Journal of
Computer Science & Systems Biology

Ahmed and Michel J Comput Sci Syst Biol 2011, 4:1
DOI: 10.4172/jcsb.1000069

Research Article Open Access

Circular Code Signal in Frameshift Genes

Ahmed Ahmed and Christian J. Michel*

Equipe de Bioinformatique Théorique, FDBT, LSIIT (UMR UdS-CNRS 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 ILLKIRCH, FRANCE

Abstract

A trinucleotide circular code is a set of trinucleotides allowing the reading frame in genes to be retrieved locally, i.e. anywhere in genes and in particular without start codon, and automatically with a window of a few nucleotides. In 1996, a common circular code X has been identified simultaneously in two large populations of eukaryotic and prokaryotic genes. The method proposed here identifies periodic signals of this code X in the two frameshift types (+1 and –1) of both eukaryotic and prokaryotic frameshift genes. As expected by the code theory, the circular code modulo 3 signals move in the same direction of translational frameshifting. Finally, in 68% of frameshift genes in the RECODE 2 database, the frameshift type (+1 and –1) is automatically identified using only this circular code periodic signal. This circular code information constitutes a new structural property of frameshift genes. It may be used directly or in association with existing methods to identify frameshift genes in genomes and their encoded proteins.

Keywords: Circular code; Signal; Periodicity; Frame; Frameshift genes; Eukaryotic genes; Prokaryotic genes; Structural property

Introduction

In 1996, a trinucleotide circular code has been identified simultaneously in eukaryotic and prokaryotic genes [1,2]. It allows their reading frame to be retrieved. Frameshift genes, by bypassing or rereading one nucleotide, shift translation. Therefore, a theoretical forecast of circular codes should lead to a shift signal in this class of genes. This hypothesis is verified in this paper. Frameshifting and circular code are briefly presented.

Frameshifting

By our convention, the reading frame in a gene established by a start codon ATG, GTG or TTG is the frame 0, and the shifted frames 1 and 2 are the reading frame 0 shifted by one and two nucleotides in the 5′ – 3′ direction, respectively.

In the reading frame of a gene, a series of three nucleotides (codons) is translated into a series of amino acids according to the genetic code. The correspondence between the nucleotide sequence and the protein sequence is determined by genetic codes that are well-conserved across species.

However, in some cases, the protein synthesized by the ribosome does not correspond to the transcribed mRNA because of a translational error [3]. These translational errors are not the consequence of a genetic code alteration but they result mainly in a change in the ribosome translation site on the mRNA. This site modification could be expressed in different ways as a frameshifting, a stop codon reading through or a ribosomal hopping [4,5,6].

There exist two types of triggers that can stimulate the translational errors: instant and programmed. Instant translational errors are very rare with a rate estimated to 3×10^{-5} [7]. Programmed translational errors are produced by different means including a special motif, a secondary structure in the mRNA or a cell lack of an amino acid. These events can increase the probability of a translational error up to near 100% [4].

Frameshifting is a translational error where the ribosome pauses the translation then bypasses one nucleotide and hence shifts translation of the reading frame from frame 0 to frame 1 (+1 frameshift) (Figure 1) or it rereads one nucleotide and hence shifts translation of the reading frame from frame 0 to frame 2 (–1 frameshift) (Figure 2). Then, the ribosome continues to translate the codons in the shifted frame until it encounters a stop codon in its frame. In some rare cases, the number of nucleotides bypassed or rereaded may vary. The protein product is then partially encoded by frame 0 and partially encoded by this shifted frame [8,9].

The frameshift prediction using computational methods is a difficult task due to the diversity of motifs and secondary or tertiary structures which stimulate frameshifting. Besides, these computational methods generate many frameshift gene candidates. However, alignment methods of frameshift genes determined experimentally constitute a classical approach to identify frameshifts [10,11,12]. For the –1 frameshift identification, some computational methods consider the primary and secondary structure of mRNA associated with the frameshift site. This structure is classically composed of a slippery heptamer X,XXY,YYZ (X, Y and Z are three nucleotides and the comma "," represents the frame 0) separated by five to nine nucleotides from a stem loop (Figure 3). The approaches based on this structural model also allowed several frameshift genes to be identified (e.g. [13,14,15,16]).

Common circular code

In 1996, two statistical studies, a codon frequency per frame and a codon correlation function per frame, showed that the 64 trinucleotides $\mathbb{T}$ = {AAA,…,TTT} are preferentially distributed in the three frames of genes [1,2]. 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): eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686

***Corresponding author:** Christian J. Michel, Equipe de Bioinformatique Théorique, FDBT, LSIIT (UMR UdS-CNRS 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 ILLKIRCH, France, E-mail: ahmed@dpt-info.u-strasbg.fr, michel@dpt-info.u-strasbg.fr

Received January 07, 2011; **Accepted** January 18, 2011; **Published** January 20, 2011

Citation: Ahmed A, Michel CJ (2011) Circular Code Signal in Frameshift Genes. J Comput Sci Syst Biol 4: 007-015. doi:10.4172/jcsb.1000069

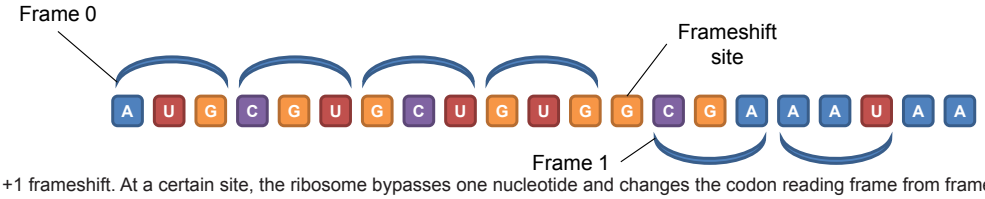

Figure 1: The +1 frameshift. At a certain site, the ribosome bypasses one nucleotide and changes the codon reading frame from frame 0 to frame 1.

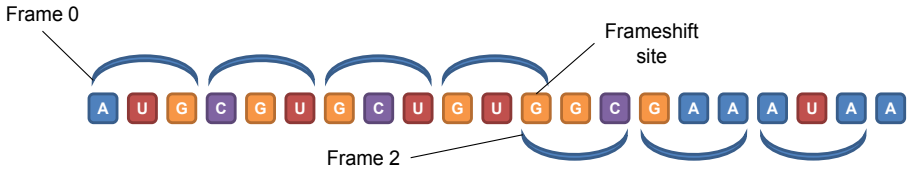

Figure 2: The –1 frameshift. At a certain site, the ribosome rereads one nucleotide and changes the codon reading frame from frame 0 to frame 2.

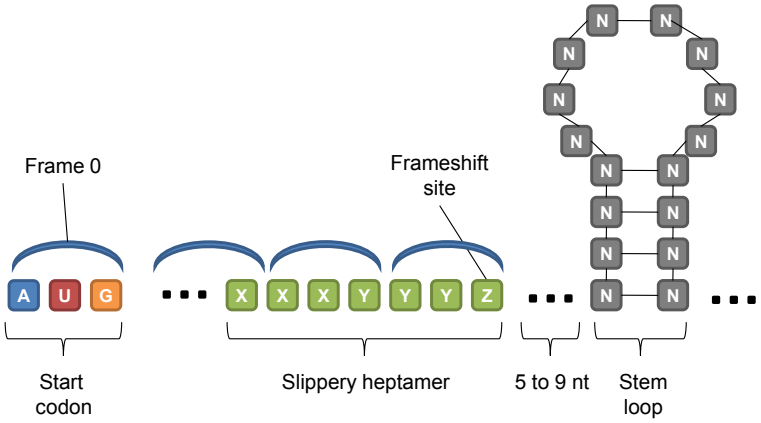

Figure 3: Frameshift slippery heptamer. Structural model of the –1 frameshift with a slippery heptamer and a stem loop separated by five to nine nucleotides.

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

Table 1: The common circular code X. 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.

sequences, 4,709,758 trinucleotides) (Table 1) [1,2]. These three trinucleotide subsets present several strong biomathematical properties, particularly the fact that they are circular codes.

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 respectively) over $\mathbb{A}_4$ is denoted by $\mathbb{A}_4^+$ ($\mathbb{A}_4^*$ respectively). Let $x_1 \ldots x_n$ be the concatenation of the words x_i for $i = 1,\ldots,n$.

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

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

Definition 3: Trinucleotide circular code: A circular code of words over $\mathbb{T}$ is called a trinucleotide circular code.

In other words, a trinucleotide circular code Y is a set of trinucleotides such that all sequences (e.g. genes) generated by concatenation of words of Y and written on a circle, where the last letter is followed by the first one, have only one decomposition (factorization) into words of Y. If a sequence has two decompositions then Y is not a circular code. As an example, let the set Y be composed of the six following trinucleotides: Y = {AAT,ATG,CCT,CTA,GCC,GGC} and the sequence w, be a series of the nine following letters: w = ATGGCCCTA. The sequence 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 (Figure 4). 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 sequence with Z, such as w for Y, and then Z is a circular code. The flower automaton is the classical algorithm in code theory to test if a set of words is a circular code or not (without

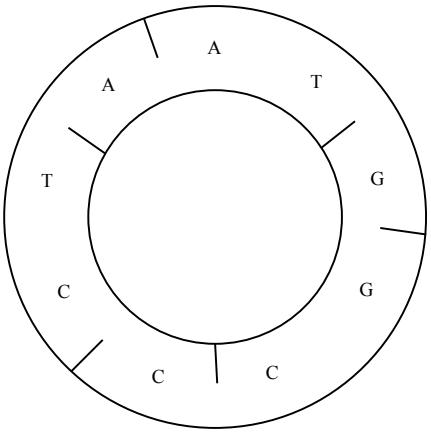

Figure 4: A counterexample of a circular code. The set Y = {AAT,ATG,CCT,CTA,GCC,GGC} is not a circular code as the sequence w = ATGGCCCTA written on a circle can be factorized into words of Y according to two different ways: ATG,GCC,CTA and AAT,GGC,CCT.

generating all the possible sequences). Furthermore, it can also determine the minimal window length for retrieving the construction frame (size of the longest ambiguous word which can be read in at least two frames, plus one letter) (detailed in [17]).

Definition 4: **Complementary map:** The complementary map $\mathcal{C}:\mathbb{A}_4^+ \to \mathbb{A}_4^+$ is defined by $\mathcal{C}(\mathrm{A})=\mathrm{T}$, $\mathcal{C}(\mathrm{C})=\mathrm{G}$, $\mathcal{C}(\mathrm{G})=\mathrm{C}$, $\mathcal{C}(\mathrm{T})=\mathrm{A}$. 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. $\mathcal{C}(\mathrm{AAC})=\mathrm{GTT}$.

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

Definition 6: **Circular permutation map:** The circular permutation map $\mathcal{P}: \mathbb{T} \to \mathbb{T}$ permutes circularly each trinucleotide $w_0 = l_0 l_1 l_2$ as follows: $\mathcal{P}(w_0) = w_1 = l_1 l_2 l_0$, e.g. $\mathcal{P}$(AAC) = ACA. The kth iterate of $\mathcal{P}$ is denoted $\mathcal{P}^k$, e.g. $\mathcal{P}^2$(AAC) = CAA.

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

Definition 8: **C^3 trinucleotide circular code:** A trinucleotide circular code Y is C^3 if $Y_1 = \mathcal{P}(Y)$ and $Y_2 = \mathcal{P}^2(Y)$ are circular codes.

Remark 1: 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 circular code gene population: A gene population $\mathcal{F}$ has the C^3 trinucleotide circular code property if the three sets $\mathcal{F}(X_0)$, $\mathcal{F}(X_1)$ and $\mathcal{F}(X_2)$ of trinucleotides with the highest occurrence frequency in frame 0, 1 and 2, respectively, of genes $\mathcal{F}$, are C^3 trinucleotide circular codes.

Definition 10: **C^3 self-complementary trinucleotide circular code:** A trinucleotide circular code Y is C^3 self-complementary if Y is a C^3 trinucleotide circular code satisfying the following properties: $Y = \mathcal{C}(Y)$ (self-complementary trinucleotide circular code), $\mathcal{C}(Y_1) = Y_2$ and $\mathcal{C}(Y_2) = Y_1$ (Y_1 and Y_2 are complementary trinucleotide circular codes).

The trinucleotide set X_0 coding the reading frames (frames 0) of eukaryotic and prokaryotic genes is a C^3 self-complementary trinucleotide circular code [1]. X_0 will be also noted C^3 code 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 C^3 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 minimal window of 13 nucleotides in each frame (Property 5 and Remark at page 121 in [1]).

A recent review of circular codes in genes details the research context, the history and the other properties of this C^3 code X (rarity, largest window length, higher frequency of "misplaced" trinucleotides in shifted frames, flexibility) [18].

Method

Circular code signal (CCS) method

We have developed circular code signal (CCS) methods to identify protein coding genes (e.g. [19]), global maximum or minimum of the C^3 code X in micro RNAs [20], etc. This approach is extended and applied here for the case of frameshift genes. It will reveal periodic signals of the C^3 code X.

Let $\mathbb{T}$ = {AAA,…,TTT} be the set of 64 trinucleotides over the 4-nucleotide alphabet $\mathcal{N}$ = {A,C,G,T}. Let $n \in \mathcal{N}$ be a nucleotide, $t \in \mathbb{T}$, a trinucleotide and $\mathcal{F}$, a frameshift gene population with $N(\mathcal{F})$ sequences s. By convention, the position i = 0 refers to the frameshift site in a frameshift gene s and is associated with the nucleotide n_0. In a +1 frameshift gene, n_0 is the bypassed nucleotide and in a −1 frameshift gene, n_0 is the reread nucleotide. By convention, all nucleotides before the frameshift nucleotide (in the 5′ direction) have a negative position $i < 0$ and all nucleotides after the frameshift nucleotide (in the 3′ direction), a positive position $i > 0$. The sequence s is considered as a series of nucleotides n_i and the method reading frame is … n_{-1}, n_0, n_1, …. The sequence s begins at position s_{min} (negative value) and ends at position s_{max} (positive value) with a length equal to $s_{max} - s_{min} + 1$.

All the sequences s of $\mathcal{F}$ are aligned and centered according to their frameshift site i = 0. Hence, due to the variations of sequence lengths, a position i may not exist with some sequences. Let $N(i,\mathcal{F})$ be the number of sequences s of the population $\mathcal{F}$ having a position i. A population $\mathcal{F}$ has a minimum i position noted $\mathcal{F}_{min} = Min_{s\in\mathcal{F}}\{s_{min}\}$ (negative value) and a maximum i position noted $\mathcal{F}_{max} = Max_{s\in\mathcal{F}}\{s_{max}\}$ (positive value). Obviously, at position i = 0, the number of sequences is maximal and $N(0,\mathcal{F}) = N(\mathcal{F})$. Let the frame 0 of a sequence s be the frame established by the trinucleotide $n_0 n_1 n_2$, and the shifted frames 1 and 2 be the frame 0 shifted by one and two nucleotides in the 5′− 3′ direction, respectively. Thus, the method reading frame is not necessarily the classical codon reading frame.

Let $w_i = n_i n_{i+1} \ldots n_{i+13}$ be a window of length $|w| = 4$ trinucleotides plus 2 nucleotides added to consider the two shifted frames in this window, where n_i is the ith nucleotide in s. Let $t_i^{l,\phi}$ be the lth trinucleotide, $l \in \{0,1,2,3\}$, in frame ϕ, $\phi \in \{0,1,2\}$, of w_i (Figure 5). Let X_f, $f \in \{0,1,2\}$, be the three codes X_0, X_1 and X_2 in the three frames f.

In a given window w_i, the function $\delta_f(t_i^{l,\phi})$ indicates whether or not the trinucleotide $t_i^{l,\phi}$ belongs to the code X_f

$$\delta_f\left(t_i^{l,\phi}\right)=\begin{cases}1 & \text{if } t_i^{l,\phi}\in X_f\\ 0 & \text{otherwise}\end{cases}. \quad (1)$$

Then, the score function $P(X_f,i,s)$ computes the occurrence of the code X_f in its associated frame, i.e. $\phi = f$, in a window w_i of a sequence s

$$P\left(X_f,i,s\right)=\frac{1}{|w|}\sum_{l=0}^{|w|-1}\delta_f\left(t_i^{l,f}\right).$$

The score function $P(i,s)$ computes the average occurrence of the C^3 code X in a window w_i of a sequence s

$$P\left(i,s\right)=\frac{1}{3}\sum_{f=0}^{2}P\left(X_f,i,s\right).$$

Finally, the score function $P(i,\mathcal{F})$ computes the occurrence of the C^3 code X in a window w_i (X_0, X_1 and X_2 in frames 0, 1 and 2 of w_i, respectively) for all sequences s having a position i in $\mathcal{F}$

$$P\left(i,\mathcal{F}\right)=\frac{1}{N\left(i,\mathcal{F}\right)}\sum_{s\in\mathcal{F}\ |\ i\in[s_{min},s_{max}]}P\left(i,s\right). \quad (2)$$

Proposition 1: If in each sequence s of $\mathcal{F}$ the trinucleotides $t_i^{l,\phi}$ of a window w_i belong to the set X_f such that $\phi = f$ then $P(i,\mathcal{F}) = 1$ and the intensity of the C^3 code X is maximum.

Proposition 2: If in each sequence s of $\mathcal{F}$ the trinucleotides $t_i^{l,\phi}$ of a window w_i belong to the set X_f such that $\phi \neq f$ then $P(i,\mathcal{F}) = 0$ and the C^3 code X is absent.

Proposition 3: $0 \leq P(i,\mathcal{F}) \leq 1$ (consequence of Propositions 1 and 2).

Circular code periodic signals revealed by the CCS method

The definitions used in this CCS method have properties which allow circular code periodic signals in genes to be identified.

Property 1: The indicator function (Formula 1) is based on a set X of 20 trinucleotides which has the following property: it contains no permuted trinucleotide $\mathcal{P}(w)$ (Definition 7), and thus, in particular, no trinucleotide $\mathbb{T}_{id}$ = {AAA,CCC,GGG,TTT}. For example, on the purine/pyrimidine alphabet {R,Y} (R={A,G}, Y={C,T}), it was already explained that the class of motifs RRR and YYY have no modulo 3 periodicity in genes (Figure 2 in [21]) as they do not occur in a preferential frame (Table 3(a) in [1]).

Property 2: The indicator function (Formula 1) is based on three sets X_f associated with the three frames f, $f \in \{0,1,2\}$, so that $X_1 = \mathcal{P}(X_0)$ and $X_2 = \mathcal{P}^2(X_0)$, i.e. the frame 1 (2 respectively) is analysed with the trinucleotides of X_0 permuted by one (two respectively) nucleotides. Note that X_1 and X_2 also do not have trinucleotides $\mathbb{T}_{id}$.

Property 3: The window length of 14 nucleotides is in correspondence with the length of the minimal window of 13 nucleotides of the three circular codes X_0, X_1 and X_2 to retrieve the frames 0, 1 and 2 in genes [18]. This sliding window length is very short compared to some classical methods of signal processing for genes ([22] and their subsequent works).

Property 4: If a gene population $\mathcal{F}$ has the C^3 code X property (Definition 9) then $P(i,\mathcal{F})$ has a modulo 3 periodicity.

Indeed, a high score value $P(i,\mathcal{F})$ in a window w_i at position i followed by two low score values at positions $i + 1$ and $i + 2$ reflect a high probability that the reading frame of trinucleotides of X_0 is in this frame i. From the other side, if the window w_i is in the reading frame then the trinucleotides of X_0 are not identified in frame 0 of w_{i+1} and w_{i+2} but in frames 2 and 1, respectively. Hence, the score values of windows w_{i+1} and w_{i+2} are low. A similar reasoning holds for the codes X_1 and X_2 which are permuted trinucleotide circular codes of X_0 (Definitions 7 and 8).

Significance level of a modulo 3 periodicity

A modulo 3 periodicity is quantified by counting the local peaks on a frame according to the following indicator function $\delta_P(i,\mathcal{F})$

$$\delta_P\left(i,\mathcal{F}\right)=\begin{cases}1 & \text{if } P\left(i-1,\mathcal{F}\right)\leq P\left(i,\mathcal{F}\right) \text{ and } P\left(i,\mathcal{F}\right)\geq P\left(i+1,\mathcal{F}\right)\\ 0 & \text{otherwise}\end{cases}.$$

The significance of this indicator function $\delta_P(i,\mathcal{F})$ can be evaluated with a binomial test. For a population $\mathcal{F}$ with i modulo 3 positions in the range $[a, b]$, let $X_i\left(\mathcal{F}\right)=\delta_P\left(i,\mathcal{F}\right)$ be the Bernoulli random variable which is equal to 1 with the probability $p_0 = ⅓$ (one chance out of 3 that $P(i,\mathcal{F})$ at position i is greater than $P(i-1,\mathcal{F})$ at position $i - 1$ and $P(i+1,\mathcal{F})$ at position $i + 1$), and to 0 with the probability $1 - p_0 = ⅔$. Its probability law is

$$Prob\left(X_i\left(\mathcal{F}\right)=x\right)=p_0{}^x\left(1-p_0\right)^{1-x} \text{ for } x=0,1.$$

In the range $[a,b]$ of a population $\mathcal{F}$, the sum of independent variables $X_i\left(\mathcal{F}\right)$ is a Binomial random variable $Y\left(\mathcal{F}\right)=\sum_{i=1}^{n}X_i\left(\mathcal{F}\right)$ of parameter $p_0 = ⅓$ and of order $n=\left\lfloor (b-a+1)/3\right\rfloor$ where $\lfloor k\rfloor$ is the largest integer not greater than k. Thus, n is the number of i modulo 3 positions in the range $[a,b]$ of $\mathcal{F}$. The variable $Y(\mathcal{F})$ counts the number y of modulo 3 maxima among the n possible values of i in $[a,b]$ of $\mathcal{F}$. Its probability law is

$$Prob\left(Y\left(\mathcal{F}\right)=y\right)=\frac{n!}{y!\left(n-y\right)!}p_0{}^y\left(1-p_0\right)^{n-y} \text{ for } y=0,1\ldots,n.$$

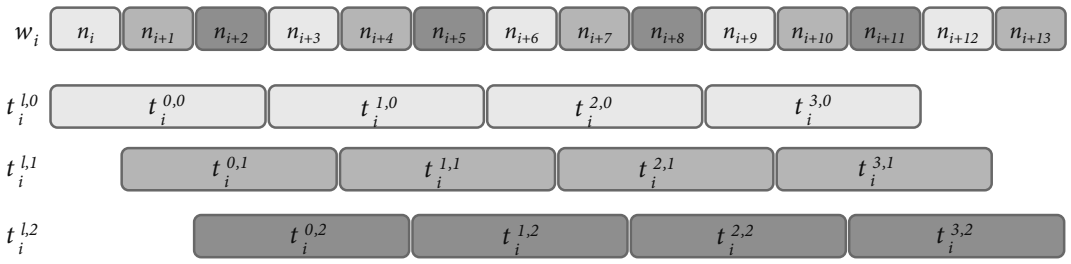

Figure 5: The structure of the sliding window w_i of 14 nucleotides analysing the three shifted frames.

Its expected value is $E(Y(\mathcal{F})) = np_0 = n/3$ and its variance, $V(Y(\mathcal{F})) = np_0(1-p_0) = 2n/9$.

Let

$$p = \frac{y}{n} \quad (3)$$

be the observed probability of modulo 3 maxima in the range $[a,b]$ among the $n = \lfloor (b - a + 1)/3 \rfloor$ possible values. For a population $\mathcal{F}$, the null hypothesis H_0: $p = p_0 = ⅓$ is tested against the alternative hypothesis H_1: $p > p_0 = ⅓$. The hypothesis H_0 is associated with a random curve in $[a,b]$. The hypothesis H_1 with $p = 1$ is associated with a (perfect) modulo 3 periodicity in $[a,b]$, and with $⅓ < p < 1$, with an incomplete modulo 3 periodicity in $[a,b]$. Knowing the number y of maxima in $[a,b]$, the significance level α of the one-tailed Binomial test is determined as follows

$$\alpha = Prob(Y(\mathcal{F}) > y) = 1 - Prob(Y(\mathcal{F}) \le y) = \sum_{i=0}^{y} \frac{n!}{i!(n-i)!} p_0^{\,i} (1-p_0)^{n-i}. \quad (4)$$

Note: When n is sufficiency large and p is not too close to 0 or 1, i.e. $n > Max\{5/p_0, 5/(1-p_0)\} = 15$, the central limit theorem applies and the approximation of the normal distribution $Z(\mathcal{F})$ to the binomial distribution $Y(\mathcal{F})$ can be used as follows

$$Z(\mathcal{F}) = \frac{Y(\mathcal{F}) - E(Y(\mathcal{F}))}{\sqrt{V(Y(\mathcal{F}))}}$$

and

$$z = \frac{\frac{y}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Knowing the number y of modulo 3 maxima in $[a,b]$, the significance level α of the one-tailed normal test is determined as follows

$$\alpha = Prob(Z(\mathcal{F}) > z) = 1 - Prob(Z(\mathcal{F}) \le z) = 1 - \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt.$$

Exact binomial tests (Formula 3) will be used to evaluate the significance levels α in Figures 6-10 and in Table 3.

In order to automate the reading frame identification, let $p(f,a,b,\mathcal{F})$ be the observed probability of modulo 3 maxima in the frame f of the range $[a,b]$ in the frameshift gene population $\mathcal{F}$. Then, the reading frame $\not{f}$, $\not{f} \in \{0,1,2\}$, in the position interval $[a,b]$ in the frameshift population $\mathcal{F}$ is

$\not{f} = f$ such that $p(f,a,b,\mathcal{F}) = Max_{i=0}^{2}\{p(i,a,b,\mathcal{F})\}$.

This statistical approach can be easily extended to evaluate any type of periodicity (modulo 2, modulo 3, etc.).

Data acquisition

The frameshift gene populations $\mathcal{F}$ used in this study are extracted from the RECODE 2 database (release 2010; [23,24, 25]). The RECODE 2 database is a compilation of programmed translational recoding events. It deals with programmed ribosomal frameshifts, codon redefinition and translational bypass occurring in a variety of organisms. Each entry includes the gene, its encoded protein for both normal and alternate decoding, the type of the recoding event involved and the *trans*-factors and *cis*-elements that influence recoding.

Our study concerns the –1 and +1 frameshifts of eukaryotic and prokaryotic genes as the common circular code X has been identified only in these populations and not, for example, in viral genes [1]. Therefore, four gene populations $\mathcal{F}$ are extracted according to the frameshift type and the organism kingdom: –1 frameshifts of eukaryotes EUK^{-1} and prokaryotes PRO^{-1}, and +1 frameshifts of eukaryotes EUK^{+1} and prokaryotes PRO^{+1}. Table 2 shows the kingdom, the shift type, the number of genes, the minimum i position $\mathcal{F}_{min}$ and the maximum i position $\mathcal{F}_{max}$ of the studied frameshift gene populations.

Results

Figures 6-9 show the graphical results of the score function $P(i, \mathcal{F})$ (Formula 2) computed on the four frameshift gene populations $\mathcal{F}$ of eukaryotes and prokaryotes EUK^{+1}, PRO^{+1}, EUK^{-1} and PRO^{-1}, respectively. The x-axis represents the position i of the sliding window in $\mathcal{F}$ and the y-axis, the score value $P(i,\mathcal{F})$. For display purposes, the

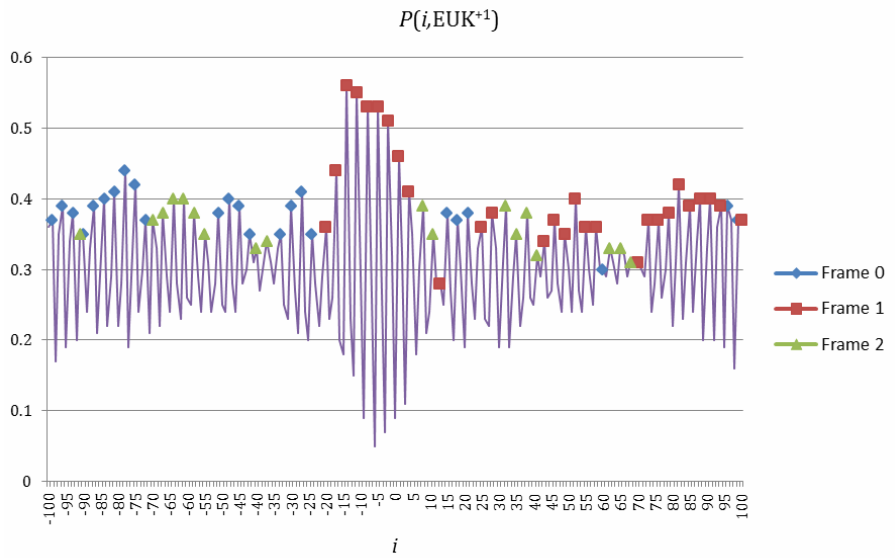

Figure 6: The score function $P(i, EUK^{+1})$ (Formula 2) computed on +1 frameshift of eukaryotic genes EUK^{+1}. Before the frameshift site, the majority of local peaks (y = 18, n = 34, y/n = 53%, α = 0.565% with Formula 4) are in frame 0 (rhombuses). After the frameshift site, the majority of local peaks (y = 21, n = 34, y/n = 62%, α = 0.018% with Formula 4) are in frame 1 (squares).

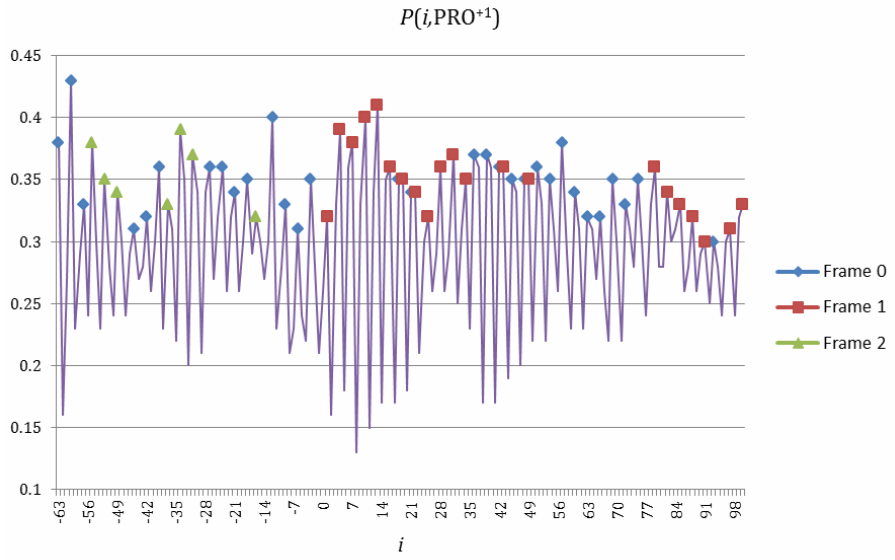

Figure 7: The score function $P(i, \mathrm{PRO}^{+1})$ (Formula 2) computed on +1 frameshift of prokaryotic genes PRO^{+1}. Before the frameshift site, the majority of local peaks (y = 14, n = 21, y/n = 67%, α = 0.040% with Formula 4) are in frame 0 (rhombuses). After the frameshift site, the majority of local peaks (y = 21, n = 34, y/n = 62%, α = 0.018% with Formula 4) are in frame 1 (squares).

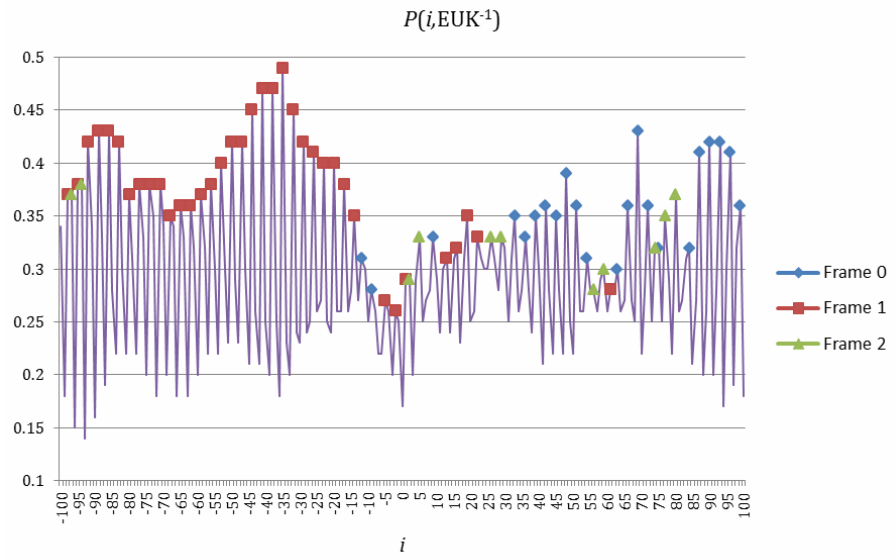

Figure 8: The score function $P(i, \mathrm{EUK}^{-1})$ (Formula 2) computed on −1 frameshift of eukaryotic genes EUK^{-1}. Before the frameshift site, the majority of local peaks (y = 32, n = 34, y/n = 94%, $\alpha \approx 10^{-15}$ with Formula 4) are in frame 1 (squares). After the frameshift site, the majority of local peaks (y = 21, n = 34, y/n = 62%, α = 0.018% with Formula 4) are in frame 0 (rhombuses).

graphical results are presented only on an interval of 200 nucleotides around the frameshift position $i = 0$.

A modulo 3 periodicity is observed in these four frameshift gene populations. It means that the C^3 code X is a main primary structure in these populations. In Figures 6-9, the local peaks are marked according to the frame of their position i. The peaks in frame $f = 0$ of the sequence s (i mod 3 = 0) are marked by rhombuses, and those in frames $f = 1$ (i mod 3 = 1) and $f = 2$ (i mod 3 = 2), by squares and triangles, respectively. It must be reported again that the frames here are established according to the frameshift position $i = 0$ and not as usual by the start codon of the sequence s, hence $f = 0$ is not necessarily the classical codon reading frame.

The frame that contains most of the local peaks before the frameshift site differs from the one after. Precisely, in Figure 6 of EUK^{+1} and Figure 7 of PRO^{+1}, before the frameshift site, the majority of peaks (53% with a significance level α = 0.565% and 67% with α = 0.040%, respectively) are in frame 0 and after the frameshift site, the majority of peaks (both 62% with α = 0.018%) are in frame 1. In Figure 8 of EUK^{-1} and Figure 9 of PRO^{-1}, before the frameshift site, the majority of peaks (94% with $\alpha \approx 10^{-15}$ and 65% with α = 0.005%, respectively) are in frame 1 and after the frameshift site, the majority of peaks (62% with α = 0.018% and 88% with $\alpha \approx 10^{-12}$, respectively) are in frame 0. This change of periodicity frame after the frameshift site is clearly observed in Figures 6-9. In genes without frameshift sites, almost all local peaks are in the

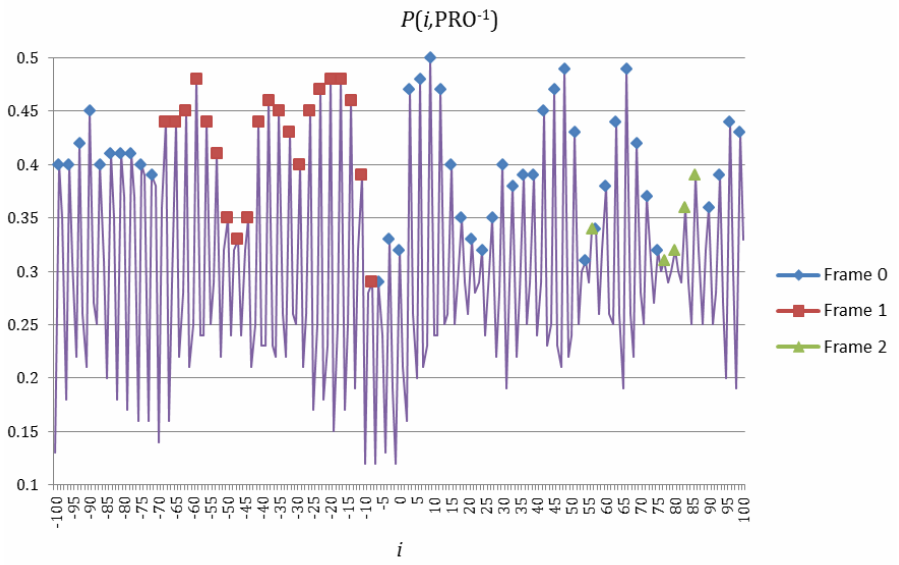

Figure 9: The score function $P(i, \mathrm{PRO}^{-1})$ (Formula 2) computed on −1 frameshift of prokaryotic genes PRO^{-1}. Before the frameshift site, the majority of local peaks (y = 22, n = 34, y/n = 65%, α = 0.005% with Formula 4) are in frame 1 (squares). After the frameshift site, the majority of local peaks (y = 30, n = 34, y/n = 88%, $\alpha \approx 10^{-12}$ with Formula 4) are in frame 0 (rhombuses).

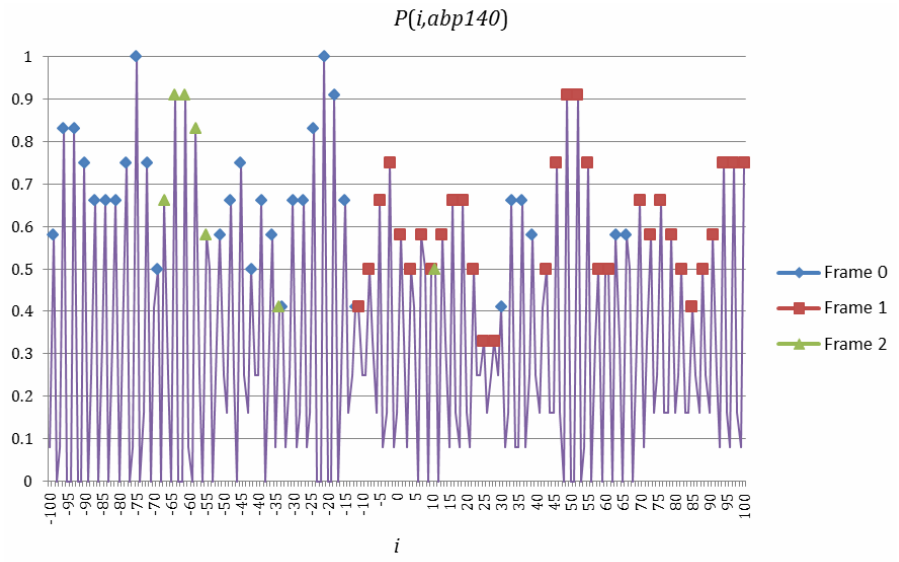

Figure 10: The score function $P(i,abp140)$ (Formula 2) computed on the +1 frameshift gene *abp140*. Before the frameshift site, the majority of local peaks (y = 25, n = 34, y/n = 74%, $\alpha \approx 10^{-7}$ with Formula 4) are in frame 0 (rhombuses). After the frameshift site, the majority of local peaks (y = 28, n = 34, y/n = 82%, $\alpha \approx 10^{-9}$ with Formula 4) are in frame 1 (squares).

same frame (see for example the pioneer papers at the gene and gene population levels published by [26,27,28] etc.).

Table 3 shows the observed probability $p(\phi,a,b,\mathcal{F})$ of modulo 3 maxima in the three frames ϕ of the ranges $[a,b]$ in the four studied frameshift gene populations $\mathcal{F}$ before and after their frameshift sites. As expected and in agreement with the figure's results, in the +1 frameshift genes, the reading frames identified are the frame $\phi = 0$ before the frameshift site and the frame $\phi = 1$ after the frameshift site. In the −1 frameshift genes, the reading frames identified are the frame $\phi = 1$ before the frameshift site and the frame $\phi = 0$ after the frameshift site. Therefore, the periodicity frame associated to the C^3 code X moves in the same direction of translational frameshifting.

This change of periodicity frame is also observed at the individual gene level. For example, Figure 10 shows the computation of $P(i,\mathcal{F})$ (Formula 2) on the gene *abp140* which is a +1 frameshift eukaryotic gene identified experimentally [29]. Before the frameshift site, the majority of local peaks (74% with $\alpha \approx 10^{-7}$) are in frame 0 while after the frameshift site, the majority of local peaks (82% with $\alpha \approx 10^{-9}$) are in frame 1. This circular code signal approach by studying the local peaks before and after the frameshift position in individual frameshift genes of the RECODE 2 database identifies successfully the frameshift type (+1 and −1) in 68% of genes.

Discussion

Periodic signals of the common circular code X are identified in the +1 and −1 frameshift genes of both eukaryotes and prokaryotes.

$\mathcal{F}$	Kingdom	Shift type	Number of genes	$\mathcal{F}_{min}$	$\mathcal{F}_{max}$
EUK^{+1}	Eukaryote	+1	37	–1302	4250
PRO^{+1}	Prokaryote	+1	50	–1146	1117
EUK^{-1}	Eukaryote	-1	27	–2849	4170
PRO^{-1}	Prokaryote	-1	27	–1451	953

Table 2: The four studied frameshift genes populations. Kingdom, shift type, number of genes, minimum i position $\mathcal{F}_{min}$ and maximum i position $\mathcal{F}_{max}$ of the studied frameshift gene populations extracted from the RECODE 2 database (release 2010) are given.

$\mathcal{F} = \mathrm{EUK}^{+1}$			
f	$p(f,\mathcal{F}_{min},0,\mathcal{F})$	$p(f,1,\mathcal{F}_{max},\mathcal{F})$	$p(f,1,\mathcal{F}_{max},\mathcal{F})-p(f,\mathcal{F}_{min},0,\mathcal{F})$
0	**54** ($\alpha \approx 10^{-19}$)	22	-32
1	20	**64** ($\alpha \approx 10^{-120}$)	+44
2	30	19	-11

$\mathcal{F} = \mathrm{PRO}^{+1}$			
f	$p(f,\mathcal{F}_{min},0,\mathcal{F})$	$p(f,1,\mathcal{F}_{max},\mathcal{F})$	$p(f,1,\mathcal{F}_{max},\mathcal{F})-p(f,\mathcal{F}_{min},0,\mathcal{F})$
0	**59** ($\alpha \approx 10^{-25}$)	47	-12
1	20	**51** ($\alpha \approx 10^{-12}$)	+31
2	31	8	-23

$\mathcal{F} = \mathrm{EUK}^{-1}$			
f	$p(f,\mathcal{F}_{min},0,\mathcal{F})$	$p(f,1,\mathcal{F}_{max},\mathcal{F})$	$p(f,1,\mathcal{F}_{max},\mathcal{F})-p(f,\mathcal{F}_{min},0,\mathcal{F})$
0	24	**73** ($\alpha \approx 10^{-198}$)	+49
1	**60** ($\alpha \approx 10^{-64}$)	13	-47
2	25	19	-6

$\mathcal{F} = \mathrm{PRO}^{-1}$			
f	$p(f,\mathcal{F}_{min},0,\mathcal{F})$	$p(f,1,\mathcal{F}_{max},\mathcal{F})$	$p(f,1,\mathcal{F}_{max},\mathcal{F})-p(f,\mathcal{F}_{min},0,\mathcal{F})$
0	13	**47** ($\alpha \approx 10^{-7}$)	+34
1	**81** ($\alpha \approx 10^{-105}$)	44	-37
2	8	16	+8

Table 3: Observed probability $p(f,a,b,\mathcal{F})$ (in % rounded) of modulo 3 maxima in the three frames f of the ranges $[a,b]$ in the four studied frameshift gene populations $\mathcal{F}$. For each population $\mathcal{F}$, this probability is computed for two position intervals. The first interval is the region preceding the frameshift site and ranging from the minimum i position $\mathcal{F}_{min} = a$ (Table 2) of the population $\mathcal{F}$ to the frameshift position $b = 0$. The second interval is the region following the frameshift site and ranging from the frameshift site $a = 1$ to the maximum i position $\mathcal{F}_{max} = b$ (Table 2) of the population $\mathcal{F}$. The values in bold indicate the identified reading frames. In all populations, the circular code periodic signals move in the same direction of translational frameshifting.

Furthermore, the circular code periodic signal shifts in the same direction of translational frameshifting. This last result confirms a theoretical forecast of circular codes. Indeed, if a circular code is associated with a unique decomposition of a reading frame, a frameshift gene with two successive reading frames should have a shift of this circular code signal.

The statistical results observed suggest two hypotheses concerning the ribosomal translational frameshifting origin. The first hypothesis assumes that the frameshift gene structure is composed of two regions (before and after the frameshift site) which have the same circular code distribution but not in the same frame. When the ribosome scans the mRNA and reaches the second region, then it detects a change in the code distribution and shifts the mRNA forward or backward in order to retrieve the common distribution. The second hypothesis assumes that the circular code distribution is the same in the entire mRNA. A particular motif or secondary structure generates causes the ribosome to shift the reading frame at a certain position. This shift drove frameshift genes to evolve their internal structure by adding or deleting a nucleotide to overcome the shifting side effect.

The common circular code is a structural property associated to genes. The statistical analysis of the RECODE 2 database (release 2010) revealed here that the frameshift genes also have this code property in their primary structure whatever the frameshift type (+1 and –1) and whatever the species kingdom (eukaryotes and prokaryotes).

The properties of this developed CCS method may lead to some considerations in signal processing of DNA sequences in the particular case of genes. Periodicities can be revealed in genes with the following window features: an unexpected short sliding window of 14 nucleotides; a window content based on a subset of trinucleotides, precisely 20 trinucleotides; a subset of 20 trinucleotides with particular properties, precisely no permuted trinucleotides; and a different subset of 20 trinucleotides for each gene frame so that the three subsets are deduced from each other by a number of permutations related to the frame shift, precisely X_1 (X_2 respectively) in frame 1 (2 respectively) is deduced by one (two respectively) permutation of 20 trinucleotides of X_0.

This circular code information may be used directly or combined with existing methods to improve the identification of frameshift genes in genomes and their encoded proteins.

References

1. Arquès DG, Michel CJ (1996) A complementary circular code in the protein coding genes. J Theor Biol 182: 45-58.
2. Arquès DG, Michel CJ (1997) A code in the protein coding genes. Biosystems 44: 107-134.
3. Gesteland RF, Weiss RB, Atkins JF (1992) Recoding: Reprogrammed genetic decoding. Science 257: 1640-1641.
4. Farabaugh PJ (1996) Programmed translational frameshifting. Annual Rev Genetics 30: 507-528.
5. Namy O, Naphtine S, Rousset JP, Brierley I (2004) Reprogrammed genetic decoding in cellular gene expression. Mol Cell 13: 157-168.
6. Cobucci-Ponzano B, Rossi M, Moracci M (2005) Recoding in archaea. Mol Microbiol 55: 339-348.
7. Parker J (1989) Errors and alternatives in reading the universal genetic code. Microbiol Rev 53: 273-298.
8. Craigen WJ, Caskey CT (1987) Translational frameshifting: Where will it stop? Cell 50: 1-2.
9. Atkins JF, Weiss RB, Gesteland RF (1990) Ribosome gymnastics - Degree of difficulty 9.5, style 10.0. Cell 62: 413-423.
10. Pook MA, Al-Mahdawi SA, Thomas NH, Appleton R, Norman A, et al. (2000) Identification of three novel frameshift mutations in patients with Friedreich's ataxia. Med Genet 37: e38.
11. Mori Y, Yin J, Rashid A, Leggett BA, Young J, et al. (2001) Instabilotyping: comprehensive identification of frameshift mutations caused by coding region microsatellite instability. Cancer Res 61: 6046-6049.
12. Hahn Y, Lee B (2005) Identification of nine human-specific frameshift mutations

by comparative analysis of the human and the chimpanzee genome sequences. Bioinformatics 21: i186-i194.

13. Hammell AB, Taylor RC, Peltz SW, Dinman JD (1999) Identification of putative programmed –1 ribosomal frameshift signals in large DNA databases. Genome Res 9: 417-427.

14. Moon S, Byun Y, Kim HJ, Jeong S, Han K (2004) Predicting genes expressed via –1 and +1 frameshifts. Nucleic Acids Res 32: 4884-4892.

15. Theis C, Reeder J, Giegerich R (2008) KnotInFrame: prediction of –1 ribosomal frameshift events. Nucleic Acids Res 36: 6013-6020.

16. Liao PY, Choi YS, Lee KH (2009) FSscan: a mechanism based program to identify +1 ribosomal frameshift hotspots. Nucleic Acids Res 37: 7302-7311.

17. Berstel J, Perrin D (1985) Theory of Codes. Academic Press, London.

18. Michel CJ (2008) A 2006 review of circular codes in genes. Comput Math Applic 55: 984-988.

19. Arquès DG, Lacan J, Michel CJ (2002) Identification of protein coding genes in genomes with statistical functions based on the circular code. Biosystems 66: 73-92.

20. Ahmed A, Michel CJ (2008) Plant microRNA detection using the circular code information. Comput Biol Chem 32: 400-405.

21. Arquès DG, Michel CJ (1993) Identification and simulation of new non-random statistical properties common to different eukaryotic gene subpopulations. Biochimie 75: 399-407.

22. Arquès DG, Michel CJ (1987) A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups. J Theor Biol 128: 457-461.

23. Baranov PV, Gurvich OL, Fayet O, Prere MF, Miller WA et al. (2001) RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. Nucleic Acids Res 29: 264-267.

24. Baranov PV, Gurvich OL, Hammer AW, Gesteland RF, Atkins JF (2003) RECODE 2003. Nucleic Acids Res 31: 87-89.

25. Bekaert M, Firth AE, Zhang Y, Gladyshev VN, Atkins JF, et al. (2010) Recode-2: new design, new search tools, and many more genes. Nucleic Acids Res 38: D69-D74.

26. Shepherd JCW (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78: 1596-1600.

27. Fickett JW (1982) Recognition of protein coding regions in DNA sequences. Nucl Acids Res 10: 5303-5318.

28. Michel CJ (1986) New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation. J Theor Biol 120: 223-236.

29. Asakura T, Sasaki T, Nagano F, Satoh A, Obaishi H, et al. (1998) Isolation and characterization of a novel actin filament-binding protein from Saccharomyces cerevisiae. Oncogene 16: 121-130.