Frameshift Signals in Genes Associated with the Circular Code

Ahmed Ahmed, Gabriel Frey and Christian J. Michel*

Equipe de Bioinformatique Théorique, LSIIT (UMR CNRS-ULP 7005), Université Louis Pasteur de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France
Tel.: +33 3 90 24 44 62; E-mail: ahmed@dpt-info.u-strasbg.fr; frey@dpt-info.u-strasbg.fr

Edited by E. Wingender; received 13 October 2006; revised 22 January 2007; accepted 18 February 2007; published 11 April 2007

ABSTRACT: Three sets of 20 trinucleotides are preferentially associated with the reading frames and their 2 shifted frames of both eukaryotic and prokaryotic genes. These 3 sets are circular codes. They allow retrieval of any frame in genes (containing these circular code words), locally anywhere in the 3 frames and in particular without start codons in the reading frame, and automatically with the reading of a few nucleotides. The circular code in the reading frame, noted X, which can deduce the 2 other circular codes in the shifted frames by permutation, is the information used for analysing frameshift genes, i. e. genes with a change of reading frame during translation. This work studies the circular code signal around their frameshift sites. Two scoring methods are developed, a function P based on this code X and a function Q based both on this code X and the 4 trinucleotides with identical nucleotides. They detect a significant correlation between the code X and the -1 frameshift signals in both eukaryotic and prokaryotic genes, and the $+1$ frameshift signals in eukaryotic genes.

KEYWORDS: Frameshift gene, frameshift signal, circular code, trinucleotide, frame, statistical method

INTRODUCTION

Frameshift

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

In the reading frame, a series of 3 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 often considered as an immutable dogma. Nevertheless, the standard rule of genetic decoding is altered in specific genes by different events that are globally termed recoding [Gesteland *et al.*, 1992]. These events are classified into 2 main groups: those occurring at the translation termination step (stop codon readthrough) and those occurring during elongation, such as frameshifts (reviewed e.g. in Farabaugh, 1996; Namy *et al.*, 2004; Cobucci-Ponzano *et al.*, 2005).

Frameshift is conceptually a simple process. At a particular position in the nucleotide sequence, the ribosome shifts its reading frame into a new one. It will continue reading in a shifted frame until

[*] Corresponding author. E-mail: michel@dpt-info.u-strasbg.fr.

Electronic publication can be found in *In Silico Biol.* **7**, 0016 <http://www.bioinfo.de/isb/2007/07/0016/>, 11 April 2007.

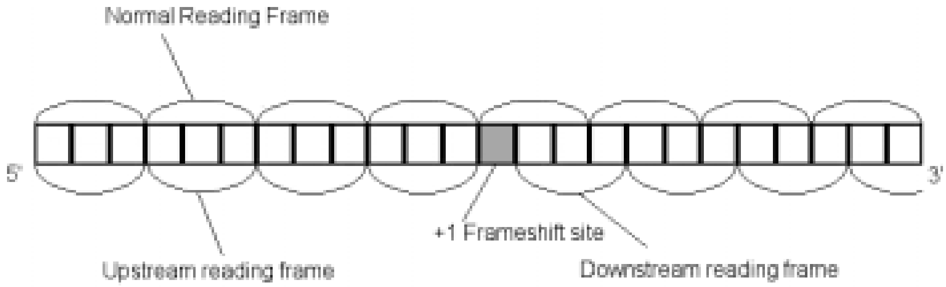

Fig. 1. The ribosomal readthrough of a single nucleotide shifting the reading frame forward by one base (5'–3' direction) is a $+1$ frameshift.

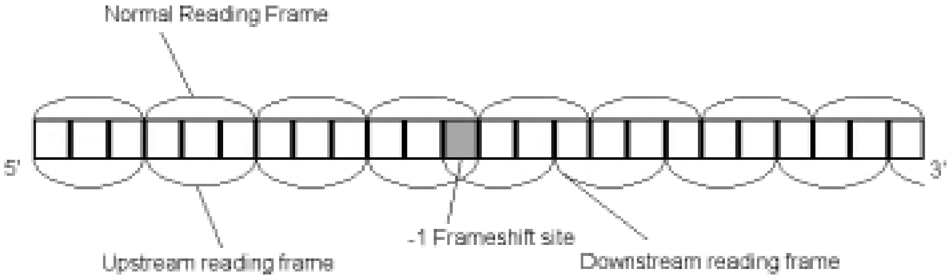

Fig. 2. The ribosomal reread of the third nucleotide of the last translated trinucleotide shifting the reading frame backward by one base (3'–5' direction) is a -1 frameshift.

it encounters a stop codon. Therefore, the protein product expressed will partly be encoded in the reading frame upstream of the frameshift site and partly in the shifted frame downstream of it. Two main categories of frameshifts are distinguished. The ribosomal readthrough of a single nucleotide shifting the reading frame forward by one base (5'–3' direction) is a $+1$ frameshift (Fig. 1). After the frameshift site, the new reading frame is the frame 1. The ribosomal reread of the third nucleotide of the last translated trinucleotide shifting the reading frame backward by one base (3'–5' direction) is a -1 frameshift (Fig. 2). After the frameshift site, the new reading frame is then the frame 2. Recently, another class of frameshifts has been identified in phage Mu. The newly observed -2 frameshift has the same overall effect on the protein sequence as the $+1$ frameshift, except for an additional amino acid encoded at the frameshift site [Xu *et al.*, 2004].

Natural frameshifts without particular slippery signals can occur but are highly improbable with a rate estimated at 3×10^{-5} per codon [Parker, 1989]. Frameshifts can be much more frequent. More than 50% of the ribosomes can shift at sites of some genes. These particular shifts are called programmed frameshifts. So, translational recoding can be in competition with normal decoding. A typical eukaryotic site triggering a -1 frameshift is the slippery heptamer X,XX.Y,YY.Z, codons in the reading frame being separated by commas and codons in the shifted frame, by dots [Baranov *et al.*, 2006]. Frameshifts generally occur in mid-passage of genes [Craigen and Caskey, 1987; Atkins *et al.*, 1990]. They are found in eukaryotes, prokaryotes, viruses and transposons and are involved in a variety of biological processes [Namy *et al.*, 2004].

Gene detection algorithms generally consider shifted frames to be sequencing errors or pseudogene signatures and only a few algorithms assign a frameshift as a possible regulatory process [Harrison *et al.*, 2002]. However, frameshifts allow expression of several polypeptides from the same mRNA [Gesteland *et al.*, 1992]. Although most of recoding events described so far have been found in small autonomous

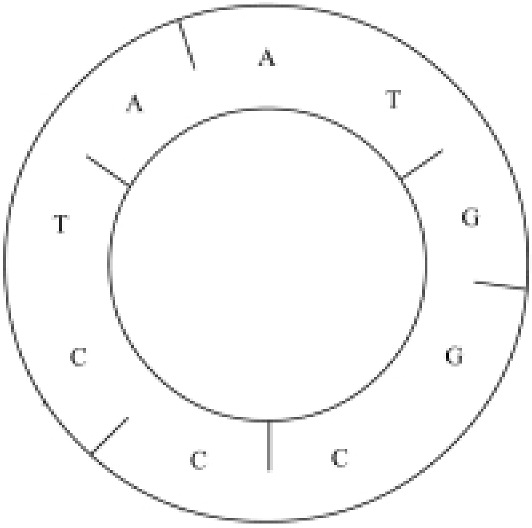

Fig. 3. The set $Y = \{$AAT,ATG,CCT,CTA,GCC,GGC$\}$ is not a circular code as the word $w =$ ATGGCCCTA, written on a circle, can be factorized into words of Y according to 2 different ways: ATG,GCC,CTA and AAT,GGC,CCT.

genetic elements [Bekaert *et al.*, 2005], a few cellular genes are known to be expressed by this mode of control [Namy *et al.*, 2004], most of them found by chance. Besides permitting the identification of new functional genes, the study of frameshifts could also be interesting for understanding the mechanisms maintaining the reading frames.

As we mentioned in our previous publications (reviewed in Michel, 2007), some particular sets of trinucleotides, called circular codes, may have a role in detecting and maintaining the reading frames in genes. Therefore, the circular code property may be altered with the -1 and $+1$ frameshifts. A possible relation between circular codes and frame alteration is investigated here by developing a new approach based on 2 score functions with the 3 circular codes X and its 2 permuted sets X_1 and X_2 of 20 trinucleotides preferentially associated with the reading frame and their 2 shifted frames 1 and 2 of both eukaryotic and prokaryotic genes. Therefore, the main code X with its properties allows retrieval of any frame in genes (containing these circular code words), locally anywhere in the 3 frames and in particular without start codons in the reading frame, and automatically with the reading of 13 nucleotides in each frame. The frameshift positions are statistically significantly identified for the -1 and $+1$ frameshifts of eukaryotic populations, and for the -1 frameshifts of prokaryotic population. This approach does not use any recoding model but only a very short sliding window of only 4 trinucleotides associated with the theoretical property of reading with the circular code.

After a brief presentation of the definition and the main properties of circular codes, the 2 score functions developed for analysing the circular code signal around the frameshift sites are detailed.

Common Circular Code

Identification

In 1996, a simple occurrence study of the 64 trinucleotides $\mathbb{T} = \{$AAA,...,TTT$\}$ in the 3 frames of genes has shown that the trinucleotides are not uniformly distributed in these 3 frames. By excluding the 4 trinucleotides with identical nucleotides $\mathbb{T}_{id} = \{$AAA,CCC,GGG,TTT$\}$ and by assigning each trinucleotide to a preferential frame (frame of its highest frequency), the same 3 subsets X_0, X_1 and

Table 1
The common C^3 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 respectively of genes

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

X_2 of 20 trinucleotides are identified in the frames 0, 1 and 2 respectively of 2 large and different gene populations of eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686 sequences, 4,708,758 trinucleotides) [Arquès and Michel, 1996]. These 3 trinucleotide subsets present several strong biomathematical properties, in particular the property of circular code. The subset $X_0 = \{$AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT, TAC,TTC$\}$ of 20 trinucleotides in frame 0 is a common circular code as it is observed in the reading frames of both eukaryotic and prokaryotic genes and furthermore, a C^3 code as its 2 permuted sets X_1 and X_2 are also circular codes (Table 1). X_0 is simply noted C^3 code X. Due to the law of large numbers, these 3 trinucleotide subsets are (obviously) retrieved in these 2 gene populations with the actual statistical studies (results not shown).

We recall the main properties of the common C^3 code X which will be involved in this paper. A recent review of circular codes in genes details the research context, the history and their different properties [Michel, 2007].

Notation 1.

$\mathbb{A}$ being a finite alphabet, $\mathbb{A}^*$ denotes the words over $\mathbb{A}$ of finite length including the empty word ε of length 0 and $\mathbb{A}^+$, the words over $\mathbb{A}$ of finite length greater or equal to 1. Let w_1w_2 be the concatenation of the 2 words w_1 and w_2.

Definition 2. The (left circular) permutation $\mathcal{P}$ of a trinucleotide $w_0 = l_0l_1l_2$, $w_0 \in \mathbb{T}$, is the permuted trinucleotide $\mathcal{P}(w_0) = w_1 = l_1l_2l_0$, e.g. $\mathcal{P}(\mathrm{AAC}) = \mathrm{ACA}$, and $\mathcal{P}(\mathcal{P}(w_0)) = \mathcal{P}(w_1) = w_2 = l_2l_0l_1$, e.g. $\mathcal{P}(\mathcal{P}(\mathrm{AAC})) = \mathrm{CAA}$. This definition is naturally extended to the trinucleotide set permutation: the permutation $\mathcal{P}$ of a set of trinucleotides is the permuted trinucleotide set obtained by the permutation $\mathcal{P}$ of all its trinucleotides.

Definition 3. The complementarity $\mathcal{C}$ of a trinucleotide $w_0 = l_0l_1l_2$, $w_0 \in \mathbb{T}$, is the complementary trinucleotide $\mathcal{C}(w_0) = \mathcal{C}(l_2)\mathcal{C}(l_1)\mathcal{C}(l_0)$ with $\mathcal{C}(\mathrm{A}) = \mathrm{T}, \mathcal{C}(\mathrm{C}) = \mathrm{G}$, $\mathcal{C}(\mathrm{G}) = \mathrm{C}$, $\mathcal{C}(\mathrm{T}) = \mathrm{A}$, e.g. $\mathcal{C}(\mathrm{AAC}) = \mathrm{GTT}$. This definition is naturally extended to the trinucleotide set complementarity.

Definition 4. A subset X of $\mathbb{A}^+$ is a circular code if $\forall\, n, m \geqslant 1$ and $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m \in X$ and $r \in \mathbb{A}^*$, $s \in \mathbb{A}^+$, the equalities $sx_2 \ldots x_nr = y_1y_2 \ldots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ and $x_i = y_i$, $1 \leqslant i \leqslant n$.

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 6 following words: $Y = \{$AAT,ATG,CCT,CTA,GCC,GGC$\}$ and the word w, be a series of the 9 following letters: $w =$ ATGGCCCTA. The word w, written on a circle, can be factorized into words of Y according to 2 different ways: ATG,GCC,CTA and AAT,GGC,CCT, the commas showing the way of decomposition (Fig. 3). Therefore, Y is not a circular code. In contrast, if the set Z obtained by

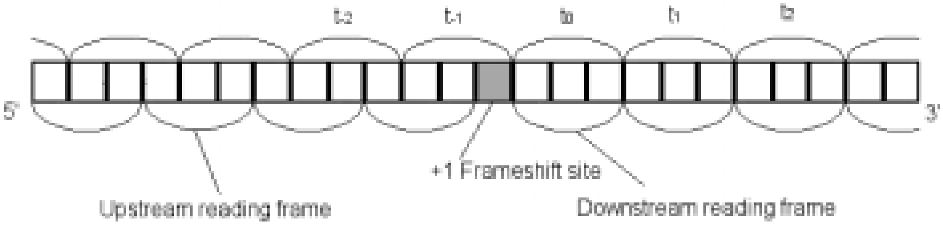

Fig. 4. In the $+1$ frameshift, the 1st nucleotide after the shifted nucleotide is the 1st nucleotide of t_0.

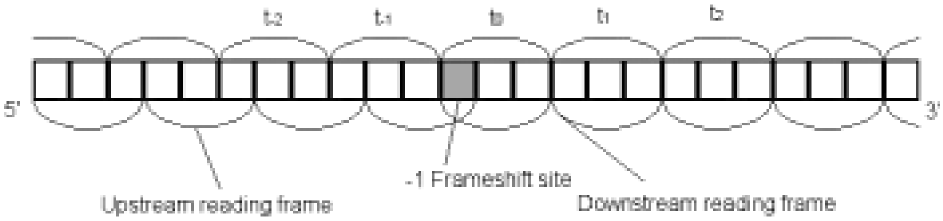

Fig. 5. In the -1 frameshift, the shifted nucleotide is the 1st nucleotide of t_0.

replacing the word GGC of Y by GTC is considered, i.e. $Z = \{$ATT,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.

The construction frame of a word generated by any concatenation of words of a circular code can be retrieved after the reading, anywhere in the generated word, of a certain number of nucleotides depending on the code. This series of nucleotides is called the window of the circular code. Then, the minimal window length is the size of the longest ambiguous word which can be read in at least 2 frames, more one letter.

Main Properties

We point out only the circular code properties related to the aim of the paper.

Property 1. Maximality: X_0 is a maximal circular code (20 trinucleotides) as it is not contained in a larger circular code, i.e. in a circular code with more words.

A circular code has several constraints:

1. The 4 trinucleotides $\mathbb{T}_{id}$ must be excluded from such a code. Indeed, the concatenation of AAA with itself, for example, does not allow the reading (original) frame to be retrieved as there are 3 possible decompositions: ...AAA,AAA,AAA,..., ...A,AAA,AAA,AA... and ...AA,AAA,AAA,A...
2. The permuted trinucleotides by the permutation $\mathcal{P}$ of a trinucleotide, for example $\mathcal{P}$ (AAC) $=$ ACA, must also be excluded from such a code. Indeed, the concatenation of AAC with itself, for example, also does not allow the reading frame to be retrieved as there are 2 possible decompositions: ...AAC,AAC,AAC,... and ...A,ACA,ACA,AC...

Therefore, by excluding the 4 trinucleotides $\mathbb{T}_{id}$ and by gathering the 60 remaining trinucleotides in 20 classes of 3 trinucleotides such that, in each class, 3 trinucleotides are deduced from each other by circular permutations, e.g. AAC, $\mathcal{P}$ (AAC) $=$ ACA and $\mathcal{P}(\mathcal{P}$ (AAC)) $=$ CAA, a circular code has only one trinucleotide per class and therefore contains at most 20 trinucleotides. Then, any 20-long circular code is maximal and there are $3^{20} \approx 3.5$ billions of potential codes of length 20.

Property 2. Permutation: $\mathcal{P}(X_0) = X_1$ and $\mathcal{P}(\mathcal{P}(X_0)) = X_2$ (X_0 generates X_1 by one permutation and X_2 by another permutation).

Property 3. Complementarity: $\mathcal{C}(X_0) = X_0$ (X_0 is self-complementary) and, $\mathcal{C}(X_1) = X_2$ and $\mathcal{C}(X_2) = X_1$ (X_1 and X_2 are complementary to each other).

Property 4. C^3 code: X_1 and X_2 obtained by permutation of X_0 (property 2) are also maximal circular codes (property 1). Therefore, if X_0, X_1 and X_2 are circular codes, then X_0, X_1 and X_2 are C^3 codes. As the code X_0 is coding for the reading frame (frame 0) in genes, i.e. the most important frame, and as it is self-complementary (property 3), it is considered as the main C^3 code and noted X simply.

Remark 1. A circular code Y_0 does not necessarily imply that Y_1 and Y_2 obtained by permutation of Y_0, are also circular codes.

Property 5. Rarity: the occurrence probability of the complementary C^3 code X is equal to $216/3^{20} \approx 6 \times 10^{-8}$.

This probability is equal to the computed number of complementary C^3 codes (216) divided by the number of potential codes ($3^{20} = 3{,}486{,}784{,}401$, see property 1).

Property 6. Largest window length: the lengths of the minimal windows of X_0, X_1 and X_2 for retrieving the frames 0, 1 and 2 respectively, are all equal to 13 nucleotides and represent the largest window length.

Property 7. Common: the C^3 code X is identified in both eukaryotic and prokaryotic genes.

Property 8. The common C^3 code X (properties 4 and 7) with its particular structure (property 5) allows retrieval of any frame in genes (containing these circular code words), locally anywhere in the 3 frames and in particular without start codons in the reading frame, and automatically with the same window length of 13 nucleotides in each frame (property 6).

METHOD

Definition of a Score Function P Based on the Common C^3 Code X

In order to identify genes with frameshift sites, we develop an algorithm based on the common C^3 code X and its property 8.

Let $\mathbb{T} = \{\text{AAA},\ldots,\text{TTT}\}$ be the set of 64 trinucleotides over the alphabet $\{\text{A,C,G,T}\}$. Let $t \in \mathbb{T}$ be a trinucleotide. Let F be a frameshift gene population with $n(F)$ sequences s. Each sequence s has a frameshift site in the nucleotide position $i = 0$. Let t_0 be the first downstream trinucleotide after the frameshift site. Precisely, in the $+1$ frameshift, the 1st nucleotide after the shifted nucleotide is the 1st nucleotide of t_0 (Fig. 4). In the -1 frameshift, the shifted nucleotide is the 1st nucleotide of t_0 (Fig. 5). By convention, all trinucleotides before the frameshift site have a negative position $i < 0$, all trinucleotides after the frameshift site have a non-negative position $i \geqslant 0$ and the method reading frame is $\ldots$, t_{-1}, t_0, t_1,$\ldots$ The sequence s is considered as a series of trinucleotides t_i. Let $w_i = t_{i_0}t_{i_1}t_{i_2}t_{i_3}$ be a window of length $|w|$ = 4 trinucleotides in the sequence s, where t_{i_0} is the ith trinucleotide in s and, t_{i_j}, the jth trinucleotide in w_i. This sliding window length is the length of the minimal windows of the

3 circular codes X_0, X_1 and X_2 for retrieving the frames 0, 1 and 2 respectively in genes (property 6). Let X_f, $f \in \{0,1,2\}$, be the 3 codes X_0, X_1 and X_2 in the 3 frames f.

In a given window w_i, the function $\delta_f(t_{i_j})$ indicates if the trinucleotide t_{i_j} belongs or not to the code X_f

$$\delta_f(t_{i_j}) = \begin{cases} 1 \text{ if } \ t_{i_j} \in X_f \\ 0 \text{ otherwise} \end{cases}$$

with $f \in \{0,1,2\}$ and $j \in \{0,1,2,3\}$. Each sequence s is associated with a frame $\mathcal{F} \in \{0,1,2\}$. In s_0 (s_1 and s_2 resp.), t_0 is the first downstream trinucleotide (more one nucleotide and 2 nucleotides resp.) after the frameshift site. Then, the score $P(X_f,\, i,\, s_\mathcal{F})$ of the code X_f in a window w_i of a given frame $\mathcal{F}$ of a sequence s is

$$P(X_f, i, s_\mathcal{F}) = \frac{1}{|w|} \sum_{j=0}^{|w|-1} \delta_f(t_{i_j}).$$

The score $P(i, s_\mathcal{F})$ of the C^3 code X in a window w_i of a given frame $\mathcal{F}$ of a sequence s measuring the frame retrieval intensity, is defined as

$$P(i, s_\mathcal{F}) = \frac{1}{2} \sum_{\substack{f, f' \in \{0,1,2\} \\ f' > f}} |P(X_f, i, s_\mathcal{F}) - P(X_{f'}, i, s_\mathcal{F})|.$$

Then, the score $P(i, s)$ of the C^3 code X in a window w_i in the average frame of a sequence s is

$$P(i, s) = \frac{1}{3} \sum_{\mathcal{F}=0}^{2} P(i, s_\mathcal{F}).$$

Finally, the score $P(i, F)$ of the C^3 code X in a window w_i in the average frame of a gene population F is

$$P(i, f) = \frac{1}{n(F)} \sum_{s \in F} P(i, s).$$

Proposition 1. If the words of the 3 codes X_0, X_1 and X_2 are uniformly distributed in the windows w_i in each frame of the sequences of a population F, then $P(X_f, i, s_\mathcal{F}) = P(X_{f'}, i, s_\mathcal{F})$ with $f' \neq f$ leading to $P(i, F) = 0$. Zero is the minimum value for $P(i, F)$.

Proposition 2. If the words of only one code X_f occur in the windows w_i in each frame of the sequences of a population F, then $P(X_f,\, i,\, s_\mathcal{F}) = 1$, $P(X_{f'},\, i,\, s_\mathcal{F}) = 0$ with $f' \neq f$ leading to $P(i, F) = 1$. One is the maximum value for $P(i, F)$.

Proposition 3. $0 \leqslant P(i, F) \leqslant 1$ (consequence of propositions 1 and 2).

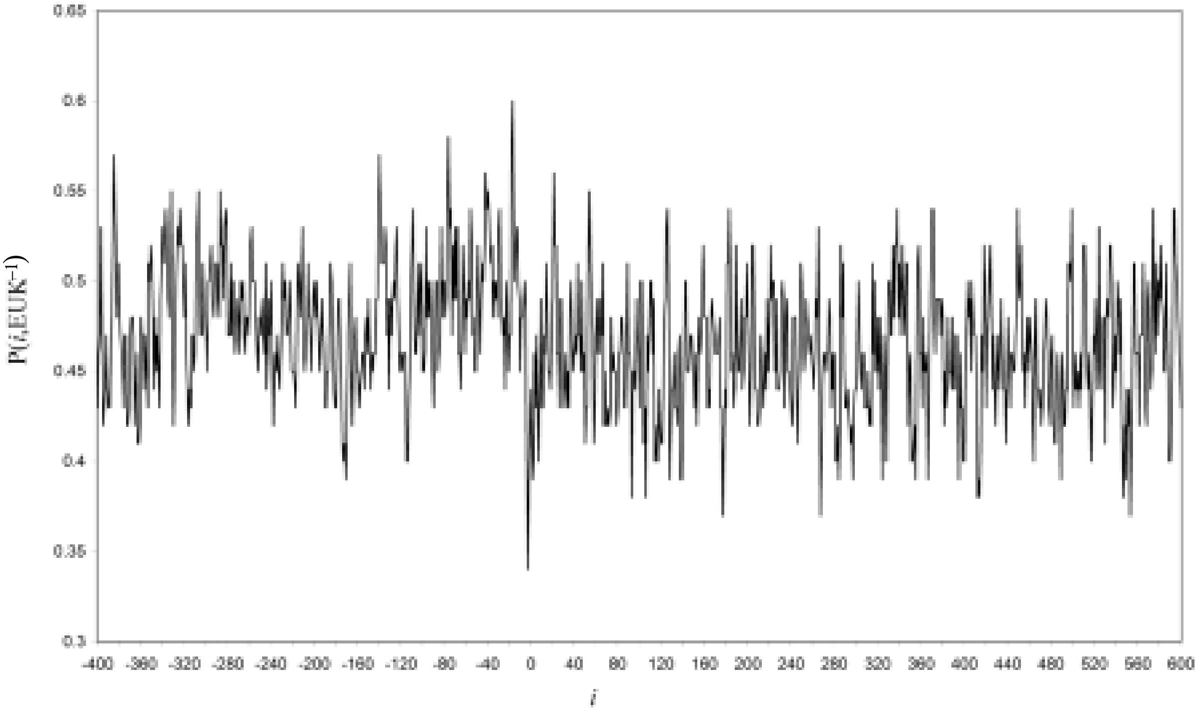

Fig. 6. Function P applied on the -1 frameshift eukaryotic population $F = EUK^{-1}$. The abscissa represents the position i in EUK^{-1} by varying i in the range $\{-400,600\}$. There is a significant lowest value around the frameshift site $i = 0$, precisely $P(-3, EUK^{-1}) = 0.34$.

Definition of a score function $\boldsymbol{Q}$ based on both the common $\boldsymbol{C}^3$ code $\boldsymbol{X}$ and the trinucleotide set $\mathbb{T}_{id}$

The score function P based only on the common C^3 code X allows detection of frameshift signals in genes. It does not consider the 4 trinucleotides with identical nucleotides $\mathbb{T}_{id} = \{\text{AAA,CCC,GGG,TTT}\}$ which are excluded in a circular code (property 1). In contrast to the circular code, the 4 trinucleotides $\mathbb{T}_{id}$ cannot identify frames in genes. This information is now considered in the score function Q, obviously in a contrary way to the circular code information. The function Q based on both the common C^3 code X and the trinucleotides $\mathbb{T}_{id}$ is developed in a way similar to the function P for detecting frameshift signals in genes.

In a given window w_i, the function $\delta_{id}(t_{i_j})$ indicates if the trinucleotide t_{i_j} belongs or not to the set $\mathbb{T}_{id}$

$$\delta_{id}(t_{i_j}) = \begin{cases} 1 \text{ if } \ t_{i_j} \in \mathbb{T}_{id} \\ 0 \text{ otherwise} \end{cases}$$

with $j \in \{0, 1, 2, 3\}$. Then, the score $Q(\mathbb{T}_{id}, i, s_{\mathcal{F}})$ of the set $\mathbb{T}_{id}$ in a window w_i of a given frame $\mathcal{F}$ of a sequence s is

$$Q(\mathbb{T}_{id}, i, s_{\mathcal{F}}) = \frac{1}{|w|} \sum_{j=0}^{|w|-1} \delta_{id}(t_{i_j}).$$

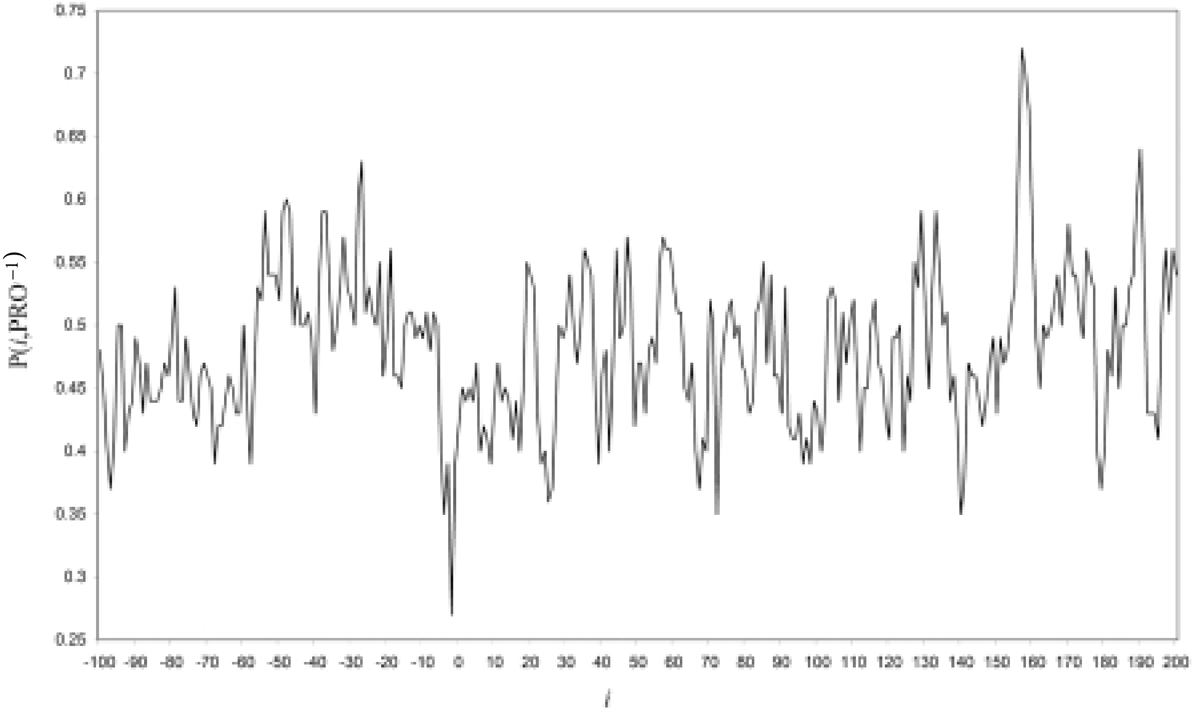

Fig. 7. Function P applied on the -1 frameshift prokaryotic population $F = PRO^{-1}$. The abscissa represents the position i in PRO^{-1} by varying i in the range $\{-100{,}200\}$. There is a significant lowest value around the frameshift site $i = 0$, precisely $P(-2, PRO^{-1}) = 0.27$.

Thus, the score $Q(i, s_{\mathcal{F}})$ of the C^3 code X in a window w_i of a given frame $\mathcal{F}$ of a sequence s by withdrawing the score $Q(\mathbb{T}_{id}, s, s_{\mathcal{F}})$ associated with the set $\mathbb{T}_{id}$ which does not allow retrieval of the frame, is

$$Q(i, s_{\mathcal{F}}) = P(i, s_{\mathcal{F}}) - Q(\mathbb{T}_{id}, i, s_{\mathcal{F}}).$$

Then, the score $Q(i, s)$ of the C^3 code X in a window w_i in the average frame of a sequence s is

$$Q(i, s) = \frac{1}{3} \sum_{\mathcal{F}=0}^{2} Q(i, s_{\mathcal{F}}).$$

Finally, the score $Q(i, F)$ of the C^3 code X in a window w_i in the average frame of a gene population F is

$$Q(i, F) = \frac{1}{n(F)} \sum_{s \in F} Q(i, s).$$

Proposition 4. If only one trinucleotide of $\mathbb{T}_{id}$ occur in the sequences of a population F, then $Q(\mathbb{T}_{id}, i, s_{\mathcal{F}}) = 1$ leading to $Q(i, F) = -1$. Minus one is the minimum value for $Q(i, F)$.

Proposition 5. If the words of only one code X_f occur in the windows w_i in each frame of the sequences of a population F, then $P(i, s_{\mathcal{F}}) = 1$ and $Q(\mathbb{T}_{id}, i, s_{\mathcal{F}}) = 0$ leading to $Q(i, F) = 1$. One is the maximum value for $Q(i, F)$.

Table 2
Type and size of the studied frameshift gene populations extracted from the RECODE database

Frameshift gene populations	Number of genes
-1 frameshift of eukaryotes EUK^{-1}	27
-1 frameshift of prokaryotes PRO^{-1}	15
$+1$ frameshift of eukaryotes EUK^{+1}	34
$+1$ frameshift of prokaryotes PRO^{+1}	48

Proposition 6. $-1 \leqslant Q(i, F) \leqslant 1$ (consequence of propositions 4 and 5).

Data Acquisition

The frameshift genes used in this study are extracted from the RECODE database [Baranov *et al.*, 2001]. The RECODE 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 C^3 code X has been identified only in these genes and not, for example, in viral genes [Arqu ès and Michel, 1996]. Therefore, 4 gene populations 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 size of the studied frameshift gene populations.

The score functions P and Q are applied on these 4 frameshift gene populations. In each population F, frameshift genes from different species and producing proteins with different functionalities are found. Therefore, the gene size and the frameshift site vary from one gene to another. This data variation implies that these 2 functions will be computed in the nucleotide interval around the frameshift site $i = 0$ where there is a sufficient number of sequences (not detailed), explaining thus the variations of abscisses i in the figures below according to the population F.

RESULTS

Results for the -1 Frameshifts

Figures 6 and 7 show the results using the function P with the frameshift gene populations $F = EUK^{-1}$ and $F = PRO^{-1}$, respectively. The abscissa represents the position i in the population F by varying i in the range $\{-400,600\}$ for EUK^{-1} and $\{-100,200\}$ for PRO^{-1}, and the ordinate, the value of the function P.

Both figures show a significant lowest value of the function P around the frameshift site $i = 0$, precisely $P(-3,\ EUK^{-1}) = 0.34$ (Fig. 6) and $P(-2,\ PRO^{-1}) = 0.27$ (Fig. 7). As expected by the circular code property, a frameshift region is associated with a mixing of frames, i.e. a random mixing of codes X_0, X_1 and X_2 which is revealed by a low value of the function P (proposition 1).

Figures 8 and 9 show the results using the function Q with the populations $F = EUK^{-1}$ and $F = PRO^{-1}$, respectively. There is a very significant lowest value of the function Q around the frameshift site

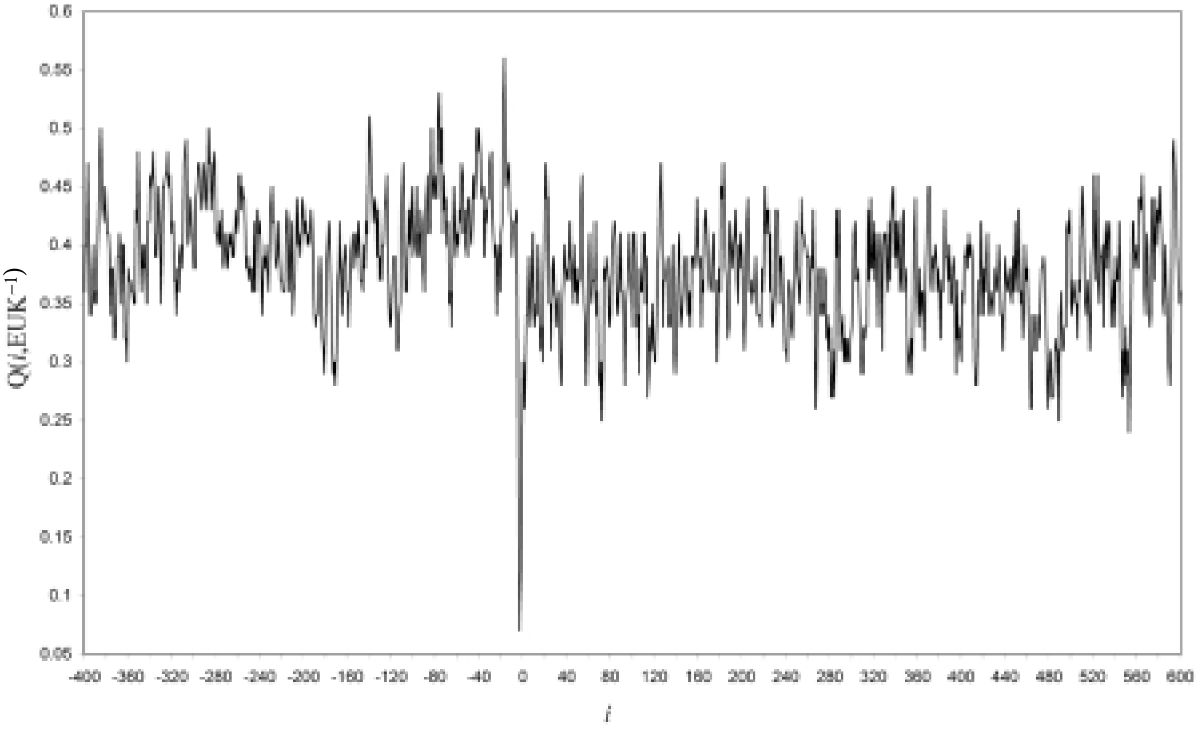

Fig. 8. Function Q applied on the -1 frameshift eukaryotic population F = *EUK*$^{-1}$. The abscissa represents the position i in *EUK*$^{-1}$ by varying i in the range $\{-400,600\}$. There is a very significant lowest value around the frameshift site $i = 0$, precisely $Q(-3,$ *EUK*$^{-1}) = 0.07$.

$i = 0$ in both figures, precisely $Q(-3,$ *EUK*$^{-1}) = 0.07$ (Fig. 8) and $Q(-2,$ *PRO*$^{-1}) = -0.05$ (Fig. 9). A better result is observed with the function Q compared to the function P as the -1 frameshift signals contain also trinucleotides $\mathbb{T}_{id}$, such as the slippery heptamer XXXYYYZ [Baranov *et al.*, 2006], which increase the loss of the circular code property.

Results for the $+1$ Frameshifts

Figure 10 shows the result for the frameshift gene population $F =$ *EUK*$^{+1}$ using the function P and by varying i in the range $\{-50,150\}$. Surprisingly, a lowest value is observed around the frameshift site $i = 0$ with $P(1,$ *EUK*$^{+1}) = 0.38$. Furthermore, the function Q does not show a lowest value around the frameshift site in *EUK*$^{+1}$. Therefore, the $+1$ frameshift eukaryotic signals are associated only with the loss of the circular code property.

No interesting result has been found until now in the frameshift gene population $F = PRO^{+1}$ using P or Q (see "Discussion").

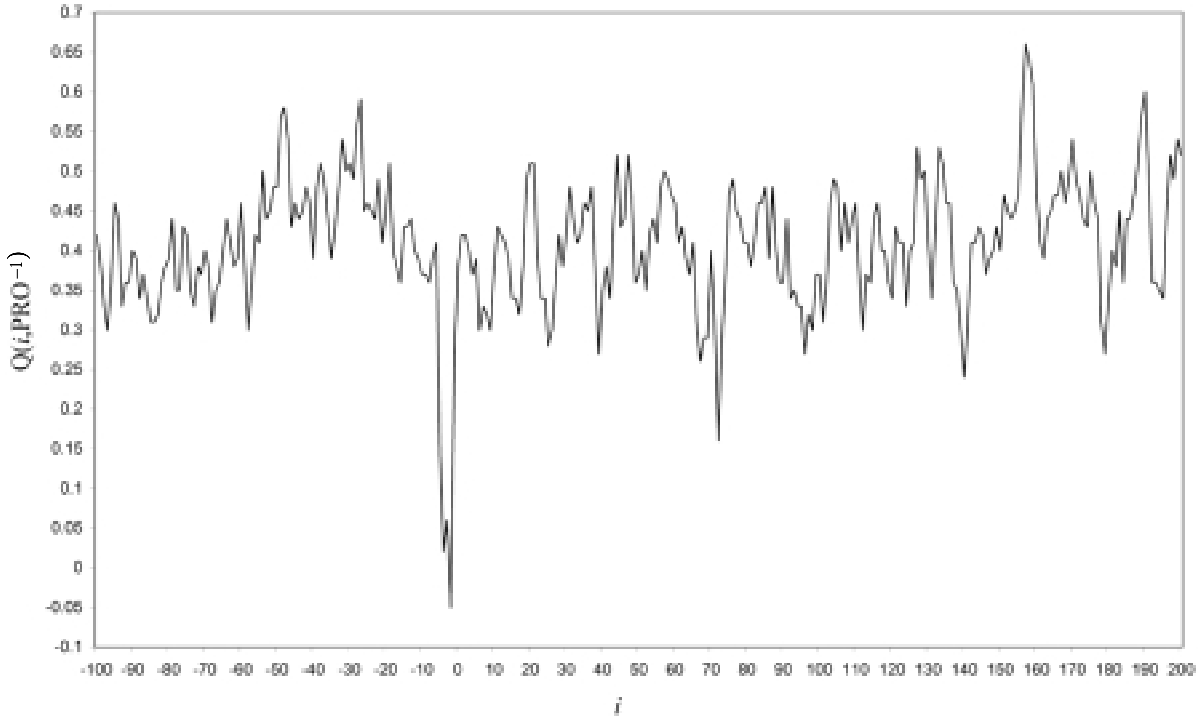

Fig. 9. Function Q applied on the -1 frameshift prokaryotic population $F = \mathit{PRO}^{-1}$. The abscissa represents the position i in PRO^{-1} by varying i in the range $\{-100{,}200\}$. There is a very significant lowest value around the frameshift site $i = 0$, precisely $Q(-2, \mathit{PRO}^{-1}) = -0.05$.

DISCUSSION

Recoding is widely distributed between organisms. It is thus likely that numerous novel recoded cellular genes remain to be discovered. However, the prediction of frameshift sites from genomic databases is currently a difficult task. Our new method shows a strong correlation between the circular code signal and the frameshift sites in genes. It can be performed without a prior knowledge of the mechanism involved in recoding. Indeed, it is based on the common C^3 code X found in eukaryotic and prokaryotic genes, i.e. a structural feature of genes, which allows retrieval of any frame in genes (containing these circular code words), locally anywhere in the 3 frames and in particular without start codons in the reading frame, and automatically with the same window length of 13 nucleotides in each frame.

This method has been applied on 4 gene populations: -1 and $+1$ frameshifts of eukaryotic and prokaryotic genes. Frameshift sites in these populations are associated with low values for the functions P and Q. The function P is directly based on the property of the C^3 code X while the function Q is based on both this property and the 4 trinucleotides with identical nucleotides which are considered in a contrary way to the circular code information. The absence of detection of frameshift signals with the $+1$ frameshift prokaryotic genes (PRO^{+1}) could be due to the special nature of these genes or to an insufficiency of the method.

Except for this population PRO^{+1}, the developed method is generic since it is applicable on both $+1$ and -1 frameshift genes even if the frameshift motifs are completely different in these various genes. It is also very efficient since frameshift signals are revealed with a window of very small size (4 trinucleotides)

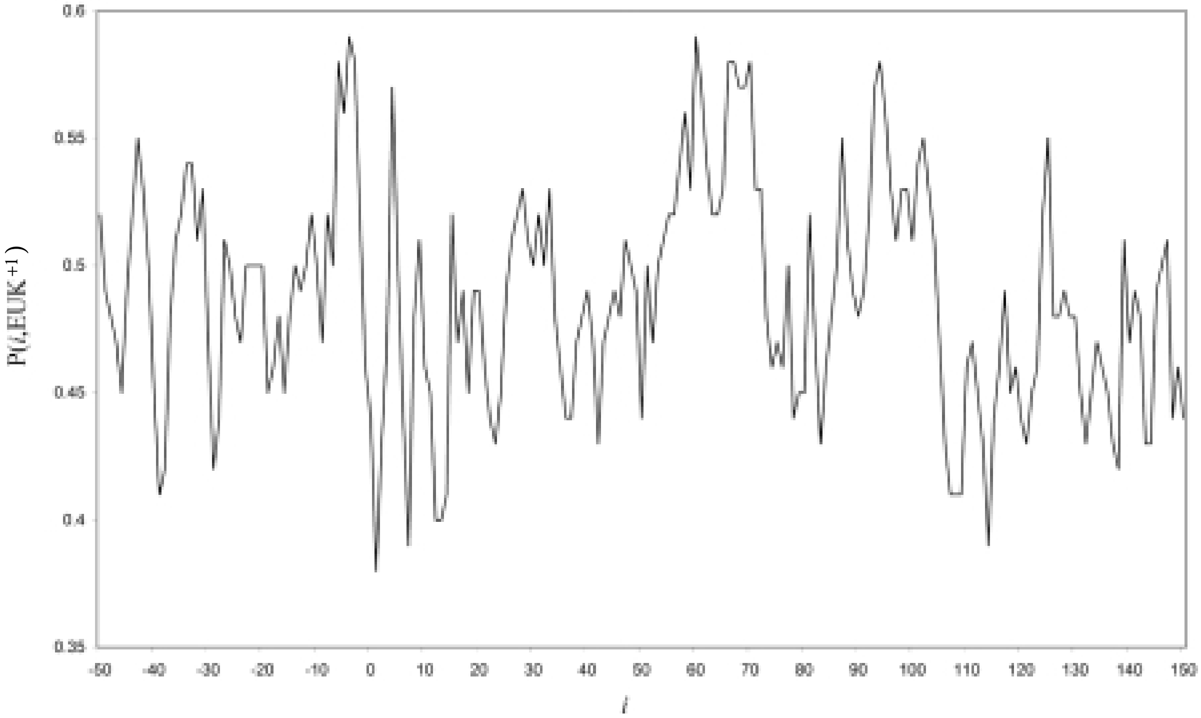

Fig. 10. Function P applied on the $+1$ frameshift eukaryotic population $F = EUK^{+1}$. The abscissa represents the position i in EUK^{+1} by varying i in the range $\{-50{,}150\}$. There is a lowest value around the frameshift site $i = 0$, precisely $P(1, EUK^{+1})$ $= 0.38$.

associated with the theoretical property of reading with the circular code. This new approach based on the circular code information is a contribution to the existing methods. We are currently investigating this approach for predicting frameshift sites at the gene level by stressing that the results have been obtained with gene populations of small sizes, e.g. 15 genes with PRO^{-1} (Table 2). Indeed, this method can obviously be improved by varying the size of the sliding window, by using circular codes specific to the genomes [Frey and Michel, 2006] and by adding recoding information.

REFERENCES

- Arquès, D. G. and Michel, C. J. (1996). A complementary circular code in the protein coding genes. J. Theor. Biol. **182**, 45-58.
- Atkins, J. F., Weiss, R. B. and Gesteland, R. F. (1990). Ribosome gymnastics – Degree of difficulty 9.5, style 10.0. Cell **62**, 413-423.
- Baranov, P. V., Gurvich, O. L., Fayet, O., Prère, M. F., Miller, W. A., Gesteland, R. F., Atkins, J. F. and Giddings, M. C. (2001). RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. Nucleic Acids Res. **29**, 264-267.
- Baranov, P. V., Fayet, O., Hendrix, R. W. and Atkins, J. F. (2006). Recoding in bacteriophages and bacterial IS elements. Trends Genet. **22**, 174-181.
- Bekaert, M., Richard, H., Prum, B. and Rousset, J.-P. (2005). Identification of programmed translational -1 frameshifting sites in the genome of *Saccharomyces cerevisiae*. Genome Res. **15**, 1411-1420.
- Cobucci-Ponzano, B., Rossi, M. and Moracci, M. (2005). Recoding in archaea. Mol. Microbiol. **55**, 339-348.
- Craigen, W. J. and Caskey, C. T. (1987). Translational frameshifting: Where will it stop? Cell **50**, 1-2.
- Farabaugh, P. J. (1996). Programmed translational frameshifting. Annual Rev. Genetics **30**, 507-528.

- Frey, G. and 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.
- Gesteland, R. F., Weiss, R. B. and Atkins, J. F. (1992). Recoding: Reprogrammed genetic decoding. Science **257**, 1640-1641.
- Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M. and Gerstein, M. (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. J. Mol. Biol. **316**, 409-419.
- Michel, C. J. (2007). A 2006 review of circular codes in genes. Comput. Math. Appl., in press.
- Namy, O., Naphtine, S., Rousset, J. P. and Brierley, I. (2004). Reprogrammed genetic decoding in cellular gene expression. Mol. Cell **13**, 157-168.
- Parker, J. (1989). Errors and alternatives in reading the universal genetic code. Microbiol. Rev. **53**, 273-298.
- Xu, J., Hendrix, R. W. and Duda, R. L. (2004). Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. Mol. Cell **16**, 11-21.