Computational Biology and Chemistry 34 (2010) 122–125

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Brief communication

Identification of all trinucleotide circular codes

Christian J. Michel[a,*], Giuseppe Pirillo[b,c]

[a] Equipe de Bioinformatique Théorique, FDBT, LSIIT (UMR CNRS-ULP 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France
[b] Consiglio Nazionale delle Ricerche, Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Unità di Firenze, Dipartimento di Matematica "U.Dini", viale Morgagni 67/A, 50134 Firenze, Italia
[c] Université de Marne-la-Vallée, 5 boulevard Descartes, 77454 Marne-la-Vallée Cedex 2, France

ARTICLE INFO

Article history:
Received 31 January 2010
Received in revised form 18 March 2010
Accepted 21 March 2010

Keywords:
Circular code
Trinucleotide
Growth function
Necklace
Gene
Genetic code

ABSTRACT

A new trinucleotide proposition is proved here and allows all the trinucleotide circular codes on the genetic alphabet to be identified (their numbers and their sets of words). This new class of genetic motifs, i.e. circular codes (or synchronizing genetic motifs), may be involved in the structure and the origin of the genetic code, and in reading frames of genes.

1. Introduction

We continue our study of the properties of trinucleotide circular codes. For 50 years, codes, comma-free codes and circular codes have been mathematical objects studied in biology, mainly to understand the structure and the origin of the genetic code as well as the reading frame (construction) of genes, see the pioneer works (Crick et al., 1957; Golomb et al., 1958a,b). In order to have an intuitive meaning of these notions, codes are written on a straight line while comma-free codes and circular codes are written on a circle, but in both cases, unique decipherability is required.

The genetic code based on 64 trinucleotides is a code in the sense of language theory, more precisely a uniform code (Berstel and Perrin, 1985), but not a circular code (Lassez, 1976) (see Remark 2 below). Before the discovery of the genetic code, Crick et al. (1957) 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 a large gene population of eukaryotes and also on a large gene population of prokaryotes (Arquès and Michel, 1996):

$$X_0 = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$$

This code X_0 has remarkable properties. For example, X_0 is self-complementary: 10 trinucleotides are complementary to the 10 other trinucleotides, e.g. *AAC* is complementary to *GTT*, *AAT* to *ATT*, etc. The two sets of 20 trinucleotides, called X_1 and X_2, obtained by a simple shift operation of X_0, one and two letters, respectively, are also maximal circular codes (Arquès and Michel, 1996). This surprising result, still mysterious, was cited/discussed in research works in mathematics/computer science and mainly in theoretical biology, e.g. (Koch and Lehman, 1997; Béal and Senellart, 1998; Bassino, 1999; Štambuk, 1999; Jolivet and Rothen, 2001; Nikolaou and Almirantis, 2003; May et al., 2004; Lassez et al., 2007; Pirillo, 2003; José et al., 2009). Its main biological consequence would be that genes have (or had) two codes: the classical genetic code to code the amino acids and a circular code to retrieve the reading frames of genes. Therefore, the computational study of trinucleotide circular codes is particularly important in biology.

The determinations of very small classes of trinucleotide circular codes, precisely the 99,320 self-complementary trinucleotide circular codes (Pirillo and Pirillo, 2005) and about 559 millions trinucleotide comma-free codes (Michel et al., 2008a), were obtained by using the classical flower automaton algorithm (Berstel and Perrin, 1985). We recently identified a relation between these two classes of trinucleotide codes by constructing a hierarchy of codes that are closed by the comma-free codes and the circular codes (Michel et al., 2008b). The whole class of all the trinucleotide circular codes is identified in this paper (their numbers and their sets of words). This problem has a computational complexity with

* Corresponding author.
E-mail addresses: michel@dpt-info.u-strasbg.fr (C.J. Michel), pirillo@math.unifi.it (G. Pirillo).

1476-9271/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2010.03.004

an order of magnitude significantly higher than the two previous cases (more than 200 times). Indeed, about 116 billion trinucleotide circular codes are identified. The proof of a new trinucleotide proposition (Proposition 3), which appears obvious afterwards, allows the computational problem associated with the general case to be solved. Thus, this short Proposition 3 which can easily be programmed, allows circular codes (synchronizing genetic motifs) on the genetic alphabet to be identified.

2. Definitions

Let $\mathcal{A}$ denote a finite alphabet, $\mathcal{A}^*$, the set of all words over $\mathcal{A}$ and $\mathcal{A}^+$, the set of all words over $\mathcal{A}$ except the empty word ε. Given a subset X of $\mathcal{A}^*$, X^n is the set of the words over $\mathcal{A}$ which is the product of n words from X, i.e. $X^n = \{x_1x_2\cdots x_n | x_i \in X\}$.

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 two (or dinucleotides or diletters) is denoted by $\mathcal{A}_4^2$. The set of the 64 words of length three (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 1. Code: A set 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 $i = 1, \ldots, n$.

Remark 1. The set $\mathcal{A}_4^3$ itself is a code. More precisely, it is a uniform code (Berstel and Perrin, 1985).

Notation 1. Consequently, any non-empty subset of $\mathcal{A}_4^3$ is a code called trinucleotide code in this paper.

Definition 2. Trinucleotide circular code: A trinucleotide code $X \in \mathcal{A}_4^3$ 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 $sx_2\cdots x_np = x'_1\cdots x'_m$ and $x_1 = ps$ imply $n = m$, $p = \varepsilon$ and $x_i = x'_i$ for $i = 1, \ldots, n$.

Remark 2. $\mathcal{A}_4^3$ is obviously not a trinucleotide circular code.

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

Remark 3. Any trinucleotide circular code with 20 words is maximal. Therefore, the lengths of trinucleotide circular codes vary between 1 and 20.

3. Propositions

Proposition 1. *The number of trinucleotide circular codes of length 1 is equal to* 60.

Proof. Obvious. □

Proposition 2. *The number of trinucleotide circular codes of length* 20 *is equal to* 12,964,440.

Proof. This number was obtained in 1996 by using the flower automaton algorithm (Table 2(d) in Arquès and Michel, 1996). □

In order to compute the growth function of trinucleotide circular codes for all lengths $l = 1, \ldots, 20$, we extend the necklace definition (Pirillo, 2003; Michel et al., 2008b). $l_1, l_2, \ldots, l_{n-1}, l_n, \ldots$ are letters in $\mathcal{A}_4$, $d_1, d_2, \ldots, d_{n-1}, d_n, \ldots$ are diletters in $\mathcal{A}_4^2$ and n is an integer satisfying $n \geq 2$.

Definition 4. Letter Diletter Continued Closed Necklaces (*LDCCN*): 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 $(n+1)$*LDCCN* for a subset $X \subset \mathcal{A}_4^3$ if $l_1d_1, l_2d_2, \ldots, l_nd_n \in X$ and $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n, d_nl_{n+1} \in X$ and $l_1 = l_{n+1}$.

Notation 2. An $(n+1)$*LDCCN* $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n, l_{n+1}$ is denoted by $[l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n]$. Accordingly: *a* 2*LDCCN*, i.e. $[l_1, d_1]$, has the form l_1, d_1, l_1; *a* 3*LDCCN*, i.e. $[l_1, d_1, l_2, d_2]$, has the form l_1, d_1, l_2, d_2, l_1; *a* 4*LDCCN*, i.e. $[l_1, d_1, l_2, d_2, l_3, d_3]$, has the form $l_1, d_1, l_2, d_2, l_3, d_3, l_1$; *a* 5*LDCCN*, i.e. $[l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4]$, has the form $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_1$.

Proposition 3. *Let X be a trinucleotide circular code. The following conditions are equivalent.*

(i) *X is a trinucleotide circular code.*
(ii) *X has no nLDCCN for any integer* $n \in \{2, 3, 4, 5\}$.

Proof. $(i) \Rightarrow (ii)$. By way of contradiction, suppose that X has some *nLDCCN* for some integer $n \in \{2, 3, 4, 5\}$.

If it is a 2*LDCCN* then $l_1, d_1, l_1, d_1, l_1, d_1, l_1, d_1, l_1$ is a 5*LDCN* for X.

If it is a 3*LDCCN* then $l_1, d_1, l_2, d_2, l_1, d_1, l_2, d_2, l_1$ is a 5*LDCN* for X.

If it is a 4*LDCCN* then $l_1, d_1, l_2, d_2, l_3, d_3, l_1, d_1, l_2$ is a 5*LDCN* for X.

If it is a 5*LDCCN* then $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_1$ is a 5*LDCN* for X.

In each of these four cases, by Proposition 1, X is not a trinucleotide circular code. Contradiction.

$(ii) \Rightarrow (i)$. By way of contradiction, suppose that X is not a trinucleotide circular code. By Proposition 1, X has a 5*LDCN*, say $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_5$. As $\mathcal{A}_4$ has four letters, then $l_i = l_j$ for some i, j, $1 \leq i \leq j \leq 5$.

If $j - i = 4$ then $l_1 = l_5$ and $[l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4]$ is a 5*LDCCN* for X.

If $j - i = 3$ then $[l_i, d_i, l_{i+1}, d_{i+1}, l_{i+2}, d_{i+2}]$ is a 4*LDCCN* for X.

If $j - i = 2$ then $[l_i, d_i, l_{i+1}, d_{i+1}]$ is a 3*LDCCN* for X.

If $j - i = 1$ then $[l_i, d_i]$ is a 2*LDCCN* for X.

In each of these four cases, by Proposition 1, there is a contradiction with (ii). □

Necklace algorithm (principle): This new Proposition 3 is used to compute all the trinucleotide circular codes (growth function for all lengths $l = 1, \ldots, 20$). The principle of this necklace algorithm is simple. If the algorithm identifies a necklace *iLDCCN* for a given $i \in \{2, 3, 4, 5\}$ in a code, then it is not circular and the algorithm stops avoiding to analyse the next necklaces *jLDCCN* for $j > i$ and $j \in \{2, 3, 4, 5\}$.

4. Results

Table 1 shows the number $\mathrm{Nb}(l)$ of trinucleotide circular codes of length l. The growth function has a minimum number $\mathrm{NbMin} = 60$ at $l = 1$ and a maximum number $\mathrm{NbMax} = 23{,}403{,}485{,}556$ at $l = 13$. Fig. 1 associated with Table 1 gives the graphical distribution of trinucleotide circular codes. The distribution is asymmetric with respect to NbMax at $l = 13$. The numbers of codes of $l = 13$ and $l = 14$ are close. There are $\mathrm{NbPot}(l) = \binom{20}{l} \times 3^l$ potential trinucleotide circular codes of length $l \in \{1, 20\}$. Therefore, the probability $\Pr(l)$ of a trinucleotide circular code of length l is equal to $\Pr(l) = \mathrm{Nb}(l)/\mathrm{NbPot}(l)$. Table 1 and Fig. 1 also show this proba-