Available online at www.sciencedirect.com
SCIENCE DIRECT
Computational Biology and Chemistry
Computational Biology and Chemistry 30 (2006) 87–101
www.elsevier.com/locate/compbiolchem

# Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes

Gabriel Frey, Christian J. Michel *

*Equipe de Bioinformatique Théorique, LSIIT (UMR CNRS-ULP 7005), Université Louis Pasteur de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France*

Received 30 September 2005; received in revised form 7 November 2005; accepted 7 November 2005

**Abstract**

We developed a statistical method that allows each trinucleotide to be associated with a unique frame among the three possible ones in a (protein coding) gene. An extensive gene study in 175 complete bacterial genomes based on this statistical approach resulted in identification of 72 new circular codes. Finding a circular code enables an immediate retrieval of the reading frame locally anywhere in a gene. No knowledge of location of the start codon is required and a short window of only a few nucleotides is sufficient for automatic retrieval. We have therefore developed a factorization method (that explores previously found circular codes) for retrieving the reading frames of bacterial genes. Its principle is new and easy to understand. Neither complex treatment nor specific information on the nucleotide sequences is necessary. Moreover, the method can be used for short regions in nucleotide sequences (less than 25 nucleotides in protein coding genes). Selected additional properties of circular codes and their possible biological consequences are also discussed.


*Keywords:* Bacterial genome; Circular code; Frame

## 1. Introduction

Each bacterial genome has its own trinucleotide distribution (Grantham et al., 1980). Indeed, the synonymous codons (codons coding for the same amino acid) do not occur with the same frequencies in bacterial genes. This synonymous codon usage is biased: a restricted subset of codons is preferred in genes. Codon usage is generally correlated with gene expressivity (Grantham et al., 1981; Ikemura, 1985; Sharp and Matassi, 1994) even if its strength varies among bacterial species (Sharp et al., 2005). A proposed explanation is that codon usage reflects the variation in the concentration of tRNAs. Major codons encoded by more abundant tRNAs should increase translational efficacy (Bulmer, 1991; Akashi and Eyre-Walker, 1998). Nevertheless, tRNA abundance could also have evolved for matching codon pattern in a genome (Fedorov et al., 2002) and then would rather be a consequence of the synonymous codon bias.

Several other processes may influence codon usage (Llopart and Aguade, 2000; Smith and Eyre-Walker, 2001; Konu and Li, 2002; Krakauer and Jansen, 2002; Rogozin et al., 2005). In particular, codon choice may depend on its context, i.e. the surrounding nucleotides (Yarus and Folley, 1984; Shpaer, 1986; Berg and Silva, 1997). These pressures might be frame independent (Antezana and Kreitman, 1999). In this line of research, we have studied the trinucleotide occurrences in the three frames of genes by computing their $3 \times 64 = 192$ frequencies. This approach has led to the identification of particular codes in genes called circular codes.

By convention, the reading frame established by a start codon (ATG, GTG and TTG) is the frame 0, and the frames 1 and 2 are the reading frame shifted by 1 and 2 nucleotides in the 5′-3′ direction, respectively. After excluding the trinucleotides with identical nucleotides (AAA, CCC, GGG and TTT) and by assigning each trinucleotide to a preferential frame, three subsets of 20 trinucleotides per frame have been identified in the gene populations of both eukaryotes EUK and prokaryotes PRO (Arquès and Michel, 1996). These three sets $X_0$(EUK_PRO), $X_1$(EUK_PRO) and $X_2$(EUK_PRO) associated with the frames 0, 1 and 2, respectively, have several strong properties, in particular the property of circular code. The circular code concept will be

* Corresponding author. Tel.: +33 3 90 24 44 62.
*E-mail addresses:* frey@dpt-info.u-strasbg.fr (G. Frey), michel@dpt-info.u-strasbg.fr (C.J. Michel).

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.11.001

briefly pointed out without mathematical notations after a short historical presentation of an another class of code which has been searched but not found in genes (over the alphabet {A,C,G,T}).

A code in genes has been proposed by Crick et al. (1957) in order to explain how the reading of a series of nucleotides could code for the amino acids constituting the proteins. The two problems stressed were: why are there more trinucleotides than amino acids and how to choose the reading frame? Crick et al. (1957) have then proposed that only 20 among 64 trinucleotides code for the 20 amino acids. Furthermore, a bijective code implies that the coding trinucleotides are found only in one frame. Such a particular code is called a comma-free code or a code without commas. However, the determination of a set of 20 trinucleotides forming a comma-free code has several constraints:

(i) A trinucleotide with identical nucleotides must be excluded from such a code. Indeed, the concatenation of AAA with itself, for example, does not allow the reading (original) frame to be retrieved as there are three possible decompositions: ...AAA,AAA,AAA,..., ...A,AAA,AAA,AA... and ...AA,AAA,AAA,A...
(ii) Two trinucleotides related to circular permutation, for example AAC and ACA, must be also excluded from such a code. Indeed, the concatenation of AAC with itself, for example, also does not allow the reading frame to be retrieved as there are two possible decompositions: ...AAC,AAC,AAC,... and ...A,ACA,ACA,AC...

Therefore, by excluding AAA, CCC, GGG and TTT and by gathering the 60 remaining trinucleotides in 20 classes of three trinucleotides such that, in each class, three trinucleotides are deduced from each other by circular permutations, e.g. AAC, ACA and CAA, a comma-free code has only one trinucleotide per class and therefore contains at most 20 trinucleotides. This trinucleotide number is identical to the amino acid one, thus leading to a comma-free code assigning one trinucleotide per amino acid without ambiguity.

The determination of comma-free codes and their properties are unrealizable without computer as there are $3^{20} \approx 3.5$ billions potential codes. A comma-free code search algorithm demonstrates in particular that there are only 408 comma-free codes of 20 trinucleotides. None of them is complementary as the maximal complementary comma-free codes contain only 16 trinucleotides (results not shown). Furthermore, in the late 1950s, the two discoveries that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes for phenylalanine (Nirenberg and Matthaei, 1961) and that genes are placed in reading frames with a particular start trinucleotide, have led to give up the concept of comma-free code over the alphabet {A,C,G,T}. For several biological reasons, in particular the interaction between mRNA and tRNA, this concept is taken up again later over the alphabet {R,Y} (R = purine = A or G, Y = pyrimidine = C or T) with two comma-free codes for primitive genes: RRY (Crick et al., 1976) and RNY (N = R or Y) (Eigen and Schuster, 1978).

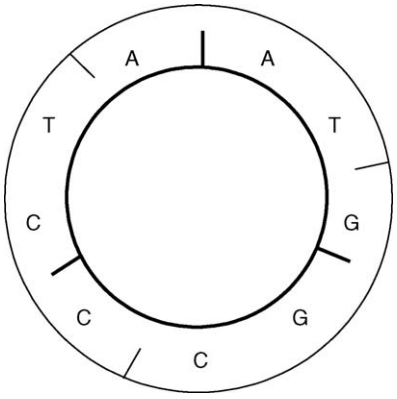


Fig. 1. The set $X = \{AAT,ATG,CCT,CTA,GCC,GGC\}$ is not a circular code as the word $w = ATGGCCCTA$, written on a circle, can be factorized into words of $X$ according to two different ways: ATG, GCC, CTA (thick line) and AAT, GGC, CCT (thin line).

A circular code also allows the reading frames of genes to be retrieved but with weaker conditions compared to a comma-free code. It is a set of words over an alphabet such that any word written on a circle (the next letter after the last letter of the word being the first letter) has at most one decomposition into words of the circular code. As an example, let the set $X$ be composed of the six following words: $X = \{AAT,ATG, CCT,CTA,GCC,GGC\}$ and the word $w$, be a series of the nine following letters: $w = ATGGCCCTA$. The word $w$, written on a circle, can be factorized into words of $X$ according to two different ways: ATG, GCC, CTA and AAT, GGC, CCT (Fig. 1). Therefore, $X$ is not a circular code. In contrast, if the set $\tilde{X}$ obtained by replacing the word GGC of $X$ by GTC is considered, i.e. $\tilde{X} = \{AAT, ATG, CCT, CTA, GCC, GTC\}$, then there never exists an ambiguous word with $\tilde{X}$, in particular $w$ is not ambiguous, and $\tilde{X}$ is a circular code. The construction frame of a word generated by any concatenation of words of a circular code can be retrieved after the reading, anywhere in the generated word, of a certain number of nucleotides depending on the code. This series of nucleotides is called the window $W$ of the circular code.

A comma free code has conditions stronger than a circular code. Indeed, the 20 trinucleotides of a comma free code are found only in one frame, i.e. in the reading frame, while some trinucleotides of a circular code can be found in the two shifted frames 1 and 2 (see below). On the other hand, the lengths of the windows $W$ of a comma free code and a circular code are less than or equal to 4 and 13 nucleotides respectively (Section 2.2.4).

Definition of the trinucleotide (left circular) permutation: the (left circular) permutation $P$ of a trinucleotide $w_0 = l_0l_1l_2$, $l_0,l_1,l_2 \in \{A,C,G,T\}$, is the permuted trinucleotide $P(w_0) = w_1 = l_1l_2l_0$, e.g. $P(AAC) = ACA$, and $P(P(w_0)) = P(w_1) = w_2 = l_2l_0l_1$, e.g. $P(P(AAC)) = CAA$. This definition is naturally extended to the trinucleotide set permutation: the permutation $P$ of a set of trinucleotides is the permuted trinucleotide set obtained by the permutation $P$ of all its trinucleotides.

The first identified circular code is the set $X_0(EUK\_PRO) = \{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,$

GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC} in the frame 0 (reading frame) of genes of eukaryotes EUK and prokaryotes PRO (Arquès and Michel, 1996). It has several important properties (some of them will be detailed in Section 2.2).

(i) Maximality: $X_0$(EUK_PRO) is a maximal circular code (20 trinucleotides).
(ii) Permutation: $X_0$(EUK_PRO) generates $X_1$(EUK_PRO) by one permutation and $X_2$(EUK_PRO) by another permutation, i.e. $P(X_0(\text{EUK\_PRO}))=X_1(\text{EUK\_PRO})$ and $P(P(X_0(\text{EUK\_PRO})))=X_2(\text{EUK\_PRO})$.
(iii) Complementarity: $X_0$(EUK_PRO) is self-complementary (10 trinucleotides of $X_0$(EUK_PRO) are complementary to 10 other trinucleotides of $X_0$(EUK_PRO)) and, $X_1$(EUK_PRO) and $X_2$(EUK_PRO) are complementary to each other (the 20 trinucleotides of $X_1$(EUK_PRO) are complementary to the 20 trinucleotides of $X_2$(EUK_PRO)).
(iv) $C^3$ code: $X_1$(EUK_PRO) and $X_2$(EUK_PRO) obtained by permutation of $X_0$(EUK_PRO) (property ii) are maximal circular codes. It is important to stress that a circular code $X_0$ does not necessarily imply that $X_1$ and $X_2$ obtained by permutation, are also circular codes.
(v) Rarity: the occurrence probability of the $C^3$ code $X_0$(EUK_PRO) is equal to $216/3^{20} \approx 6 \times 10^{-8}$, i.e. the computed number of complementary $C^3$ codes (216) divided by the number of potential codes ($3^{20}$ = 3,486,784,401).
(vi) Flexibility:

(via) The lengths of the minimal windows of $X_0$(EUK_PRO), $X_1$(EUK_PRO) and $X_2$(EUK_PRO) for retrieving automatically the frames 0, 1 and 2, respectively, are all equal to 13 nucleotides and represent the largest window length among the 216 $C^3$ codes.
(vib) The frequencies of "misplaced" trinucleotides in the shifted frames 1 and 2 are both equal to 24.6%. If the trinucleotides of $X_0$(EUK_PRO) are randomly concatenated, for example as follows: ...GAA,GAG,GTA,GTA,ACC, AAT,GTA,CTC,TAC,TTC,ACC,ATC... then, the trinucleotides in frame 1: ...G,AAG,AGG,TAG,TAA, CCA,ATG,TAC,TCT,ACT,TCA,CCA,TC... and the trinucleotides in frame 2: ...GA,AGA,GGT,AGT,AAC, CAA,TGT,ACT,CTA,CTT,CAC,CAT,C... mainly belong to $X_1$(EUK_PRO) and $X_2$(EUK_PRO), respectively. A few trinucleotides are misplaced in the shifted frames. With this example, in frame 1, nine trinucleotides belong to $X_1$(EUK_PRO), one trinucleotide (TAC) to $X_0$(EUK_PRO) and one trinucleotide (TAA) to $X_2$(EUK_PRO). In frame 2, eight trinucleotides belong to $X_2$(EUK_PRO), two trinucleotides (GGT, AAC) to $X_0$(EUK_PRO) and one trinucleotide (ACT) to $X_1$(EUK_PRO). By computing exactly, the frequencies of misplaced trinucleotides in frame 1 are 11.9% for $X_0$(EUK_PRO) and 12.7% for $X_2$(EUK_PRO). In frame 2, the frequencies of misplaced trinucleotides are 11.9% for $X_0$(EUK_PRO) and 12.7% for $X_1$(EUK_PRO). The complementarity property (iii) explains on the one hand, the identical frequencies of $X_0$(EUK_PRO) in frames 1 and 2 (such words are impossible with a comma free code), and on the other hand, the identical frequencies of $X_2$(EUK_PRO) in frame 1 and $X_1$(EUK_PRO) in frame 2. Then, the frequency sum of misplaced trinucleotides in frame 1 ($X_0$(EUK_PRO) and $X_2$(EUK_PRO)) is equal to the one of misplaced trinucleotides in frame 2 ($X_0$(EUK_PRO) and $X_1$(EUK_PRO)) and is equal to 24.6%. This value is close to the highest frequency (27.9%) of misplaced trinucleotides among the 216 $C^3$ codes.
(vic) The four types of nucleotides occur in the three trinucleotide sites with $X_0$(EUK_PRO), and also obviously by the permutation property (ii) with $X_1$(EUK_PRO) and $X_2$(EUK_PRO). It is important to stress that $C^3$ codes can have missing nucleotides in trinucleotide sites.

The circular code information for retrieving reading frames coexists with the classical genetic code for coding amino acids. Similarly to the existence of variant genetic codes and different codon usage, several circular codes exist in genes. Circular codes have been identified in mitochondria (Arquès and Michel, 1997) and archaea (Frey and Michel, 2003), and now in bacterial genomes by using a quantitative and sensitive statistical method.

A necessary but not sufficient condition for a code to be circular is the absence of two permuted words in the code, otherwise there is no unique decomposition. Then, the 60 trinucleotides (without AAA, CCC, GGG and TTT) are gathered in 20 classes of three trinucleotides invariant by permutation. The developed method, called frame permuted trinucleotide frequency (FPTF), considers both the preferential frame of a trinucleotide by comparing its occurrence frequencies in the three frames and the preferential permuted trinucleotide in a frame by comparing the occurrence frequencies of the three permuted trinucleotides in a same frame. A statistical function based on these two parameters allows each trinucleotide to be associated with a unique frame.

By analysing an extensive data set of 175 complete bacterial genomes, the method FPTF will identify 72 new $C^3$ codes. Several properties and biological consequences of these new codes will also be described.

## 2. Methods

### *2.1. Assignment of a preferential trinucleotide set to each frame of genes in a genome*

In order to have a general and automatic approach for the trinucleotide assignment to a frame, the quantitative and sensitive method FPTF considers the occurrence frequencies of the three permuted trinucleotides in their three frames. It will identify several new circular codes in genes of bacterial genomes.

Over the genetic alphabet {A,C,G,T}, there are 60 trinucleotides with non-identical nucleotides $w \in \{\text{AAA},\ldots,\text{TTT}\} - \{\text{AAA,CCC,GGG,TTT}\}$ which can be gathered in 20 sets $S_j$, $j \in \{0,\ldots,19\}$, of three trinucleotides invariant by permutation: $S_0 = \{\text{AAC,ACA,CAA}\}$, $S_1 = \{\text{AAG,AGA,GAA}\},\ldots,$

$\mathcal{S}_{19}$ = {GTT,TTG,TGT}. The $i$th, $i \in \{0,1,2\}$, trinucleotide $w$ in a set $\mathcal{S}$ is noted $w_i$. Therefore, $w_1 = \mathcal{P}(w_0)$ and $w_2 = \mathcal{P}(\mathcal{P}(w_0))$. For example in $\mathcal{S}_0$, AAC, ACA and CAA are noted $w_0$, $w_1$ and $w_2$, respectively. In genes, there are three frames $p \in \{0,1,2\}$, $p = 0$ is the reading frame established by a start trinucleotide, and $p = 1$ and $p = 2$ are the shifted frames 1 and 2 by one and two nucleotides in the 5′-3′ direction, respectively. Let $w^p$ be a trinucleotide $w$ read in the frame $p$. A trinucleotide $w_i$, $i \in \{0,1,2\}$, in a set $\mathcal{S}$ read in a frame $p \in \{0,1,2\}$, is noted $w_i^p$. Therefore, a group $G_j$ associated with a set $\mathcal{S}_j$, $j \in \{0,\ldots,19\}$, has $3 \times 3 = 9$ trinucleotides $w_i^p$, $i,p \in \{0,1,2\}$. For example, the group $G_0$ associated with $\mathcal{S}_0$ is $G_0 = \{\mathrm{AAC}^0, \mathrm{AAC}^1, \mathrm{AAC}^2, \mathrm{ACA}^0, \mathrm{ACA}^1, \mathrm{ACA}^2, \mathrm{CAA}^0, \mathrm{CAA}^1, \mathrm{CAA}^2\}$. With 20 groups $G$, there are $20 \times 9 = 180$ trinucleotides $w_i^p$. The occurrence probability of a trinucleotide $w_i^p$, $i,p \in \{0,1,2\}$, in a group $G$ will be compared simultaneously to the two occurrence probabilities of $w_i^{p'}$ and $w_i^{p''}$ in the two other frames $p'$ and $p''$, and to the two occurrence probabilities of its two permuted trinucleotides $w_{i'}^p$ and $w_{i''}^p$ in the same frame $p$. Let $o(w_i^p)$ be the observed occurrence probability of a trinucleotide $w_i^p$ in a frame $p$ of genes in a genome. Then, in a group $G$, the function $P(w_i^p)$ of a trinucleotide $w_i^p$ computes the average probability of $w_i$ in the three frames $p \in \{0,1,2\}$ as follows:

$$P(w_i^p) = \frac{o(w_i^p)}{\sum_{p=0}^{2} o(w_i^p)}. \tag{1}$$

Similarly, in a group $G$, the function $Q(w_i^p)$ of a trinucleotide $w_i^p$ computes the average probability of the three permuted trinucleotides $w_0$, $w_1$ and $w_2$ in the frame $p$ as follows:

$$Q(w_i^p) = \frac{o(w_i^p)}{\sum_{i=0}^{2} o(w_i^p)}. \tag{2}$$

**Remark.** In a genome with hundreds of genes, the denominators $\mathrm{DEN}(P(w_i^p))$ and $\mathrm{DEN}(Q(w_i^p))$ of the two previous functions are different from 0. Indeed, each stop codon $w_s \in \{\mathrm{TAA}, \mathrm{TAG}, \mathrm{TGA}\}$ occurs in a different set $\mathcal{S}$, precisely $\mathrm{TAA} \in \mathcal{S}_2$, $\mathrm{TAG} \in \mathcal{S}_8$ and $\mathrm{TGA} \in \mathcal{S}_{10}$. Furthermore, a stop codon $w_s$ does not occur in frame 0 of genes, i.e. $o(w_s^0) = 0$, but in frames 1 and 2, i.e. $o(w_s^1) > 0$ and $o(w_s^2) > 0$, then $\mathrm{DEN}(P(w_s^p)) = \sum_{p=1}^{2} o(w_s^p) > 0$. On the other hand, as the two permuted trinucleotides $\mathcal{P}(w_s^0)$ and $\mathcal{P}(\mathcal{P}(w_s^0))$ of $w_s$ occur in frame 0, i.e. $o(\mathcal{P}(w_s^0)) > 0$ and $o(\mathcal{P}(\mathcal{P}(w_s^0))) > 0$, then $\mathrm{DEN}(Q(w_s^0)) = o(\mathcal{P}(w_s^0)) + o(\mathcal{P}(\mathcal{P}(w_s^0))) > 0$. These two inequalities could obviously not be verified with one gene of short length, case which never exists in a genome.

A trinucleotide $w_i$ occurring with the highest (or lowest) probability in a frame $p$ compared to the two other frames, can have a probability lower (or higher) than the probabilities of its two permuted trinucleotides in this frame $p$. In order to evaluate a trinucleotide simultaneously compared to its two other frames and its two other permuted trinucleotides, the function $M(w_i^p)$ of a trinucleotide $w_i^p$ is defined as the mean of the functions $P(w_i^p)$ and $Q(w_i^p)$

$$M(w_i^p) = \frac{1}{2}(P(w_i^p) + Q(w_i^p)). \tag{3}$$

The higher the value $M(w_i^p)$ of a trinucleotide $w_i^p$, the stronger its weight simultaneously in its frame and in its permutation set. Therefore, a trinucleotide $w_i^p$ with the highest value $M(w_i^p)$ occurs preferentially in the frame $p$, i.e. $w_i^p$ does not occur preferentially in the two other frames $p'$ and $p''$, and the two other permuted trinucleotides $w_{i'}^p$ and $w_{i''}^p$ do not occur preferentially in the frame $p$.

The next step of the method FPTF consists in selecting a set $S$ of three trinucleotides $w_i^p$ in a group $G_j$, $j \in \{0,\ldots,19\}$, according to their values $M(w_i^p)$. As a group $G$ has nine trinucleotides $w_i^p$, there are $\binom{9}{3} = 84$ possible sets $S_k$, $k \in \{0,\ldots,83\}$, of three trinucleotides. These 84 sets $S$ are defined as follows: $\{\{w_0^0, w_0^1, w_0^2\}, \ldots, \{w_0^0, w_0^2, w_0^0\}, \ldots, \{w_0^0, w_0^1, w_0^1\}, \ldots, \{w_0^0, w_0^1, w_0^2\}, \ldots, \{w_0^0, w_0^2, w_0^0\}, \ldots, \{w_0^0, w_0^2, w_0^1\}, \ldots, \{w_0^0, w_0^2, w_0^2\}, \ldots, \{w_0^1, w_0^2, w_0^0\}, \ldots, \{w_0^2, w_0^0, w_0^1\}, \ldots, \{w_0^1, w_0^1, w_0^2\}, \ldots, \{w_0^1, w_0^2, w_0^0\}, \ldots, \{w_1^2, w_2^0, w_2^1\}, \ldots, \{w_2^0, w_2^1, w_2^2\}\} = \{S_0, \ldots, S_{83}\}$.

Three sets among these 84 ones associate each trinucleotide with a frame and each frame with a permuted trinucleotide by respecting the definition of trinucleotide (left circular) permutation (see Section 1). These three interesting sets are $S_{21} = \{w_0^0, w_1^1, w_2^2\}$, $S_{43} = \{w_0^1, w_1^2, w_2^0\}$ and $S_{52} = \{w_0^2, w_1^0, w_2^1\}$. Therefore, in these three sets, one relation between a trinucleotide and its frame allows the two others relations between the permuted trinucleotides and their frames to be deduced by permutation.

In order to quantify a set $S = \{w_{i_0}^{p_0}, w_{i_1}^{p_1}, w_{i_2}^{p_2}\}$, the statistical function $F(S)$ is defined as being the mean of the function $M(w_i^p)$ with the three words $w_{i_0}^{p_0}$, $w_{i_1}^{p_1}$ and $w_{i_2}^{p_2}$

$$\begin{aligned} F(S) &= F(\{w_{i_0}^{p_0}, w_{i_1}^{p_1}, w_{i_2}^{p_2}\}) \\ &= \frac{1}{3}(M(w_{i_0}^{p_0}) + M(w_{i_1}^{p_1}) + M(w_{i_2}^{p_2})). \end{aligned} \tag{4}$$

**Property.** If the nine probabilities $o(w_i^p)$ in a group $G_j$ associated with a set $\mathcal{S}_j$, $j \in \{0,\ldots,19\}$, are identical (random case), i.e. $o(w_i^p) = o(w_{i'}^{p'})$ $\forall i,i',p,p' \in \{0,1,2\}$, then the 84 functions $F(S)$ are identical and equal to $F(S_k) = 1/3$ $\forall k \in \{0,\ldots,83\}$ (proof obvious). Therefore, the 20 sets $\mathcal{S}_j$ can be compared to this random value 1/3. This interesting property allows the method FPTF to be sensitive.

The set $S_{\max}$ having the highest value with the function $F(S)$ among the 84 sets $S$, i.e. with the first rank $\mathrm{Rk} = 1$, is defined by

$$S_{\max} = S_{k'} \quad \text{such that } F(S_{k'}) = \underset{k=0}{\overset{83}{\mathrm{MAX}}}\{F(S_k)\}. \tag{5}$$

Very unexpectedly, in the majority of the cases with the $20 \times 175 = 3500$ groups $G$ in the 175 genomes $\mathcal{G}$, the set $S_{\max}$ is one of the three interesting sets $S_{21}$, $S_{43}$ and $S_{52}$ (see the results

in Section 3). Otherwise, the preferential set $S_{\text{pref}}$ among $S_{21}$, $S_{43}$ and $S_{52}$ is chosen such that

$$S_{\text{pref}} = S_{k'} \quad \text{such that } F(S_{k'}) = \underset{k=21,43,52}{\text{MAX}}\{F(S_k)\} \tag{6}$$

and the rank Rk associated with its value $F(S_{\text{pref}})$ among the 84 values $F(S)$ is determined. $S_{\text{pref}}$ has the first rank Rk = 1 when $F(S_{\text{pref}}) = F(S_{\max})$.

The method FPTF allows the identification of 20 preferential sets $S_{\text{pref}}$ of three trinucleotides in a genome such that in each set $S_{\text{pref}}$, three permuted trinucleotides are assigned to three different frames. Therefore, three sets $X_0(G)$, $X_1(G)$ and $X_2(G)$ of 20 trinucleotides can be associated with the frames 0, 1 and 2, respectively, of genes in a genome $G$. Each set $X_0(G)$, $X_1(G)$ and $X_2(G)$ is a potential circular code.

### 2.2. Circular code

#### 2.2.1. Definition

**Notations.** $A$ being a finite alphabet, $A^*$ denotes the words over $A$ of finite length including the empty word of length 0 and $A^+$, the words over $A$ of finite length $\geq 1$. Let $w_1 w_2$ be the concatenation of the two words $w_1$ and $w_2$.

A subset $X$ of $A^+$ is a circular code if $\forall n,m \geq 1$, $x_1,x_2,\ldots,x_n$, $y_1,y_2,\ldots,y_m \in X$, $r \in A^*$ and $s \in A^+$, the equalities $sx_2x_3\ldots x_n$ $r = y_1y_2\ldots y_m$ and $x_1 = rs$ imply $n = m$, $r = 1$ and $x_i = y_i$, $1 \leq i \leq n$ (Béal, 1993; Berstel and Perrin, 1985). In other terms, every word over $A$ "written on a circle" has at most one decomposition (factorization) over $X$. Therefore, the construction frame of any word generated by a circular code $X$ (precisely, of any concatenation of words of a circular code $X$) can be retrieved as the generated word has a unique decomposition over $X$. In the following, $X$ will be a set of words of length 3 over $A = \{\text{A,C,G,T}\}$ as genes are concatenations of trinucleotides.

By excluding the four trinucleotides $w = lll$, $l \in A$, and by gathering the 60 remaining trinucleotides in 20 sets of three trinucleotides such that, in each set, the three trinucleotides are deduced from each other by permutation, a potential circular code has at most one trinucleotide per set. Therefore, there are $3^{20} \approx 3.5$ billions potential circular codes.

#### 2.2.2. Maximal circular code

A finite circular code is defined to be maximal if it is not contained in a larger finite circular code, i.e. in a circular code with more words. For words of length 3 over a four-letter alphabet, a circular code has at most 20 words (Béal, 1993; Berstel and Perrin, 1985). Then, any 20-long circular code is maximal.

#### 2.2.3. Flower automaton

In order to verify that a set $X(G)$ of trinucleotides identified by the method FPTF in a genome $G$ is a circular code, its associated flower automaton must be constructed (Béal, 1993; Berstel and Perrin, 1985). The flower automaton $F(X(G))$ associated with a set $X(G)$ has a particular state labelled 1 and cycles issued from this state 1 labelled by words of $X(G)$. Fig. 2 gives an example of flower automaton with the bacterial genome Fusobacterium nucleatum (AE009951) (associated with the $C^3$ code $C_{37}$ in Table 3a). Therefore, to prove that "$X(G)$ is a circular code" is equivalent to prove that $F(X(G))$ does not contain two cycles labelled with the same word.

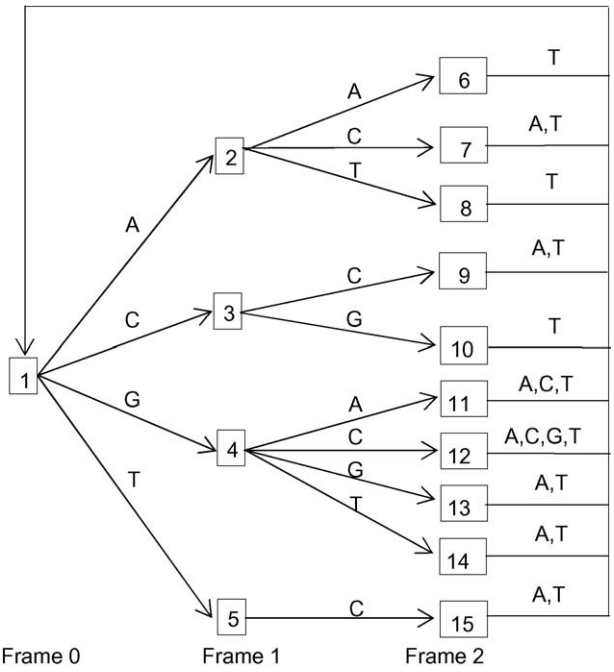


Fig. 2. Flower automaton of the bacterial genome Fusobacterium nucleatum (AE009951) (associated with the $C^3$ code $C_{37}$ in Table 3a).

#### 2.2.4. Window of a circular code

The decomposition of a word $w$ into words of a circular code $X$ is unique. Then, its construction frame formed by a concatenation of words over $X$ has to be decided. This decomposition can still be ambiguous after the reading of a few letters. For example, the bi-infinite word $w = \ldots\text{ACTGTTC}\ldots$ can be factorized in several ways: …,ACT,GTT,C… or …A,CTG,TTC,… If $X$ contains the two words {CTG,TTC}, then only the second factorization of $w$ is possible. However, some additional constraints must be also considered, in particular $X$ must contain a word finishing by A and not simultaneously the two words ACT and GTT which occur in a shifted frame of $w$. In contrast, if $X$ contains the four words {ACT,GTT,CTG,TTC}, then two factorizations of $w$ are possible. However, as the decomposition into words of a circular code is unique, more letters must be read. The window $W$ of a circular code $X$ is the series of letters which must be read in order to retrieve the construction frame of any word generated by $X$. Then, the minimal window length $|W|$ of $X$ is the size of the longest ambiguous word more one letter. This length $|W|$ depends on the code $X$.

In general, a window cannot be defined for a circular code. However, the circular codes which will be identified in bacterial genomes are all finite and uniform as all their words are trinucleotides, i.e. words with the same length of three letters. Therefore, it exists a window $W$ for each code found. A finite uniform circular code is also a finite interpreting delay code (Guesnet, 2000). The delay is the minimal number of words which must be read for retrieving the construction frame. The

notion of delay is similar to the window length. The delay is a number of words while the window length, a number of letters. With circular codes composed of trinucleotides, i.e. words of three letters over a four-letter alphabet, it is equal to the ceil of the window length divided by three and the minimal window length $|W|$ is less than or equal to 13 letters, i.e. $|W| \leq 13$ (four cycles of length 3 in the flower automaton more one letter). Therefore, only the window lengths, more precise than the delays, will be determined with the identified circular codes.

### 2.2.5. $C^3$ codes

The three sets $X_0(G)$, $X_1(G)$ and $X_2(G)$ of 20 words in the frames 0, 1 and 2, respectively, of genes in a genome $G$ which will be identified by the method FPTF, are invariant by permutation, i.e. $P(X_0(G)) = X_1(G)$, $P(X_1(G)) = X_2(G)$ and $P(X_2(G)) = X_0(G)$. A $C^3$ code is a particular circular code such that the three sets obtained by permutation, are also circular codes. Therefore, if $X_0(G)$, $X_1(G)$ and $X_2(G)$ are circular codes, then $X_0(G)$, $X_1(G)$ and $X_2(G)$ are $C^3$ codes. As the circular code $X_0(G)$ is coding for the reading frame (frame 0) in genes, i.e. the most important frame, it is considered as the main $C^3$ code.

### 2.2.6. Data acquisition

Circular codes are searched in 175 complete bacterial genomes $G$ sequenced at the time of writing this article, i.e. in 483,926 genes representing 523,375 kb. In all these genomes, the genes extracted from both DNA strands begin obligatorily with a start codon ATG, GTG and TTG, and end with a stop codon TAA, TAG and TGA. Genes containing frameshifts are eliminated. These large gene populations allow having stable frequencies leading to significant statistical results.

## 3. Results

### 3.1. Identification of three subsets of 20 trinucleotides in the three frames of genes in bacterial genomes

The trinucleotide occurrence frequencies $o(w_i^p)$ are computed in the three frames of genes in the 175 bacterial genomes $G$. As an example, Table 1 gives these frequencies $o(w_i^p)$ in the genome Fusobacterium nucleatum (AE009951).

**Remark.** The frequencies of the three stop codons TAA, TAG and TGA in frame 0 are equal to 0 in all genomes (see also the example in Table 1).

For each genome $G$ and for each group $G$, among 20, of nine trinucleotides $w_i^p$, the function $F$ (4) using the frequencies $o(w_i^p)$ is computed for the 84 sets $S$, i.e. $F(S_0)$,. . .,$F(S_{83})$, and the preferential set $S_{\text{pref}}$ and its rank Rk among 84 are determined by formula (6). With the previous genome AE009951, Table 2 gives, for each group $G$, the values of the function $F$ with the three sets $S_{21}$, $S_{43}$ and $S_{52}$, i.e. $F(S_{21})$, $F(S_{43})$ and $F(S_{52})$, and the selected set $S_{\text{pref}}$ with its rank among the 84 sets $S$. Thus, 20 preferential sets $S_{\text{pref}}$ of three permuted trinucleotides are identified in each genome $G$.

On the whole, there are $20 \times 175 = 3500$ groups $G$ in the 175 genomes $G$. The preferential sets $S_{\text{pref}}$ with the first rank Rk = 1,

Table 1
Trinucleotide occurrence frequencies (%) per frame in the bacterial genome Fusobacterium nucleatum (AE009951)

| | | Fusobacterium nucleatum (AE009951) | | |
|---|---|---|---|---|
| | | Frame 0 | Frame 1 | Frame 2 |
| $S_0$ | AAC | 0.71 | 1.3 | **2.47** |
| | ACA | **2.36** | 0.71 | 1.5 |
| | CAA | 1.97 | **3.36** | 0.71 |
| $S_1$ | AAG | 1.58 | **5.55** | 2.62 |
| | AGA | 2.79 | 1.06 | **6.72** |
| | GAA | **6.99** | 2.89 | 1.36 |
| $S_2$ | AAT | **5.68** | 3.25 | 5.33 |
| | ATA | 4.89 | **6.1** | 1.92 |
| | TAA | 0 | 5.53 | **5.68** |
| $S_3$ | ACC | 0.13 | 0.28 | **0.94** |
| | CCA | **1.21** | 0.21 | 0.26 |
| | CAC | 0.17 | **0.73** | 0.26 |
| $S_4$ | ACG | 0.05 | **0.12** | 0.03 |
| | CGA | 0.02 | 0.08 | **0.13** |
| | GAC | **0.56** | 0.62 | 0.38 |
| $S_5$ | ACT | **2.31** | 0.82 | 1.32 |
| | CTA | 0.78 | **2.65** | 0.58 |
| | TAC | 0.5 | 1.14 | **1.74** |
| $S_6$ | AGC | 0.26 | 0.28 | **2.62** |
| | GCA | **2.65** | 0.17 | 0.3 |
| | CAG | 0.19 | **2.84** | 0.24 |
| $S_7$ | AGG | 0.23 | 0.8 | **2.79** |
| | GGA | **3.76** | 0.5 | 1.09 |
| | GAG | 0.89 | **2.22** | 0.39 |
| $S_8$ | AGT | 1.68 | 0.53 | **2.91** |
| | GTA | **2.25** | 1.41 | 0.35 |
| | TAG | 0 | **4.44** | 1.72 |
| $S_9$ | ATC | 0.59 | 1.17 | **1.29** |
| | TCA | **1.93** | 0.52 | 1.29 |
| | CAT | 1.01 | **1.2** | 0.4 |
| $S_{10}$ | ATG | 2.31 | **4.76** | 0.44 |
| | TGA | 0 | 1.47 | **5.34** |
| | GAT | **4.83** | 0.85 | 0.97 |
| $S_{11}$ | ATT | **4.47** | 3.42 | 4.16 |
| | TTA | 5.68 | **4.28** | 1.87 |
| | TAT | 3.93 | 2.5 | **5.31** |
| $S_{12}$ | CCG | 0.02 | **0.04** | 0.01 |
| | CGC | 0.01 | 0.02 | **0.06** |
| | GCC | **0.27** | 0.08 | 0.2 |
| $S_{13}$ | CCT | **1.26** | 0.26 | 0.15 |
| | CTC | 0.07 | **0.85** | 0.34 |
| | TCC | 0.11 | 0.22 | **1.18** |
| $S_{14}$ | CGG | 0 | **0.06** | 0.07 |
| | GGC | 0.21 | 0.18 | **0.5** |
| | GCG | **0.07** | 0.04 | 0.02 |
| $S_{15}$ | CGT | **0.14** | 0.04 | 0.05 |
| | GTC | 0.21 | **0.33** | 0.25 |
| | TCG | 0.06 | 0.1 | **0.11** |
| $S_{16}$ | CTG | 0.11 | **2.36** | 0.13 |
| | TGC | 0.08 | 0.37 | **2.29** |
| | GCT | **2.47** | 0.27 | 0.33 |
| $S_{17}$ | CTT | 1.82 | **2.11** | 0.97 |
| | TTC | 0.64 | 1.22 | **2.16** |
| | TCT | **1.94** | 0.67 | 0.98 |
| $S_{18}$ | GGT | **1.91** | 0.22 | 0.48 |
| | GTG | 0.43 | **1.48** | 0.1 |
| | TGG | 0.62 | 1.26 | **3.01** |
| $S_{19}$ | GTT | **3.22** | 1.2 | 0.77 |
| | TTG | 0.94 | **4.66** | 0.76 |
| | TGT | 0.69 | 0.69 | **2.64** |
| | AAA | 8.54 | 7.4 | 8.75 |
| | CCC | 0.06 | 0.07 | 0.23 |
| | GGG | 0.46 | 0.42 | 0.47 |
| | TTT | 4.28 | 3.62 | 5.56 |

Three trinucleotides invariant by permutation are gathered in a set $S$. The frequencies in bold are the values selected by the function $F$ (4) given in Table 2.

Table 2
Preferential sets $S_{pref}$ in the bacterial genome Fusobacterium nucleatum (AE009951)

| | | Fusobacterium nucleatum (AE009951) | |
|---|---|---|---|
| | | Function, $F$ | Rank, Rk |
| $G_0$ | $AAC^0$; $ACA^1$; $CAA^2$ | 0.143 | |
| | $AAC^1$; $ACA^2$; $CAA^0$ | 0.316 | |
| | $AAC^2$; $ACA^0$; $CAA^1$ | **0.541** | 1 |
| $G_1$ | $AAG^0$; $AGA^1$; $GAA^2$ | 0.127 | |
| | $AAG^1$; $AGA^2$; $GAA^0$ | **0.609** | 1 |
| | $AAG^2$; $AGA^0$; $GAA^1$ | 0.264 | |
| $G_2$ | $AAT^0$; $ATA^1$; $TAA^2$ | **0.461** | 1 |
| | $AAT^1$; $ATA^2$; $TAA^0$ | 0.124 | |
| | $AAT^2$; $ATA^0$; $TAA^1$ | 0.415 | |
| $G_3$ | $ACC^0$; $CCA^1$; $CAC^2$ | 0.147 | |
| | $ACC^1$; $CCA^2$; $CAC^0$ | 0.171 | |
| | $ACC^2$; $CCA^0$; $CAC^1$ | **0.682** | 1 |
| $G_4$ | $ACG^0$; $CGA^1$; $GAC^2$ | 0.287 | |
| | $ACG^1$; $CGA^2$; $GAC^0$ | **0.467** | 9 |
| | $ACG^2$; $CGA^0$; $GAC^1$ | 0.246 | |
| $G_5$ | $ACT^0$; $CTA^1$; $TAC^2$ | **0.565** | 1 |
| | $ACT^1$; $CTA^2$; $TAC^0$ | 0.159 | |
| | $ACT^2$; $CTA^0$; $TAC^1$ | 0.276 | |
| $G_6$ | $AGC^0$; $GCA^1$; $CAG^2$ | 0.07 | |
| | $AGC^1$; $GCA^2$; $CAG^0$ | 0.081 | |
| | $AGC^2$; $GCA^0$; $CAG^1$ | **0.849** | 1 |
| $G_7$ | $AGG^0$; $GGA^1$; $GAG^2$ | 0.091 | |
| | $AGG^1$; $GGA^2$; $GAG^0$ | 0.222 | |
| | $AGG^2$; $GGA^0$; $GAG^1$ | **0.687** | 1 |
| $G_8$ | $AGT^0$; $GTA^1$; $TAG^2$ | 0.325 | |
| | $AGT^1$; $GTA^2$; $TAG^0$ | 0.057 | |
| | $AGT^2$; $GTA^0$; $TAG^1$ | **0.617** | 1 |
| $G_9$ | $ATC^0$; $TCA^1$; $CAT^2$ | 0.161 | |
| | $ATC^1$; $TCA^2$; $CAT^0$ | 0.373 | |
| | $ATC^2$; $TCA^0$; $CAT^1$ | **0.466** | 1 |
| $G_{10}$ | $ATG^0$; $TGA^1$; $GAT^2$ | 0.224 | |
| | $ATG^1$; $TGA^2$; $GAT^0$ | **0.714** | 1 |
| | $ATG^2$; $TGA^0$; $GAT^1$ | 0.062 | |
| $G_{11}$ | $ATT^0$; $TTA^1$; $TAT^2$ | **0.398** | 7 |
| | $ATT^1$; $TTA^2$; $TAT^0$ | 0.259 | |
| | $ATT^2$; $TTA^0$; $TAT^1$ | 0.342 | |
| $G_{12}$ | $CCG^0$; $CGC^1$; $GCC^2$ | 0.304 | |
| | $CCG^1$; $CGC^2$; $GCC^0$ | **0.523** | 4 |
| | $CCG^2$; $CGC^0$; $GCC^1$ | 0.174 | |
| $G_{13}$ | $CCT^0$; $CTC^1$; $TCC^2$ | **0.739** | 1 |
| | $CCT^1$; $CTC^2$; $TCC^0$ | 0.162 | |
| | $CCT^2$; $CTC^0$; $TCC^1$ | 0.099 | |
| $G_{14}$ | $CGG^0$; $GGC^1$; $GCG^2$ | 0.172 | |
| | $CGG^1$; $GGC^2$; $GCG^0$ | **0.479** | 8 |
| | $CGG^2$; $GGC^0$; $GCG^1$ | 0.349 | |
| $G_{15}$ | $CGT^0$; $GTC^1$; $TCG^2$ | **0.458** | 4 |
| | $CGT^1$; $GTC^2$; $TCG^0$ | 0.259 | |
| | $CGT^2$; $GTC^0$; $TCG^1$ | 0.283 | |
| $G_{16}$ | $CTG^0$; $TGC^1$; $GCT^2$ | 0.095 | |
| | $CTG^1$; $TGC^2$; $GCT^0$ | **0.849** | 1 |
| | $CTG^2$; $TGC^0$; $GCT^1$ | 0.056 | |
| $G_{17}$ | $CTT^0$; $TTC^1$; $TCT^2$ | 0.317 | |
| | $CTT^1$; $TTC^2$; $TCT^0$ | **0.5** | 1 |
| | $CTT^2$; $TTC^0$; $TCT^1$ | 0.182 | |
| $G_{18}$ | $GGT^0$; $GTG^1$; $TGG^2$ | **0.678** | 1 |
| | $GGT^1$; $GTG^2$; $TGG^0$ | 0.095 | |
| | $GGT^2$; $GTG^0$; $TGG^1$ | 0.227 | |
| $G_{19}$ | $GTT^0$; $TTG^1$; $TGT^2$ | **0.67** | 1 |
| | $GTT^1$; $TTG^2$; $TGT^0$ | 0.172 | |
| | $GTT^2$; $TTG^0$; $TGT^1$ | 0.159 | |

The values of the function $F$ (4) with the three sets $S_{21}$, $S_{43}$ and $S_{52}$ are given for each group $G$. The selected set $S_{pref}$ (in bold) is associated with its rank among the 84 sets $S$.

i.e. the highest value with the function $F$, (the first three ranks $Rk \leq 3$ resp.) among 84 occur in 2285 (2804 resp.) groups $G$, i.e. 65% (80% resp.). With the given example, 15 sets $S_{pref}$ have the first rank (Table 2).

The $20 \times 175 = 3500$ preferential sets $S_{pref}$ in the 175 genomes $\mathcal{G}$ lead to 175 sets of 3 subsets $X_0(\mathcal{G})$, $X_1(\mathcal{G})$ and $X_2(\mathcal{G})$ of 20 trinucleotides associated with the frames 0, 1 and 2, respectively. All these $3 \times 175 = 525$ trinucleotide sets $X(\mathcal{G})$ are potential maximal circular codes.

### 3.2. Identification of 72 new $C^3$ codes in bacterial genomes

The flower automaton algorithm (not described here) testing if a set of words is a circular code or not, shows, very unexpectedly, that 405 identified sets $X(\mathcal{G})$ among 525, i.e. 77%, are directly maximal circular codes. These 405 codes are distributed per frame in the following way: 143 among 175 are in frame 0, i.e. 82% of the sets $X_0(\mathcal{G})$ are maximal circular codes, 138 among 175 are in frame 1, i.e. 79% of the sets $X_1(\mathcal{G})$, and 124 among 175 are in frame 2, i.e. 71% of the sets $X_2(\mathcal{G})$. Furthermore, 99 sets $X_0(\mathcal{G})$ (57%) are directly $C^3$ codes, i.e. $X_0(\mathcal{G})$, $X_1(\mathcal{G})$ and $X_2(\mathcal{G})$ are simultaneously maximal circular codes.

In other words, 99 among 175 (57%) bacterial genomes contain directly $C^3$ codes. This result is very unexpected as the occurrence probability of a $C^3$ code is very low and equal to $6.3 \times 10^{-5}$ (see Section 3.3).

For the $175 - 99 = 76$ (43%) bacterial genomes which have partial $C^3$ codes, 46 (61%) genomes have two maximal circular codes and one non-maximal circular code, 16 (21%) genomes have one maximal circular code and two non-maximal circular codes, and 14 (18%) genomes have no maximal circular code.

For the $525 - 405 = 120$ (23%) sets $X(\mathcal{G})$ which are not maximal circular codes, almost all (117, i.e. 98%) are 19-long circular codes (the last three sets being 18-long circular codes).

Such partial $C^3$ codes in 76 genomes are still unexpected (see Section 3.3).

These partial $C^3$ codes are the consequence of a random set $\mathcal{S}_{rand}$ among 20 of three permuted trinucleotides with similar frequencies in two or three frames of genes in a genome. This random case is very rare as it represents 2% of the 3500 analysed sets $\mathcal{S}$ in the 175 genomes. A random set $\mathcal{S}_{rand}$ leads to a preferential set $S_{pref}$ with a value $F(S_{pref})$ close to the random one (1/3, see Property in Section 2.1). Therefore, the determination of this particular set $S_{pref}$ becomes less decisive for identifying a complete $C^3$ code in a genome. In order to take account this rare random case, two preferential sets $S_{pref}$ and $S'_{pref}$ are considered. They lead to two sets of three subsets $X_0(\mathcal{G})$, $X_1(\mathcal{G})$ and $X_2(\mathcal{G})$ which differ then by one permuted trinucleotide per frame and which are both tested as potential $C^3$ codes. This approach allows the identification of (complete) $C^3$ codes in the 76 genomes.

The method FPTF identifies 175 $C^3$ codes in the 175 analysed bacterial genomes. Several $C^3$ codes are identical with different genomes (see Section 4). Therefore, 72 new $C^3$ codes $C_i$, $i \in \{0,\ldots,71\}$, are identified in bacterial genomes (Tables 3a and 3b). Remember that the two maximal circular codes $X_1(\mathcal{G})$ and $X_2(\mathcal{G})$ in frames 1 and 2, respectively, can be deduced from a $C^3$ code $C_i$ by permutation.

Table 3a
List of the 175 bacterial genomes $G$ associated with the 72 $C^3$ codes

| $C^3$ code | Nb of genomes | Name of genomes (EMBL identification, number of genes, size in kb) |
|---|---|---|
| $C_0$ | 17 | Bordetella bronchiseptica RB50 (BX470250, 5018 g, 5339 kb), Bordetella parapertussis12822 (BX470249, 4627 g, 4774 kb), Bordetella pertussis Tohama I (BX470248, 4083 g, 4086 kb), Bradyrhizobium japonicum USDA110 (BA000040, 8317 g, 9106 kb), Caulobacter crescentus CB15 (AE005673, 3737 g, 4017 kb), Chromobacterium violaceum ATCC12472 (AE016825, 4407 g, 4751 kb), Desulfovibrio vulgaris subsp vulgaris Hildenborough (AE017285, 3380 g, 3571 kb), Leifsonia xyli subsp xyli CTCB07 (AE016822, 2030 g, 2584 kb), Mesorhizobium loti MAFF303099 (BA000012, 6752 g, 7036 kb), Mycobacterium avium subsp. paratuberculosisk10 (AE016958, 4350 g, 4830 kb), Pseudomonas aeruginosa PAO1 (AE004091, 5566 g, 6264 kb), Ralstonia solanacearum GMI1000 (AL646052, 3442 g, 3716 kb), Rhodopseudomonas palustris CGA009 (BX571963, 4845 g, 5459 kb), Streptomyces avermitilis (BA000030, 7575 g, 9026 kb), Streptomyces coelicolor (AL645882, 7851 g, 8668 kb), Xanthomonas axonopodis pv. citri306 (AE008923, 4312 g, 5176 kb), Xanthomonas campestris pv. campestris ATCC33913 (AE008922, 4181 g, 5076 kb) |
| $C_1$ | 14 | Erwinia carotovora subsp. atroseptica SCRI1043 (BX950851, 4519 g, 5064 kb), *Escherichia coli* CFT073 (AE014075, 5380 g, 5231 kb), *Escherichia coli* K12 MG1655 (U00096, 4255 g, 4640 kb), *Escherichia coli* O157 H7 EDL933 (AE005174, 5350 g, 5529 kb), *Escherichia coli* O157 H7 (BA000007, 5362 g, 5498 kb), Nitrosomonas europaea ATCC19718 (AL954747, 2574 g, 2812 kb), Salmonella enterica CT18 (AL513382, 4606 g, 4809 kb), Salmonella enterica subsp. enterica serovar Typhi Ty2 (AE014613, 4324 g, 4792 kb), Salmonella typhimurium LT2 (AE006468, 4453 g, 4857 kb), Shigella flexneri 2a2457T (AE014073, 4074 g, 4599 kb), Shigella flexneri 2a301 (AE005674, 4434 g, 4607 kb), Thermosynechococcus elongatus BP-1 (BA000039, 2476 g, 2594 kb), Treponema pallidum subsp. pallidum Nichols (AE000520, 1031 g, 1138 kb), Wolinella succinogenes DSM1740 (BX571656, 2044 g, 2110 kb) |
| $C_2$ | 12 | Campylobacter jejuni subsp jejuni NCTC11168 (AL111168, 1654 g, 1641 kb), Chlamydophila caviae GPIC (AE015925, 998 g, 1173 kb), Haemophilus influenzae RdKW20 (L42023, 1709 g, 1830 kb), Onion yellows phytoplasma OY-M (AP006628, 754 g, 861 kb), Staphylococcus aureus MRSA252 (BX571856, 2834 g, 2903 kb), Staphylococcus aureus MSSA476 (BX571857, 2649 g, 2800 kb), Staphylococcus aureus Mu50 (BA000017, 2699 g, 2879 kb), Staphylococcus aureus MW2 (BA000033, 2632 g, 2820 kb), Staphylococcus aureus N315 (BA000018, 2593 g, 2815 kb), Staphylococcus epidermidis ATCC12228 (AE015929, 2419 g, 2499 kb), Ureaplasma parvum serovar 3ATCC700970 (AF222894, 611 g, 752 kb), Yersinia pseudotuberculosis IP32953 (BX936398, 3983 g, 4745 kb) |
| $C_3$ | 9 | Chlamydia muridarum Nigg (AE002160, 904 g, 1073 kb), Chlamydia pneumoniae CWL029 (AE001363, 1052 g, 1230 kb), Chlamydia trachomatis D/UW-3/CX (AE001273, 896 g, 1043 kb), Chlamydophila pneumoniae AR39 (AE002161, 1110 g, 1230 kb), Chlamydophila pneumoniae J138 (BA000008, 1069 g, 1227 kb), Chlamydophila pneumoniae TW-183 (AE009440, 1113 g, 1226 kb), Haemophilus ducreyi 35000HP (AE017143, 1717 g, 1699 kb), Nostoc sp. PCC7120 (BA000019, 5372 g, 6414 kb), Parachlamydia sp. UWE25 (BX908798, 2031 g, 2414 kb) |
| $C_4$ | 7 | Bacillus anthracis Ames (AE016879, 5313 g, 5227 kb), Bacillus anthracis Ames Ancestor (AE017334, 5311 g, 5227 kb), Bacillus anthracis Sterne (AE017225, 5288 g, 5229 kb), Bacillus cereus ATCC10987 (AE017194, 5606 g, 5224 kb), Bacillus cereus ATCC14579 (AE016877, 5234 g, 5412 kb), Bacillus cereus ZK (CP000001, 5134 g, 5301 kb), Bacillus thuringiensis serovar konkukian97-27 (AE017355, 5117 g, 5238 kb) |
| $C_5$ | 6 | Lactobacillus johnsonii NCC533 (AE017198, 1821 g, 1993 kb), Mycoplasma mycoides subsp. mycoides SCPG1 (BX293980, 1016 g, 1212 kb), Mycoplasma pulmonis UABCTIP (AL445566, 782 g, 964 kb), Rickettsia prowazekii Madrid E (AJ235269, 835 g, 1112 kb), Rickettsia typhi Wilmington (AE017197, 841 g, 1111 kb), Wolbachia endosymbiont of Drosophila melanogaster (AE017196, 1195 g, 1268 kb) |
| $C_6$ | 4 | Mycobacterium bovis AF2122/97 (BX248333, 3953 g, 4345 kb), Mycobacterium tuberculosis CDC1551 (AE000516, 4187 g, 4404 kb), Mycobacterium tuberculosis H37Rv (AL123456, 3999 g, 4412 kb), Pseudomonas putida KT2440 (AE015451, 5350 g, 6182 kb) |
| $C_7$ | 4 | Bartonella quintana Toulouse (BX897700, 1308 g, 1581 kb), Lactococcus lactis subsp. lactis IL1403 (AE005176, 2266 g, 2366 kb), Streptococcus agalactiae 2603V/R (AE009948, 2124 g, 2160 kb), Streptococcus agalactiae NEM316 (AL732656, 2134 g, 2211 kb) |
| $C_8$ | 4 | Brucella melitensis 16 M chromosome I (AE008917, 2059 g, 2117 kb), Brucella melitensis 16 M chromosome II (AE008918, 1139 g, 1178 kb), Brucella suis 1330 chromosome I (AE014291, 2124 g, 2108 kb), Brucella suis 1330 chromosome II (AE014292, 1148 g, 1207 kb) |
| $C_9$ | 4 | Leptospira interrogans serovar Copenhageni Fiocruz L1-130 chromosome I (AE016823, 3394 g, 4277 kb), Leptospira interrogans serovar Copenhageni Fiocruz L1-130 chromosome II (AE016824, 264 g, 350 kb), Leptospira interrogans serovar lai56601 chromosome I (AE010300, 4358 g, 4332 kb), Leptospira interrogans serovar lai56601 chromosome II (AE010301, 367 g, 359 kb) |
| $C_{10}$ | 4 | Streptococcus pyogenes M1GAS (AE004092, 1696 g, 1852 kb), Streptococcus pyogenes MGAS315 (AE014074, 1865 g, 1901 kb), Streptococcus pyogenes MGAS8232 (AE009949, 1845 g, 1895 kb), Streptococcus pyogenes SSI-1 (BA000034, 1861 g, 1894 kb) |
| $C_{11}$ | 3 | Agrobacterium Tumefaciens C58 circular Washington (AE008688, 2785 g, 2841 kb), Agrobacterium tumefaciens C58 linear chromosome (AE008689, 1876 g, 2076 kb), Sinorhizobium meliloti 1021 (AL591688, 3341 g, 3654 kb) |
| $C_{12}$ | 3 | Borrelia burgdorferi B31 (AE000783, 850 g, 911 kb), Borrelia garinii PBi (CP000013, 832 g, 904 kb), Prochlorococcus marinus CCMP1986 (BX548174, 1717 g, 1658 kb) |
| $C_{13}$ | 3 | Neisseria meningitidis MC58 (AE002098, 2025 g, 2272 kb), Neisseria meningitides Z2491 (AL157959, 2121 g, 2184 kb), Pirellula sp.1 (BX119912, 7325 g, 7146 kb) |
| $C_{14}$ | 3 | Photorhabdus luminescens subsp. laumondii TTO1 (BX470251, 4905 g, 5689 kb), Streptococcus pneumoniae R6 (AE007317, 2043 g, 2039 kb), Streptococcus pneumoniae TIGR4 (AE005672, 2094 g, 2161 kb) |
| $C_{15}$ | 3 | Vibrio parahaemolyticus RIMD2210633 chromosome 2 (BA000032, 1752 g, 1877 kb), Vibrio vulnificus CMCP6 chromosome I (AE016795, 2972 g, 3282 kb), Vibrio vulnificus YJ016 chromosome I (BA000037, 3262 g, 3355 kb) |
| $C_{16}$ | 3 | Yersinia pestis biovar Medievalis91001 (AE017042, 3895 g, 4595 kb), Yersinia pestis CO92 (AL590842, 4034 g, 4654 kb), Yersinia pestis KIM (AE009952, 4090 g, 4601 kb) |
| $C_{17}$ | 3 | Mesoplasma florum L1 (AE017263, 683 g, 793 kb), Mycoplasma mobile 163K (AE017308, 633 g, 777 kb), Mycoplasma penetrans HF-2 (BA000026, 1037 g, 1359 kb) |

Table 3a (*Continued*)

| $C^3$ code | Nb of genomes | Name of genomes (EMBL identification, number of genes, size in kb) |
|---|---|---|
| $C_{18}$ | 3 | Buchnera aphidicola Sg (AE013218, 545 g, 641 kb), Buchnera aphidicola APS (BA000003, 564 g, 641 kb), Buchnera aphidicola Bp (AE016826, 504 g, 616 kb) |
| $C_{19}$ | 2 | Deinococcus radiodurans R1 chromosome 1 (AE000513, 2579 g, 2649 kb), Deinococcus radiodurans R1 chromosome 2 (AE001825, 357 g, 412 kb) |
| $C_{20}$ | 2 | Helicobacter pylori 26695 (AE000511, 1566 g, 1668 kb), Helicobacter pylori J99 (AE001439, 1505 g, 1644 kb) |
| $C_{21}$ | 2 | Xylella fastidiosa 9a5c (AE003849, 2767 g, 2679 kb), Xylella fastidiosa Temecula1 (AE009442, 2034 g, 2520 kb) |
| $C_{22}$ | 2 | Vibrio cholerae O1 biovar N16961 chromosome I (AE003852, 2736 g, 2961 kb), Vibrio cholerae O1 biovar N16961 chromosome II (AE003853, 1092 g, 1072 kb) |
| $C_{23}$ | 2 | Vibrio vulnificus CMCP6 chromosome II (AE016796, 1565 g, 1845 kb), Vibrio vulnificus YJ016 chromosome II (BA000038, 1697 g, 1857 kb) |
| $C_{24}$ | 2 | Chlorobium tepidum TLS (AE006470, 2252 g, 2155 kb), Geobacter sulfurreducens PCA (AE017180, 3447 g, 3814 kb) |
| $C_{25}$ | 2 | Enterococcus faecalis V583 (AE016830, 3113 g, 3218 kb), Rickettsia conorii Malish 7 (AE006914, 1375 g, 1269 kb) |
| $C_{26}$ | 2 | Helicobacter hepaticus ATCC51449 (AE017125, 1875 g, 1799 kb), Streptococcus mutans UA159 (AE014133, 1960 g, 2031 kb) |
| $C_{27}$ | 2 | Tropheryma whipplei TW08/27 (BX072543, 788 g, 926 kb), Tropheryma whipplei Twist (AE014184, 808 g, 927 kb) |
| $C_{28}$ | 2 | Gloeobacter violaceus PCC7421 (BA000045, 4430 g, 4659 kb), Pseudomonas syringae pv. tomato DC3000 (AE016853, 5471 g, 6397 kb) |
| $C_{29}$ | 2 | Listeria monocytogenes EGD-e (AL591824, 2855 g, 2945 kb), Listeria monocytogenes 4bF2365 (AE017262, 2822 g, 2905 kb) |
| $C_{30}$ | 2 | Corynebacterium glutamicum ATCC13032 (BA000036, 3099 g, 3309 kb), Corynebacterium glutamicum ATCC13032 4-5 (BX927147, 3058 g, 3283 kb) |
| $C_{31}$ | 2 | Burkholderia pseudomallei K96243 chr. 1 (BX571965, 3503 g, 4075 kb), Burkholderia pseudomallei K96243 chr. 2 (BX571966, 2445 g, 3173 kb) |
| $C_{32}$ | 1 | Thermotoga maritima MSB8 (AE000512, 1846 g, 1861 kb) |
| $C_{33}$ | 1 | Aquifex aeolicus VF5 (AE000657, 1522 g, 1551 kb) |
| $C_{34}$ | 1 | Clostridium acetobutylicum ATCC824 (AE001437, 3672 g, 3941 kb) |
| $C_{35}$ | 1 | Pasteurella multocida PM70 (AE004439, 2014 g, 2257 kb) |
| $C_{36}$ | 1 | Thermoanaerobacter tengcongensis MB4 (AE008691, 2588 g, 2689 kb) |
| $C_{37}$ | 1 | Fusobacterium nucleatum subsp. nucleatum ATCC25586 (AE009951, 2068 g, 2174 kb) |
| $C_{38}$ | 1 | Bifidobacterium longum NCC2705 (AE014295, 1727 g, 2257 kb) |
| $C_{39}$ | 1 | Shewanella oneidensis MR-1 (AE014299, 4630 g, 4970 kb) |
| $C_{40}$ | 1 | Mycoplasma gallisepticum R (AE015450, 726 g, 996 kb) |
| $C_{41}$ | 1 | Porphyromonas gingivalis W83 (AE015924, 1909 g, 2343 kb) |
| $C_{42}$ | 1 | Clostridium tetani E88 (AE015927, 2373 g, 2799 kb) |
| $C_{43}$ | 1 | Bacteroides thetaiotaomicron VPI-5482 (AE015928, 4778 g, 6260 kb) |
| $C_{44}$ | 1 | Coxiella burnetii RSA493 (AE016828, 2010 g, 1995 kb) |
| $C_{45}$ | 1 | Prochlorococcus marinus subsp. marinus CCMP1375 (AE017126, 1882 g, 1751 kb) |
| $C_{46}$ | 1 | Thermus thermophilus HB27 (AE017221, 1982 g, 1895 kb) |
| $C_{47}$ | 1 | Treponema denticola ATCC35405 (AE017226, 2767 g, 2843 kb) |
| $C_{48}$ | 1 | Propionibacterium acnes KPA171202 (AE017283, 2297 g, 2560 kb) |
| $C_{49}$ | 1 | Bacillus subtilis168 (AL009126, 4109 g, 4215 kb) |
| $C_{50}$ | 1 | Mycobacterium leprae TN (AL450380, 2720 g, 3268 kb) |
| $C_{51}$ | 1 | Listeria innocua Clip11262 (AL592022, 2981 g, 3011 kb) |
| $C_{52}$ | 1 | Lactobacillus plantarum WCFS1 (AL935263, 3051 g, 3308 kb) |
| $C_{53}$ | 1 | Bacillus halodurans C-125 (BA000004, 4066 g, 4202 kb) |
| $C_{54}$ | 1 | Clostridium perfringens13 (BA000016, 2660 g, 3031 kb) |
| $C_{55}$ | 1 | Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis (BA000021, 615 g, 698 kb) |
| $C_{56}$ | 1 | Synechocystis sp. PCC6803 (BA000022, 3171 g, 3573 kb) |
| $C_{57}$ | 1 | Oceanobacillus iheyensis (BA000028, 3496 g, 3631 kb) |
| $C_{58}$ | 1 | Vibrio parahaemolyticus RIMD2210633 chromosome 1 (BA000031, 3080 g, 3289 kb) |
| $C_{59}$ | 1 | Corynebacterium efficiens YS-314 (BA000035, 2942 g, 3147 kb) |
| $C_{60}$ | 1 | Corynebacterium diphtheriae NCTC13129 (BX248353, 2400 g, 2489 kb) |
| $C_{61}$ | 1 | Blochmannia floridanus (BX248583, 589 g, 706 kb) |
| $C_{62}$ | 1 | Synechococcus sp. WH8102 (BX548020, 2527 g, 2434 kb) |
| $C_{63}$ | 1 | Prochlorococcus marinus MIT9313 (BX548175, 2274 g, 2411 kb) |
| $C_{64}$ | 1 | Bdellovibrio bacteriovorus HD100 (BX842601, 3583 g, 3783 kb) |
| $C_{65}$ | 1 | Bartonella henselaeHouston-1 (BX897699, 1612 g, 1931 kb) |
| $C_{66}$ | 1 | Photobacterium profundum SS9 chromosome 1 (CR354531, 3416 g, 4085 kb) |
| $C_{67}$ | 1 | Photobacterium profundum SS9 chromosome 2 (CR354532, 1997 g, 2238 kb) |
| $C_{68}$ | 1 | Desulfotalea psychrophila LSv54 (CR522870, 3118 g, 3523 kb) |
| $C_{69}$ | 1 | Acinetobacter sp. ADP1 (CR543861, 3325 g, 3599 kb) |
| $C_{70}$ | 1 | Mycoplasma genitalium G-37 (L43967, 480 g, 580 kb) |
| $C_{71}$ | 1 | Mycoplasma pneumoniae M129 (U00089, 688 g, 816 kb) |

Table 3b
List of the 72 $C^3$ codes in the 175 bacterial genomes $G$

| $C^3$ codes | Nb of genomes | List of the 20 trinucleotides | $\|W_0\|$ | $\|W_1\|$ | $\|W_2\|$ |
|---|---|---|---|---|---|
| $C_0$ | 17 | AAC AAG AAT ACC GAC TAC CAG GAG TAG ATC ATG TAT GCC CTC GGC GTC CTG TTC GTG TTG | 9 | 10 | 11 |
| $C_1$ | 14 | AAC GAA AAT ACC GAC TAC CAG GAG GTA ATC GAT ATT GCC CTC GCG GTC CTG CTT GTG GTT | 10 | 11 | 11 |
| $C_2$ | 12 | CAA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GGC CGT GCT TCT GGT GTT | 11 | 10 | 7 |
| $C_3$ | 9 | CAA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GCG CGT GCT TCT GGT GTT | 13 | 10 | 10 |
| $C_4$ | 7 | ACA GAA AAT CCA ACG ACT GCA GGA GTA CAT GAT ATT CCG CCT GCG CGT GCT TCT GGT GTT | 9 | 10 | 8 |
| $C_5$ | 6 | ACA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GGC CGT GCT TCT GGT GTT | 11 | 10 | 9 |
| $C_6$ | 4 | AAC AAG AAT ACC GAC TAC CAG GAG TAG ATC ATG TAT GCC CTC GGC GTC CTG TTC GTG GTT | 10 | 10 | 11 |
| $C_7$ | 4 | AAC AAG AAT ACC GAC TAC CAG GAG GTA ATC ATG ATT GCC CTC GGC GTC CTG TTC GTG TTG | 10 | 10 | 13 |
| $C_8$ | 4 | ACA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GCG CGT GCT CTT GGT GTT | 11 | 7 | 10 |
| $C_9$ | 4 | AAC AAG AAT ACC GAC TAC CAG GAG GTA ATC GAT TAT GCC CTC GGC GTC CTG CTT GTG GTT | 10 | 10 | 13 |
| $C_{10}$ | 4 | ACA GAA AAT ACC GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GCG GTC GCT TCT GGT GTT | 10 | 10 | 10 |
| $C_{11}$ | 3 | ACA GAA AAT ACC GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GCG CGT GCT CTT GGT GTT | 11 | 9 | 10 |
| $C_{12}$ | 3 | ACA GAA AAT CCA GAC ACT GCA GGA GTA TCA GAT ATT GCC CCT GGC GTC GCT TCT GGT GTT | 8 | 10 | 6 |
| $C_{13}$ | 3 | AAC GAA AAT ACC GAC TAC AGC GAG GTA ATC ATG ATT GCC CTC GGC GTC CTG TTC GTG TTG | 13 | 10 | 10 |
| $C_{14}$ | 3 | CAA GAA AAT CAC GAC ACT GCA GAG GTA CAT GAT ATT GCC CCT GCG CGT GCT CTT GGT GTT | 11 | 9 | 7 |
| $C_{15}$ | 3 | AAC GAA AAT ACC GAC ACT GCA GAG GTA ATC GAT ATT GCC CCT GCG GTC GCT CTT GGT GTT | 11 | 9 | 7 |
| $C_{16}$ | 3 | CAA GAA AAT CAC GAC CTA CAG GAG GTA ATC GAT ATT GCC CTC GCG GTC CTG TTC GTG GTT | 10 | 8 | 9 |
| $C_{17}$ | 3 | ACA GAA AAT CCA GAC ACT GCA GGA GTA TCA GAT ATT GCC CCT GGC CGT GCT TCT GGT GTT | 11 | 10 | 9 |
| $C_{18}$ | 3 | CAA GAA AAT CCA CGA ACT GCA GGA GTA CAT GAT ATT GCC CCT GGC CGT GCT TCT GGT GTT | 11 | 10 | 9 |
| $C_{19}$ | 2 | AAC AAG AAT ACC GAC TAC CAG GAG TAG ATC ATG ATT GCC CTC GGC GTC CTG TTC GTG TTG | 8 | 10 | 9 |
| $C_{20}$ | 2 | AAC GAA AAT ACC GAC ACT AGC GAG GTA ATC GAT ATT GCC TCC GGC GTC GCT TCT GTG TTG | 10 | 11 | 13 |
| $C_{21}$ | 2 | AAC GAA AAT ACC GAC CTA CAG GAG GTA ATC GAT ATT GCC CTC GGC GTC CTG TTC GTG GTT | 7 | 10 | 13 |
| $C_{22}$ | 2 | AAC GAA AAT ACC GAC CTA CAG GAG GTA ATC GAT ATT GCC CTC GCG GTC CTG TTC GTG GTT | 7 | 11 | 11 |
| $C_{23}$ | 2 | AAC GAA AAT ACC GAC TAC CAG GAG GTA ATC GAT ATT GCC CTC GCG GTC CTG TTC GTG GTT | 10 | 11 | 8 |
| $C_{24}$ | 2 | AAC AAG AAT ACC GAC TAC CAG GAG GTA ATC ATG ATT GCC CTC GGC GTC CTG TTC GTG GTT | 10 | 10 | 10 |
| $C_{25}$ | 2 | ACA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GGC GTC GCT TCT GGT GTT | 8 | 10 | 9 |
| $C_{26}$ | 2 | CAA GAA AAT CCA GAC ACT GCA GAG GTA CAT GAT ATT GCC CCT GGC CGT GCT CTT GGT GTT | 11 | 7 | 10 |
| $C_{27}$ | 2 | ACA GAA ATA ACC GAC ACT GCA GAG GTA ATC GAT ATT GCC CTC GCG GTC GCT CTT GGT GTT | 9 | 11 | 7 |
| $C_{28}$ | 2 | AAC GAA AAT ACC GAC TAC CAG GAG GTA ATC ATG ATT GCC CTC GGC GTC CTG TTC GTG TTG | 13 | 10 | 13 |
| $C_{29}$ | 2 | ACA GAA AAT CCA GAC ACT GCA GGA GTA ATC GAT ATT GCC CCT GGC GTC GCT CTT GGT GTT | 9 | 11 | 7 |
| $C_{30}$ | 2 | AAC GAA AAT ACC GAC TAC GCA GAG GTA ATC GAT ATT GCC CTC GGC GTC GCT TTC GTG GTT | 13 | 10 | 7 |
| $C_{31}$ | 2 | AAC AAG AAT CAC GAC TAC CAG GAG TAG ATC ATG TAT GCC CTC GGC GTC CTG TTC GTG TTG | 9 | 7 | 8 |
| $C_{32}$ | 1 | AAC GAA ATA ACC GAC TAC GCA GAG GTA ATC ATG TTA GCC CTC GCG GTC CTG TTC GTG GTT | 10 | 8 | 10 |
| $C_{33}$ | 1 | AAC GAA ATA CAC GAC TAC GCA GAG GTA ATC ATG ATT GCC CTC GCG GTC GCT TTC GTG GTT | 7 | 8 | 9 |
| $C_{34}$ | 1 | ACA GAA AAT CCA GAC ACT GCA GGA GTA TCA GAT ATT GCC CCT GCG GTC GCT CTT GGT GTT | 10 | 10 | 10 |
| $C_{35}$ | 1 | CAA GAA AAT CCA ACG ACT GCA GAG GTA CAT GAT ATT GCC CCT GCG CGT GCT CTT GGT GTT | 8 | 10 | 10 |
| $C_{36}$ | 1 | ACA GAA ATA ACC GAC ACT GCA GGA GTA ATC GAT ATT GCC CCT GCG GTC GCT TCT GTG GTT | 10 | 10 | 10 |
| $C_{37}$ | 1 | ACA GAA AAT CCA GAC ACT GCA GGA GTA TCA GAT ATT GCC CCT GCG CGT GCT TCT GGT GTT | 13 | 10 | 10 |
| $C_{38}$ | 1 | AAC AAG AAT ACC GAC TAC CAG GAG GTA ATC ATG ATT GCC TCC GGC GTC CTG TTC GTG TTG | 10 | 10 | 13 |
| $C_{39}$ | 1 | ACA GAA AAT ACC GAC ACT GCA GAG GTA ATC GAT ATT GCC TCC GCG GTC GCT TTC GGT GTT | 11 | 9 | 7 |
| $C_{40}$ | 1 | CAA GAA AAT CAC GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GGC GTC GCT CTT GGT GTT | 6 | 9 | 7 |
| $C_{41}$ | 1 | AAC GAA AAT ACC GAC TAC CAG GAG GTA ATC ATG ATT GCC CTC GGC GTC CTG TTC GTG GTT | 13 | 10 | 13 |
| $C_{42}$ | 1 | ACA GAA ATA CCA GAC ACT GCA GGA GTA TCA GAT ATT GCC CCT GCG CGT GCT TCT GGT GTT | 10 | 10 | 10 |
| $C_{43}$ | 1 | AAC GAA AAT ACC GAC ACT GCA GAG GTA ATC GAT ATT GCC CCT GGC GTC GCT CTT GTG GTT | 11 | 9 | 7 |
| $C_{44}$ | 1 | CAA GAA AAT CAC GAC ACT GCA GAG GTA CAT GAT ATT GCC CCT GCG GTC GCT CTT GTG GTT | 8 | 7 | 7 |
| $C_{45}$ | 1 | ACA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GGC CGT GCT CTT GGT GTT | 9 | 7 | 9 |
| $C_{46}$ | 1 | AAC AAG ATA CAC GAC TAC CAG GAG TAG ATC ATG ATT GCC CTC GCG GTC CTG TTC GTG TTG | 7 | 8 | 6 |
| $C_{47}$ | 1 | ACA GAA ATA ACC GAC ACT GCA GAG GTA ATC GAT ATT GCC CCT GGC GTC GCT CTT GTG GTT | 8 | 9 | 7 |
| $C_{48}$ | 1 | AAC AAG AAT ACC GAC TAC CAG GAG GTA ATC GAT ATT GCC CTC GGC GTC CTG TTC GTG GTT | 10 | 10 | 13 |
| $C_{49}$ | 1 | AAC GAA AAT ACC GAC CTA GCA GAG GTA ATC GAT ATT GCC CTC GGC GTC CTG CTT GTG GTT | 9 | 11 | 10 |
| $C_{50}$ | 1 | AAC GAA AAT ACC GAC TAC CAG GAG GTA ATC ATG ATT GCC CTC GCG GTC CTG TTC GTG TTG | 13 | 11 | 11 |
| $C_{51}$ | 1 | ACA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GGC GTC GCT CTT GGT GTT | 6 | 7 | 6 |
| $C_{52}$ | 1 | ACA GAA AAT ACC ACG ACT GCA GAG GTA ATC GAT ATT GCC TCC GCG GTC GCT TTC GGT GTT | 11 | 9 | 10 |
| $C_{53}$ | 1 | CAA GAA ATA CCA ACG CTA GCA GAG GTA ATC GAT ATT GCC CTC GCG GTC GCT CTT GTG GTT | 10 | 9 | 11 |
| $C_{54}$ | 1 | ACA GAA ATA CCA GAC ACT GCA GGA GTA TCA GAT TTA GCC CCT GGC GTC GCT TCT GGT GTT | 5 | 10 | 9 |
| $C_{55}$ | 1 | ACA GAA ATA CCA GAC ACT GCA GGA GTA TCA GAT ATT GCC CCT GGC CGT GCT TCT GGT GTT | 8 | 7 | 9 |
| $C_{56}$ | 1 | AAC GAA AAT ACC GAC ACT CAG GAG GTA ATC GAT ATT GCC CTC GCG GTC GCT CTT GTG GTT | 7 | 9 | 8 |
| $C_{57}$ | 1 | CAA GAA AAT CCA ACG ACT GCA GGA GTA CAT GAT ATT GCC CCT GCG CGT GCT TCT GGT GTT | 10 | 10 | 13 |
| $C_{58}$ | 1 | AAC GAA AAT CAC GAC CTA GCA GAG GTA CAT GAT ATT GCC CCT GCG GTC GCT CTT GGT GTT | 11 | 11 | 10 |
| $C_{59}$ | 1 | AAC AAG AAT ACC GAC TAC CAG GAG TAG ATC GAT TAT GCC CTC GCG GTC CTG TTC GTG GTT | 7 | 11 | 9 |
| $C_{60}$ | 1 | AAC GAA AAT ACC GAC ACT GCA GAG GTA ATC GAT ATT GCC CTC GGC GTC GCT CTT GTG GTT | 9 | 8 | 7 |
| $C_{61}$ | 1 | CAA GAA AAT CCA CGA ACT GCA GGA GTA CAT GAT ATT CCG CCT GCG CGT GCT TCT GGT GTT | 9 | 10 | 8 |

Table 3b (*Continued*)

| $C^3$ codes | Nb of genomes | List of the 20 trinucleotides | $\|W_0\|$ | $\|W_1\|$ | $\|W_2\|$ |
|---|---|---|---|---|---|
| $C_{62}$ | 1 | AAC AAG AAT ACC GAC TAC CAG GAG GTA ATC GAT TAT GCC CTC GGC GTC CTG TTC GTG GTT | 10 | 10 | 13 |
| $C_{63}$ | 1 | AAC GAA AAT ACC GAC CTA GCA GAG GTA ATC GAT ATT GCC CTC GGC GTC GCT CTT GTG GTT | 9 | 11 | 7 |
| $C_{64}$ | 1 | AAC GAA AAT ACC GAC TAC CAG GAG GTA ATC ATG ATT GCC TCC GCG GTC CTG TTC GTG GTT | 13 | 11 | 11 |
| $C_{65}$ | 1 | ACA GAA AAT CCA CGA ACT GCA GGA GTA CAT GAT ATT GCC CCT GCG CGT GCT CTT GGT GTT | 8 | 7 | 7 |
| $C_{66}$ | 1 | CAA GAA AAT CCA GAC ACT GCA GAG GTA CAT GAT ATT GCC CCT GCG CGT GCT CTT GGT GTT | 8 | 7 | 7 |
| $C_{67}$ | 1 | ACA GAA AAT ACC GAC ACT GCA GAG GTA ATC GAT ATT GCC CCT GCG GTC GCT CTT GGT GTT | 11 | 9 | 7 |
| $C_{68}$ | 1 | AAC GAA AAT ACC GAC TAC GCA GAG GTA ATC GAT ATT GCC CTC GGC GTC GCT CTT GGT GTT | 13 | 10 | 10 |
| $C_{69}$ | 1 | CAA GAA AAT CCA GAC CTA GCA GAG GTA CAT GAT ATT GCC CCT GCG CGT GCT CTT GGT GTT | 8 | 9 | 7 |
| $C_{70}$ | 1 | CAA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT CGC CCT GGC CGT GCT CTT GGT GTT | 9 | 7 | 11 |
| $C_{71}$ | 1 | AAC GAA AAT CAC GAC ACT CAG GAG GTA ATC GAT ATT GCC CTC GGC GTC GCT TTC GGT GTT | 13 | 13 | 10 |

For each $C^3$ code, the minimal window lengths $|W_0|$, $|W_1|$ and $|W_2|$ of $X_0(G)$, $X_1(G)$ and $X_2(G)$ in frames 0, 1 and 2, respectively, are given.

The minimal window lengths $|W_0|$, $|W_1|$ and $|W_2|$ are computed for each code $X_0(G)$, $X_1(G)$ and $X_2(G)$, respectively (Table 3b). Their lengths with the $3 \times 72 = 216$ identified codes vary between 5 ($X_0(G)$ of $C_{54}$) and 13.

### 3.3. Statistical significance of these results

The occurrence probability of a $C^3$ code is theoretically very rare: $221{,}544/3^{20} \approx 6.3 \times 10^{-5}$. This probability is obtained by computing the number of $C^3$ codes (221,544) among the $3^{20}$ potential sets of 20 trinucleotides (algorithm not described here; Arquès and Michel, 1996; Lacan and Michel, 2001).

Furthermore, the significance of the $3 \times 175 = 525$ bacterial circular codes identified in the three frames of genes in the 175 genomes, i.e. precisely the 175 sets of three subsets $X_0(G)$, $X_1(G)$ and $X_2(G)$ before the statistical treatment of the partial $C^3$ codes, is also evaluated as follows. The complete set $\mathcal{G}$ of the 175 bacterial genomes $G$ is associated with a set $\mathcal{R}$ of 175 random genomes $R$. A random genome $R$ has a number of genes identical to that of its associated genome $G$. Three sets $\mathcal{R}_N$, $\mathcal{R}_D$ and $\mathcal{R}_T$ of random genomes are generated by keeping the basic gene constraints according to the distributions of nucleotides, dinucleotides and trinucleotides respectively. The set $\mathcal{R}_N$ ($\mathcal{R}_D$ resp.) of random genomes $R_N$ ($R_D$ resp.) is constructed such that each random genome $R_N$ ($R_D$ resp.) has identical nucleotide (dinucleotide resp.) frequencies with its associated genome $G$. In order to obtain different random trinucleotide compositions from different genes, the set $\mathcal{R}_T$ of random genomes $R_T$ is constructed such that each trinucleotide in a random genome $R_T$ has a frequency randomly chosen among the 64 ones of its associated genome $G$.

**Remark.** In order to get very stable statistical results, 20 random genomes $R$ are in fact generated for one genome $G$.

For each random genome $R$ in a given set $\mathcal{R}$, the three trinucleotide sets $X_0(R)$, $X_1(R)$ and $X_2(R)$ in the three frames are determined with the statistical method FPTF. As with the bacterial genomes, the circular codes in random genomes are identified with the flower automaton algorithm. Then, for each set $\mathcal{R}$, the average lengths of the 175 codes, i.e. the average numbers of words in the codes, in each frame in the 175 random genomes $R$ are determined. Furthermore, for each set $\mathcal{R}$, the average length in the average frame (frames 0, 1 and 2), i.e. the average length of the 525 codes, is also computed. These numbers are compared with those of bacterial codes.

Table 4 shows the average lengths per frame and in the average frame for the codes in the bacterial genomes and in the three random genomes $\mathcal{R}_N$, $\mathcal{R}_D$ and $\mathcal{R}_T$ according to the nucleotide, dinucleotide and trinucleotide distributions.

In the sets $\mathcal{G}$, $\mathcal{R}_N$, $\mathcal{R}_D$ and $\mathcal{R}_T$, the average lengths of codes are almost identical in each frame and very close to the average length in the average frame which is thus a representative parameter (Table 4).

The average lengths of codes in random genomes are significantly shorter than those in bacterial genomes. Indeed, the average length in the average frame for the bacterial codes is 19.77 words and only approximately 17.6 words in random genomes, precisely 17.45, 17.41 and 17.91 words for the set $\mathcal{R}_N$, $\mathcal{R}_D$ and $\mathcal{R}_T$, respectively (Table 4).

This difference between the code lengths is even greater by considering the $C^3$ codes (three codes related by permutation): 19.5 words for the bacterial codes and only approximately 16

Table 4
Average lengths of circular codes and $C^3$ codes in the 175 bacterial genomes and in the random genomes with distributions depending on nucleotides, dinucleotides and trinucleotides, respectively

| | Average length of circular codes | Average length of $C^3$ codes |
|---|---|---|
| Set $\mathcal{G}$ of 175 bacterial genomes | 19.77 (19.81, 19.79 and 19.70 in frames 0, 1 and 2 resp.) | 19.5 |
| Set $\mathcal{R}_N$ of random genomes with a distribution depending on nucleotides | 17.45 (17.42, 17.46 and 17.47 in frames 0, 1 and 2 resp.) | 15.78 |
| Set $\mathcal{R}_D$ of random genomes with a distribution depending on dinucleotides | 17.41 (17.40, 17.42 and 17.42 in frames 0, 1 and 2 resp.) | 15.72 |
| Set $\mathcal{R}_T$ of random genomes with a distribution depending on trinucleotides | 17.91 (18.15, 17.84 and 17.83 in frames 0, 1 and 2 resp.) | 16.48 |

words for the random genomes, precisely 15.78, 15.72 and 16.48 words for the set $\mathcal{R}_N$, $\mathcal{R}_D$ and $\mathcal{R}_T$, respectively (Table 4).

These statistical evaluations demonstrate that the computed differences between the code lengths in bacterial genomes and random ones are strongly significant:

(i) the 525 bacterial codes are close to the maximality of 20 words (19.77 in Table 4),
(ii) the codes in random genomes are far from being maximal (17.6),
(iii) the bacterial partial $C^3$ codes of 19 words are still unexpected compared to the codes in random genomes.

**Remark.** As a subcode of a circular code is necessary a circular code, the probability of a trinucleotide set to be a circular code increases when its length decreases, i.e. when its number of words decreases. The number of potential subcodes of length $n$ among 20 trinucleotides increases according to $\binom{n}{20}$, thus explaining the rarity of maximal circular codes.

### *3.4. A new factorization method for retrieving the reading frames of bacterial genes by using the identified circular codes*

Genes are not "pure" circular codes as their reading frames are not only composed of 20 trinucleotides with the property of circular code. Nevertheless, as the identified bacterial $C^3$ codes contain the most important information about the trinucleotide occurrences in the three frames of genes, i.e. $3 \times 20$ trinucleotides, in each bacterial genome and as they have a particular algebraic structure, they could have a biological function in the reading synchronisation of bacterial genes. For nucleotide sequences which can be completely factorized into words of a circular code, the reading frame can be retrieved without any ambiguity after the reading of a few nucleotides (window of the code), to the maximum 13 nucleotides (Section 2.2.4). Some nucleotide regions and sites are pure circular codes (results not shown). It is (obviously) not the general case as the actual genes contain 61 codons (coding the amino acids) which could have evolved from substitutions of the common circular code (Frey and Michel, 2006). In order to apply the concept of frame retrieval with nucleotide sequences which are not pure circular codes, we have developed a simple factorization method FRM (frame retrieving method) giving the average probability of retrieving the reading frame of any words located anywhere in genes by using the bacterial $C^3$ code information, in particular its trinucleotide preferential positioning per frame.

In order to get stable and significant statistical results, the method FRM is applied to a great number of words of various lengths extracted at random positions from different genes randomly chosen in a given genome. By convention, genes begin with a start codon at position 0. Let $w_0$ be a word extracted in the reading frame (frame 0) of a gene in a genome $G$. A $C^3$ code $X_0(G)$ is associated with each genome $G$ (Section 3.2). The permutations of the code $X_0(G)$ in frame 0 lead to the codes $X_1(G)$ and $X_2(G)$ in frames 1 and 2, respectively. The two other words of $w_0$ in the two shifted frames modulo 3 are $w_1$ ($w_0$ minus its first letter) and $w_2$ ($w_0$ minus its first two letters). The endings of $w_0$, $w_1$ and $w_2$ are truncated such that their lengths are 0 modulo 3. Then, $w_0$ is factorized into words of the codes $X_0(G)$, $X_1(G)$ and $X_2(G)$, and similarly for $w_1$ and $w_2$. Therefore, a proposed frame can be inferred from the location of the words of $X_0(G)$, $X_1(G)$ and $X_2(G)$ in the words $w_0$, $w_1$ and $w_2$.

Let $N(w, X(G))$ be the number of words of a code $X(G)$ in the factorization of a word $w$. Three values $V(G)$ will be compared

$$V_0(G) = N(w_0, X_0(G)) + N(w_1, X_1(G)) + N(w_2, X_2(G)),$$

$$V_1(G) = N(w_0, X_1(G)) + N(w_1, X_2(G)) + N(w_2, X_0(G)),$$

$$V_2(G) = N(w_0, X_2(G)) + N(w_1, X_0(G)) + N(w_2, X_1(G)).$$

As the words of the code $X_0(G)$, ($X_1(G)$ and $X_2(G)$ resp.) are associated with the frame 0 (1 and 2 resp.), then a high value $V_0(G)$ ($V_1(G)$ and $V_2(G)$ resp.) suggests that the word $w$ is in frame 0 (1 and 2 resp.)

$$\text{proposed frame } i' \text{ such that } V_{i'}(G) = \mathop{\mathrm{MAX}}_{i=0}^{2}\{V_i(G)\}.$$

In order to evaluate this method FRM simply, the proposed frame of $w$ is compared to its real one. The proposed frame is considered to be retrieved correctly when it is identical to the real one and when there is no ambiguity in the choice of the highest value $V(G)$, i.e. the two highest values $V(G)$ are different. Two identical highest values may occur when $w$ has very few trinucleotides.

Finally, for the words of a given length, the average probability of retrieving the correct frame is equal to the ratio of the number of words with correct frames by the total number of words studied. The lengths of the studied words vary between 5 (one trinucleotide for $w_0$, $w_1$ and $w_2$) and 50 nucleotides.

More than 35 millions words per length were examined in the 175 bacterial genomes with this frame retrieving method FRM. Fig. 3 shows that the correct frame is retrieved with the short-

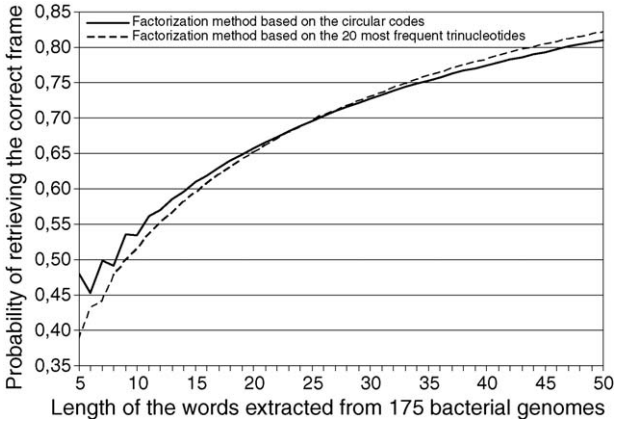


Fig. 3. Probability of retrieving the correct frame of words extracted at random positions from different genes randomly chosen in 175 bacterial genomes as a function of their length in nucleotides. Two factorization methods are studied, one based on the circular codes (thick line) and the other on the 20 most frequent trinucleotides per frame (dash line).

est words of five nucleotides in approximately half of the cases (48.0%), i.e. with a probability which is significantly higher than the random one 1/3 (one among three possibilities for choosing randomly a frame). The switchback aspect of the curve at its beginning is related to the modulo 3 truncation effect with the shorter words. The reading frame of nucleotide sequences completely factorized into words of a circular code, is retrieved in all cases with words of 13 nucleotides (Section 2.2.4), i.e. with a probability equal to 1. The average probability for finding the correct frame with words of 13 nucleotides in bacterial genes is 58.6% with this factorization method FRM. It increases as a function of the word length. For the largest studied words of 50 nucleotides, it reaches 81.0%.

These probabilities have been also computed per genome (results not shown). They present variations depending on the strength of the circular codes, i.e. according to the statistical function $F(S)$ (Section 2.1). For example, the probability for retrieving the correct frame with words of five nucleotides extracted from the 175 bacterial genomes varies between 40.4% and 64.0%, and with words of 50 nucleotides, between 61.9% and 97.8%.

Finally, these probabilities based on the bacterial $C^3$ codes are compared to those obtained with the three sets composed of the 20 most frequent trinucleotides per frame in each bacterial genome. The principle of using the most frequent trinucleotides seems a priori more powerful for retrieving the correct frame in genes. Very surprisingly, the method FRM using the circular code information leads to better results with short words less than 25 nucleotides compared to the usage of the most frequent trinucleotides (Fig. 3). The frequent trinucleotides are not circular codes as they can contain, in particular, permuted trinucleotides. The property of circular code with words greater than 25 nucleotides becomes less interesting.

## 4. Discussion

Genes in 175 bacterial genomes (483,926 genes, 523,375 kb) have been analysed with the statistical method FPTF which considers both the preferential frame of a trinucleotide and the preferential permuted trinucleotide in a frame. This approach has identified 72 new $C^3$ codes in these bacterial genomes (Table 3b). These $C^3$ codes are specific to genes as they are not significant in randomly generated genomes (Section 3.3). They may be related to variant genetic codes and different codon usage.

They occur with a great disparity in bacterial genomes. Indeed, 11 $C^3$ codes are found in half of the genomes (Table 3a). Nevertheless, several $C^3$ codes only occur once. This distribution may reflect biological interest in the choice of sequencing. Organisms widely studied are more represented (multiple lineages) and so are the corresponding codes. Codes appearing only once are often related to specific organisms and are generally strongly similar to the codes of other bacteria (results not shown).

$C^3$ codes have been searched for different chromosomes of a species (Table 3a). Nine species have two chromosomes. Six of them have identical $C^3$ codes in their two chromosomes: Brucella melitensis (AE008917 and AE008918) with a code $C_8$, Brucella suis (AE014291 and AE014292) with a code $C_8$, Deinoccocus radiodurans (AE000513 and AE001825) with a code $C_{19}$, Leptospira interrogans (AE010300, AE010301, AE016823 and AE016824) with a code $C_9$, Burkholderia pseudomallei (BX571965 and BX571966) with a code $C_{31}$, and Vibrio cholerae (AE003852 and AE003853) with a code $C_{22}$. For the last three species, the $C^3$ codes in the two chromosomes are different: Vibrio parahaemolyticus (BA000031 and BA000032) with the codes $C_{58}$ and $C_{15}$, respectively, Vibrio vulnificus (AE016795 and BA000037, and AE016796 and BA000038) with the codes $C_{15}$ and $C_{23}$, respectively, and Photobacterium profundum (CR354531 and CR354532) with the codes $C_{66}$ and $C_{67}$, respectively.

Similarly, 22 species have genomes corresponding to diverse strains or subspecies. For 21 species, the $C^3$ codes associated with different genomes of a same species are identical. Prochlorococcus marinus is the only species with different codes ($C_{12}$, $C_{45}$, $C_{63}$) associated with its different strains (AE017126, BX548174, BX548175).

Several bacterial $C^3$ codes are closed to the complementary $C^3$ code $X_0$(EUK_PRO) found in eukaryotic and prokaryotic genes (Arquès and Michel, 1996) (results not shown). Furthermore, the average code $X_0$(PRO) in the frame 0 of the 175 bacterial genomes and $X_0$(EUK_PRO) differs only from one trinucleotide: GTG in $X_0$(PRO) is replaced by GGT in $X_0$(EUK_PRO). Therefore, several bacterial $C^3$ codes could have derived by mutation from the $C^3$ code $X_0$(EUK_PRO) which is the only code with the strong property of complementarity. Such an evolutionary model has been recently proposed with archaeal circular codes (Frey and Michel, 2006).

The common and rare codons in the 72 bacterial $C^3$ codes, i.e. the trinucleotides belonging to the 72 sets $X_0(G)$ in frame 0, are the following ones (from Table 3b):

- 10 codons are absent, codon number $Nb = 0$ in these 72 codes: AGA, AGG, AGT, CGG, TAA, TCG, TGA, TGC, TGG, TGT,
- 18 codons are very rare, $0 < Nb \leq 18$ (in the first quarter): AAG, ACG, AGC, ATA, ATG, CAA, CAC, CCG, CGA, CGC, CTA, TAG, TAT, TCA, TCC, TCT, TTA, TTG,
- 18 codons are rare, $18 < Nb \leq 37$ (in the second quarter): AAC, ACA, ACC, CAG, CAT, CCA, CCT, CGT, CTC, CTG, CTT, GCG, GGA, GGC, GGT, GTG, TAC, TTC,
- six codons are common, $37 < Nb \leq 55$ (in the third quarter): ACT, ATC, GAG, GCA, GCT, GTC,
- eight codons are very common, $Nb > 55$ (in the last quarter): AAT, ATT, GAA, GAC, GAT, GCC, GTA, GTT.

The four types of nucleotides in the codons of the 72 bacterial $C^3$ codes occur in (from Table 3b):

- the 1st trinucleotide site except C in three codes and T in 25 codes,
- the 2nd trinucleotide site except G in 13 codes,
- the 3rd trinucleotide site except A in five codes, C in two codes and G in 13 codes, respectively.

In the three $C^3$ codes $C_{44}$, $C_{53}$ and $C_{56}$, there is neither T in the 1st site nor G in the 2nd site. In the five $C^3$ codes $C_{29}$, $C_{40}$, $C_{45}$, $C_{51}$ and $C_{70}$, there is neither T in the 1st site nor G in the 3rd site. In one $C^3$ code $C_{59}$, there is neither G in the 2nd site nor A in the 3rd site. Similar rules can obviously be deduced with the 72 bacterial codes $X_1(G)$ and $X_2(G)$ by permutation.

The three circular codes $X_0(G)$, $X_1(G)$ and $X_2(G)$ in a bacterial genome $G$ have 20 trinucleotides in the frames 0, 1 and 2, respectively. Therefore, a preferential frame for the 8 R/Y trinucleotides, i.e. {RRR,. . .,YYY}, over the alphabet {R,Y} (R = purine = {A,G}, Y = pyrimidine = {C,T}) can be deduced by considering for each R/Y trinucleotide, the average frame associated with the eight A/C/G/T trinucleotides specified on the R/Y trinucleotide and belonging to the codes $X_0(G)$ in frame 0, $X_1(G)$ in frame 1 and $X_2(G)$ in frame 2. For example with the code $C_0$, RRY is associated with four trinucleotides AAC, AAT, GAC and GGC in frame 0, AGC and AGT in frame 1 (CAG and TAG are in frame 0), and GAT and GGT in frame 2 (ATG and GTG are in frame 0) (Table 3b). Then, the average frame of RRY in $C_0$ is 0. All the 72 bacterial $C^3$ codes have the trinucleotide RYY in frame 0, like in the two $C^3$ codes $X_0$(EUK_PRO) and $X_0$(MIT) of mitochondria (Arquès and Michel, 1996, 1997). Its permuted trinucleotides YYR and YRY occur obviously in frames 1 and 2, respectively. 45 bacterial $C^3$ codes have the trinucleotide RRY in frame 0 and its permuted trinucleotides RYR and YRR in frames 1 and 2, respectively. 16 $C^3$ codes have the trinucleotide RYR, instead of RRY, in frame 0. For the 11 remaining codes, RRY and RYR occur identically in frame 0. There is no preferential frame for RRR and YYY. Therefore, most of the bacterial $C^3$ codes follow the pattern RNY = {RRY,RYY} (N = {R,Y}) (Eigen and Schuster, 1978) which is found in the complementary $C^3$ code $X_0$(EUK_PRO) (Arquès and Michel, 1996).

Two amino acids (AA) are never coded by the codons in the 72 bacterial $C^3$ codes: Cys and Trp (from Table 3b). These two AA have a complex chemical structure in terms of their numbers of atoms or cycles. Indeed, Cys can form disulfide linkages by reaction with another Cys and Trp is the single AA with two cycles. Six AA are always coded by these bacterial codes: Ala, Asp, Glu, Ile, Thr (except for the six codes $C_{16}$, $C_{29}$, $C_{31}$, $C_{46}$, $C_{58}$ and $C_{69}$) and Val. Ala and Asp represent the complete group of negatively charged (acidic) polar AA. These six AA are equally represented in the two classes of aminoacyl-tRNA synthetases with a class I containing Glu, Ile and Val, and a class II holding Ala, Asp and Thr (rewieved in Schimmel et al., 1993; Hartman, 1995; Saks and Sampson, 1995). These bacterial codes code for a number of AA varying from eight AA with $C_{27}$ to 15 AA with $C_{38}$.

The property of circular code in genes presents several advantages. In particular, an interpreting delay, i.e. the reading of a few nucleotides, anywhere in the sequence, permits the deciphering of the construction frame. Then, the beginning of the reading of a sequence at a start codon is no more necessary to retrieve the reading frame. The window lengths of the $3 \times 72 = 216$ bacterial codes corresponding to the longest ambiguous words more one nucleotide, vary between 5 and 13 (Table 3b). But even for codes with large windows, the long ambiguous words are rare. In the majority of the cases, the reading frame can be retrieved after the reading of about two trinucleotides only. In other words, the deciphering delay is very short.

Even for nucleotide sequences which cannot be completely factorized into words of a circular code, short words generated by a circular code distributed along genes could permit the frame synchronisation. Moreover, the $C^3$ code contains information about trinucleotides for each frame. Therefore, the words of a $C^3$ code could also mark the two other frames or amplify words in order to synchronize the current reading frame. An infinity of such words exist as they need to be generated only by words of a code. Their polymorphism makes them adaptable for a large variety of nucleotide sequences constrained by the amino acid composition. Indeed, a codon which does not belong to a circular code could be substituted by a synonymous one of a circular code.

A new factorization method FRM based on bacterial $C^3$ codes has been developed for retrieving reading frames in bacterial genes (Section 3.4). Very surprisingly, it is more powerful with short words less than 25 nucleotides than the 20 most frequent trinucleotides per frame in each bacterial genome. Furthermore, there is no constraint with the position of these short words which can be located anywhere in the sequences. Other methods can retrieve reading frames in a more reliable way but with a more complex treatment and more information on the structure of sequences. The proposed method FRM only depends on a $C^3$ code of 20 trinucleotides associated with the reading frame, the two other circular codes in the shifted frames being automatically deduced from the code in reading frame. Its principle is new and should be investigated for improving the algorithms for searching reading frames, e.g. by considering a series of short words of circular codes at different locations in a way similar to the particular sites (CAAT and TATA boxes) existing in nucleotide sequences and used for finding reading frames.

There are hints that circular codes could be issued from primitive genes, in particular their "universal" presence in genes of various genomes (archaea, prokaryotes, eukaryotes, mitochondria), their strong properties, in particular for retrieving reading frames, and their biological consequences (see above). However, it is still not known to date which biological apparatus could have used these circular codes and if their words still have a function in actual genes.

## Acknowledgement

We thank Dr A.K. Konopka for his advice.

## References

Akashi, H., Eyre-Walker, A., 1998. Translational selection and molecular evolution. Curr. Opin. Genet. Dev. 8, 688–693.

Antezana, M.A., Kreitman, M., 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J. Mol. Evol. 49, 36–43.

Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182, 45–58.

Arquès, D.G., Michel, C.J., 1997. A circular code in the protein coding genes of mitochondria. J. Theor. Biol. 189, 273–290.

Béal, M.-P., 1993. Codage Symbolique. Masson, Paris.

Berstel, J., Perrin, D., 1985. Theory of Codes. Academic Press, New York.

Berg, O.G., Silva, P.J.N., 1997. Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. Nucl. Acids Res. 25, 1397–1404.

Bulmer, M., 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.

Crick, F.H.C., Brenner, S., Klug, A., Pieczenik, G., 1976. A speculation on the origin of protein synthesis. Origins Life 7, 389–397.

Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Natl. Acad. Sci. 43, 416–421.

Eigen, M., Schuster, P., 1978. The hypercycle. A principle of natural self-organization. Part C: the realistic hypercycle. Naturwissenschaften 65, 341–369.

Fedorov, A., Saxonov, S., Gilbert, W., 2002. Regularities of context-dependent codon bias in eukaryotic genes. Nucl. Acids Res. 30, 1192–1197.

Frey, G., Michel, C.J., 2003. Circular codes in archaeal genomes. J. Theor. Biol. 223, 413–431.

Frey, G., Michel, C.J., 2006. An analytical model of gene evolution with six mutation parameters: an application to archaeal circular codes. J. Comput. Biol. Chem. 30, 1–11.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucl. Acids Res. 8, r49–r62.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucl. Acids Res. 9, r43–r74.

Guesnet, Y., 2000. On codes with finite interpreting delay: a defect theorem. Theor. Inform. Appl. 34, 47–59.

Hartman, H., 1995. Speculations on the origin of the genetic code. J. Mol. Evol. 40, 541–544.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 12–34.

Lacan, J., Michel, C.J., 2001. Analysis of a circular code model. J. Theor. Biol. 213, 159–170.

Konu, O., Li, M.D., 2002. Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. J. Mol. Evol. 54, 35–41.

Krakauer, D.C., Jansen, A.A., 2002. Red queen dynamics of protein translation. J. Theor. Biol. 218, 97–109.

Llopart, A., Aguade, M., 2000. Nucleotide polymorphism at the RpII215 gene in Drosophila subobscura: weak selection on synonymous mutations. Genetics 155, 1245–1252.

Nirenberg, M.W., Matthaei, J.H., 1961. The dependance of cell-free protein synthesis in *E. Coli* upon naturally occurring or synthetic polyribonucleotides. Proc. Natl. Acad. Sci. 47, 1588–1602.

Rogozin, I.B., Malyarchuk, B.A., Pavlov, Y.I., Milanesi, L., 2005. From context-dependance of mutations to molecular mechanisms of mutagenesis. Pac Symp. Biocomput., 409–420.

Saks, M.E., Sampson, J.R., 1995. Evolution of tRNA recognition systems and tRNA gene sequences. J. Mol. Evol. 40, 509–518.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., Sockett, R.E., 2005. Variation in the strength of selected codon usage bias among bacteria. Nucl. Acids Res. 33, 1141–1153.

Sharp, P.M., Matassi, G., 1994. Codon usage and genome evolution. Curr. Opin. Genet. Dev. 4, 851–860.

Shpaer, E.G., 1986. Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation. J. Mol. Biol. 188, 555–564.

Schimmel, P., Giegé, R., Moras, D., Yokoyama, S., 1993. An operational RNA code for amino acids and possible relationship to genetic code. Proc. Natl. Acad. Sci. 90, 8763–8768.

Smith, N.G.C., Eyre-Walker, A., 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. 18, 982–986.

Yarus, M., Folley, L.S., 1984. Sense codons are found in specific contexts. J. Mol. Biol. 182, 529–540.