Journal of Theoretical Biology 264 (2010) 613–622

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Essential molecular functions associated with the circular code evolution

Ahmed Ahmed, Gabriel Frey, Christian J. Michel *

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

ARTICLE INFO

Article history:
Received 24 July 2009
Received in revised form
17 January 2010
Accepted 5 February 2010
Available online 11 February 2010

Keywords:
Circular code
Evolution
Evolutionary trinucleotides
Circular code stability
Trinucleotide stability
Comparative genomics
Database
Molecular functions
Essential genes

ABSTRACT

A circular code is a set of trinucleotides allowing the reading frames in genes to be retrieved locally, i.e. anywhere in genes and in particular without start codons, and automatically with a window of few nucleotides. In 1996, a common circular code, called *X*, was identified in large populations of eukaryotic and prokaryotic genes. Hence, it is believed to be an ancestral structural property of genes. A new computational approach based on comparative genomics is developed to identify essential molecular functions associated with circular codes. It is based on a quantitative and sensitive statistical method (FPTF) to identify three permuted trinucleotide sets in the three frames of genes, a flower automaton algorithm to determine if a trinucleotide set is a circular code or not, and an integrated Gene Ontology and Taxonomy (iGOT) database. By carrying out automatic circular code analyses on a huge number of gene populations where each population is associated with a particular molecular function, it identifies 266 gene populations having circular codes close to *X*. Surprisingly, their molecular functions include 98% of those covered by the essential genes of the DEG database (Database of Essential Genes). Furthermore, three trinucleotides GTG, AAG and GCG, replacing three trinucleotides of the code *X* and called "evolutionary" trinucleotides, significantly occur in these 266 gene populations. Finally, a new method developed to analyse and quantify the stability of a set of trinucleotides demonstrates that these evolutionary trinucleotides are associated with a significant increase of the stability of the common circular code *X*. Indeed, its stability increases from the 1502th rank to the 16th rank after the replacement of the three evolutionary trinucleotides among 9920 possible trinucleotide replacement sets.

1. Introduction

1.1. Minimal gene set

The study of essential genes and the analysis of their features are obviously of great interest in basic and applied researches. We recall a common definition of a minimal gene set. A minimal gene set, or essential genes, is a smallest possible group of genes that would be sufficient to sustain a functioning cellular life form under the most favourable conditions imaginable, that is, in the presence of a full complement of essential nutrients and in the absence of environmental stress (Cho et al., 1999; Hutchison et al., 1999; Mushegian, 1999; Nelson et al., 1999). The upper bound of a minimal gene set is the number of genes of the smallest known genome (Koonin, 2003).

The sequencing of the *Mycoplasma genitalium* which is the smallest known genome until now with a size of 580 kb and 470 predicted genes, allowed an upper bound to be determined (Fraser et al., 1995). This bacterium is capable of living independent of its host.

The analysis of a minimal gene set aims to identify those genes that are essential to support the cell life using comparative and experimental methods. However, a minimal gene set depends on environmental conditions of the cell (Koonin, 2000; Gerdes et al., 2006). Some researchers defined a set of functions that should be covered by a minimal gene set to maintain the cell integrity, including translation, transcription, replication, membrane-transport and energy conversion (Alberts et al., 2002; Gil et al., 2004).

A minimal gene set is an essential factor for further experimental and theoretical researches as the full-synthesize and semi-synthesize of a functional cell (Luisi et al., 2006; Forster and Church, 2006; Gibson et al., 2008) or the reconstruction of the last universal common ancestor (Lazcano and Forterre, 1999; Koonin, 2003). A specialized DEG database (Database of Essential Genes) is developed to gather all published essential genes. It organizes the genes according to its kingdom and allows search using gene criteria or BLAST (Zhang et al., 2004; Zhang and Lin, 2009).

1.1.1. Comparative genomic approach

The comparative genomic is a bioinformatics approach to identify essential genes. It consists of selecting those genes that

* Corresponding author.
E-mail addresses: ahmed@dpt-info.u-strasbg.fr (A. Ahmed), g.frey@dpt-info.u-strasbg.fr (G. Frey), michel@dpt-info.u-strasbg.fr (C.J. Michel).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.02.006

are shared between distantly related organisms (Mushegian and Koonin, 1996). Evidently, genes are not identically shared between genomes. Hence, the term "shared genes" is redefined by orthologue or homologue genes (Fitch, 2000). This comparative approach was used when the first two bacterial genome sequences of *H. influenzae* and *M. genitalium* were completed (Fraser et al., 1995). This comparison revealed 241 direct orthologue genes but although this gene set cannot maintain a viable cell. For certain substantial essential molecular functions, different organisms do not use genes that are orthologue or even homologue. Such genes are not identified by orthologue comparative approach, rather they are detected by examining non-orthologue genes for missing essential functions. Such genes between genomes are called non-orthologue gene displacement NOGD (Koonin et al., 1996). By adding these NODG genes of *H. influenzae* and *M. genitalium*, this minimal gene set reaches 256 genes.

Although only 15 NODG genes were identified in this bacterial study, large genome comparisons shown that the NODG genes are associated with most essential genes including transcription, translation and replication (Koonin and Galperin, 2002). These genetic findings show that the concept of essential gene functions is more appropriate than that of essential genes.

1.1.2. Experimental approach

The experimental approach to identify essential genes existed before the comparative genomic one. It is simply based on gene-knockouts to determine lethal genes. The first experimental study generated 79 random gene-knockouts in the bacterial genome *B. subtilis* and identified 73 genes which did not kill the cell and six lethal genes (Itaya, 1995). It did not list the essential genes but only their proportion regarding to the total number of genes. This proportion (six essential genes out of 79 genes, i.e. about 8%) is very close to that obtained by comparative genomic approach (256 essential genes out of 4100 genes of the studied organism, i.e. about 6%).

Genetic methods used by this approach include transposon-insertion mutagenesis (Judson and Mekalanos, 2000), plasmid-insertion mutagenesis (Vagner et al., 1998) and gene inactivation using antisense RNAs (Ji et al., 2002).

The study of full wide genomes using experimental approaches showed many limitations. A first problem appeared because about half of gene-knockouts tend to be interrupted during experiences, and hence, a full list of essential genes is difficult to obtain. A second problem is the difficulty to identify essential cell functions performed by multiple proteins. Indeed, the disruption of one gene may not be lethal as gene redundancy exists, even if the function is essential to the cell. Another limitation of this experimental approach is that it does not consider the notion of gene evolution at all, in contrast to the comparative approach which is based on common orthologue genes.

The common circular code which was identified in large populations of prokaryotic and eukaryotic genes, is considered as an ancestral structural property of genes, and hence, as an essential property. It satisfies the principle of the comparative genomic approach. In this paper, we study the essential molecular functions using a comparative genomic approach based on this common circular code and its evolution. Hence, essential molecular functions will be analysed according to this circular code property.

In the next section, we recall the definition of the common circular code identified in genes of eukaryotes and prokaryotes and its main properties which will be used in the different methods developed.

1.2. The common circular code

1.2.1. Identification

In 1996, a simple occurrence study of the 64 trinucleotides $\mathbb{T} = \{AAA, \ldots, TTT\}$ in the three frames of genes showed that the trinucleotides are not uniformly distributed in these three frames. By excluding the four trinucleotides with identical nucleotides $\mathbb{T}_{id} = \{AAA, CCC, GGG, TTT\}$ and by assigning each trinucleotide to a preferential frame (frame of its highest occurrence frequency), the same three subsets X_0, X_1 and X_2 of 20 trinucleotides are found in the frames 0, 1 and 2, respectively, of two large and different gene populations (protein coding regions) of eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686 sequences, 4,709,758 trinucleotides) (Arquès and Michel, 1996) (Table 1). Note: A 2010 statistical study of archaeal genes (150 genomes, 85,804 sequences, 70,411 kb) shows that this set X_0 is only partially retrieved in these genes (AAT is replaced by ATA, CAG by GCA, CTG by GCT, GGT by GTG and TTC by CTT). By convention, the reading frame established by a start codon {ATG, GTG, TTG} is the frame 0, and the frames 1 and 2 are the reading frame 0 shifted by 1 and 2 nucleotides in the 5′–3′ direction, respectively. These three trinucleotide subsets present several strong biomathematical properties, particularly the fact that they are circular codes.

Before to give a few basic definitions of circular codes, the principle and the biological importance of circular codes which could construct the coding sequences are described. In present-day genes, the principle of decoding of a DNA sequence is "not complicated": by knowing the beginning of a sequence (a start codon), a complex translation apparatus (ribosome with ARNs and proteins) allows a reliable reading of nucleotides three by three, each trinucleotide being a word coding an amino acid (except the stop codons). This efficient translation process has no frameshifting (except for the particular cases of frameshift genes). In contrast, reading a DNA sequence from a random initial position in the sequence is a "more complicated" problem. Indeed, the correct reading frame must be retrieved among the three potential frames. Is the randomly selected nucleotide the 1st, the 2nd or the 3rd nucleotide of a codon?

Circular codes are sets of words. Trinucleotide circular codes, i.e. circular codes which are subsets of the genetic code, have interesting synchronizing properties. Any DNA sequence composed of a concatenation of words of a circular code can be read (decoded) from any position randomly chosen in the sequence as well as automatically (no need for a translation apparatus). Indeed, circular words sufficiently long always synchronize (at least 13 nucleotides with the common circular code). In other words, the reading frame of the sequence can always be retrieved. In addition to this random position property, circular words have polymorphic patterns. Indeed, a maximal trinucleotide circular code (see Definition 3 and Remark 3 below) leads to 20^n synchronizing (circular) words of codon length $n > 4$ (≥ 13 nucleotides). Note: there also exist short synchronizing words of codon length $n \leq 4$ (not detailed).

Table 1
The common circular code X identified in both eukaryotic and prokaryotic genes: X_0, X_1 and X_2 are the preferential sets of 20 trinucleotides in frames 0, 1 and 2 of genes, respectively.

X_0	AAC AAT ACC ATC ATT CAG CTC CTG GAA GAC GAG GAT GCC GGC GGT GTA GTC GTT TAC TTC
X_1	ACA ATA CCA TCA TTA AGC TCC TGC AAG ACG AGG ATG CCG GCG GTG TAG TCG TTG ACT TCT
X_2	CAA TAA CAC CAT TAT GCA CCT GCT AGA CGA GGA TGA CGC CGG TGG AGT CGT TGT CTA CTT

Synchronizing words (or circular words) in present-day genes could be markers of specific DNA boxes. Their multiple localization in the DNA sequence, even if some of them were destroyed by random mutations, and their polymorphic patterns make their identification specific and quick according to a parallel process (no need for a sequential reading of the DNA sequence from a start codon). Thus, they could help the ribosome to maintain the reading frame during translation (by checking the position of the ribosome with respect to the correct frame). As there are numerous and polymorphic, they can potentially be superposed to the genetic code. Thus, genes could 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.

1.2.2. Definitions and main properties of the common circular code

We recall the definitions and the main properties of this common circular code which will be involved in this paper.

Notation 1. The letters (or nucleotides or bases) define the genetic alphabet $\mathbb{A}_4 = \{A, C, G, T\}$. The set of non-empty words (words resp.) over $\mathbb{A}_4$ is denoted by $\mathbb{A}_4^+$ ($\mathbb{A}_4^*$ resp.). Let $x_1 \dots x_n$ be the concatenation of the words x_i for $i = 1, \dots, n$.

Definition 1. *Code*: A set X of words is a code if for each $x_1, \dots, x_n$, $y_1, \dots, y_m \in X, n, m \geq 1$, the condition $x_1 \dots x_n = y_1 \dots y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \dots, n$.

Remark 1. The set $\mathbb{T}$ itself is a code, more precisely a uniform code. Consequently, its non-empty subsets are codes. In this paper, we call them trinucleotide codes.

Definition 2. *Trinucleotide circular code*: A trinucleotide code X is circular if for each $x_1, \dots, x_n, y_1, \dots, y_m \in X, n, m \geq 1$, $r \in \mathbb{A}_4^*$, $s \in \mathbb{A}_4^+$, the conditions $sx_2 \dots x_n r = y_1 \dots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word) and $x_i = y_i$ for $i = 1, \dots, n$ (Fig. 1).

Remark 2. The set $\mathbb{T}$ is obviously not a trinucleotide circular code.

A circular code allows the reading frames in genes to be retrieved. 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 a unique decomposition into words of the circular code. As an example, let the set Y be composed of the six following words: $Y = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ 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 Y according to two different ways: ATG,GCC,CTA and AAT,GGC,CCT, the commas showing the way of decomposition (Fig. 2). Therefore, Y is not a circular code. In contrast, if the set Z obtained by replacing the word GGC of Y by GTC is considered, i.e. $Z = \{AAT, ATG, CCT, CTA, GCC, GTC\}$, then there never exists an ambiguous word with Z, such as w for Y, and then Z is a circular code.

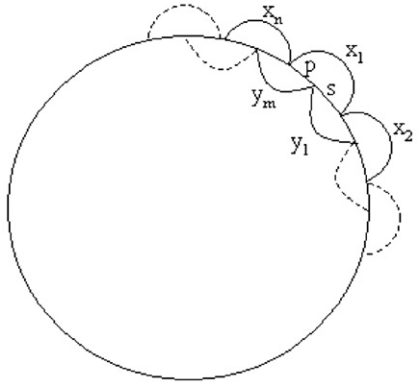

Fig. 1. A graphical representation of the circular code definition. A set of words X is a circular code if any word generated by a concatenation of words of X and written on a circle has a unique decomposition into words of X.

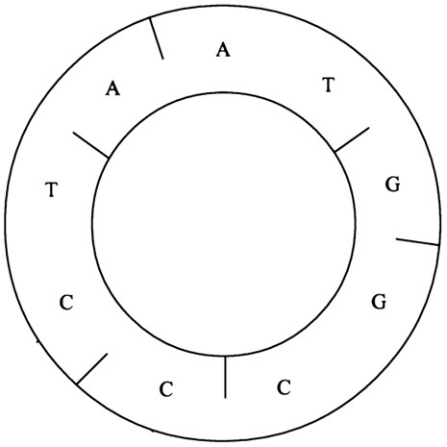

Fig. 2. The set $Y = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ is not a circular code as the word $w = ATGGCCCTA$ written on a circle can be factorized into words of Y according to two different ways: ATG,GCC,CTA and AAT,GGC,CCT.

Fig. 3. Retrieval of the reading frame of the word $w = \dots$AGGTAATTACCAA$\dots$ located anywhere in a sequence generated by a concatenation of words of the common circular code $X = X_0$ (Table 1).

The construction frame of a word generated by any concatenation of words of a circular code can be retrieved anywhere in the generated word after the reading of a certain number of nucleotides depending on the code. This series of nucleotides is called the window of the circular code. Then, the minimal window length to retrieve the construction frame is the size of the longest ambiguous word which can be read in at least two frames, plus one letter. Fig. 3 shows an example with the word $w = \dots$AGGTAATTACCAA$\dots$ located anywhere in a sequence generated by a concatenation of words of the common circular code X_0 (Table 1). Is the letter A the 1st, the 2nd or the 3rd nucleotide of a codon? By trying the three possible factorizations (frames) w^0, w^1 and w^2 of w into words of X_0, only w^2 is possible, i.e. the first letter A of w is the 2nd letter of a codon. Indeed, all the codons NAG, GTA, ATT, ACC and AAN (N being an appropriate letter) belong to X_0. The factorization w^0 is impossible as no word of X_0 starts with AG (Table 1). The factorization w^1 is also impossible as CAA is not a word of X_0 (Table 1). The chosen word AGGTAATTACCA (w without the last A) is one of the longest ambiguous words with the circular code X_0 (12 letters).

Definition 3. *Maximal trinucleotide circular code*: A trinucleotide circular code X is maximal if for each $y \in \mathbb{T}, X \cup \{y\}$ is not a trinucleotide circular code.

Remark 3. Any trinucleotide circular code with 20 words is maximal.

Definition 4. *Complementary map*: The complementary $\mathcal{C}: \mathbb{A}_4^+ \to \mathbb{A}_4^+$ is a map of $\mathbb{A}_4^+$ given by $\mathcal{C}(A) = T$, $\mathcal{C}(C) = G$, $\mathcal{C}(G) = C$, $\mathcal{C}(T) = A$. Furthermore, according to the property of the complementary and antiparallel double helix, $\mathcal{C}(uv) = \mathcal{C}(v)\mathcal{C}(u)$, $\forall u, v \in \mathbb{A}_4^+$, e.g. $\mathcal{C}(AAC) = GTT$.

Definition 5. *Self-complementary trinucleotide circular code*: A trinucleotide circular code X is self-complementary if for each $x \in X, \mathcal{C}(x) \in X$.

Definition 6. *Circular permutation map*: The circular permutation $\mathcal{P}: \mathbb{T} \to \mathbb{T}$ permutes circularly each trinucleotide $l_0l_1l_2$ as follows $\mathcal{P}(w_0) = l_1l_2l_0$, e.g. $\mathcal{P}(\mathrm{AAC}) = \mathrm{ACA}$. The kth iterate of $\mathcal{P}$ is denoted by $\mathcal{P}^k$, e.g. $\mathcal{P}^2(\mathrm{AAC}) = \mathrm{CAA}$.

Definition 7. *Permuted trinucleotide circular code*: A permuted trinucleotide circular code $\mathcal{P}(X)$ is the circular permutation of a trinucleotide circular code X so that for each $x \in X$, $\mathcal{P}(x) \in \mathcal{P}(X)$.

Definition 8. *C^3 trinucleotide circular code*: A trinucleotide circular code X is C^3 if $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ are trinucleotide circular codes.

Remark 4. A trinucleotide circular code Y does not necessarily imply that $Y_1 = \mathcal{P}(Y)$ and $Y_2 = \mathcal{P}^2(Y)$ are also trinucleotide circular codes.

Definition 9. *C^3 self-complementary trinucleotide circular code*: A trinucleotide circular code X is C^3 self-complementary if X is a C^3 trinucleotide circular code satisfying the following properties $X = \mathcal{C}(X)$ (self-complementary trinucleotide circular code), $\mathcal{C}(X_1) = X_2$ and $\mathcal{C}(X_2) = X_1$ (X_1 and X_2 are complementary codes).

Result 1. (*Arquès and Michel, 1996*; *Table 1*). The trinucleotide set X_0 coding the reading frames (frames 0) of eukaryotic and prokaryotic genes is a maximal C^3 self-complementary trinucleotide circular code. X_0 will be also noted X and simply called common circular code.

Therefore, the common circular code X and its two permuted circular codes $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ can exist in a DNA double helix simultaneously: X in a given DNA strand can be paired with X in the antiparallel complementary DNA (cDNA) strand, X_1 (X shifted by one nucleotide in the 5′–3′ direction) in a given DNA strand can be paired with X_2 (X shifted by two nucleotides in the 5′–3′ direction) in the cDNA strand and X_2 in a given DNA strand can similarly be paired with X_1 in the cDNA strand. Furthermore, the code X allows retrieval of any frame in genes, locally anywhere in the three frames and in particular without start codons in reading frames, and automatically with a series of a few nucleotides (13 nucleotides in each frame; Michel, 2008, property (via)).

A recent review of circular codes in genes details the research context, the history and the other properties of this circular code X (rarity, largest window length, higher frequency of "misplaced" trinucleotides in shifted frames, flexibility) (Michel, 2008).

Result 2. (*Ahmed and Michel, 2008*). Circular codes exist not only in genes (protein coding regions) but also in plant microRNAs which inherit the same structure of their target genes.

Result 3. (*Ahmed et al., 2007*). Circular codes are absent in the frameshift sites of genes where the ribosome shifts the reading frame to a shifted frame 1 or 2.

2. Method

We present here a new computational approach based on comparative genomics to identify essential molecular functions associated with circular codes. It will also allow an evolution study of the common circular code. It is based on a quantitative and sensitive statistical method (FPTF) to identify three permuted trinucleotide sets in the three frames of genes, a flower automaton algorithm to determine if a trinucleotide set is a circular code or not, and an integrated Gene Ontology and Taxonomy (iGOT) database. Finally, a new method is developed to analyse and quantify the stability of a set of trinucleotides. It is then applied to the circular code X.

2.1. Recall of the main steps of the FPTF (Frame Permuted Trinucleotide Frequency) method

The FPTF method published earlier is a quantitative and sensitive statistical analysis of the occurrence frequencies of the three permuted trinucleotides in their three frames which is developed in order to have an automatic approach for the trinucleotide assignment to a frame (Frey and Michel, 2006).

The 60 trinucleotides with non-identical nucleotides $w \in \mathbb{T} - \mathbb{T}_{id}$ can be gathered in 20 sets E_j, $j \in \{0, \ldots, 19\}$, of three trinucleotides invariant by permutation: $E_0 = \{\mathrm{AAC}, \mathrm{ACA}, \mathrm{CAA}\}$, $E_1 = \{\mathrm{AAG}, \mathrm{AGA}, \mathrm{GAA}\}, \ldots, E_{19} = \{\mathrm{GTT}, \mathrm{TTG}, \mathrm{TGT}\}$. The ith trinucleotide w in a set E is noted w_i, $i \in \{0, 1, 2\}$. Therefore, $w_1 = \mathcal{P}(w_0)$ and $w_2 = \mathcal{P}^2(w_0)$. For example in E_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 0, and $p = 1$ and $p = 2$ are the shifted frames 1 and 2 by 1 and 2 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 E read in a frame $p \in \{0, 1, 2\}$ is noted w_i^p. Therefore, a trinucleotide group G_j associated with a set E_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 E_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 is 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 a population of genes. 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:

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

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 $Q(w_i^p)$

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

Higher is the value $M(w_i^p)$ of a trinucleotide w_i^p, stronger is 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 FPTF method 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 $M(w_i^p)$. In a group G with nine trinucleotides w_i^p, there are three sets $S_k, k \in \{0,1,2\}$, of three trinucleotides associating each trinucleotide with a frame and each frame with a permuted trinucleotide by respecting Definition 6: $S_0 = \{w_0^0, w_1^1, w_2^2\}$, $S_1 = \{w_0^1, w_1^2, w_2^0\}$ and $S_2 = \{w_0^2, w_1^0, w_2^1\}$. 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 $S_{p_0} = \{w_{i_0}^{p_0}, w_{i_1}^{p_1}, w_{i_2}^{p_2}\}$, $p_0 \in \{0,1,2\}$, the statistical function $F(S_{p_0})$ 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 $w_{i_2}^{p_2}$

$$F(S_{p_0}) = F\left(\left\{w_{i_0}^{p_0}, w_{i_1}^{p_1}, w_{i_2}^{p_2}\right\}\right) = \tfrac{1}{3}\left(M(w_{i_0}^{p_0}) + M(w_{i_1}^{p_1}) + M(w_{i_2}^{p_2})\right). \quad (4)$$

The set S_{pref} having the highest value with the function $F(S)$ among these three sets S is defined by

$$S_{pref} = S_{p_0} \text{ such that } F(S_{p_0}) = \max_{k=0,1,2}\{S_k\}. \quad (5)$$

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

2.2. Application example of the FPTF method

As an application example to comment on the FPTF method, we consider the population $\mathcal{G} = D$ of drug binding genes. Drug binding in the gene ontology terms is a molecular function interacting selectively with a drug, any naturally occurring or synthetic substance. This population D contains 217 genes from different species. Table 2 presents the occurrence probabilities $o(w_i^p)$ of nine trinucleotides w_i^p, $i,p \in \{0,1,2\}$, in each group G_j associated with the 20 permuted trinucleotide sets $E_j, j \in \{0,\ldots,19\}$, of genes D. Table 3 presents the function $M(w_i^p)$ (Formula 3) of nine trinucleotides w_i^p, $i,p \in \{0,1,2\}$, in the group G_0 associated with the set $E_0 = \{\text{AAC}, \text{ACA}, \text{CAA}\}$ of genes D in order to evaluate a trinucleotide simultaneously compared to its two other frames and its two other permuted trinucleotides. For the rest of the application example, the computations will be shown only for this set E_0, the other sets being computed in the same manner. A trinucleotide w_i^p with the highest value $M(w_i^p)$ occurs preferentially in the frame p. Table 4 computes the function $F(S_{p_0})$, $p_0 \in \{0,1,2\}$ (Formula 4), evaluating the set $E_0 = \{\text{AAC}, \text{ACA}, \text{CAA}\}$ of genes D. As $F(S_0)$ has the highest value (among the three $F(S_{p_0})$ values), then $S_{pref} = S_0 = \{\text{AAC}^0, \text{ACA}^1, \text{CAA}^2\}$. Therefore, the preferential frames for the three trinucleotides AAC, ACA and CAA are the frames 0, 1 and 2, respectively. In the same way, the preferential frames for the other permuted trinucleotide sets $E_1,\ldots,E_{19}$ are computed. Table 5 presents the three preferential sets of 20 trinucleotides $X_0(D)$, $X_1(D)$ and $X_2(D)$ in frames 0, 1 and 2, respectively, of genes D.

2.3. Flower automaton algorithm

A flower automaton algorithm is developed to determine if a set of trinucleotides is a circular code or not (not detailed; for the mathematical concept see Berstel and Perrin, 1985).

2.4. The iGOT database (integrated Gene Ontology and Taxonomy)

An integrated Gene Ontology and Taxonomy (iGOT) database is developed and associated with the FPTF (Frame Permuted

Table 2
Occurrence probabilities $o(w_i^p)$ of nine trinucleotides $w_i^p, i,p \in \{0,1,2\}$, in each group G_j associated with the 20 permuted trinucleotide sets E_j, $j \in \{0,\ldots,19\}$, for the example of drug binding genes D.

G_j	E_j	$o(w_i^0)$	$o(w_i^1)$	$o(w_i^2)$	G_j	E_j	$o(w_i^0)$	$o(w_i^1)$	$o(w_i^2)$
G_0	AAC	0.0197	0.0140	0.0120	G_1	AAG	0.0312	0.0286	0.0121
	ACA	0.0133	0.0255	0.0109		AGA	0.0165	0.0250	0.0305
	CAA	0.0155	0.0154	0.0351		GAA	0.0369	0.0084	0.0260
G_2	AAT	0.0117	0.0157	0.0109	G_3	ACC	0.0209	0.0196	0.0070
	ATA	0.0035	0.0091	0.0050		CCA	0.0161	0.0315	0.0239
	TAA	0.0000	0.0053	0.0132		CAC	0.0161	0.0058	0.0238
G_4	ACG	0.0055	0.0158	0.0022	G_5	ACT	0.0148	0.0176	0.0107
	CGA	0.0048	0.0036	0.0156		CTA	0.0069	0.0068	0.0132
	GAC	0.0242	0.0064	0.0105		TAC	0.0185	0.0046	0.0082
G_6	AGC	0.0131	0.0209	0.0174	G_7	AGG	0.0082	0.0289	0.0098
	GCA	0.0167	0.0173	0.0204		GGA	0.0071	0.0095	0.0322
	CAG	0.0242	0.0216	0.0165		GAG	0.0368	0.0118	0.0133
G_8	AGT	0.0084	0.0174	0.0147	G_9	ATC	0.0215	0.0112	0.0075
	GTA	0.0079	0.0109	0.0080		TCA	0.0083	0.0247	0.0134
	TAG	0.0000	0.0105	0.0042		CAT	0.0128	0.0116	0.0237
G_{10}	ATG	0.0240	0.0344	0.0102	G_{11}	ATT	0.0168	0.0101	0.0141
	TGA	0.0011	0.0277	0.0469		TTA	0.0087	0.0123	0.0067
	GAT	0.0274	0.0029	0.0159		TAT	0.0130	0.0066	0.0135
G_{12}	CCG	0.0054	0.0144	0.0096	G_{13}	CCT	0.0147	0.0171	0.0219
	CGC	0.0076	0.0055	0.0086		CTC	0.0147	0.0173	0.0123
	GCC	0.0286	0.0129	0.0156		TCC	0.0158	0.0209	0.0133
G_{14}	CGG	0.0086	0.0083	0.0112	G_{15}	CGT	0.0054	0.0031	0.0139
	GGC	0.0163	0.0129	0.0211		GTC	0.0191	0.0059	0.0089
	GCG	0.0050	0.0057	0.0075		TCG	0.0046	0.0133	0.0073
G_{16}	CTG	0.0349	0.0327	0.0152	G_{17}	CTT	0.0164	0.0138	0.0296
	TGC	0.0139	0.0292	0.0324		TTC	0.0242	0.0149	0.0118
	GCT	0.0292	0.0151	0.0250		TCT	0.0118	0.0205	0.0153
G_{18}	GGT	0.0188	0.0043	0.0207	G_{19}	GTT	0.0179	0.0075	0.0156
	GTG	0.0282	0.0203	0.0094		TTG	0.0237	0.0329	0.0066
	TGG	0.0138	0.0368	0.0174		TGT	0.0119	0.0171	0.0236

Table 3
Function $M(w_i^p)$ of nine trinucleotides $w_i^p, i,p \in \{0,1,2\}$, in the group G_0 associated with the set E_0={AAC, ACA, CAA} for the example of drug binding genes D.

E_0	$M(w_i^0)$	$M(w_i^1)$	$M(w_i^2)$
AAC	0.419	0.280	0.235
ACA	0.270	0.489	0.203
CAA	0.277	0.257	0.568

Table 4
Function $F(S_{p0}), p_0 \in \{0,1,2\}$, evaluating the set E_0={AAC,ACA,CAA} for the example of drug binding genes D.

S_k	$F(S_k)$
S_0	0.492
S_1	0.253
S_2	0.254

Table 5
The three preferential sets of 20 trinucleotides $X_0(D)$, $X_1(D)$ and $X_2(D)$ in frames 0, 1 and 2, respectively, of the example of drug binding genes D.

$X_0(D)$	AAC GAA AAT ACC GAC TAC CAG GAG GTA ATC GAT ATT GCC CTC GGC GTC GCT TTC GTG GTT
$X_1(D)$	ACA AAG ATA CCA ACG ACT AGC AGG TAG TCA ATG TTA CCG TCC GCG TCG CTG TCT TGG TTG
$X_2(D)$	CAA AGA TAA CAC CGA CTA GCA GGA AGT CAT TGA TAT CGC CCT CGG CGT TGC CTT GGT TGT

Trinucleotide Frequency) method (see above) to analyse massive gene populations in order to identify automatically circular codes close to the common circular code X of eukaryotic and prokaryotic genes (Fig. 4). It is freely available upon request. The iGOT database is based on the Gene Ontology (GO) database (Gene Ontology Consortium, 2000, 2008). The GO database describes genes and gene products in a species independent manner. It covers the molecular function of gene products, their roles in multi-step biological processes and their localization in cellular components. For example, the genes responsible for the drug binding are found under the binding function which covers all binding functions and which is localized under the molecular functions. In this paper, we excluded the biological processes and the cellular components which are not directly related to a circular code function.

The GO database focuses on the gene product itself rather than the nucleic sequence of genes, thus the genes are presented as a sequence of amino acids. Therefore, we applied some modifications on the table structures and imported the nucleic sequences of their genes. Fig. 4 shows the main entities of the iGOT database which inherits some entities of the GO database. iGOT has been updated with the last GO database release.

In the GO role hierarchy, the first four levels cover 1183 populations of genes while the first five levels cover 5566 populations of genes. Due to the exponential increase in the data-processing computing time, the number of populations in our statistical analysis has been limited to the first four levels, i.e. 1183 gene populations $\mathcal{G}$ representing in total about 300,000 genes (312,000 kb).

The TERM and TERM2TERM tables store the molecular functions as well as their hierarchical organizations. The GO terms are organized in a network like structure, i.e. a term may be assigned to many parent term groups. The GENE_PRODUCT table stores the gene entity itself rather than its sequence which is stored in the SEQ table for the amino acid sequences and in the SEQ_EMBL table for the nucleic sequences. The SPECIES table stores the species and their organization. The DBXREF table stores all the external references. An extension to store the FPTF calculation results is developed (not shown in Fig. 4).

2.5. Circular code stability method

2.5.1. Principle

We present here a new method to analyse and quantify the stability of a set of trinucleotides. This general concept applied to a circular code X is related to the notion of ambiguity of circular words. Short sequences of circular words (necessarily shorter than the code window length of 13 nucleotides) can occur in non-reading frames (shifted frames) of words generated by a concatenation of circular words (e.g. the words GGT, AAT and TAC in the factorization w^1 of the example in Fig. 3). The stability function $L(X)$ of a code X will measure the proportion of words of X which occur in shifted frames when concatenating words of X. Then, circular codes which are less ambiguous for the frame retrieval (codes with few words in shifted frames) have a higher stability function $L(X)$. A circular code X with a stability value of $L(X)=100\%$ only contains words of X in the reading frame (no words of X in the shifted frames). This class of codes with stronger synchronization properties are called comma-free codes (Berstel et al., 2005). More generally, the more the $L(X)$ value of a code X is raised, the more the probability to retrieve the reading frame increases with shorter words of X. We apply here this concept to the common circular code X.

2.5.2. Definition

As the code X occurs in the reading frame (frame 0) of genes, we denote by $t_0=n_0n_1,n_2$, n_i being a nucleotide {A,C,G,T}, a trinucleotide of X in frame 0. The cardinality of X is $\mathrm{card}(X)=20$. Let the di-trinucleotide w be a concatenation of the two trinucleotides t_0 and $t'=n'_0n'_1n'_2$ of X, i.e. $w=t_0t'\in X^2$. By assuming that the trinucleotides of X are equiprobable, then there are 400 possible di-trinucleotides $w\in X^2$ leading to an occurrence probability $P(w)=1/\mathrm{card}(X)^2$. We denote by $t_0(w)$, $t_1(w)$ and $t_2(w)$

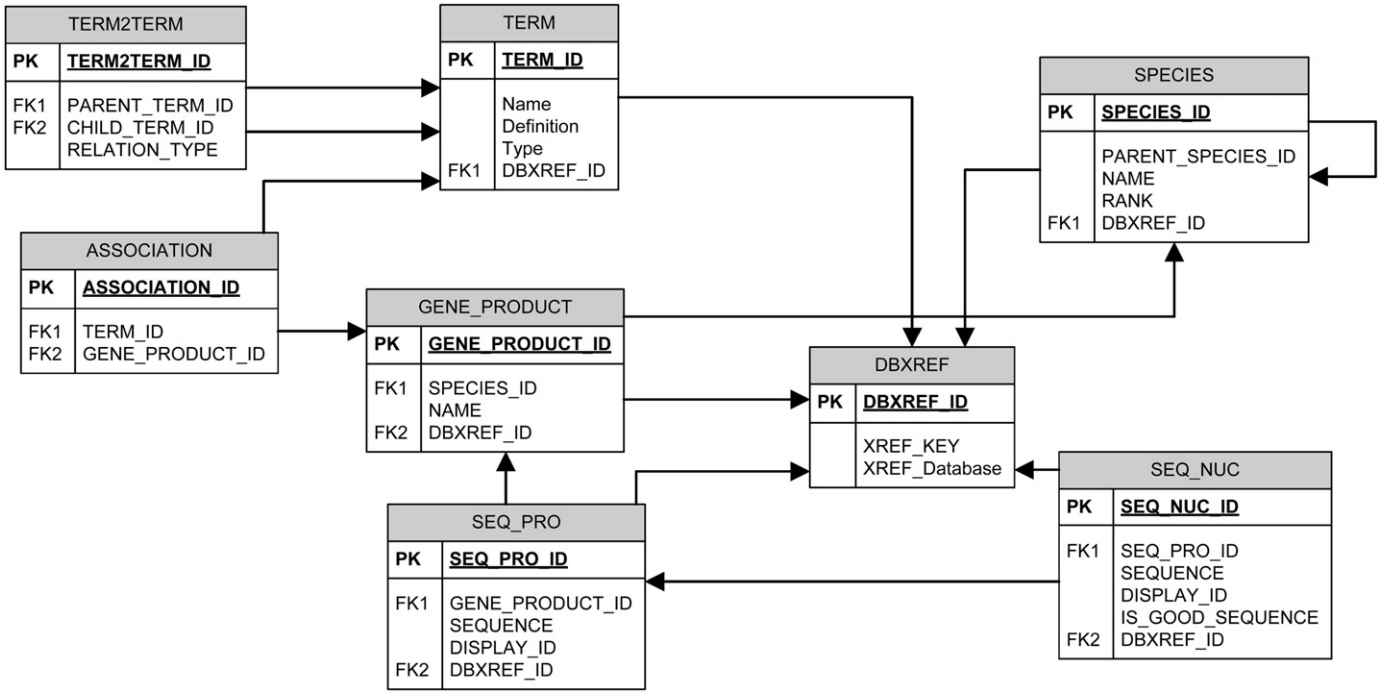

Fig. 4. Main entities of the iGOT database (integrated Gene Ontology and Taxonomy).

the trinucleotides $n_0n_1n_2$ in frame 0, $n_1n_2n'_0$ in frame 1 and $n_2n'_0n'_1$ in frame 2 of a di-trinucleotide $w \in X^2$, respectively. The concatenation of the two trinucleotides $t_0 \in X$ and $t' \in X$ may yield a trinucleotide $t''_f(w) \in X$ but in a frame $f \neq 0$. This property exists with the circular codes but not with the comma-free codes (see Definitions 4 and 5 in Michel et al., 2008). For example, the concatenation of the trinucleotides $t_0 = \mathrm{TAC} \in X$ and $t' = \mathrm{CTC} \in X$ (Table 1), i.e. w=TACCTC, leads to the trinucleotide $t''_1(w) = \mathrm{ACC} \in X$ (Table 1) which thus occurs in frame 0 but also in frame 1.

Our objective is the study of the stability of trinucleotides of X in the reading frame after the mixing of all possible trinucleotides of X (with equiprobability) or in other words, the tendency of trinucleotides of X to occur in other frames than the reading frame. Let the indicator function $\delta(t, t_f(w))$ be equal to 1 if a trinucleotide $t \in X$ is equal to the trinucleotide $t_f(w)$ in the frame f of w

$$\delta(t, t_f(w)) = \begin{cases} 1 & \text{if } t = t_f(w) \text{ with } w \in X^2, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

Then, the stability function $L_f(t)$ (in %) of a trinucleotide $t \in X$ in a frame $f \in \{0,1,2\}$ is measured by

$$L_f(t) = 100 \frac{\sum_{w \in X^2} \delta(t, t_f(w))}{\sum_{f' \in \{0,1,2\}} \sum_{w \in X^2} \delta(t, t_{f'}(w))}. \quad (7)$$

Remark 5. In the frame $f = 0$, $\sum_{w \in X^2} \delta(t, t_0(w)) = 20$, whatever the trinucleotide $t \in X$.

The more the $L_0(t)$ value is raised, the more the trinucleotide t in the frame 0 is stable.

Then, the stability function $L_f(X)$ (in %) of the common circular code X in a frame $f \in \{0,1,2\}$ is naturally measured by the average values of $L_f(t)$

$$L_f(X) = \frac{1}{\mathrm{card}(X)} \sum_{t \in X} L_f(t). \quad (8)$$

Similarly, the more the $L_0(X)$ value is raised, the more the common circular code X in the frame 0 is stable.

3. Results

3.1. Essential molecular functions associated with circular codes

The first four levels of molecular functions of the iGOT database are scanned and 1183 gene populations $\mathcal{G}$ sharing the same molecular function are thus constructed. Then, the FPTF method and the flower automaton algorithm are applied on these gene populations $\mathcal{G}$ in order to identify the gene populations $\tilde{\mathcal{G}}$ which are maximal C^3 trinucleotide circular codes (Definitions 2, 3 and 8) close to the common circular code X. These maximal C^3 trinucleotide circular codes close to X are noted $\tilde{X}$.

The common circular code X is considered as a common ancestral property of genes, and hence, as an essential structural property of genes. However, in small and particular gene populations, such as $\mathcal{G}$ sharing the same molecular function, the code X and the codes $\tilde{X}$ derived by evolution of X are considered. This code evolutionary process is considered with a word distance d giving the number of different trinucleotides between the two codes X and $\tilde{X}$: $d = 20 - |X \cap X(\mathcal{G})|$ with $0 \leq d \leq 20$. As mainly three "evolutionary" trinucleotides, GTG, AAG and GCG, are identified in the statistical analysis of 1183 gene populations $\mathcal{G}$ (see below Section 3.2), a code $\tilde{X}$ is considered to be close to the code X if $d \leq 3$, i.e. there are at least 17 identical trinucleotides between X and $\tilde{X}$. In general, a code $\tilde{X}$ lost the self-complementary property (Definition 9).

This approach identifies 266 gene populations $\tilde{\mathcal{G}}$ among the 1183 ones (about 22.5%) which are codes $\tilde{X}$. Fig. 5 shows the identified molecular functions encoded by these gene populations $\tilde{\mathcal{G}}$. In this figure, each cell represents a population of genes at a certain level. Since the terms in iGOT are organized in a network like structure and presented in this figure in a tree like structure, a term may appear twice in a figure. Three values are presented in each cell: the size, the T and V values. The size (in percentage) is the dimension of the gene population relative to its sibling gene populations, the T value, the total number of children gene populations $\mathcal{G}$ and the V value, the number of children gene populations $\tilde{\mathcal{G}}$, $|\tilde{\mathcal{G}}| \leq |\mathcal{G}|$. For displaying purposes, Fig. 5 shows only the first three levels and a grey cell indicates a gene population $\tilde{\mathcal{G}}$.

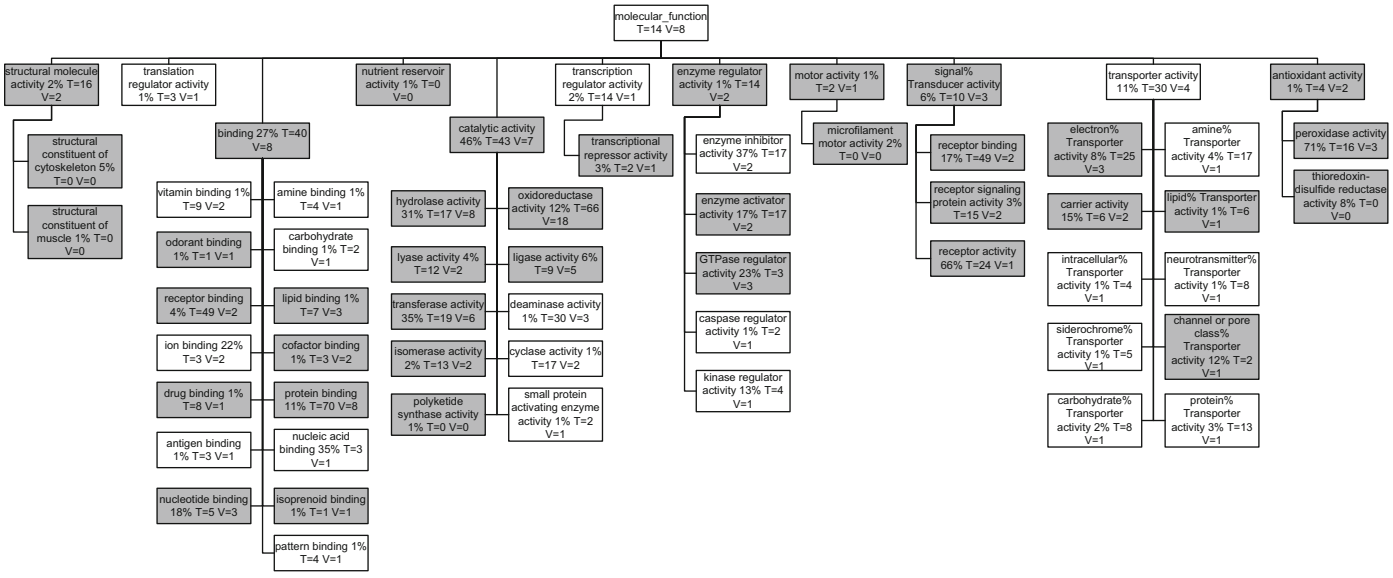

Fig. 5. The molecular functions encoded by the 266 gene populations $\tilde{\mathcal{G}}$ (maximal C^3 trinucleotide circular codes $\tilde{X}$ close to the common circular code X; grey cells) in the iGOT database (integrated Gene Ontology and Taxonomy). The percentage value is the size of the gene population relative to its sibling gene populations, the T value, the total number of children gene populations and the V value, the number of children gene populations $\tilde{\mathcal{G}}$.

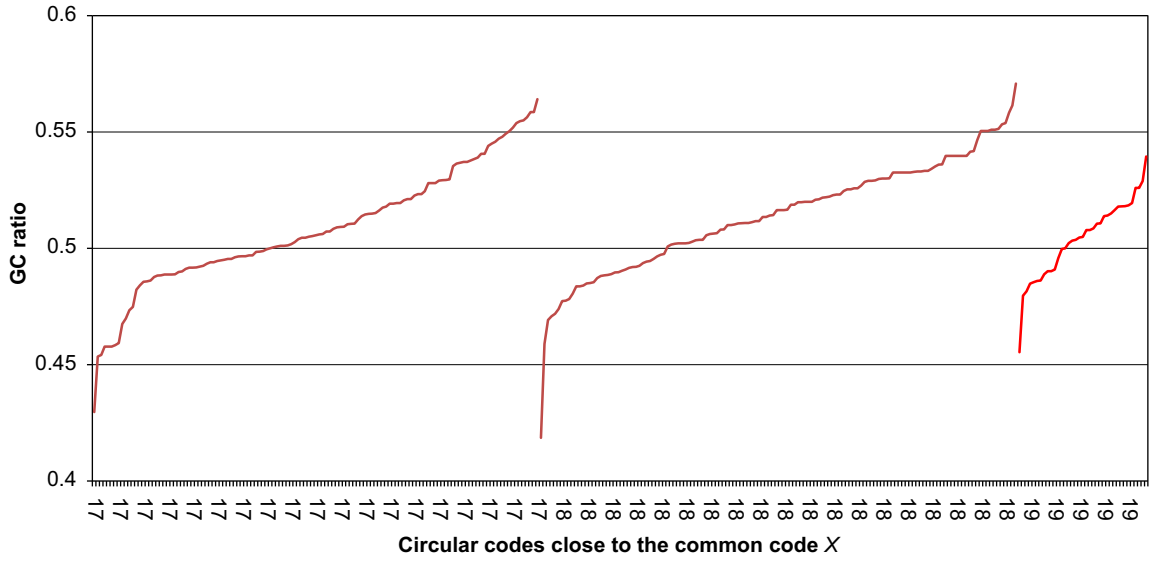

Fig. 6. Absence of correlation between the maximal C^3 trinucleotide circular codes $\tilde{X}$ (close to the common circular code X) and the GC ratio of their 266 gene populations $\tilde{\mathcal{G}}$. The circular codes are distributed on both sides of the random GC ratio 0.5. The X-axis represents the circular codes ordered by their distance to the common circular code X (17, 18 and 19 trinucleotides in common), and for a given distance, ordered by increasing values of gene GC ratio, and the Y-axis, the gene GC ratio.

Surprisingly, by comparing the 266 gene populations $\tilde{\mathcal{G}}$ with the essential genes of the DEG database, the molecular functions of the 266 gene populations $\tilde{\mathcal{G}}$ include 98% of the functions covered by the essential genes of the DEG database.

Furthermore, a statistical study is being performed to evaluate if the circular code property is affected by gene GC content. Thus, the identified circular codes $\tilde{X}$ are analysed according to the GC ratio of their 266 gene populations $\tilde{\mathcal{G}}$. Fig. 6 shows that they are distributed on both sides of the random GC ratio 0.5, i.e. with values greater or less than 0.5. Therefore, according to this statistical study, there is no-correlation between the circular code property and the gene GC content.

3.2. Evolutionary trinucleotides in the common circular code

For most of the 266 gene populations $\tilde{\mathcal{G}}$, the evolutionary trinucleotides are mainly the three trinucleotides GTG, AAG and GCG among the 40 possible trinucleotides occurring in frames 1 or 2. Belonging to frame 1 with X (belonging to X_1, Table 1), they occur in frame 0 (reading frame) with 81.3%, 40.0% and 24.6% of these 266 gene populations $\tilde{\mathcal{G}}$, respectively (Table 6). Remark: This statistical result explains the previous choice to set the word distance $d \leq 3$ (see above).

Very surprisingly, these evolutionary trinucleotides may be explained by the stability of the common circular code X. Table 7 presents the stability function $L_f(t)$ (Formula 8; in %) of the 20 trinucleotides t of the common circular code X in the three frames $f \in \{0,1,2\}$ ordered by increasing values of stability.

Result 4. The most stable trinucleotides of X are CAG, CTG, CTC and GAG with a value $L_0(t)=100.0\%$, i.e. with no occurrence in the shifted frames (Table 7). The least stable trinucleotides of X are ACC and GGT with a value $L_0(t)=69.0\%$ (Table 7).

Remark 5. $L_1(t)=L_2(\mathcal{C}(t))$ and $L_2(t)=L_1(\mathcal{C}(t))$ whatever $t \in X$ (Result 1 and consequence of Definition 9).

Result 5. The stability of the common circular code X has a value $L_0(X)=82.5\%$ (Table 7).

Remark 6. $L_1(X)=L_2(X)$ (Result 1 and consequence of Definition 9).

Table 6
Main evolutionary trinucleotides and their occurrence percentages in the 266 gene populations $\tilde{\mathcal{G}}$.

Evolutionary trinucleotides	Occurrence (%)
GTG	81.3
AAG	40.0
GCG	24.6

In order to determine if a relation between the evolutionary trinucleotides GTG, AAG and GCG and the stability of the common circular code X exists, we develop an iterative procedure which replaces a trinucleotide of X by another trinucleotide of the same permuted set E_j, $j \in \{0,\ldots,19\}$, (defined in Section 2.1) and computes the stability $L_f(Y)$ of this new set Y (Formula 8). For example, if the least stable trinucleotide GGT of X which tends to occur in frame 2 (Table 7) is replaced by GTG in frame 0 (as $\mathcal{P}^2(\mathrm{GTG})=\mathrm{GGT}$) then the stability computation of this new set Y leads to a value $L_0(Y)=84.7\%$ which is unexpectedly greater than $L_0(X)=82.5\%$ (Result 5). Therefore, this set Y is more stable than X. As there are (mainly) three evolutionary trinucleotides, the procedure analyses all the trinucleotides sets until a maximum of three trinucleotide replacements in the common circular code X. There are

$$\sum_{i=1}^{3} 2^i \binom{\mathrm{card}(X)}{i} = 9920$$

such sets Y. Fig. 7 shows these 9920 sets Y sorted according to their decreasing values $L_0(Y)$ of stability. This function $L_0(Y)$ has a maximal value equals to 89.3%, a minimal value equals to 67.8% and an average value equal to 78.6% at rank 4768 (Fig. 7). The common circular code X occurs at the 1502th rank (Fig. 7).

Notation 2. The maximal C^3 circular code $\tilde{X}$ is now defined precisely as being the common circular code X (maximal C^3 self-complementary trinucleotide circular code) in which the three evolutionary trinucleotides GTG, AAG and GCG replace the three trinucleotides GGT, GAA and GGC, respectively (Table 8). Note: $\tilde{X}$ is not self-complementary.

Result 6. The stability of the C^3 circular code $\tilde{X}$ has a value $L_0(\tilde{X})=87.8\%$ and occurs at the 16th rank among the 9920 sets, i.e. among the first 0.15% sets (Fig. 7).

This surprising result agrees with a relation between the evolutionary trinucleotides and the stability of the common circular code X.

A comparative statistical study of the common circular code X between genes coding membrane proteins (GO term 0016020: membrane, 46,182 sequences, 71,000 kb) and genes coding soluble proteins (GO term 0005625: soluble fraction, 567 sequences, 983 kb) is also performed. In genes of membrane proteins, the common circular code X has four evolutionary trinucleotides (CAG is replaced by GCA, CTG by GCT, GGC by GCG and GGT by GTG). In contrast, it is more conserved in genes of soluble proteins with two evolutionary trinucleotides (GAA is replaced by AAG and GGT by GTG). Both classes of genes have the main evolutionary trinucleotide GTG (Table 6) and each class contains one to the two other identified evolutionary trinucleotide AAG or GCG (Table 6).

Table 7
Stability value $L_f(t)$ (in %) of the 20 trinucleotides t of the common circular code X in the three frames $f \in \{0,1,2\}$ ordered by increasing values of stability and stability function $L_f(X)$ (in %) of the common circular code X in the three frames $f \in \{0,1,2\}$ (last row).

t	$L_0(t)$	$L_1(t)$	$L_2(t)$
ACC	69.0	31.0	0.0
GGT	69.0	0.0	31.0
GTA	71.4	17.9	10.7
TAC	71.4	10.7	17.9
GAT	76.9	0.0	23.1
ATC	76.9	23.1	0.0
GAA	76.9	0.0	23.1
TTC	76.9	23.1	0.0
AAT	76.9	7.7	15.4
ATT	76.9	15.4	7.7
AAC	80.0	12.0	8.0
GTT	80.0	8.0	12.0
GCC	87.0	13.0	0.0
GGC	87.0	0.0	13.0
GAC	87.0	0.0	13.0
GTC	87.0	13.0	0.0
CAG	100	0.0	0.0
CTG	100	0.0	0.0
CTC	100	0.0	0.0
GAG	100	0.0	0.0
$L_f(X)$	82.5	8.7	8.7

4. Conclusion

In 1996, a common circular code X was identified in large populations of eukaryotic and prokaryotic genes (Result 1 and Table 1). It allows the reading frames in genes to be retrieved locally, i.e. anywhere in genes and in particular without start codons, and automatically with a window of a few nucleotides. It is believed to be an ancestral structural property of genes (Arquès and Michel, 1996 and their subsequent papers).

We developed here a new computational approach based on comparative genomics to identify molecular functions associated with this common circular code and its evolution in close C^3 circular codes (Definition 8). In order to carry out automatic circular code analyses on a huge number of gene populations, it is based on a quantitative and sensitive statistical method (FPTF) which allows three permuted trinucleotide sets in the three frames of genes to be identified (Section 2.1), a flower automaton algorithm to determine if a trinucleotide set is a circular code or not (Section 2.3), and an integrated Gene Ontology and Taxonomy database (iGOT) (Section 2.4 and Fig. 4). It identifies 266

Table 8
The common circular code X identified in both eukaryotic and prokaryotic genes (identical to X_0 of Table 1) and the circular code $\tilde{X}$ with the three evolutionary trinucleotides GTG, AAG and GCG (underlined) (last row).

X	AAC AAT ACC ATC ATT CAG CTC CTG GAA GAC GAG GAT GCC GGC GGT GTA GTC GTT TAC TTC
$\tilde{X}$	AAC AAT ACC ATC ATT CAG CTC CTG AAG GAC GAG GAT GCC GCG GTG GTA GTC GTT TAC TTC

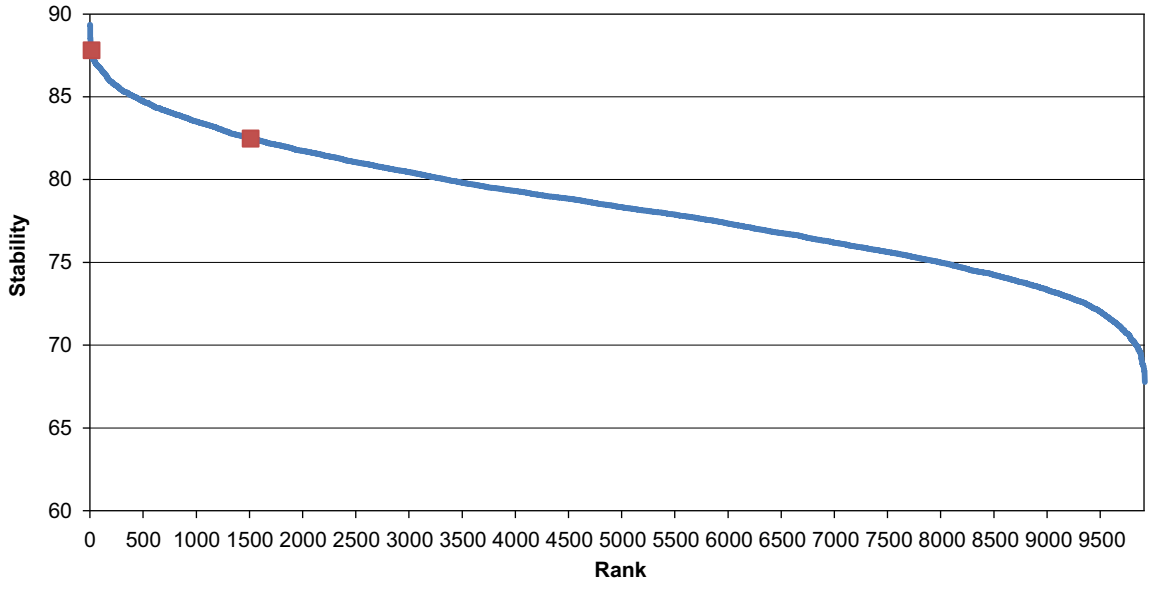

Fig. 7. Stability function $L_0(Y)$ (in %) of 9920 sets Y associated with a maximum of three trinucleotide replacements in the common circular code X and sorted according to their decreasing values of stability. The common circular code X has a stability $L_0(X)=82.5\%$ (Table 7) and occurs at the 1502th rank (right square) among these 9920 sets. The circular code $\tilde{X}$ with the three evolutionary trinucleotides GTG, AAG and GCG has a stability equal to $L_0(Y)=87.8\%$ and occurs at the 16th rank (left square) among these 9920 sets.

molecular functions (Fig. 5) which surprisingly include 98% of the functions covered by the essential genes of the DEG database. This observation suggests that the common circular code X and its evolution in close C^3 circular codes $\tilde{X}$ are associated with essential molecular functions.

This approach also identifies three (main) evolutionary trinucleotides in the common circular code X: GTG, AAG and GCG (Table 6). Very surprisingly, these three evolutionary trinucleotides of X may be explained by the stability of X. Indeed, the trinucleotide $\mathrm{GGT}=\mathcal{P}^2(\mathrm{GTG})$ of X (31.0% in frame 2) with a value $L_0(\mathrm{GGT})=69.0\%$ is the least stable trinucleotide of X (Table 7). Similarly, the trinucleotide $\mathrm{GAA}=\mathcal{P}^2(\mathrm{AAG})$ of X (23.1% in frame 2, i.e. the second highest value) is also one of the most unstable trinucleotides of X (Table 7). Therefore, a new C^3 circular code $\tilde{X}$ is identified in gene populations. $\tilde{X}$ is the common circular code X in which the three evolutionary trinucleotides GTG, AAG and GCG replace the three trinucleotides GGT, GAA and GGC, respectively (Table 8). A new method is developed to analyse and quantify the stability of a circular code, precisely to study the tendency of trinucleotides of X to stay in reading frame or to expand in shifted frames (Section 2.5). Furthermore, an iterative procedure analysing the stability of the 9920 trinucleotide sets (until a maximum of three trinucleotide replacements in X) allows the stability of the two codes X and $\tilde{X}$ to be compared. The common circular code X has a stability value of 82.5% (Table 7) and occurs at the 1502th rank among the 9920 trinucleotide sets (Fig. 7). Surprisingly, the C^3 circular code $\tilde{X}$ has a significant stability increase with a value of 87.8% associated with the 16th rank among these 9920 trinucleotide sets (Fig. 7). This stability increase of $\tilde{X}$ could be related to a better reading of gene frames from a point of view of accuracy and speed.

Acknowledgement

We thank a reviewer for his advices and suggestions.

References

Ahmed, A., Frey, G., Michel, C.J., 2007. Frameshift signals in genes associated with the circular code. Silico Biol. 7, 155–168.
Ahmed, A., Michel, C.J., 2008. Plant microRNA detection using the circular code information. Comput. Biol. Chem. 32, 400–405.
Alberts, B., et al., 2002. Molecular Biology of the Cell. Garland Science, New York.
Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182, 45–58.
Berstel, J., Perrin, D., 1985. Theory of Codes. Academic Press, London.
Berstel, J., Perrin, D., Reutenauer, C., 2005. Codes and Automata. Cambridge University Press.
Cho, M.K., Magnus, D., Caplan, A.L., McGee, D., 1999. Policy forum: genetics—ethical considerations in synthesizing a minimal genome. Science 286, 2087–2090.
Fitch, W.M., 2000. Homology: a personal view on some of the problems. Trends Genet. 16, 227–231.
Forster, A.C., Church, G.M., 2006. Towards synthesis of a minimal cell. Mol. Syst. Biol. 2, 45.
Fraser, C.M., et al., 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270, 397–403.
Frey, G., Michel, C.J., 2006. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Comput. Biol. Chem. 30, 87–101.
Gene Ontology Consortium, 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29.
Gene Ontology Consortium, 2008. The gene ontology project in 2008. Nucleic Acids Res. 36, D440–D444.
Gerdes, S., Edwards, R., Kubal, M., Fonstein, M., Stevens, R., Osterman, A., 2006. Essential genes on metabolic maps. Curr. Opin. Biotechnol. 17, 448–456.
Gibson, D.G., et al., 2008. Complete chemical synthesis, assembly, and cloning of a *Mycoplasma Genitalium* genome. Science 319, 1215–1220.
Gil, R., et al., 2004. Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev. 68, 518–537.
Hutchison, C.A., Peterson, S.N., Gil, S.R., Cline, R.T., White, O., et al., 1999. Global transposon mutagenesis and a minimal mycoplasma genome. Science 286, 2165–2169.
Itaya, M., 1995. An estimation of minimal genome size required for life. FEBS Lett. 362, 257–260.
Ji, Y., Woodnutt, G., Rosenberg, M., Burnham, M.K., 2002. Identification of essential genes in *Staphylococcus aureus* using inducible antisense RNA. Methods Enzymol. 358, 123–128.
Judson, N., Mekalanos, J.J., 2000. Transposon-based approaches to identify essential bacterial genes. Trends Microbiol. 8, 521–526.
Koonin, E.V., 2000. How many genes can make a cell: the minimal-gene-set concept. Annu. Rev. Genomics Hum. Genet. 1, 99–116.
Koonin, E.V., 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat. Rev. Microbiol. 1, 127–136.
Koonin, E.V., Galperin, M.Y., 2002. Sequence–Evolution–Function. Computational Approaches in Comparative Genomics. Kluwer Academic, New York.
Koonin, E.V., Mushegian, A.R., Bork, P., 1996. Non-orthologous gene displacement. Trends Genet. 12, 334–336.
Lazcano, A., Forterre, P., 1999. The molecular search for the last common ancestor. Mol. Evol. 49, 411–412.
Luisi, P.L., Ferri, F., Stano, P., 2006. Approaches to semi-synthetic minimal cells. Naturwissenschaften 93, 1–13.
Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.
Michel, C.J., Pirillo, G., Pirillo, M.A., 2008. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theor. Comput. Sci. 401, 17–25.
Mushegian, A., 1999. The minimal genome concept. Curr. Opin. Genet. Dev. 9, 709–714.
Mushegian, A.R., Koonin, E.V., 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Nat. Acad. Sci. USA 93, 10268–10273.
Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., et al., 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399, 323–329.
Vagner, V., Dervyn, E., Ehrlich, S.D., 1998. A vector for systematic gene inactivation in *Bacillus subtilis*. Microbiology 144, 3097–3104.
Zhang, R., Ou, H., Zhang, C., 2004. DEG, a database of essential genes. Nucleic Acids Res. 32, D271–D272.
Zhang, R., Lin, Y., 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res. 37, D455–D458.