. *Trinucleotide circular code: A trinucleotide code $X \subset \mathcal{A}_4^3$ 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 $sx_2 \cdots x_n p = x'_1 \cdots x'_m$ and $x_1 = ps$ imply $n = m$, $p = \varepsilon$ and $x_i = x'_i$ for $i = 1, \ldots, n$.*

Remark 1. *$\mathcal{A}_4^3$ is obviously not a circular code and even less a comma-free code. However, several subsets of $\mathcal{A}_4^3$ are trinucleotide circular codes (e.g. Propositions 1 and 2).*

Definition 2.4. *Maximal trinucleotide circular code: A trinucleotide circular code $X \subset \mathcal{A}_4^3$ is* maximal *if for each $x \in \mathcal{A}_4^3$, $x \notin X$, $X \cup \{x\}$ is not a trinucleotide circular code.*

Definition 2.5. *A trinucleotide circular code containing exactly k elements is called a k-*trinucleotide circular code.

Remark 2. *A 20-trinucleotide circular code is*
- maximal (in the sense that it cannot be contained in a trinucleotide circular code with more words);
- maximum (in the sense that no trinucleotide circular code can contain more than 20 words).

We now recall some definitions and previous results related to the trinucleotide circular code necklaces. In the sequel, $l_1, l_2, \ldots, l_n$ are letters in $\mathcal{A}_4$, $d_1, d_2, \ldots, d_n$ are diletters in $\mathcal{A}_4^2$ and n is an integer satisfying $n \geq 2$.

Definition 2.6. *Letter Diletter Necklaces (LDN): We say that the ordered sequence $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n$* is an *$nLDN$* for a subset *$X \subset \mathcal{A}_4^3$ if $l_1d_1, l_2d_2, \ldots, l_nd_n \in X$ and $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n \in X$.*

Definition 2.7. *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 $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n, d_nl_{n+1} \in X$.*

Definition 2.8. *Diletter Letter Necklaces (DLN): We say that the ordered sequence $d_1, l_1, d_2, l_2, \ldots, l_{n-1}, d_n, l_n$* is an *$nDLN$* for a subset *$X \subset \mathcal{A}_4^3$ if $d_1l_1, d_2l_2, \ldots, d_nl_n \in X$ and $l_1d_2, l_2d_3, \ldots, l_{n-1}d_n \in X$.*

Definition 2.9. *Diletter Letter Continued Necklaces (DLCN): We say that the ordered sequence $d_1, l_1, d_2, l_2, \ldots, l_{n-1}, d_n, l_n, d_{n+1}$* is an *$(n+1)DLCN$* for a subset *$X \subset \mathcal{A}_4^3$ if $d_1l_1, d_2l_2, \ldots, d_nl_n \in X$ and $l_1d_2, l_2d_3, \ldots, l_{n-1}d_n, l_nd_{n+1} \in X$.*

Proposition 2.10. *[20]. Let X be a trinucleotide code. The following conditions are equivalent.*
(i) X is circular code.
(ii) X has no $5LDCN$.

Proposition 2.11. *[17]. Let X be a trinucleotide code. The following conditions are equivalent.*
(i) X is a comma-free code.
(ii) X has no $2LDN$ and no $2DLN$.

Definition 2.12. *Let X be a trinucleotide code. For any integer $n \in \{2,3,4,5\}$, we say that X* belongs to the class C^{nLDN} *if X has no $nLDN$ and that X* belongs to the class C^{nDLN} *if X has no $nDLN$. Similarly, for any integer $n \in \{3,4,5\}$, we say that X* belongs to the class C^{nLDCN} *if X has no $nLDCN$ and that X* belongs to the class C^{nDLCN} *if X has no $nDLCN$.*

Notation 1. *For any integer $n \in \{2,3,4,5\}$, $I^n = C^{nLDN} \cap C^{nDLN}$ and $U^n = C^{nLDN} \cup C^{nDLN}$. Similarly, for any integer $n \in \{3,4,5\}$, $I^nC = C^{nLDCN} \cap C^{nDLCN}$ and $U^nC = C^{nLDCN} \cup C^{nDLCN}$.*

Proposition 2.13. *[18]. The following chains of inclusions hold.*
(i) $C^{2LDN} \subset C^{3LDCN} \subset C^{3LDN} \subset C^{4LDCN} \subset C^{4LDN} \subset C^{5LDCN} \subset C^{5LDN}$.
(ii) $C^{2DLN} \subset C^{3DLCN} \subset C^{3DLN} \subset C^{4DLCN} \subset C^{4DLN} \subset C^{5DLCN} \subset C^{5DLN}$.
(iii) $C^{2LDN} \subset C^{3DLCN} \subset C^{3LDN} \subset C^{4DLCN} \subset C^{4LDN} \subset C^{5DLCN} \subset C^{5LDN}$.
(iv) $C^{2DLN} \subset C^{3LDCN} \subset C^{3DLN} \subset C^{4LDCN} \subset C^{4DLN} \subset C^{5LDCN} \subset C^{5DLN}$.
(v) $I^2 \subset I^3C \subset I^3 \subset I^4C \subset I^4 \subset I^5C \subset I^5$.
(vi) $U^2 \subset U^3C \subset U^3 \subset U^4C \subset U^4 \subset U^5C \subset U^5$.

Remark 3. *By Proposition 2, the chain of inclusions of Proposition 3 (v) begins with I^2 which is the class of comma-free codes.*

Proposition 2.14. *With 20-trinucleotide circular codes, the following chains of inclusions and equalities hold*
$I^2 \subset U^2 = I^3C \subset U^3C = I^3 \subset U^3 = I^4C \subset U^4C = I^4 \subset U^4 = I^5C \subset U^5C = I^5 = U^5$.

3. Strong trinucleotide circular codes

We introduce new definitions which impose very strong conditions on the words of a subset of $\mathcal{A}_4^3$. These word subsets, strongly constrained, are indeed new circular codes which are stronger than the trinucleotide comma-free codes according to the following propositions.

Definition 3.1. *A subset X of $\mathcal{A}_4^3$* has the DLD property *if, for any $l_1, l_2, l_3, l'_1, l'_2, l'_3 \in \mathcal{A}_4$, the conditions $l_1l_2l_3 \in X$ and $l'_1l'_2l'_3 \in X$ imply $l_1 \neq l'_3$.*

No letter of $\mathcal{A}_4$ can occur in the first position of a trinucleotide of X when it is also in the last position of another trinucleotide of X.

Definition 3.2. *A subset X of $\mathcal{A}_4^3$* has the LDL property *if, for any $l_1, l'_1 \in \mathcal{A}_4$, $d_1, d'_1 \in \mathcal{A}_4^2$, the conditions $l_1d_1 \in X$, $d'_1l'_1 \in X$ imply $d_1 \neq d'_1$.*

No diletter of $\mathcal{A}_4^2$ can occur as a prefix of a trinucleotide of X when it is also a suffix of another trinucleotide of X.

Remark 4. *The trinucleotide code $\{ACG, GTA\}$ is not a DLD-strong trinucleotide circular code but it is a LDL-strong trinucleotide circular code. The trinucleotide code $\{ACG, CGT\}$ is not a LDL-strong trinucleotide circular code but it is a DLD-strong trinucleotide circular code.*

Therefore, the class of DLD-strong trinucleotide circular codes is different from the class of LDL-strong trinucleotide circular codes. However, both are very particular cases of comma-free codes according to the following propositions.

Proposition 3.3. *A DLD-strong trinucleotide circular code over $\mathcal{A}_4$ is comma-free.*

Proof. Suppose that X is a DLD-strong trinucleotide circular code X and by way of contradiction, that it is not comma-free. Then, there exist two trinucleotides $xyz, x'y'z' \in X$ such that either yzx' or $zx'y'$ are in X. In the first case, x' is a prefix of $x'y'z'$ and a suffix of yzx' while in the second case, z is a prefix of $zx'y'$ and a suffix of xyz. In both cases, X is not a DLD-strong circular code X. Contradiction. ∎

Proposition 3.4. *A LDL-strong trinucleotide circular code over $\mathcal{A}_4$ is comma-free.*

Proof. Suppose that X is a LDL-strong trinucleotide circular code X and by way of contradiction, that it is not comma-free. Then, there exist two trinucleotides $xyz, x'y'z' \in X$ such that either yzx' or $zx'y'$ are in X. In the first case, yz is a prefix of yzx' and a suffix of xyz while in the second case, $x'y'$ is a prefix of $x'y'z'$ and a suffix of $zx'y'$. In both cases, X is not a LDL-strong circular code X. Contradiction. ∎

Remark 5. *There are trinucleotide comma-free codes which are not DLD-strong trinucleotide circular codes. Example: $\{ACA\}$.*

Remark 6. *There are trinucleotide comma-free codes which are not LDL-strong trinucleotide circular codes. Example: $\{ACG, CGT\}$.*

The two following propositions are obvious.

Proposition 3.5. *For any letters $x, y, z \in \mathcal{A}_4$, a trinucleotide singleton $xyz \in \mathcal{A}_4^3$ is a DLD-strong trinucleotide circular code over $\mathcal{A}_4$ if and only if $x \neq z$.*

Proposition 3.6. *For any letters $x, y, z \in \mathcal{A}_4$, a trinucleotide singleton $xyz \in \mathcal{A}_4^3$ is a LDL-strong trinucleotide circular code over $\mathcal{A}_4$ if and only if at least two of its letters are different.*

Remark 4 showed that DLD-strong and LDL-strong trinucleotide circular codes are different classes. The following propositions give more information about their difference.

Proposition 3.7. *No 20-trinucleotide circular code can be a DLD-strong trinucleotide circular code.*

Proof. Suppose, by way of contradiction, that a 20-trinucleotide circular code X is also a DLD-strong trinucleotide circular code. Let P (resp. S) be the set containing the letters l_1 (resp. l_3) of the trinucleotides $l_1l_2l_3$ of X. We have: $P \cap S = \emptyset$ (otherwise, X has not the DLD property), $|P| > 1$ (otherwise, X has at most 16 elements) and $|S| > 1$ (otherwise, X has at most 16 elements). Using Pigeon Hole Principle, it follows that $\mathcal{A}_4$ has two disjoint subsets, say $\{a, b\}$ and $\{c, d\}$, such that $P = \{a, b\}$ and $S = \{c, d\}$. Consequently, X has at most the following elements: aAc, aCc, aGc, aTc, aAd, aCd, aGd, aTd, bAc, bCc, bGc, bTc, bAd, bCd, bGd, bTd, i.e. again at most 16 elements. Contradiction. ∎

Proposition 3.8. *A 20-trinucleotide circular code is comma-free if and only if it is a LDL-strong trinucleotide circular code.*

Proof. **If**. By Proposition 3.4, any LDL-strong trinucleotide circular code X is also comma-free.
Only if. Suppose that X is comma-free and, by way of contradiction, that it is not a LDL-strong trinucleotide circular code X. Then, there exist two letters $a, b \in \mathcal{A}_4$ and a diletter $d_1 \in \mathcal{A}_4^2$ such that $ad_1, d_1b \in X$. As X cannot contain two elements in the same conjugation class, the condition $a \neq b$ holds. So, $\mathcal{A}_4 - \{a, b\}$ contains exactly two elements, say c and d.
X being a comma-free code, X must contain exactly one trinucleotide in each of the 20 conjugation classes. By considering the conjugation class $\{aac, aca, caa\}$, only aac can belong to X. Indeed, $\{aca, ad_1, d_1b\}$ and $\{caa, ad_1, d_1b\}$ are not comma-free codes as the concatenations $aca.d_1b$ and $caa.d_1b$ lead to ad_1 in contradiction with Definition 2. With the conjugation class $\{bbc, bcb, cbb\}$, only cbb can belong to X. Indeed, $\{bbc, ad_1, d_1b\}$ and $\{bcb, ad_1, d_1b\}$ are not comma-free codes as the concatenations $ad_1.bbc$ and $ad_1.bcb$ lead to d_1b in contradiction with Definition 2. Similarly, aad and dbb must belong to X. Moreover, with the conjugation class $\{acb, cba, bac\}$, only acb can belong to X.
Now, we have:
- $acd \notin X$ (otherwise $\{aac, acd, dbb\}$ is not a comma-free code);
- $cda \notin X$ (otherwise $\{cda, acb, cbb\}$ is not a comma-free code);
- $dac \notin X$ (otherwise $\{aad, dac, acb\}$ is not a comma-free code).
So, no element in the conjugation class $\{acd, cda, dac\}$ belongs to X. Contradiction. ∎

4. Extended hierarchy

The previous hierarchy of trinucleotide circular codes [18] is now extended with these new DLD and LDL codes. By Proposition 3.7, the set of DLD-strong 20-trinucleotide circular codes is empty. Moreover, by Proposition 3.8, the set of LDL-strong 20-trinucleotide circular codes coincide with the set of trinucleotide comma-free codes (set I^2). With the notations I^n and U^n (Notation 1), the hierarchy of the above recalled Proposition 2.14 is extended with these new strong trinucleotide circular codes as follows.

Proposition 4.1. *With the* 20*-trinucleotide circular codes, the following chains of inclusions and equalities hold*
$\emptyset = LDL \cap DLD \subset LDL \cup DLD = LDL = I^2 \subset U^2 = I^3C \subset U^3C = I^3 \subset U^3 = I^4C \subset U^4C = I^4 \subset U^4 = I^5C \subset U^5C = I^5 = U^5$.

5. Coding of trinucleotide circular codes with the symmetric group Σ_n

We use the symmetric group Σ_n (e.g. the review [13]) to develop a coding of trinucleotide circular codes.

A *permutation* of a set X is a bijection σ from X into itself. Given a positive integer n, $[n]$ denotes the set $\{0, 1, \ldots, n-1\}$. As $[n]$ has a natural total order $0 < 1 < \cdots < n-1$, a permutation σ of $[n]$ is the word $\sigma(0)\sigma(1)\ldots\sigma(n-1)$ giving the successive images of the elements of $[n]$. Analogously, $\{a_{[n]}\}$ denotes any totally ordered set $\{a_0, a_1, \ldots, a_{n-1}\}$, $a_0 < a_1 < \cdots < a_{n-1}$, of n elements. Also as a consequence of the total order, a permutation σ of $\{a_{[n]}\}$ is the word $a_{\sigma(0)}a_{\sigma(1)}\ldots a_{\sigma(n-1)}$ and by abuse of language, σ can also be considered as a permutation of $[n]$. The symmetric group Σ_n denotes all the permutations of $\{a_{[n]}\}$.

Recall that $|X|$ denotes the number of elements of a set X. Recall that if $w = w(0)w(1)\ldots\ w(k-1)$ is a word of length k on the alphabet $\mathcal{A}$, then $Alph(w) = \{w(0), w(1), ..., w(k-1)\}$. So, $Alph(w)$ is the set of the letters of $\mathcal{A}$ having at least one occurrence in w.

A permutation of $\{a_{[n]}\}$ can be represented by a word of length $n-1$. Clearly, the prefix of length $n-1$ of the word $a_{\sigma(0)}\ldots a_{\sigma(n-1)}$ uniquely determines σ. There are also four other cases to represent the elements of Σ_n by words of length $n-1$: $i < j$ and $\sigma(i) < \sigma(j)$; $i < j$ and $\sigma(i) > \sigma(j)$; $i > j$ and $\sigma(i) < \sigma(j)$; $i > j$ and $\sigma(i) > \sigma(j)$. We begin with the case $i < j$ and $\sigma(i) > \sigma(j)$.

For a given $h \in [n-1]$, $\{a_{[h]}\}$ denotes the subset of $[n-1]$ containing its first h elements $a_0, \ldots, a_{h-1}$. For a given $i \in [n]$ and for a permutation σ of $\{a_{[n]}\}$, the set r_i^σ is defined as follows: $r_i^\sigma = \{a_{[\sigma(i)]}\} \cap Alph(a_{\sigma(i+1)}\ldots a_{\sigma(n-1)})$ contains the elements of $\{a_{[\sigma(i)]}\} = \{a_0, \ldots, a_{\sigma(i)-1}\}$ having one occurrence in $a_{\sigma(i+1)}\ldots a_{\sigma(n-1)}$, the suffix of length $n-i-1$ of $a_{\sigma(0)}\ldots a_{\sigma(n-1)}$. Consequently, $|r_i^\sigma|$ counts the number of elements j of $[n]$ such that $i < j$ and $a_{\sigma(i)} > a_{\sigma(j)}$. In other words, $|r_i^\sigma|$ counts the number of elements a_k of $\{a_{[n]}\}$ such that $a_k < a_{\sigma(i)}$ and a_k is on the right of $a_{\sigma(i)}$ in the word $a_{\sigma(0)}\ldots a_{\sigma(n-1)}$. Put $r(i) = |r_i^\sigma|$ and let the *code* of σ be the word $r(0)r(1)\ldots r(n-1)$ denoted by $r(\sigma)$.

For a given permutation σ, $r(0)$ is the number of the letters of $a_{\sigma(0)}\ldots a_{\sigma(n-1)}$ that are strictly smaller than $a_{\sigma(0)}$ or equivalently, the number of the elements of the alphabet $\{a_{[n]}\}$ that are strictly smaller than the leftmost letter $a_{\sigma(0)}$, and by the choice of the alphabet, this number is exactly $\sigma(0)$ and belongs to $[n-0] = [n]$. Then, $r(1)$ is the number of the letters of $a_{\sigma(0)}\ldots a_{\sigma(n-1)}$ that are strictly smaller than $a_{\sigma(1)}$ and on the right of $a_{\sigma(1)}$ or equivalently, the number of the elements of the alphabet $\{a_{[n]}\} - \{\sigma(0)\}$ that are strictly smaller than $a_{\sigma(1)}$ and this number belongs to $[n-1]$. And so on until $r(n-2)$ which is the number of the letters of $a_{\sigma(0)}\ldots a_{\sigma(n-1)}$ that are strictly smaller than $a_{\sigma(n-2)}$ and on the right of $a_{\sigma(n-2)}$ or equivalently, the number of the elements of the two-letter alphabet $\{a_{[n]}\} - \{\sigma(0), \ldots, \sigma(n-3)\} = \{\sigma(n-2), \sigma(n-1)\}$ that are strictly smaller than $a_{\sigma(n-2)}$ and this number belongs to $[n-(n-2)] = [2]$, i.e. with only values 0 or 1. Finally, $r(n-1)$ is the number of the letters of $a_{\sigma(0)}\ldots a_{\sigma(n-1)}$ that are strictly smaller than $a_{\sigma(n-1)}$ and on the right of $a_{\sigma(n-1)}$ or equivalently, the number of the elements of the one-letter alphabet $\{a_{[n]}\} - \{\sigma(0), \ldots, \sigma(n-2)\} = \{\sigma(n-1)\}$ that are strictly smaller than $a_{\sigma(n-1)}$ and this number belongs to $[n-(n-1)] = [1]$, i.e. with value equal to 0. Thus, $r(0) \in [n], r(1) \in [n-1], \ldots, r(i) \in [n-i]$ and $r(0)r(1)\ldots r(n-1)$ belongs to a set of cardinality $n!$ which is exactly the cardinality of Σ_n.

Clearly, if σ and τ are two different permutations of $\{a_{[n]}\}$, then $r(\sigma) \neq r(\tau)$. Indeed, let k the maximum integer such that $a_{\sigma(k)} = a_{\tau(k)}$. Without loss of generality, suppose that $a_{\sigma(k+1)} < a_{\tau(k+1)}$. As $Alph(a_{\sigma(k+1)} \dots a_{\sigma(n-1)}) = Alph(a_{\tau(k+1)} \dots a_{\tau(n-1)})$, then $|r^{\sigma}_{k+1}| < |r^{\tau}_{k+1}|$. So, $r(\sigma)$ is different from $r(\tau)$.

Example 1. *The code of the permutation* $\sigma = a_4a_6a_2a_1a_3a_0a_5$ *of* $\{a_{[7]}\}$ *is* $r(\sigma) = 452110$.

The correspondence $\rho : \sigma \to r(\sigma)$ is an injective map between two finite sets of same cardinality ($n!$). So ρ is a bijection and to each $r(\sigma)$ corresponds a unique σ. The following algorithm allows the permutation σ from the code $r(\sigma)$ to be retrieved.

Algorithm 1 (principle) $a_{\sigma(0)} = a_{r(0)}$;
only one element, say a_α, *in* $\{a_{[n]}\} - \{a_{\sigma(0)}\}$ *can verify* $r(1) = \left|\{a_\zeta \in \{a_{[n]}\} - \{a_{\sigma(0)}\} \mid a_\alpha > a_\zeta\}\right|$, *so* $a_{\sigma(1)} = a_\alpha$;
only one element, say a_β, *in* $\{a_{[n]}\} - \{a_{\sigma(0)}, a_{\sigma(1)}\}$ *can verify*
$r(2) = \left|\{a_\zeta \in \{a_{[n]}\} - \{a_{\sigma(0)}, a_{\sigma(1)}\} \mid a_\beta > a_\zeta\}\right|$, *so* $a_{\sigma(2)} = a_\beta$;
repeat this procedure until all the elements $\{a_{\sigma(0)}, \dots, a_{\sigma(n-2)}\}$ *are found;*
finally $a_{\sigma(n-1)}$ *is the unique value in* $\{a_{[n]}\} - \{a_{\sigma(0)}, \dots, a_{\sigma(n-2)}\}$.

Remark 7. *In general,* $r(i+1) \dots r(n-1)$ *is the code of the permutation* $a_{\sigma(i+1)} \dots a_{\sigma(n-1)}$ *on the totally ordered alphabet* $\{a_{\sigma(i+1)}, \dots, a_{\sigma(n-1)}\}$.

Example 2. *Consider the previous example with the permutation* σ *of* $\{a_{[7]}\}$ *having the code* $r(\sigma) = 452110$. *As* $r(0) = 4$, *then* $a_{\sigma(0)} = a_4$; $a_{\sigma(1)} = a_6$ *as* $\{a_0, a_1, a_2, a_3, a_5, a_6\}$ *contains* $r(1) = 5$ *elements strictly smaller;* $a_{\sigma(2)} = a_2$ *as* $\{a_0, a_1, a_2, a_3, a_5\}$ *contains* $r(2) = 2$ *elements strictly smaller;* $a_{\sigma(3)} = a_1$ *as* $\{a_0, a_1, a_3, a_5\}$ *contains* $r(3) = 1$ *element strictly smaller;* $a_{\sigma(4)} = a_3$ *as* $\{a_0, a_3, a_5\}$ *contains* $r(4) = 1$ *element strictly smaller;* $a_\sigma(5) = a_0$ *as* $\{a_0, a_5\}$ *contains* $r(5) = 0$ *element strictly smaller; finally,* $a_{\sigma(6)} = a_5$ *as* $\{a_{[7]}\} - \{a_{\sigma(0)}, a_{\sigma(1)}, a_{\sigma(2)}, a_{\sigma(3)}, a_{\sigma(4)}, a_{\sigma(5)}\} = \{a_{[7]}\} - \{a_4, a_6, a_2, a_1, a_3, a_0\} = \{a_5\}$. *So, the permutation* σ *is* $a_4a_6a_2a_1a_3a_0a_5$.

For a given a permutation σ, we can also define the sets $l^\sigma_i = \{a_{[\sigma(i)]}\} \cap Alph\left(a_{\sigma(0)} \dots a_{\sigma(i)-1}\right)$, $R^\sigma_i = \left(\{a_{[n]}\} - \{a_{[\sigma(i)+1]}\}\right) \cap Alph\left(a_{\sigma(i+1)} \dots a_{\sigma(n-1)}\right)$ and $L^\sigma_i = \left(\{a_{[n]}\} - \{a_{[\sigma(i)+1]}\}\right) \cap Alph\left(a_{\sigma(0)} \dots a_{\sigma(n-1)}\right)$. The set l^σ_i consists of the elements of $\{a_{[\sigma(i)]}\} = \{a_{\sigma(0)}, \dots, a_{\sigma(i)-1}\}$ that have one occurrence in the prefix of length i of $a_{\sigma(0)} \dots a_{\sigma(n-1)}$. Its cardinality $|l^\sigma_i|$ counts the number of elements j of $[n]$ such that $j < i$ and $\sigma(j) < \sigma(i)$ or, in other words, $|l^\sigma_i|$ counts the number of elements a_k of $\{a_{[n]}\}$ such that $a_k < a_{\sigma(i)}$ and a_k is on the left of $a_{\sigma(i)}$ in $a_{\sigma(0)} \dots a_{\sigma(n-1)}$. Similarly, the set R^σ_i consists of the elements of $\{a_{[n]}\} - \{a_{[\sigma(i)+1]}\} = \{a_{\sigma(i)+1}, \dots, a_{n-1}\}$ that have one occurrence in $a_{\sigma(i+1)} \dots a_{\sigma(n-1)}$, the suffix of length $n - i - 1$ of $a_{\sigma(0)} \dots a_{\sigma(n-1)}$. Its cardinality $|R^\sigma_i|$ counts the number of elements j of $[n]$ such that $j > i$ and $\sigma(j) > \sigma(i)$ or, in other words, $|R^\sigma_i|$ counts the number of elements a_k of $\{a_{[n]}\}$ such that $a_k > a_{\sigma(i)}$ and a_k is on the right of $a_{\sigma(i)}$ in $a_{\sigma(0)} \dots a_{\sigma(n-1)}$. Finally, the set L^σ_i consists of the elements of $\{a_{[n]}\} - \{a_{[\sigma(i)+1]}\} = \{a_{\sigma(i)+1}, \dots, a_{n-1}\}$ that have one occurrence in the prefix of length i of $a_{\sigma(0)} \dots a_{\sigma(n-1)}$. Its cardinality $|L^\sigma_i|$ counts the number of elements j of $[n]$ such that $j < i$ and $\sigma(j) > \sigma(i)$ or, in other words, $|L^\sigma_i|$ counts the number of elements a_k of $\{a_{[n]}\}$ such that $a_k > a_{\sigma(i)}$ and a_k is on the left of $a_{\sigma(i)}$ in $a_{\sigma(0)} \dots a_{\sigma(n-1)}$.

There are trivial relations

$$\begin{aligned} l^\sigma_i + L^\sigma_i &= i, \\ r^\sigma_i + R^\sigma_i &= n - i - 1, \\ r^\sigma_i + l^\sigma_i &= \sigma(i), \\ R^\sigma_i + L^\sigma_i &= n - \sigma(i) - 1. \end{aligned}$$

For a given permutation, l^σ_i, R^σ_i and L^σ_i allow the construction of three other codes, namely $l(0)l(1) \dots l(n-1)$, $R(0)R(1) \dots R(n-1)$ and $L(0)L(1) \dots L(n-1)$, which have similar properties to the code $r(0)r(1) \dots r(n-1)$. These relations can retrieve more efficiently the permutation σ from the code $r(\sigma)$. For the interesting case $n = 4$ of this paper, an efficient algorithm is given.

Algorithm 2 (principle) $a_{\sigma(0)} = a_{r(0)}$.
Consider $\{\sigma(1), \sigma(2), \sigma(3)\}$ *and let* $\{\sigma(1), \sigma(2), \sigma(3)\} = \{\alpha, \beta, \gamma\}$ *with* $\alpha < \beta < \gamma$.
If $r(1) = 2$, *then* $a_{\sigma(1)} = a_\gamma$ *and, if* $r(2) = 1$, *then* $a_{\sigma(2)}a_{\sigma(3)} = a_\beta a_\alpha$ *or, if* $r(2) = 0$, *then* $a_{\sigma(2)}a_{\sigma(3)} = a_\alpha a_\beta$.
If $r(1) = 1$, *then* $a_{\sigma(1)} = a_\beta$ *and, if* $r(2) = 1$, *then* $a_{\sigma(2)}a_{\sigma(3)} = a_\gamma a_\alpha$ *or, if* $r(2) = 0$, *then* $a_{\sigma(2)}a_{\sigma(3)} = a_\alpha a_\gamma$.
If $r(1) = 0$, *then* $a_{\sigma(1)} = a_\alpha$ *and, if* $r(2) = 1$, *then* $a_{\sigma(2)}a_{\sigma(3)} = a_\gamma a_\beta$ *or, if* $r(2) = 0$, *then* $a_{\sigma(2)}a_{\sigma(3)} = a_\beta a_\gamma$.

The number $r(1)r(2)$ is the code of the permutation $a_{\sigma(\alpha)}a_{\sigma(\beta)}a_{\sigma(\gamma)}$ on $\{a_\alpha, a_\beta, a_\gamma\}$.

Example 3. *Consider the permutation* σ *of* $\{a_{[4]}\}$ *having* 111 *as its code. Clearly,* $a_{\sigma(0)} = a_1$. *Then, the considered set* $\{\sigma(1), \sigma(2), \sigma(3)\} = \{\alpha, \beta, \gamma\}$ *is* $\{a_0, a_2, a_3\}$. *As* $r(1) = 1$, *then* $a_{\sigma(1)} = a_\beta = a_2$ *and as* $r(2) = 1$, *then* $a_{\sigma(2)}a_{\sigma(3)} = a_\gamma a_\alpha = a_3 a_0$. *So, the permutation* σ *is* $a_1a_2a_3a_0$.

Finally, the code of a permutation $\sigma(A)\sigma(C)\sigma(G)\sigma(T)$ on the genetic alphabet $\mathcal{A}_4$ $(A < C < G < T)$ can easily be computed by putting $A = a_0$, $C = a_1$, $G = a_2$ and $T = a_3$. Similarly, for the totally ordered alphabet $\{a, b, c, d\}$ $(a < b < c < d)$ in Section 5, the code of a permutation is obtained by putting $a = a_0$, $b = a_1$, $c = a_2$ and $d = a_3$.

Permutation σ	Code $r(\sigma)$
a_0a_1	0
a_1a_0	1

Table 1a. The first column contains the permutations σ of the symmetric group Σ_2 on the alphabet $\{a_{[2]}\} = \{a_0, a_1\}$ and the second column contains their codes $r(\sigma)$.

Permutation σ	Code $r(\sigma)$
$a_0a_1a_2$	00
$a_0a_2a_1$	01
$a_1a_0a_2$	10
$a_1a_2a_0$	11
$a_2a_0a_1$	20
$a_2a_1a_0$	21

Table 1b. The first column contains the permutations σ of the symmetric group Σ_3 on the alphabet $\{a_{[3]}\} = \{a_0, a_1, a_2\}$ and the second column their codes $r(\sigma)$.

Permutation σ	Code $r(\sigma)$
$a_0a_1a_2a_3$	000
$a_0a_1a_3a_2$	001
$a_0a_2a_1a_3$	010
$a_0a_2a_3a_1$	011
$a_0a_3a_1a_2$	020
$a_0a_3a_2a_1$	021
$a_1a_0a_2a_3$	100
$a_1a_0a_3a_2$	101
$a_1a_2a_0a_3$	110
$a_1a_2a_3a_0$	111
$a_1a_3a_0a_2$	120
$a_1a_3a_2a_0$	121
$a_2a_0a_1a_3$	200
$a_2a_0a_3a_1$	201
$a_2a_1a_0a_3$	210
$a_2a_1a_3a_0$	211
$a_2a_3a_0a_1$	220
$a_2a_3a_1a_0$	221
$a_3a_0a_1a_2$	300
$a_3a_0a_2a_1$	301
$a_3a_1a_0a_2$	310
$a_3a_1a_2a_0$	311
$a_3a_2a_0a_1$	320
$a_3a_2a_1a_0$	321

Table 1c. The first column contains the permutations σ of the symmetric group Σ_4 on the alphabet $\{a_{[4]}\} = \{a_0, a_1, a_2, a_3\}$ and the second column contains their codes $r(\sigma)$. This table easily allows to determine the codes for permutations on any other totally ordered four-letter alphabet, in particular the alphabet $[4] = \{0, 1, 2, 3\}$ $(0 < 1 < 2 < 3)$, the genetic alphabet $\mathcal{A}_4$ $(A < C < G < T)$ and the alphabet $\{a, b, c, d\}$ $(a < b < c < d)$. For example, 211 is the code for 2130 on the alphabet [4], for $GCTA$ on the alphabet $\mathcal{A}_4$ and for $cbda$ on the alphabet $\{a, b, c, d\}$.

6. Role of the symmetric group Σ_4

We put $a = A$, $b = C$, $c = G$ and $d = T$ and identify the elements of the symmetric group Σ_4 over $\{a, b, c, d\}$ ($a < b < c < d$) with the 24 permutations of the word *abcd*. We denote the permutations by their codes (Table 1*c*).

We wish to point out that a computer calculus confirms that the 20-trinucleotide comma-free codes are exactly the *LDL*-strong 20-trinucleotide circular codes. These codes are partitioned into 28 classes: $C_1, C_2, \ldots, C_{28}$. There are four classes containing six codes each (Table 2*a*), 16 classes containing 12 codes each (Table 2*b*) and eight classes containing 24 codes each (Table 2*c*). For each class, we give explicitely the list (in lexicographical order) of trinucleotides: the first (in lexicographical order) *LDL*-strong 20-trinucleotide circular code X (pattern of the class) and the codes of the permutations of Σ_4 (Table 1*c*) on X giving the other *LDL*-strong 20-trinucleotide circular codes of the class. The classes are lexicographically ordered according to the patterns of classes.

Moreover, a computer calculus describes the properties of prefixes and suffixes for the 28 classes of *LDL*-strong 20-trinucleotide circular codes X. The set L_1 is formed by the letters l_1 in the first position of the trinucleotides of X and the set L_3, by the letters l_3 in the last position of the trinucleotides of X. The set D_1 is formed by the diletters d_1 in prefix position of the trinucleotides of X and the set D_2, by the diletters d_2 in suffix position of the trinucleotides of X. Eight classes have both four letters in L_1 and L_2 (C_1-C_4, C_8, C_{10}, C_{17}, C_{20}). Ten classes have four letters in L_1 and three letters in L_2 (C_5-C_7, C_9, C_{11}, C_{12}, C_{21}-C_{24}). Reciprocally, ten classes have four letters in L_2 and three letters in L_1 (C_{13}-C_{16}, C_{18}, C_{19}, C_{25}-C_{28}). Three classes have nine diletters in D_1 (C_5, C_7, C_{21}) and similarly, three classes have nine diletters in D_2 (C_{16}, C_{18}, C_{28}). Only the class C_{28} has six diletters in D_1 and nine diletters in D_2 and similarly, only the class C_{21} has six diletters in D_2 and nine diletters in D_1. All the sets $D_1 \cap D_2$ are obviously empty.

These tables and the other symmetric relations identified before (e.g. Proposition 6 of [18]) suggest us that the symmetric group Σ_4 can have a very important role in the study of these trinucleotide circular codes.

$C_1 : aab, aac, aad, bab, bac, bad, bbc, bda, bdb, bdc, cab, cac, cad, ccb, cda, cdb, cdc, dda, ddb, ddc$ Codes of $C_1 : 211, 220, 221, 301, 320$
$C_2 : aab, aac, aad, bab, bac, bad, bca, bcb, bcd, bdd, cca, ccb, ccd, dab, dac, dad, dbb, dca, dcb, dcd$ Codes of $C_2 : 201, 221, 311, 320, 321$
$C_3 : aab, aca, acb, acc, ada, adb, add, bba, bca, bcb, bcc, bda, bdb, bdd, cda, cdb, cdd, dca, dcb, dcc$ Codes of $C_3 : 120, 121, 310, 311, 321$
$C_4 : aba, abb, abc, acc, ada, adc, add, bda, bdc, bdd, caa, cba, cbb, cbc, cda, cdc, cdd, dba, dbb, dbc$ Codes of $C_4 : 201, 221, 311, 320, 321$

Table 2a. The four classes having each six *LDL*-strong 20-trinucleotide circular codes. Each class is described by its pattern and the five codes of the permutations of the symmetric group Σ_4 on the pattern allow the other five *LDL*-strong 20-trinucleotide circular codes of the class to be deduced.

$C_5 : aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, ccd, dab, dac, dad, dbc, dbd, ddc$ Codes of $C_5 : 020, 021, 101, 120, 121, 300, 301, 310, 311, 320, 321$
$C_6 : aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, cdb, cdc, dab, dac, dad, ddb, ddc$ Codes of $C_6 : 011, 020, 201, 220, 221, 300, 301, 310, 311, 320, 321$
$C_7 : aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, cdd, dab, dac, dad, dbc, dbd, dcc$ Codes of $C_7 : 020, 021, 101, 120, 121, 300, 301, 310, 311, 320, 321$
$C_8 : aab, aac, aad, bab, bac, bad, bbc, bda, bdb, bdc, cab, cac, cad, cbc, cda, cdb, cdc, dda, ddb, ddc$ Codes of $C_8 : 111, 120, 121, 211, 220, 221, 300, 301, 310, 320, 321$
$C_9 : aab, aac, aad, bab, bac, bad, bbc, bdb, bdc, bdd, cab, cac, cad, ccb, cdb, cdc, cdd, dab, dac, dad$ Codes of $C_9 : 011, 021, 200, 201, 210, 211, 220, 221, 301, 320, 321$
$C_{10} : aab, aac, aad, bab, bac, bad, bca, bcb, bcd, bdb, bdd, cca, ccb, ccd, dab, dac, dad, dca, dcb, dcd$ Codes of $C_{10} : 110, 111, 121, 200, 201, 211, 220, 221, 311, 320, 321$
$C_{11} : aab, aac, aad, bab, bac, bad, bcb, bcc, bcd, bdd, cab, cac, cad, dab, dac, dad, dbb, dcb, dcc, dcd$ Codes of $C_{11} : 011, 020, 201, 220, 221, 300, 301, 310, 311, 320, 321$
$C_{12} : aab, aac, aad, bab, bac, bad, bcb, bcc, bdb, bdd, cab, cac, cad, cdb, cdd, dab, dac, dad, dcb, dcc$ Codes of $C_{12} : 020, 021, 101, 120, 121, 300, 301, 310, 311, 320, 321$
$C_{13} : aab, aac, ada, adb, adc, add, bab, bac, bbc, bda, bdb, bdc, bdd, cab, cac, ccb, cda, cdb, cdc, cdd$ Codes of $C_{13} : 011, 021, 200, 201, 210, 211, 220, 221, 301, 320, 321$
$C_{14} : aab, aac, ada, adb, adc, add, bab, bac, bca, bcb, bda, bdb, bdc, bdd, cca, ccb, cda, cdb, cdc, cdd$ Codes of $C_{14} : 110, 111, 121, 200, 201, 211, 220, 221, 311, 320, 321$
$C_{15} : aab, aac, ada, adb, adc, add, bab, bac, bcc, bda, bdb, bdc, bdd, cab, cac, cbb, cda, cdb, cdc, cdd$ Codes of $C_{15} : 011, 021, 200, 201, 210, 211, 220, 221, 301, 320, 321$
$C_{16} : aab, aca, acb, acc, acd, ada, adb, add, bba, bca, bcb, bcc, bcd, bda, bdb, bdd, dca, dcb, dcc, dcd$ Codes of $C_{16} : 101, 110, 111, 120, 121, 210, 211, 221, 310, 311, 321$
$C_{17} : aab, aca, acb, acc, ada, adb, add, bab, bca, bcb, bcc, bda, bdb, bdd, cda, cdb, cdd, dca, dcb, dcc$ Codes of $C_{17} : 020, 021, 101, 120, 121, 300, 301, 310, 311, 320, 321$
$C_{18} : aba, abb, abc, abd, aca, acc, acd, add, cba, cbb, cbc, cbd, daa, dba, dbb, dbc, dbd, dca, dcc, dcd$ Codes of $C_{18} : 111, 120, 121, 211, 220, 221, 300, 301, 310, 320, 321$
$C_{19} : aba, abb, abc, abd, aca, acc, ada, add, cba, cbb, cbc, cbd, cda, cdd, dba, dbb, dbc, dbd, dca, dcc$ Codes of $C_{19} : 020, 021, 101, 120, 121, 300, 301, 310, 311, 320, 321$
$C_{20} : aba, abb, abc, aca, acc, ada, adc, add, bda, bdc, bdd, cba, cbb, cbc, cda, cdc, cdd, dba, dbb, dbc$ Codes of $C_{20} : 011, 020, 201, 220, 221, 300, 301, 310, 311, 320, 321$

Table 2b. The 16 classes having each 12 LDL-strong 20-trinucleotide circular codes. Each class is described by its pattern and the 11 codes of the permutations of the symmetric group Σ_4 on the pattern allows the other 11 LDL-strong 20-trinucleotide circular codes of the class to be deduced.

$C_{21} : aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, ccd, dab, dac, dad, dbc, dbd, dcd$
$C_{22} : aab, aac, aad, bab, bac, bad, bbc, bbd, cab, cac, cad, cbc, cbd, cdc, cdd, dab, dac, dad, dbc, dbd$
$C_{23} : aab, aac, aad, bab, bac, bad, bbc, bdb, bdc, bdd, cab, cac, cad, cbc, cdb, cdc, cdd, dab, dac, dad$
$C_{24} : aab, aac, aad, bab, bac, bad, bcb, bcc, bcd, bdb, bdd, cab, cac, cad, dab, dac, dad, dcb, dcc, dcd$
$C_{25} : aab, aac, ada, adb, adc, add, bab, bac, bbc, bda, bdb, bdc, bdd, cab, cac, cbc, cda, cdb, cdc, cdd$
$C_{26} : aab, aac, ada, adb, adc, add, bab, bac, bcb, bcc, bda, bdb, bdc, bdd, cab, cac, cda, cdb, cdc, cdd$
$C_{27} : aab, aca, acb, acc, acd, ada, adb, add, bab, bca, bcb, bcc, bcd, bda, bdb, bdd, dca, dcb, dcc, dcd$
$C_{28} : aba, abb, abc, abd, aca, acc, acd, ada, add, cba, cbb, cbc, cbd, dba, dbb, dbc, dbd, dca, dcc, dcd$

Table 2c. The eight classes having each 24 LDL-strong 20-trinucleotide circular codes. In this case, each class is only described by its pattern as the other 23 LDL-strong 20-trinucleotide circular codes are obtained with the 23 permutations of the symmetric group Σ_4.

Codes	L_1	L_3	D_1	D_2
C_1	a, b, c, d	a, b, c, d	$aa, ba, bb, bd, ca, cc, cd, dd$	$ab, ac, ad, bc, cb, da, db, dc$
C_2	a, b, c, d	a, b, c, d	$aa, ba, bc, bd, cc, da, db, dc$	$ab, ac, ad, bb, ca, cb, cd, dd$
C_3	a, b, c, d	a, b, c, d	$aa, ac, ad, bb, bc, bd, cd, dc$	$ab, ba, ca, cb, cc, da, db, dd$
C_4	a, b, c, d	a, b, c, d	$ab, ac, ad, bd, ca, cb, cd, db$	$aa, ba, bb, bc, cc, da, dc, dd$
C_5	a, b, c, d	b, c, d	$aa, ba, bb, ca, cb, cc, da, db, dd$	$ab, ac, ad, bc, bd, cd, dc$
C_6	a, b, c, d	b, c, d	$aa, ba, bb, ca, cb, cd, da, dd$	$ab, ac, ad, bc, bd, db, dc$
C_7	a, b, c, d	b, c, d	$aa, ba, bb, ca, cb, cd, da, db, dc$	$ab, ac, ad, bc, bd, cc, dd$
C_8	a, b, c, d	a, b, c, d	$aa, ba, bb, bd, ca, cb, cd, dd$	$ab, ac, ad, bc, da, db, dc$
C_9	a, b, c, d	b, c, d	$aa, ba, bb, bd, ca, cc, cd, da$	$ab, ac, ad, bc, cb, db, dc, dd$
C_{10}	a, b, c, d	a, b, c, d	$aa, ba, bc, bd, cc, da, dc$	$ab, ac, ad, ca, cb, cd, db, dd$
C_{11}	a, b, c, d	b, c, d	$aa, ba, bc, bd, ca, da, db, dc$	$ab, ac, ad, bb, cb, cc, cd, dd$
C_{12}	a, b, c, d	b, c, d	$aa, ba, bc, bd, ca, cd, da, dc$	$ab, ac, ad, cb, cc, db, dd$
C_{13}	a, b, c	a, b, c, d	$aa, ad, ba, bb, bd, ca, cc, cd$	$ab, ac, bc, cb, da, db, dc, dd$
C_{14}	a, b, c	a, b, c, d	$aa, ad, ba, bc, bd, cc, cd$	$ab, ac, ca, cb, da, db, dc, dd$
C_{15}	a, b, c	a, b, c, d	$aa, ad, ba, bc, bd, ca, cb, cd$	$ab, ac, bb, cc, da, db, dc, dd$
C_{16}	a, b, d	a, b, c, d	$aa, ac, ad, bb, bc, bd, dc$	$ab, ba, ca, cb, cc, cd, da, db, dd$
C_{17}	a, b, c, d	a, b, c, d	$aa, ac, ad, ba, bc, bd, cd, dc$	$ab, ca, cb, cc, da, db, dd$
C_{18}	a, c, d	a, b, c, d	$ab, ac, ad, cb, da, db, dc$	$aa, ba, bb, bc, bd, ca, cc, cd, dd$
C_{19}	a, c, d	a, b, c, d	$ab, ac, ad, cb, cd, db, dc$	$ba, bb, bc, bd, ca, cc, da, dd$
C_{20}	a, b, c, d	a, b, c, d	$ab, ac, ad, bd, cb, cd, db$	$ba, bb, bc, ca, cc, da, dc, dd$
C_{21}	a, b, c, d	b, c, d	$aa, ba, bb, ca, cb, cc, da, db, dc$	ab, ac, ad, bc, bd, cd
C_{22}	a, b, c, d	b, c, d	$aa, ba, bb, ca, cb, cd, da, db$	$ab, ac, ad, bc, bd, dc, dd$
C_{23}	a, b, c, d	b, c, d	$aa, ba, bb, bd, ca, cb, cd, da$	$ab, ac, ad, bc, db, dc, dd$
C_{24}	a, b, c, d	b, c, d	$aa, ba, bc, bd, ca, da, dc$	$ab, ac, ad, cb, cc, cd, db, dd$
C_{25}	a, b, c	a, b, c, d	$aa, ad, ba, bb, bd, ca, cb, cd$	$ab, ac, bc, da, db, dc, dd$
C_{26}	a, b, c	a, b, c, d	$aa, ad, ba, bc, bd, ca, cd$	$ab, ac, cb, cc, da, db, dc, dd$
C_{27}	a, b, d	a, b, c, d	$aa, ac, ad, ba, bc, bd, dc$	$ab, ca, cb, cc, cd, da, db, dd$
C_{28}	a, c, d	a, b, c, d	ab, ac, ad, cb, db, dc	$ba, bb, bc, bd, ca, cc, cd, da, dd$

Table 3. The prefixes and suffixes of the pattern of the 28 classes of LDL-strong 20-trinucleotide circular codes (Tables 2a-c). For each pattern X, the subset L_1 (resp. L_3) of $\{a, b, c, d\}$ consists of the letters l_1 (resp. l_3) that appear at least once in prefix (resp. suffix) position of the trinucleotides of X. Similarly, the subset D_1 (resp. D_2) of $\{a, b, c, d\}^2$ consists of the diletters d_1 (resp. d_2) that appear at least once in prefix (resp. suffix) position of the trinucleotides of X.

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

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

[7] 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).

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

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

[10] 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).

[11] 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).

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

[13] A. Lascoux, The symmetric group, June 2010.

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

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

[16] 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).

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

[18] 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).

[19] 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).

[20] 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).

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

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

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

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