Computational Biology and Chemistry 37 (2012) 24–37

Contents lists available at SciVerse ScienceDirect

Computational Biology and Chemistry

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

Circular code motifs in transfer and 16S ribosomal RNAs: A possible translation code in genes

Christian J. Michel

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

ARTICLE INFO

Article history:
Received 22 September 2011
Accepted 1 October 2011

Keywords:
Circular code motif
Trinucleotide
Messenger RNA
Transfer RNA
Ribosomal RNA
Translation code

ABSTRACT

In 1996, a common trinucleotide circular code, called *X*, is identified in genes of eukaryotes and prokaryotes (Arquès and Michel, 1996). This circular code *X* is a set of 20 trinucleotides allowing the reading frames in genes to be retrieved locally, i.e. anywhere in genes and in particular without start codons. This reading frame retrieval needs a window length *l* of 12 nucleotides ($l \geq 12$). With a window length strictly less than 12 nucleotides ($l < 12$), some words of *X*, called ambiguous words, are found in the shifted frames (the reading frame shifted by one or two nucleotides) preventing the reading frame in genes to be retrieved. Since 1996, these ambiguous words of *X* were never studied.

In the first part of this paper, we identify all the ambiguous words of the common trinucleotide circular code *X*. With a length *l* varying from 1 to 11 nucleotides, the type and the occurrence number (multiplicity) of ambiguous words of *X* are given in each shifted frame. Maximal ambiguous words of *X*, words which are not factors of another ambiguous words, are also determined. Two probability definitions based on these results show that the common trinucleotide circular code *X* retrieves the reading frame in genes with a probability of about 90% with a window length of 6 nucleotides, and a probability of 99.9% with a window length of 9 nucleotides (100% with a window length of 12 nucleotides, by definition of a circular code).

In the second part of this paper, we identify *X* circular code motifs (shortly *X* motifs) in transfer RNA and 16S ribosomal RNA: a tRNA *X* motif of 26 nucleotides including the anticodon stem-loop and seven 16S rRNA *X* motifs of length greater or equal to 15 nucleotides. Window lengths of reading frame retrieval with each trinucleotide of these *X* motifs are also determined. Thanks to the crystal structure 3I8G (Jenner et al., 2010), a 3D visualization of *X* motifs in the ribosome shows several spatial configurations involving mRNA *X* motifs, A-tRNA and E-tRNA *X* motifs, and four 16S rRNA *X* motifs. Another identified 16S rRNA *X* motif is involved in the decoding center which recognizes the codon–anticodon helix in A-tRNA. From a code theory point of view, these identified *X* circular code motifs and their mathematical properties may constitute a translation code involved in retrieval, maintenance and synchronization of reading frames in genes.

1. Introduction

1.1. Transfer and 16S ribosomal RNAs

In order to understand our theory of circular code motifs in a translation code, we briefly recall the structure and the function of the transfer RNA (tRNA) and ribosomal RNA (rRNA) which are involved for translating the genetic information into proteins. For detail, we refer the reader to a recent review of Zaher and Green (2009b).

Protein synthesis in actual genes is a complex molecular process. Ribosome (complex of RNAs and ribosomal proteins), tRNA and its aminoacyl tRNA synthetase, and messenger RNA (mRNA) are the main biological elements involved in the molecular process of translating. In prokaryotes, the ribosome (70S) is composed of a 50S large subunit and a 30S small subunit. The 50S subunit is the active site involved in the formation of peptide bonds and the elongation of the nascent polypeptide. The 30S subunit is the decoding site containing the codon–anticodon interaction, i.e. the interaction between mRNA and tRNA. Thus, it has a critical function in decoding mRNA by monitoring base pairing between the codon on mRNA and the anticodon on tRNA. It is composed of 16S rRNA of about 1500 nucleotides and 21 ribosomal proteins (labeled S1 to S21). During protein synthesis, a tRNA moves through three distinct binding sites of the ribosome: the aminoacyl site (A-site), the peptidyl site (P-site) and the exit site (E-site). The tRNAs at these ribosomal binding sites are called aminoacyl-tRNA (A-tRNA), peptidyl-tRNA (P-tRNA) and exit-tRNA (E-tRNA), respectively.

E-mail address: michel@dpt-info.u-strasbg.fr

1476-9271/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2011.10.002

During translation, the tRNA enters the ribosome and binds to a codon of mRNA in the A-site. After accepting transfer of the growing peptide from the preceding tRNA, it translocates to the P-site, donates the peptide to the succeeding tRNA and moves to the E-site before dissociating from the ribosome.

The stability of codon–anticodon interactions in solution is weak (Lipsett et al., 1960). The 30S subunit stabilizes the association between mRNA and tRNA (Gorini and Kataja, 1964; McLaughlin et al., 1966). The decoding site for A-site tRNA binding involved conserved nucleotides of 16S rRNA, in *E. coli*: G at position 529 (G529 in helix 18), G at position 530 (G530 in helix 18), A at position 1492 (A1492 in helix 44) and A at position 1493 (A1493 in helix 44) (Moazed and Noller, 1990; Ogle et al., 2001). An overview of the 16S RNA secondary structures of *E. coli* and *T. thermophilus* with their corresponding numbering is given in Brodersen et al. (2002). The E-site tRNA binding is also involved in frame maintenance (Devaraj et al., 2009). Indeed, perturbations of the E-site codon–anticodon pairing promotes frameshifting (Márquez et al., 2004). The P-site tRNA binding is also associated to fidelity during codon recognition in the A-site (Sundararajan et al., 1999; Zaher and Green, 2009a). Thus, there are several ribosomal regions which are implicated in codon recognition and reading frame maintenance.

The ability of all living organisms to efficiently and accurately translate genomic information into functional proteins is a fascinating molecular function. The ribosome must correctly associate, according to the genetic code, the amino-acid attached to the tRNA with the trinucleotide in reading frame (codon) of mRNA. It must decode only successive codons and not trinucleotides in shifted frames. However, a mRNA lacks punctuation (or comma) which could be used by the transfer and ribosomal RNAs to identify the trinucleotides in reading frame. Errors in translation occur with a frequency between 10^{-3} and 10^{-4} per codon (Kurland et al., 1996). In contrast to missense errors, nearly all frameshift errors are detrimental to the synthesis of a functional protein as the residue sequence after the frameshift is incorrect and typically a stop codon is encountered in the frameshift frame. The frequency of frameshift errors is estimated to be less than one event per 30,000 residues incorporated (Jorgensen and Kurland, 1990). Post-transcriptional modifications of tRNAs are important for maintaining the reading frame and decreasing the frequency of frameshifting (reviewed in Gustilo et al., 2008). More than 80 different modified nucleotides have been characterized in tRNAs (Rozenski et al., 1999). They are found in tRNAs from all organisms and at many different positions within the tRNAs. However, the majority are located in the anticodon loop, in particular at positions 34 (wobble position) and 37 (Rozenski et al., 1999).

The maintenance of the correct reading frame of genes is believed to be a complex process from a conceptual point of view. I say the opposite from a theoretical point of view. Indeed, there are sets of trinucleotides called circular codes Y which have the property of reading frame retrieval, synchronization and maintenance. Furthermore, there are circular codes Y which have in addition the $\mathcal{C}$ self-complementary property, i.e. the trinucleotides of Y are complementary to each other, i.e. $Y = \mathcal{C}(Y)$. Finally, there are self-complementary circular codes Y which have in addition the C^3 property, i.e. the permuted trinucleotide sets $\mathcal{P}(Y)$ and $\mathcal{P}^2(Y)$ of Y by one and two nucleotides, respectively, are also trinucleotide circular codes and complementary to each other, i.e. $\mathcal{C}(Y_1) = Y_2$ and $\mathcal{C}(Y_2) = Y_1$. In 1996, a C^3 self-complementary trinucleotide circular code X has been identified in genes (reading frame of mRNAs) simultaneously in eukaryotes and prokaryotes (Arquès and Michel, 1996). In this paper, motifs of this circular code X, called X circular code motifs or shortly X motifs, are searched in transfer and ribosomal RNAs. The $\mathcal{C}$ self-complementary property and the C^3 property enable, from a coding theory, spatially closed X motifs to pair according to different configurations: mRNA-mRNA, tRNA-tRNA, rRNA-rRNA, mRNA-tRNA, mRNA-rRNA and tRNA-rRNA. These elementary configurations could also be combined, e.g. mRNA-rRNA-tRNA. Thus, these X motifs may constitute a translation code for retrieving and maintaining the reading frame in genes.

In the next section, we briefly recall the definitions and properties of circular codes which are involved in this paper.

1.2. Common trinucleotide circular code of prokaryotes and eukaryotes

In 1996, an occurrence frequency study of the 64 trinucleotides $T = \{\text{AAA}, \ldots, \text{TTT}\}$ in the three frames of genes of both prokaryotes and eukaryotes showed that the trinucleotides are not uniformly distributed in these three frames (Arquès and Michel, 1996). By convention here, 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 one and two nucleotides in the 5′-3′ direction, respectively. By excluding the four trinucleotides with identical nucleotides $T_{id} = \{\text{AAA, CCC, GGG, TTT}\}$ and by assigning each trinucleotide to a preferential frame (frame of its highest occurrence frequency), three subsets X, 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) as well as prokaryotes (13,686 sequences, 4,709,758 trinucleotides) (Arquès and Michel, 1996). This set X contains the 20 following trinucleotides

$$X = \{\text{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG,} \\ \text{GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC}\} \quad (1)$$

The two sets X_1 and X_2, of 20 trinucleotides each, in the frames 1 and 2 can be deduced from X by circular permutation (see below). These three trinucleotide subsets present several strong mathematical properties, particularly the fact that they are circular codes.

A circular code Y is a particular set of words which allows the retrieval of the construction (reading frame) of any word generated by Y. Furthermore, this reading frame retrieval can be obtained anywhere in any generated word by Y but with a window of a few letters (see below). A circular code with words composed of trinucleotides will be called here a trinucleotide circular code.

The decoding of the reading frame in genes is a theoretical problem that was raised several years ago and still an intriguing but difficult subject of current research. Over 50 years ago, before the discovery of the genetic code, a class of trinucleotide circular codes, called comma-free codes (or codes without commas), was proposed by Crick et al. (1957) for explaining how the reading of a series of trinucleotides among 64 could code the 20 amino acids. However, no trinucleotide comma-free code was identified in genes, theoretically or statistically. Furthermore, in the late fifties, the discovery that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes for phenylalanine (Nirenberg and Matthaei, 1961), led to the abandonment of the concept of a comma-free code for gene translation.

We briefly recall a few properties of the common trinucleotide circular code X (1) which may be involved in a translation code in genes.

Notation 1. The letters (or nucleotides or bases) define the genetic alphabet $A_4 = \{\text{A, C, G, T}\}$. The set of non-empty words (words resp.) over A_4 is denoted by A_4^+ (A_4^* resp.). Let $x_1 \ldots x_n$ be the concatenation of the words x_i for $i = 1, \ldots, n$.

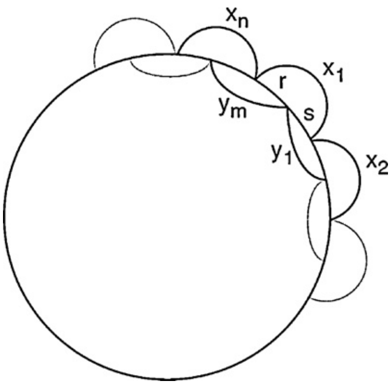

Fig. 1. A graphical representation of the circular code definition (Definition 2). A set of words Y is a circular code if any word (sequence) generated by a concatenation of words of Y and written on a circle has a unique decomposition into words of Y.

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 $i = 1, \ldots, n$.

Remark 1. The set $T = \{\text{AAA}, \ldots, \text{TTT}\}$ itself is a code. Consequently, its non-empty subsets are codes. In this paper, we call them trinucleotide codes.

Definition 2. *Trinucleotide circular code*: A trinucleotide code Y is circular if, for each $x_1, \ldots, x_n, y_1, \ldots, y_m \in Y$, $n, m \geq 1$, $r \in A_4^*$, $s \in A_4^+$, 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 $i = 1, \ldots, n$.

Fig. 1 gives a graphical representation of the trinucleotide circular code definition.

Remark 2. A set containing a trinucleotide with identical nucleotides, e.g. AAA, cannot be a circular code (detail, e.g. in Michel, 2008). Thus, the set T is obviously not a trinucleotide circular code.

Remark 3. A set containing two trinucleotides related to circular permutation $\mathcal{P}$, e.g. AAC and $\mathcal{P}(\text{AAC}) = \text{ACA}$, cannot be a circular code (detail, e.g. in Michel, 2008). Thus, the set $T \backslash T_{id}$ is also not a trinucleotide circular code.

A trinucleotide circular code allows the reading frame 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: $Y = \{\text{AAT, ATG, CCT, CTA, GCC, GGC}\}$ and the word w, be a series of the nine following letters: $w = \text{ATGGCCCTA}$. 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. $Z = \{\text{AAT, ATG, CCT, CTA, GCC, GTC}\}$, then there never exists an ambiguous word with Z, such as w for Y, and then Z is a circular code.

Definition 3. *Window of a circular code*: The construction frame (reading frame) of a word w generated by any concatenation of words of a circular code Y (or shortly, a word w of a circular code Y or even more simply a word w of Y) can be retrieved anywhere in the word w after the reading of a certain number of letters. This series of letters is called the window of the circular code Y. Then, the window length to retrieve the construction frame of the word w is the letter length of the longest ambiguous word which can be read in at least two frames, plus one letter. The window length depends on the circular code Y. The longest window length with the 12,964,440 trinucleotide circular codes is equal to 13 nucleotides. The common trinucleotide circular code X has a window length of 13 nucleotides (Arquès and Michel, 1996) and so, it belongs to the trinucleotide circular codes among 12,964,440 with a maximum window length. The classical window length $\tilde{l}$ is defined for a biinfinite word $\tilde{w} = \ldots l_{-1} l_0 l_1 \ldots$ of a circular code Y.

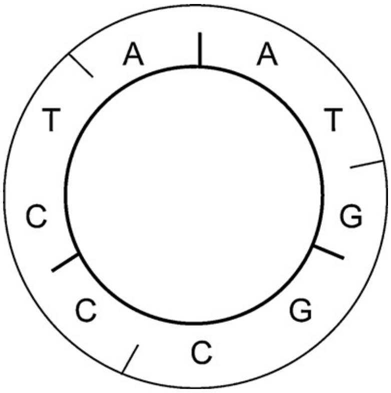

Fig. 2. The set $Y = \{\text{AAT, ATG, CCT, CTA, GCC, GGC}\}$ is not a circular code as the word $w = \text{ATGGCCCTA}$ written on a circle can be factorized into words of Y according to two different ways: ATG, GCC, CTA and AAT, GGC, CCT.

In order to analyse trinucleotide circular code motifs read in only one direction and to investigate the reading frame retrieval of their trinucleotides, in the same way as the trinucleotides (and nucleotides) are read only in one direction in DNA and RNA, i.e. the 5′-3′ direction, we introduce a new concept based on a window length l for a right infinite word $w = l_0 l_1 l_2 l_3 l_4 \ldots$ of a trinucleotide circular code Y such that $l_0 l_1 l_2 \in Y$, i.e. w begins with a trinucleotide belonging to Y and ends with either a proper prefix of a trinucleotide of Y or with a trinucleotide of Y.

We first consider the classical case of the window length $\tilde{l}$ for a biinfinite word $\tilde{w} = \ldots l_{-1} l_0 l_1 \ldots$ of a circular code Y. Fig. 3 shows an example with the biinfinite word $\tilde{w} = \ldots \text{AGGTAATTACCAG} \ldots$ of the common circular code X. Is the first nucleotide of $\tilde{w}$, i.e. A, the 1st, the 2nd or the 3rd nucleotide of a trinucleotide of X? By trying the three possible factorizations (frames) $\tilde{w}_0$, $\tilde{w}_1$ and $\tilde{w}_2$ ($\tilde{w}_1$ and $\tilde{w}_2$ being $\tilde{w}_0$ shifted by one and two nucleotides, respectively) into trinucleotides of X, only one factorization, i.e. $\tilde{w}_1$, is possible. Thus, the first nucleotide A of $\tilde{w}$ is the 3rd nucleotide of a trinucleotide of X. Indeed, the factorization $\tilde{w}_1$ leads to the trinucleotides NNA, GGT, AAT, TAC and CAG (N being any appropriate letter of X) which belong to X (1). The factorizations $\tilde{w}_0$ and $\tilde{w}_2$ are impossible as no trinucleotide of X starts with the prefix AG (1). This case occurs immediately for $\tilde{w}_0$ and after 11 letters for $\tilde{w}_2$ (Fig. 3). Thus, the unique factorization of $\tilde{w}$ is $\tilde{w}_1 = \ldots \text{A, GGT, AAT, TAC, CAG}, \ldots$. This word $\tilde{w}$ can be located anywhere in a sequence of X, i.e. the sequence of X does not require an initiator codon (or a stop codon) to retrieve the reading frame. The (finite) word $\tilde{w}_a = \text{AGGTAATTACCA}$ ($\tilde{w}$ without the last G) with a length of 12 nucleotides is ambiguous as it has two factorizations $\tilde{w}_1$ and $\tilde{w}_2$ into trinucleotides of X (Fig. 3). The word $\tilde{w}_a$ is

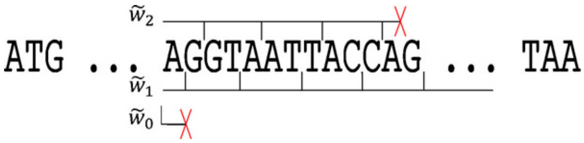

Fig. 3. Retrieval of the reading frame of the biinfinite word $\tilde{w} = \ldots \text{AGGTAATTACCAG} \ldots$ of the common circular code X. Among the three factorizations $\tilde{w}_0$, $\tilde{w}_1$ and $\tilde{w}_2$, only the unique factorization $\tilde{w}_1$ in words of X is possible leading to ... A, GGT, AAT, TAC, CAG, ... Thus, the first letter A of $\tilde{w}$ is the 3rd letter of a trinucleotide of X.

called ambiguous word of X. By definition of a circular code, all the ambiguous words are finite words. We will prove that $\tilde{w}_a$, taken as an illustration example here, is one of the two longest ambiguous words of X (Section 3.1). Thus, the window length $\tilde{l}$ to retrieve the construction frame of any biinfinite word of a circular code Y is the letter length of the longest ambiguous words $\tilde{w}_a$, plus one letter. Thus, with the common circular code X, $\tilde{l} = 12 + 1 = 13$ nucleotides (Michel, 2008). The window lengths $\tilde{l}$ for the trinucleotide circular codes $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ are also equal to $\tilde{l} = 13$ nucleotides (Michel, 2008). In conclusion, the retrieval of the reading frame (frame 1 and frame 2, respectively) of the common circular code X (X_1 and X_2, respectively) needs a window length $\tilde{l}$ of 13 nucleotides ($\tilde{l} \geq 13$) in each frame.

The new concept of a window length l to retrieve the construction frame for any right infinite word $w = l_0l_1l_2l_3l_4\ldots$ of a circular code Y such that $l_0l_1l_2 \in Y$ is the letter length of the longest ambiguous words w_a, plus one letter. The window length l for a right infinite word w of Y and the window length $\tilde{l}$ for a biinfinite word $\tilde{w}$ of Y depend on the same ambiguous words of Y. With the previous example of the common circular code X, the right infinite word w associated to the biinfinite word $\tilde{w} = \ldots$AGGTAATTACCAG$\ldots$ is w = GGTAATTACCAG$\ldots$ as GGT $\in X$. Its ambiguous (finite) word w_a = GGTAATTACCA (w without the last G) has a length of 11 nucleotides. Similarly to $\tilde{w}_a$, w_a is one of the two longest ambiguous words of X (Section 3.1). Thus, with the common circular code X, $l = 11 + 1 = 12$ nucleotides. In the following of the paper, only the window length l associated with any right infinite word is considered.

Property 1. *At window length $l \geq 12$ nucleotides (with any right infinite word), there is no ambiguous word of the common circular code X.*

Definition 4. *Complementary map $\mathcal{C}$*: The complementary map $\mathcal{C} : A_4^+ \rightarrow A_4^+$ is defined by $\mathcal{C}(\mathrm{A}) = \mathrm{T}$, $\mathcal{C}(\mathrm{C}) = \mathrm{G}$, $\mathcal{C}(\mathrm{G}) = \mathrm{C}$, $\mathcal{C}(\mathrm{T}) = \mathrm{A}$ and, according to the property of the complementary and antiparallel double helix, by $\mathcal{C}(uv) = \mathcal{C}(v)\mathcal{C}(u)$ for all $u, v \in A_4^+$, e.g. $\mathcal{C}(\mathrm{AAC}) = $ GTT. This map on words is naturally extended to word sets: a complementary trinucleotide set is obtained by applying the complementary map $\mathcal{C}$ to all its trinucleotides.

Definition 5. *Circular permutation map $\mathcal{P}$*: The circular permutation map $\mathcal{P} : T \rightarrow T$ permutes circularly each trinucleotide $l_0l_1l_2$ as follows $\mathcal{P}(l_0l_1l_2) = l_1l_2l_0$, e.g. $\mathcal{P}(\mathrm{AAC}) = \mathrm{ACA}$. The kth iterate of $\mathcal{P}$ is denoted $\mathcal{P}^k$, e.g. $\mathcal{P}^2(\mathrm{AAC}) = \mathrm{CAA}$. This map on words is also naturally extended to word sets: a permuted trinucleotide set is obtained by applying the circular permutation map $\mathcal{P}$ (or the kth iterate of $\mathcal{P}$) to all its trinucleotides.

Definition 6. *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 7. *Permutation trinucleotide set*: If Y is a trinucleotide circular code, we denote by Y_1 the permuted trinucleotide set $\mathcal{P}(Y)$, i.e. for each $y \in Y$, $\mathcal{P}(y) \in \mathcal{P}(Y)$, and similarly, by Y_2 the permuted trinucleotide set $\mathcal{P}^2(Y)$.

Definition 8. *C^3 trinucleotide circular code*: A trinucleotide circular code Y is C^3 if the permuted trinucleotide sets $Y_1 = \mathcal{P}(Y)$ and $Y_2 = \mathcal{P}^2(Y)$ 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 Y is C^3 self-complementary if Y, $Y_1 = \mathcal{P}(Y)$ and $Y_2 = \mathcal{P}^2(Y)$ are trinucleotide circular codes satisfying the following properties $Y = \mathcal{C}(Y)$ (self-complementary), $\mathcal{C}(Y_1) = Y_2$ and $\mathcal{C}(Y_2) = Y_1$ (Y_1 and Y_2 are complementary).

Result 1 (Arquès and Michel, 1996). The common trinucleotide set X coding the reading frames (frames 0) of eukaryotic and prokaryotic genes is a C^3 self-complementary trinucleotide circular code.

Property 2. *As a consequence of Result 1, the self-complementary circular code $X = \mathcal{C}(X)$ and the complementary circular codes $X_1 = \mathcal{C}(X_2)$ and $X_2 = \mathcal{C}(X_1)$ 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 circular codes X, $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ allow retrieval of the reading frame and the shifted frames 1 and 2 in genes, locally anywhere in the three frames and in particular without start codon in reading frame, and with a window length l of 12 nucleotides in each frame.*

A review of this trinucleotide circular code X details its additional properties (Michel, 2008).

Property 3. *By extending Property 2 with the cDNA strand to RNAs (messenger, transfer and ribosomal RNAs), X circular code motifs in RNAs can also pair according to different configurations: mRNA-mRNA, tRNA-tRNA, rRNA-rRNA, mRNA-tRNA, mRNA-rRNA and tRNA-rRNA with the property of "frame" retrieval.*

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

Result 3 (Ahmed and Michel, 2011). In frameshift genes, the X circular code signal shifts in the same direction of translational frameshifting.

Result 4 (Gonzalez et al., 2011). A new definition of a statistical function analysing the covering capability of a circular code has showed on a recent gene data set that the common circular code X has, on average, the best covering capability among the whole class of the 216 C^3 self-complementary trinucleotide circular codes. Furthermore, permutation tests of bases in the codon sites of X also suggest a reading frame synchronization property of X.

From a coding theory point of view, the common circular code X has a function of retrieval, maintenance and synchronization in mRNAs (reading frames in genes). The aim of this paper is to identify X circular code motifs in transfer and ribosomal RNAs (molecular structures involved in translation) and to visualize their spatial relations.

The paper is organized into two parts. The first part is a theoretical study of the ambiguous words of the common trinucleotide circular code X (Sections 3.1 and 3.2). It was never done since the identification of this circular code X in 1996. The type, the length and the occurrence number (multiplicity) of the ambiguous words of X in the two shifted frames are described. These results allow to define some probabilities of reading frame retrieval of X (Section 3.3). In the second part, X circular code motifs (X motifs) are identified in tRNA and 16S rRNA (Section 3.4). Their properties, i.e. type, position and reading frame retrieval, are analysed thanks to the previous determination of ambiguous words of X. The X-ray crystallography structure 3I8G of the elongation complex of the 70S ribosome of *T. thermophilus* containing the 16S rRNA, three tRNAs (A-tRNA, P-tRNA, E-tRNA) and the mRNA allows a 3D representation of the identified X motifs (Section 3.5). Several spatial regions in the ribosome involve mRNA X motifs, A-tRNA and E-tRNA X motifs and four rRNA X motifs. Thus, from a theoretical point of

Table 1
Ambiguous words of the common circular code X in frames 1 and 2 at nucleotide length l, $1 \leq l \leq 11$. The multisets $\mathcal{M}_1^l$ and $\mathcal{M}_2^l$ gives the type and the occurrence number (multiplicity) of ambiguous words of X in frames 1 and 2, respectively, at nucleotide length l. For example, at length $l = 1$, the ambiguous nucleotide A occurs eight times in frame 1. The ambiguous words of X in bold occur in both frames, however with different occurrences.

	Ambiguous words of the circular code X in frame 1			Ambiguous words of the circular code X in frame 2		
Length l	$\mathcal{M}_1^l$	$\mathrm{card}(M_1^l)$	$\mathrm{card}(\mathcal{M}_1^l)$	$\mathcal{M}_2^l$	$\mathrm{card}(M_2^l)$	$\mathrm{card}(\mathcal{M}_2^l)$
1	**A:8 C:2 G:2 T:8**	4	20	**A:2 C:10 G:3 T:5**	4	20
2	**AA:1 AC:3 AT:2 GC:1 GT:1 TA:1 TT:2**	7	11	**AA:10 AC:6 AT:4 GC:9 GT:6 TA:25 TT:10** CA:50 CT:20 GA:15 GG:30	11	185
3	**AAC:3 AAT:2 ATT:4 GTA:5 GTT:2 TAC:3** ACC:9 ATC:6 GCC:3 GTC:3 TTC:6	11	46	**AAC:2 AAT:4 ATT:2 GTA:3 GTT:3 TAC:5** GAA:6 GAC:3 GAT:6 GGC:3 GGT:9	11	46
4	**AATT:1 GTAC:1** AACA:1 AACT:2 AATA:1 ACCA:3 ACCT:6 ATCA:2 ATCT:4 ATTA:2 ATTT:2 GCCA:1 GCCT:2 GTAA:2 GTAT:2 GTCA:1 GTCT:2 GTTA:1 GTTT:1 TACA:1 TACT:2 TTCA:2 TTCT:4	23	46	**AATT:2 GTAC:3** AACC:2 AATC:2 ATTC:2 GAAC:3 GAAT:3 GACC:3 GATC:3 GATT:3 GGCC:3 GGTA:3 GGTC:3 GGTT:3 GTTC:3 TACC:5	16	46
5	AATAC:1 ATTAC:2 GTAAC:1 GTAAT:1 GTATT:1 GTTAC:1	6	7	AACCA:10 AACCT:4 AATCA:10 AATCT:4 AATTA:10 AATTT:4 ATTCA:10 ATTCT:4 GAACA:15 GAACT:6 GAATA:15 GAATT:6 GACCA:15 GACCT:6 GATCA:15 GATCT:6 GATTA:15 GATTT:6 GGCCA:15 GGCCT:6 GGTAA:15 GGTAC:9 GGTAT:6 GGTCA:15 GGTCT:6 GGTTA:15 GGTTT:6 GTACA:15 GTACT:6 GTTCA:15 GTTCT:6 TACCA:25 TACCT:10	33	331
6	AATACC:3 ATTACC:6 GTAACC:3 GTAATC:3 GTAATT:2 GTATTC:3 GTTACC:3	7	23	AATTAC:2 GAATAC:3 GATTAC:3 GGTAAC:3 GGTAAT:6 GGTATT:3 GGTTAC:3	7	23
7	AATACCA:1 AATACCT:2 ATTACCA:2 ATTACCT:4 GTAACCA:1 GTAACCT:2 GTAATCA:1 GTAATCT:2 GTAATTA:1 GTAATTT:1 GTATTCA:1 GTATTCT:2 GTTACCA:1 GTTACCT:2	14	23	AATTACC:2 GAATACC:3 GATTACC:3 GGTAACC:3 GGTAATC:3 GGTAATT:3 GGTATTC:3 GGTTACC:3	8	23
8	GTAATTAC:1	1	1	AATTACCA:10 AATTACCT:4 GAATACCA:15 GAATACCT:6 GATTACCA:15 GATTACCT:6 GGTAACCA:15 GGTAACCT:6 GGTAATCA:15 GGTAATCT:6 GGTAATTA:15 GGTAATTT:6 GGTATTCA:15 GGTATTCT:6 GGTTACCA:15 GGTTACCT:6	16	161
9	GTAATTACC:3	1	3	GGTAATTAC:3	1	3
10	GTAATTACCA:1 GTAATTACCT:2	2	3	GGTAATTACC:3	1	3
11		0	0	GGTAATTACCA:15 GGTAATTACCT:6	2	21

Table 2
Maximal ambiguous words of the common circular code *X*.

Length *l*	Maximal ambiguous words of the circular code *X*
5	GAACA GAACT GAATT GACCA GACCT GATCA GATCT GATTT GGCCA GGCCT GGTAC GGTCA GGTCT GGTTT GTACA GTACT GTTCA GTTCT
8	GAATACCA GAATACCT GATTACCA GATTACCT GGTAACCA GGTAACCT GGTAATCA GGTAATCT GGTAATTT GGTATTCA GGTATTCT GGTTACCA GGTTACCT
11	GGTAATTACCA GGTAATTACCT

view, the identified *X* circular code motifs and their mathematical properties may be a translation code for retrieving, maintaining and synchronizing the reading frame in genes.

2. Method

2.1. Identification of the ambiguous words of the common trinucleotide circular code *X*

The ambiguous words of the common trinucleotide circular code *X*, i.e. words of *X* occurring in at least one shifted frame, are identified by algorithm. No combinatorial approach is possible so far.

A set is a collection of distinct elements without repetition and without order. It is written here with an open face font, e.g. S. A multiset is a generalization of a set. It is an unordered collection of elements with multiple but finite occurrences of any element. It is written here with a script font, e.g. $\mathcal{S}$. The multiplicity $m_{\mathcal{S}}(e)$ of an element e in a multiset $\mathcal{S}$ is its occurrence number. In our context, and for readability reason, a multiset is represented as follows, e.g. $\mathcal{S} = \{\text{A} : m_{\mathcal{S}}(\text{A}), \text{C} : m_{\mathcal{S}}(\text{C}), \text{G} : m_{\mathcal{S}}(\text{G}), \text{T} : m_{\mathcal{S}}(\text{T})\}$. We briefly recall the definitions of intersection and union for multisets. Let $\mathcal{S}$ and $\mathcal{T}$ be two multisets. The union $\mathcal{S} \cup \mathcal{T}$ of $\mathcal{S}$ and $\mathcal{T}$ is the multiset defined by $m_{\mathcal{S}\cup\mathcal{T}}(e) = \max(m_{\mathcal{S}}(e), m_{\mathcal{T}}(e))$, i.e. the multiplicity of an element in $\mathcal{S} \cup \mathcal{T}$ is equal to the maximum of the multiplicities of the element in $\mathcal{S}$ and $\mathcal{T}$. For example, if $\mathcal{S} = \{\text{A} : 3, \text{G} : 1, \text{T} : 2\}$ and $\mathcal{T} = \{\text{A} : 2, \text{C} : 1, \text{G} : 2\}$ then $\mathcal{S} \cup \mathcal{T} = \{\text{A} : 3, \text{C} : 1, \text{G} : 2, \text{T} : 2\}$. The intersection $\mathcal{S} \cap \mathcal{T}$ of $\mathcal{S}$ and $\mathcal{T}$ is the multiset defined by $m_{\mathcal{S}\cap\mathcal{T}}(e) = \min(m_{\mathcal{S}}(e), m_{\mathcal{T}}(e))$, i.e. the multiplicity of an element in $\mathcal{S} \cap \mathcal{T}$ is equal to the minimum of the multiplicities of the element in $\mathcal{S}$ and $\mathcal{T}$. With the previous example, $\mathcal{S} \cap \mathcal{T} = \{\text{A} : 2, \text{G} : 1\}$. Finally, a subset S of a multiset $\mathcal{S}$ is called the support of $\mathcal{S}$ if for every element e such that $m_{\mathcal{S}}(e) > 0$ this implies that $e \in S$, and for every element e such that $m_{\mathcal{S}}(e) = 0$ this implies that $e \notin S$. For example, the set $S = \{\text{A, G, T}\}$ is the support of $\mathcal{S} = \{\text{A} : 3, \text{G} : 1, \text{T} : 2\}$. For simplification in the writing of the algorithm, the same operators of intersection and union are used for sets and multisets. Thus, the intersection $S \cap \mathcal{T}$ of a set S and a multiset $\mathcal{T}$ leads to a set (the support T of $\mathcal{T}$ replacing $\mathcal{T}$). The union $S \cup \mathcal{T}$ of a set S and a multiset $\mathcal{T}$ leads to a multiset (the multiset $\mathcal{S}$ of multiplicity 1 replacing S).

Let the set X be the common trinucleotide circular code defined in (1). Let a word of X be the three letters $l_1 l_2 l_3$. Let S_1 (S_2 and S_3, respectively) be the set containing the letters l_1 (l_2 and l_3, respectively) of X. Then,

$$S_1 = S_2 = S_3 = A_4 = \{\text{A, C, G, T}\} \tag{2}$$

Let $\mathcal{S}_1$ ($\mathcal{S}_2$ and $\mathcal{S}_3$, respectively) be the multiset containing the letters l_1 (l_2 and l_3, respectively) of X. Then,

$$\mathcal{S}_1 = \{\text{A} : 5, \text{C} : 3, \text{G} : 10, \text{T} : 2\} \tag{3}$$

$$\mathcal{S}_2 = \{\text{A} : 8, \text{C} : 2, \text{G} : 2, \text{T} : 8\} \tag{4}$$

$$\mathcal{S}_3 = \{\text{A} : 2, \text{C} : 10, \text{G} : 3, \text{T} : 5\} \tag{5}$$

Remark 5. As the trinucleotide set X is self-complementary, $\mathcal{C}(\mathcal{S}_1) = \{\mathcal{C}(\text{A}) : 5, \mathcal{C}(\text{C}) : 3, \mathcal{C}(\text{G}) : 10, \mathcal{C}(\text{T}) : 2\} = \{\text{T} : 5, \text{G} : 3, \text{C} : 10, \text{A} : 2\} = \mathcal{S}_3$. Similarly, $\mathcal{C}(\mathcal{S}_3) = \mathcal{S}_1$. Also, $\mathcal{C}(\mathcal{S}_2) = \{\mathcal{C}(\text{A}) : 8, \mathcal{C}(\text{C}) : 2, \mathcal{C}(\text{G}) : 2, \mathcal{C}(\text{T}) : 8\} = \{\text{T} : 8, \text{G} : 2, \text{C} : 2, \text{A} : 8\} = \mathcal{S}_2$.

Let S_{12} ($\mathcal{S}_{12}$, respectively) be the set (multiset, respectively) containing the prefix $l_1 l_2$ of X. Then,

$$S_{12} = \{\text{AA, AC, AT, CA, CT, GA, GC, GG, GT, TA, TT}\} \tag{6}$$

$$\begin{aligned}\mathcal{S}_{12} = \{&\text{AA} : 2, \text{AC} : 1, \text{AT} : 2, \text{CA} : 1, \text{CT} : 2, \text{GA} : 4,\\ &\text{GC} : 1, \text{GG} : 2, \text{GT} : 3, \text{TA} : 1, \text{TT} : 1\}\end{aligned} \tag{7}$$

Let S_{23} ($\mathcal{S}_{23}$, respectively) be the set (multiset, respectively) containing the suffix $l_2 l_3$ of X. Then,

$$S_{23} = \{\text{AA, AC, AG, AT, CC, GC, GT, TA, TC, TG, TT}\} \tag{8}$$

$$\begin{aligned}\mathcal{S}_{23} = \{&\text{AA} : 1, \text{AC} : 3, \text{AG} : 2, \text{AT} : 2, \text{CC} : 2, \text{GC} : 1, \text{GT} : 1,\\ &\text{TA} : 1, \text{TC} : 4, \text{TG} : 1, \text{TT} : 2\}\end{aligned} \tag{9}$$

Remark 6. $\text{card}(S_{12}) = \text{card}(S_{23}) = 11$ (among 16 dinucleotides). $S_{12} \neq S_{23}$ and $S_{12} \cap S_{23} = \{\text{AA, AC, AT, GC, GT, TA, TT}\}$. $\mathcal{C}(S_{12}) = S_{23}$, $\mathcal{C}(S_{23}) = S_{12}$ and $\mathcal{C}(S_{12} \cap S_{23}) = S_{12} \cap S_{23}$.

Let A and B be two sets of words. $A{\cdot}B$ is the set of words which are the products (concatenation $\cdot$) of one word of A and one word of B, i.e. $A{\cdot}B = \{a_i{\cdot}b_j | a_i \in A, b_j \in B\}$. Thus, $A^n = \underbrace{A \cdot A \cdot \ldots \cdot A}_{n}$ is the set of words which are the products of n, $n \geq 0$, words of A,

Table 3
Probability of reading frame retrieval of the common trinucleotide circular code *X*. The probability $p(l)$ of reading frame retrieval is based on the sets M_f^l of ambiguous words of *X* in frames 1 and 2. The probability $q(l)$ of reading frame retrieval is based on the multisets $\mathcal{M}_f^l$ of ambiguous words of *X* in frames 1 and 2. Both definitions show that the common trinucleotide circular code *X* retrieves the reading frame with a probability of about 90% with a window length of 6 nucleotides, and a probability of 99.9% with a window length of 9 nucleotides (100% with a window length of 12 nucleotides, Property 1).

Nucleotide window length l of the circular code X	1	2	3	4	5	6	7	8	9	10	11
card (W^l) of the sets W^l of words of X	4	11	20	80	220	400	1600	4400	8000	32,000	88,000
$\text{card}(M^l) = \text{card}(M_1^l) + \text{card}(M_2^l)$ of the sets M_f^l of ambiguous words of X in frames 1 and 2	8	18	22	39	39	14	22	17	2	3	2
Probability $p(l) = \text{card}(W^l)/(\text{card}(W^l) + \text{card}(M^l))$ (%) of reading frame retrieval of the circular code X	33.333	37.931	47.619	67.227	84.942	96.618	98.644	99.615	99.975	99.991	99.998
$\text{card}(\mathcal{W}^l)$ of the multisets $\mathcal{W}^l$ of words of X	20	20	20	400	400	400	8000	8000	8000	160,000	160,000
$\text{card}(\mathcal{M}^l) = \text{card}(\mathcal{M}_1^l) + \text{card}(\mathcal{M}_2^l)$ of the multisets $\mathcal{M}_f^l$ of ambiguous words of X in frames 1 and 2	40	196	92	92	338	46	46	162	6	6	21
Probability $q(l) = \text{card}(\mathcal{W}^l)/(\text{card}(\mathcal{W}^l) + \text{card}(\mathcal{M}^l))$ (%) of reading frame retrieval of the circular code X	33.333	9.259	17.857	81.301	54.200	89.686	99.428	98.015	99.925	99.996	99.987

i.e. $A^n = \{a_1 \cdot a_2 \cdot \ldots \cdot a_n | a_i \in A\}$, A^0 being the empty set. For example, X^2 is the set of all concatenations of two words of X, i.e. {AACAAC, AACAAT, . . ., TTCTTC} (1). All these definitions on sets are naturally extended on multisets.

The algorithm AmbiguousWords$X$ gives the ambiguous words of the common circular code X in the shifted frames (f) 1 and 2 with a length varying from 1 to 11 nucleotides. Remember that there is no ambiguous word of X with a length $l \geq 12$ nucleotides (Property 1).

```
Algorithm_AmbiguousWords$X$
// Ambiguous words of $X$ in frames 1 and 2
    for $l \leftarrow$ 1 to 11 step +1 do // Nucleotide length $l$

        // Set $W^l$ of words of $X$ at length $l$
        $W^l \leftarrow$ wordsX($l$)

        // Multiset $\mathcal{W}_1^l$ of words of $X$ in frame 1 at length $l$
        $\mathcal{W}_1^l$ ←wordsXFrame1($l$)
        // Multiset $\mathcal{W}_2^l$ of words of $X$ in frame 2 at length $l$
        $\mathcal{W}_2^l$ ←wordsXFrame2($l$)

        for $f \leftarrow$ 1 to 2 step +1 do // Frame $f$
            // Set $M_f^l$ of ambiguous words of $X$ in frame $f$ at length $l$
            $M_f^l \leftarrow W^l \cap \mathcal{W}_f^l$
            // Multiset $\mathcal{M}_f^l$ of ambiguous words of $X$ in frame $f$ at length $l$
            $\mathcal{M}_f^l \leftarrow M_f^l \cup \mathcal{W}_f^l$
```

```
wordsX($l$)
// Determination of the set $W$ of words of $X$
    $W \leftarrow \{\}$
    if $l = 1[3]$ then $W \leftarrow X^{\lfloor \frac{l}{3} \rfloor} \cdot S_1$
    else if $l = 2[3]$ then $W \leftarrow X^{\lfloor \frac{l}{3} \rfloor} \cdot S_{12}$
        else $W \leftarrow X^{\lfloor \frac{l}{3} \rfloor}$
    return $W$
```

```
wordsXFrame1($l$)
// Determination of the multiset $\mathcal{W}$ of words of $X$ in frame 1
    $\mathcal{W} \leftarrow \{\}$
    if $l = 1$ then $\mathcal{W} \leftarrow \mathcal{S}_2$
    else if $l = 2[3]$ then $\mathcal{W} \leftarrow \mathcal{S}_{23} \cdot X^{\lfloor \frac{l}{3} \rfloor}$
        else if $l = 0[3]$ then $\mathcal{W} \leftarrow \mathcal{S}_{23} \cdot X^{\lfloor \frac{l}{3} \rfloor - 1} \cdot \mathcal{S}_1$
            else $\mathcal{W} \leftarrow \mathcal{S}_{23} \cdot X^{\lfloor \frac{l}{3} \rfloor - 1} \cdot \mathcal{S}_{12}$
    return $\mathcal{W}$
```

```
wordsXFrame2($l$)
// Determination of the multiset $\mathcal{W}$ of words of $X$ in frame 2
    $\mathcal{W} \leftarrow \{\}$
    if $l = 1[3]$ then $\mathcal{W} \leftarrow \mathcal{S}_3 \cdot X^{\lfloor \frac{l}{3} \rfloor}$
    else if $l = 2[3]$ then $\mathcal{W} \leftarrow \mathcal{S}_3 \cdot X^{\lfloor \frac{l}{3} \rfloor} \cdot \mathcal{S}_1$
        else $\mathcal{W} \leftarrow \mathcal{S}_3 \cdot X^{\lfloor \frac{l}{3} \rfloor} \cdot \mathcal{S}_{12}$
    return $\mathcal{W}$
```

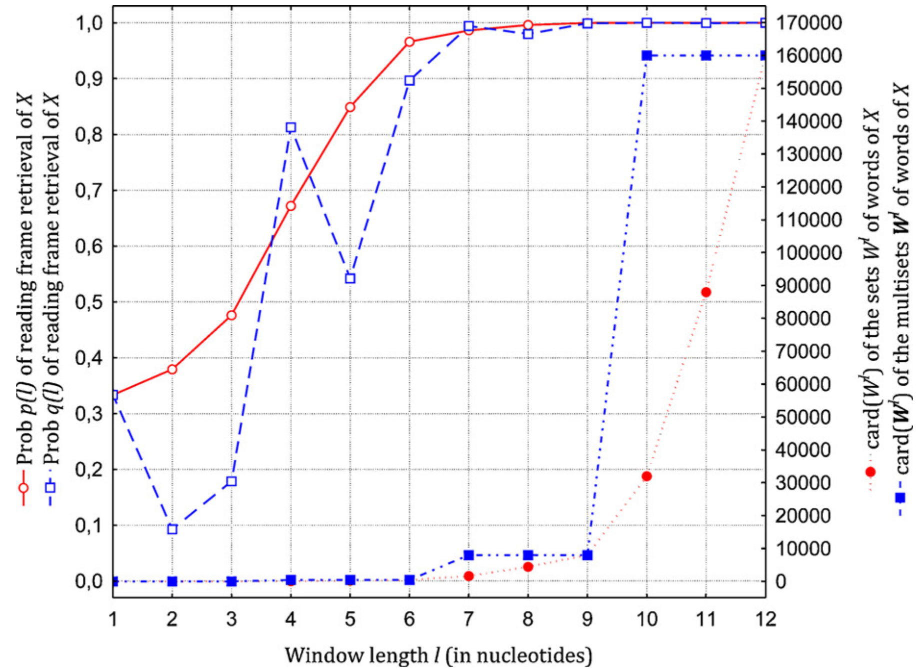

Fig. 4. Probability of reading frame retrieval of the common trinucleotide circular code X. Graphical representation of the probabilities $p(l)$ and $q(l)$ (Table 3). Both definitions show that the common trinucleotide circular code X retrieves the reading frame with a probability of about 90% with a window length of 6 nucleotides, and a probability of 99.9% with a window length of 9 nucleotides (100% with a window length of 12 nucleotides, Property 1).

Remark 7. $\lfloor x \rfloor = \max\{n \in \mathbb{Z} | n \leq x\}$ gives the greatest integer less than or equal to x.

Remark 8. M_f^l is the support of $\mathcal{M}_f^l$.

Example 1. We give a step of the algorithm AmbiguousWords$X$ for determining the ambiguous words of X in frame 1 at length 2. The set W^2 of words of X at length 2 is W^2 = {AA, AC, AT, CA, CT, GA, GC, GG, GT, TA, TT}. Note that $W^2 = S_{12}$. The multiset $\mathcal{W}_1^2$ of words of X in frame 1 at length 2 is $\mathcal{W}_1^2$ = {AA : 1, AC : 3, AG : 2, AT : 2, CC : 2, GC : 1, GT : 1, TA : 1, TC : 4, TG : 1, TT : 2}. The set M_1^2 of ambiguous words of X in frame 1 at length 2 is $M_1^2 = W^2 \cap \mathcal{W}_1^2$ = {AA, AC, AT, GC, GT, TA, TT} which is the support of $\mathcal{M}_1^2$ = {AA : 1, AC : 3, AT : 2, GC : 1, GT : 1, TA : 1, TT : 2}.

2.2. *Definition of probabilities of reading frame retrieval with the common trinucleotide circular code X*

Let l, $1 \leq l \leq 11$, be the nucleotide length of words of the common circular code X. The number card(W^l) of the set W^l of words of X at length l is equal to

$$\begin{aligned}\mathrm{card}(W^l) &= \begin{cases} \mathrm{card}(S_1) \times 20^{\lfloor l/3 \rfloor} & \text{if} \quad l = 1[3] \\ \mathrm{card}(S_{12}) \times 20^{\lfloor l/3 \rfloor} & \text{if} \quad l = 2[3] \\ \mathrm{card}(X)^{\lfloor l/3 \rfloor} & \text{if} \quad l = 0[3] \end{cases} \\ &= \begin{cases} 4 \times 20^{\lfloor l/3 \rfloor} & \text{if} \quad l = 1[3] \\ 11 \times 20^{\lfloor l/3 \rfloor} & \text{if} \quad l = 2[3] \\ 20^{\lfloor l/3 \rfloor} & \text{if} \quad l = 0[3] \end{cases}\end{aligned} \tag{10}$$

The number card($\mathcal{W}^l$) of the multiset $\mathcal{W}^l$ of words of X at length l is equal to

$$\mathrm{card}(\mathcal{W}^l) = \mathrm{card}(X)^{\lceil l/3 \rceil} = 20^{\lceil l/3 \rceil} \tag{11}$$

Remark 9. $\lceil x \rceil = \min\{n \in \mathbb{Z} | n \geq x\}$ gives the smallest integer greater than or equal to x.

Remark 10. $\mathrm{card}(W^l) = \mathrm{card}(\mathcal{W}^l)$ if $l = 0[3]$.

We propose two definitions to evaluate the probability of reading frame retrieval of the common circular code X.

The first definition is based on the sets M_f^l of ambiguous words of X in frames 1 and 2 which are computed by the algorithm AmbiguousWords$X$. Let $\mathrm{card}(M^l) = \mathrm{card}(M_1^l) + \mathrm{card}(M_2^l)$ be the total number of ambiguous words of X in frames 1 and 2 at length l among card(W^l) sets W^l of words of X at length l. Then, the probability $p(l)$ of reading frame retrieval at nucleotide length l is defined by

$$p(l) = \frac{\mathrm{card}(W^l)}{\mathrm{card}(W^l) + \mathrm{card}(M^l)} \tag{12}$$

The second definition is based on the multisets $\mathcal{M}_f^l$ of ambiguous words of X in frames 1 and 2 which are computed by the algorithm AmbiguousWords$X$. Let $\mathrm{card}(\mathcal{M}^l) = \mathrm{card}(\mathcal{M}_1^l) + \mathrm{card}(\mathcal{M}_2^l)$ be the total number of ambiguous words of X in frames 1 and 2 at length l among card($\mathcal{W}^l$) sets $\mathcal{W}^l$ of words of X at length l. Then, the probability $q(l)$ of reading frame retrieval at nucleotide length l is defined by

$$q(l) = \frac{\mathrm{card}(\mathcal{W}^l)}{\mathrm{card}(\mathcal{W}^l) + \mathrm{card}(\mathcal{M}^l)} \tag{13}$$

Remark 11. At nucleotide length $l \geq 12$, there is no ambiguous words of X in shifted frames (Property 1), thus $\mathrm{card}(M^l) = \mathrm{card}(\mathcal{M}^l) = 0$ and $p(l) = q(l) = 1$, i.e. the probability of reading frame retrieval of X is equal to 1.

2.3. *Crystallographic data*

The crystal structure of the elongation complex of the 70S ribosome with three tRNAs (A-tRNA, P-tRNA, E-tRNA) and the mRNA of *T. thermophilus* is obtained from the entry 3I8G (www.pdb.org/pdb/explore.do?structureId=3I8G, DOI:10.2210/pdb3i8g/pdb, NDB ID: NA0100; Jenner et al., 2010) of the Protein Data Bank (PDB, www.rcsb.org/pdb/home/home.do). We remove all ribosomal proteins in 3I8G in order to visualize the X circular code motifs in the mRNA, the three tRNAs (A-tRNA, P-tRNA, E-tRNA) and the 16S rRNA. Thus, the studied entry 3I8G contains:

- the 16S rRNA, polymer 1, type: polyribonucleotide, length: 1516, chains: A;
- the tRNA-Phe (with unmodified nucleotides except for MIA37), polymer: 22, type: polyribonucleotide, length: 76, chains: B, C, D;
- the mRNA, polymer: 23, type: polyribonucleotide, length: 30, chains: 1.

2.4. Scripts in Jmol language

Jmol is a Java molecular viewer for three-dimensional chemical structures. Features include reading a variety of file types and output from quantum chemistry programs, and animation of multi-frame files and computed normal modes from quantum programs.

We write several scripts in Jmol language (version 12.2: http://chemapps.stolaf.edu/jmol/docs) for a 3D visualization of the *X* circular code motifs in the messenger, transfer and ribosomal RNAs complex of 3I8G. These Jmol scripts are not detailed.

3. Results

3.1. Identification of the ambiguous words of the common trinucleotide circular code X

Table 1 shows the ambiguous words of the common circular code *X* in the shifted frames at nucleotide length *l*, $1 \le l \le 11$. There are 76 ambiguous words in frame 1 and 110 in frame 2, i.e. a total of 186 ambiguous words of *X*.

At length $l = 1$, the four nucleotides occur in both shifted frames.

At length $l = 2$, seven dinucleotides $D_{1,2}$ = {AA, AC, AT, GC, GT, TA, TT} occur in both shifted frames. The set $D_{1,2} = S_{12} \cap S_{23}$ is self-complementary (Remark 6). Four additional dinucleotides D_2 = {CA, CT, GA, GG} occur in frame 2.

At length $l = 3$, six trinucleotides $T_{1,2}$ = {AAC, AAT, ATT, GTA, GTT, TAC} occur in both shifted frames. The set $T_{1,2}$ is also self-complementary, i.e. $T_{1,2} = \mathcal{C}(T_{1,2})$. There are 11 trinucleotides in frame 1: T_1 = {ACC, ATC, GCC, GTC, TTC} $\cup\, T_{1,2}$ and 11 trinucleotides in frame 2: T_2 = {GAA, GAC, GAT, GGC, GGT} $\cup\, T_{1,2}$. The sets T_1 and T_2 are complementary, $T_1 = \mathcal{C}(T_2)$ and $T_2 = \mathcal{C}(T_1)$. Thus, there are four trinucleotides of the common circular code *X* which are not ambiguous: $\tilde{X}$ = {CAG, CTC, CTG, GAG}. This subset $\tilde{X}$ constitutes the most stable trinucleotides of *X* (see also Result 4 in Ahmed et al., 2010). This subset $\tilde{X}$ is a C^3 self-complementary trinucleotide comma-free code. The proof can be done by hand without formalism. Indeed, the words $l_1 l_2 l_3$ of $\tilde{X}$ have the following property: $l_1, l_3 \in$ {C, G} and $l_2 \in$ {A, T}. Thus, any concatenation $l_1 l_2 l_3 l'_1 l'_2 l'_3$ of two words of $\tilde{X}$ has the following structure: {C, G}{A, T}{C, G}{C, G}{A, T}{C, G}. Any word in frame 1 (starting from l_2) begins with {A, T} but no words of $\tilde{X}$ begin with {A, T}. The decomposition of any word of $\tilde{X}$ in frame 1 is impossible. Any word in frame 2 (starting from l_3) ends with {A, T} but no word of $\tilde{X}$ ends with {A, T}. The decomposition of any word of $\tilde{X}$ in frame 2 is also impossible. As $\tilde{X}$ has a unique decomposition so that no words of $\tilde{X}$ is in a shifted frame, $\tilde{X}$ is, by definition, a comma-free code. $\tilde{X}$ is also self-complementary: $\tilde{X} = \mathcal{C}(\tilde{X})$. The codon CAG codes for glutamine (Gln, Q), CTC and CTG for leucine (Leu, L) and GAG for glutamic acid (Glu, E). The sets $\tilde{X}_1 = \mathcal{P}(\tilde{X})$ = {AGC, AGG, TCC, TGC} and $\tilde{X}_2 = \mathcal{P}^2(\tilde{X})$ = {CCT, GCA, GCT, GGA} are also comma free-codes (proof similar) and complementary, i.e. $\mathcal{C}(\tilde{X}_1) = \tilde{X}_2$ and $\mathcal{C}(\tilde{X}_2) = \tilde{X}_1$.

At length $l = 4$, two tetranucleotides $\Gamma_{1,2}$ = {AATT, GTAC} occur in both shifted frames. The set $\Gamma_{1,2}$ is also self-complementary.

At length $l > 4$, no ambiguous word of *X* occur in both shifted frames. At length $l = 11$, there is no ambiguous word in frame 1 but two ambiguous words in frame 2: m_1 = GGTAATTACCA and m_2 = GGTAATTACCT. Note that $m_1 = w_a$ is the ambiguous word of the example in Definition 3.

3.2. Identification of maximal ambiguous words of the common trinucleotide circular code X

Table 2 gives the maximal ambiguous words of the common trinucleotide circular code *X*. A maximal ambiguous word is a word which is not a factor of another ambiguous word. There are 18 maximal ambiguous words of length 5, 13 maximal ambiguous words of length 8 and two maximal ambiguous words of length 11, i.e. a total of 33 among 186 (about 18%). All these maximal ambiguous words begin with the nucleotide G. None begins with the dinucleotide GC.

3.3. Probability of reading frame retrieval of the common trinucleotide circular code X

The probabilities $p(l)$ and $q(l)$ of reading frame retrieval of the common trinucleotide circular code *X* are computed from Eqs. (12) and (13), respectively (Table 3). Fig. 4 gives a graphical representation of these two probabilities $p(l)$ and $q(l)$ as well as the card(W^l) of the sets W^l of words of *X* and the card($\mathcal{W}^l$) of the multisets $\mathcal{W}^l$ of words of *X*. With a window length $l = 1$, the probability of reading frame retrieval of *X* is equal to 1/3. It is the random case as there is one chance out of three to retrieve the reading frame among the three frames (reading frame and the two shifted frames). Indeed, the four nucleotides occur in the three frames with equiprobability (20 times in each trinucleotide site). With both probability definitions, the common trinucleotide circular code *X* retrieves the reading frame in genes with a probability of about 90% with a window length of 6 nucleotides, and a probability of 99.9% with a window length of 9 nucleotides (100% with a window length of 12 nucleotides, Property 1). From a biological point of view, already two trinucleotides of *X* retrieve in average the reading frame with a good confidence.

3.4. Identification of X circular code motifs in transfer and 16S ribosomal RNAs

Longest *X* circular code motifs (*X* motifs) are searched in tRNA-Phe (76 nucleotides) and 16S rRNA (1516 nucleotides) of 3I8G. A *X* motif $m_s(b, e, l)$ is described by its begin position *b*, its end position *e* in the sequence *s* and its nucleotide length $l = e - b + 1$.

Table 4a

Identification of an almost perfect *X* motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ = G, GTA, GAG, CAG, GGG, ATT, GAA, AAT, CCC, C (blue and blueviolet in bold) of 26 nucleotides in tRNA-Phe (76 nucleotides) of the crystal structure 3I8G. The anticodon GAA (black in bold) of tRNA-Phe also belongs to the circular code *X* and the two nucleotides G30 and C40 (blue in bold and italics) are substituted by C30 and G40, respectively (see Remark 12 and Table 4b).

```
TRNA-Phe (1-76) Anticodon: GAA at 34-36
              *    |    *    |    *    |    *    |    *    |    *    |    *    |    *
Seq: GCCCGGATAGCTCAGTCGGTAGAGCAGGGGATTGAAAATCCCCGTGtCCTTGGTTCGATTCCGAGTCCGGGCACCA
Str: >>>>>>>..>>>>........<<<<.>>>>>.......<<<<<.....>>>>>.......<<<<<<<<<<<<....
```

Table 4b
Nucleotide window length of reading frame retrieval and its associated non-ambiguous word (from Table 1) of each trinucleotide of the *X* motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ (Table 4a and with the same color convention) identified in tRNA-Phe of the crystal structure 3I8G. For example, the window length of reading frame retrieval of the trinucleotide GTA is four nucleotides as GTA is an ambiguous word of the common circular code *X* in frames 1 and 2 (Table 1) but GTAG (G being the 1st nucleotide of the following trinucleotide GAG) is a non-ambiguous word of *X* (absent in Table 1). The trinucleotides CAG and GAG belong to the set $\tilde{X}$ of non-ambiguous words of *X* and their window lengths of reading frame retrieval are three nucleotides. The *X* motif $m_{\text{tRNA-Phe}}$ has a strong capacity of reading frame retrieval as six trinucleotides (eight trinucleotides if Remark 12 is considered) retrieve the reading frame with only three or four nucleotides.

Trinucleotides of the *X* motif $m_{\text{tRNA-Phe}}(18,43,26)$	**GTA**	**GAG**	**CAG**	***GGC*** G30 → C30	**ATT**	**GAA**	**AAT**	***GCC*** C40 → G40
Nucleotide window length of reading frame retrieval	4	3	3	4	4	4	4	4
Non-ambiguous word of the circular code *X*	**GTAG**	**GAG** $\in \tilde{X}$	**CAG** $\in \tilde{X}$	**GGC*A***	**ATTG**	**GAAA**	**AAT*G***	***GCC*C**

3.4.1. Identification of a X circular code motif in tRNA

An almost perfect *X* motif $m_{\text{tRNA-Phe}}(18,43,26)$ = G, GTA, GAG, CAG, GGG, ATT, GAA, AAT, CCC, C of 26 nucleotides is identified in tRNA-Phe of 3I8G (Table 4a). This *X* motif $m_{\text{tRNA-Phe}}$ is also represented in the secondary structure of tRNA-Phe (Fig. 5). This tRNA-Phe secondary structure was obtained with the program tRNAscan Search Server (v.1.21, http://lowelab.ucsc.edu/tRNAscan-SE; Lowe and Eddy, 1997). In the case of this tRNA-Phe, the anticodon GAA is also a trinucleotide of *X*. The *X* motif $m_{\text{tRNA-Phe}}$ is located in the anticodon stem-loop, the D-stem and partially in the D-loop with a *X* motif $m_{\text{tRNA-Phe-5'}}$ = G, GTA, GAG, CAG, GGG, ATT of 16 nucleotides before the anticodon GAA and a *X* motif $m_{\text{tRNA-Phe-3'}}$ = AAT, CCC, C of seven nucleotides after the anticodon GAA. Very unexpectedly, these two *X* motifs $m_{\text{tRNA-Phe-5'}}$ and $m_{\text{tRNA-Phe-3'}}$ are "in frame" with the anticodon.

Table 4b gives the nucleotide length of the reading frame retrieval and its associated non-ambiguous word of each trinucleotide of this *X* motif $m_{\text{tRNA-Phe}}$. Without loss of generality, this *X* motif $m_{\text{tRNA-Phe}}$ is formalized as follows $m_{\text{tRNA-Phe}} = l_{-1}, l_0 l_1 l_2, l_3 l_4 l_5, \ldots$, with $l_0 l_1 l_2, l_3 l_4 l_5 \in X$ and l_{-1} a suffix of *X*. Let $t = l_i l_{i+1} l_{i+2}$ with $i = 0 \bmod 3$ be a trinucleotide of $m_{\text{tRNA-Phe}}$. If $t \in \tilde{X} = \{\text{CAG, CTC, CTG, GAG}\}$, i.e. *t* belongs to the non-ambiguous set $\tilde{X}$, then the reading frame retrieval of *t* has three nucleotides. Otherwise, *t* is an ambiguous word of the circular code *X* which is identified in Table 1. Then, the next letters l_{i+3}, l_{i+4}, l_{i+5}, ... of $m_{\text{tRNA-Phe}}$ are concatenated to *t*, i.e. $t{\cdot}l_{i+3}$, $t{\cdot}l_{i+3}l_{i+4}$, $t{\cdot}l_{i+3}l_{i+4}l_{i+5}$, until a non-ambiguous word of *X* is obtained. By definition of a circular code, this concatenation process converges (Property 1). This procedure is repeated for all trinucleotides of $m_{\text{tRNA-Phe}}$, i.e. $t' = l_{i+3}l_{i+4}l_{i+5}$, etc.

Remark 12. The trinucleotide GGG and its complementary trinucleotide CCC in the anticodon-stem of this tRNA-Phe belong to the set T_{id} (Section 1.2) and not to *X* (1). As the nucleotide C40 is often a modified nucleotide (5-methyl-C), i.e. a 5-letter alphabet, we make an appropriate substitution of GGG into GGC and CCC into GCC as GGC and GCC belong to *X* (Table 4b).

The *X* motif $m_{\text{tRNA-Phe}}$ in tRNA-Phe has a strong capacity of reading frame retrieval. Indeed, six trinucleotides (eight trinucleotides

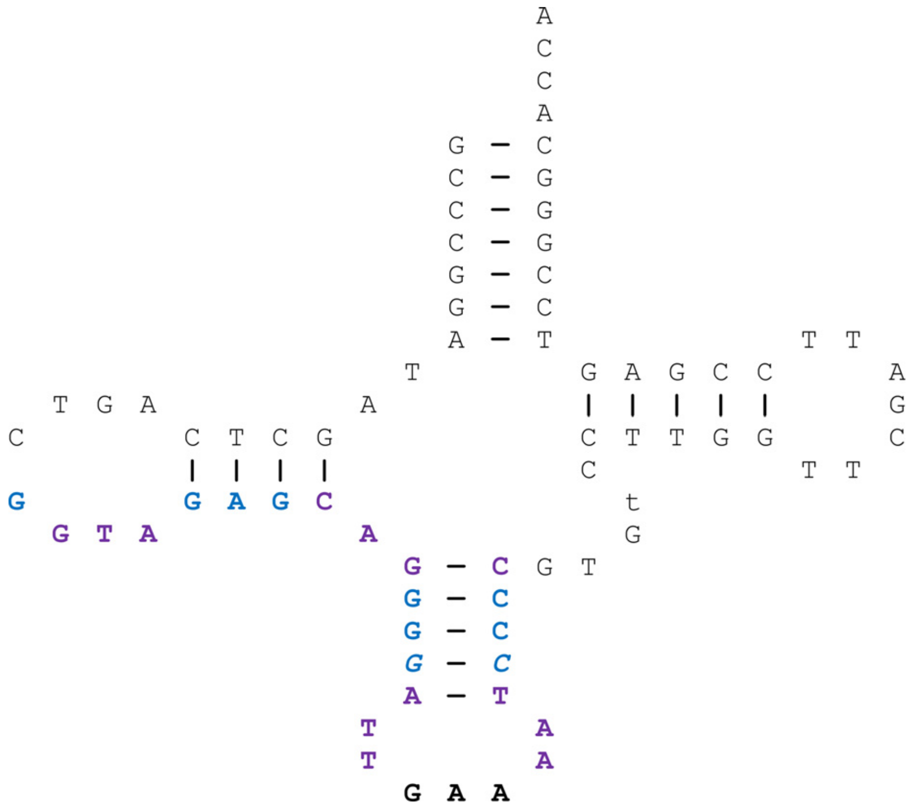

Fig. 5. The almost perfect *X* motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ = G, GTA, GAG, CAG, GGG, ATT, GAA, AAT, CCC, C (blue and blueviolet in bold with the anticodon GAA in black and bold) of 26 nucleotide length in the secondary structure of tRNA-Phe (76 nucleotides) of the crystal structure 3I8G (detail in Table 4a). This *X* motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ is located in the anticodon stem-loop, the D-stem and partially in the D-loop and is "in frame" with the anticodon.

Table 5a

Identification of seven X circular code motifs of length greater or equal to 15 nucleotides in 16S rRNA (1516 nucleotides) of the crystal structure 3I8G: $m_{16SrRNA\text{-}1}$(694, 713, 20) = G, AAC, GCC, GAT, GGC, GAA, GGC, A (red and pink in bold), $m_{16SrRNA\text{-}2}$(1189, 1206, 18) = T, TAC, GGC, CTG, GGC, GAC, AC (red and yellow in bold), $m_{16SrRNA\text{-}3}$(559, 574, 16) = GT, GTA, GGC, GGC, CTG, GG (red and orange in bold), $m_{16SrRNA\text{-}4}$(813, 827, 15) = T, CTG, GGT, CTC, CTG, GG (red and violet in bold), $m_{16SrRNA\text{-}5}$(1461, 1475, 15) = G, GGC, GAA, GTC, GTA, AC (red and fuchsia in bold), $m_{16SrRNA\text{-}6}$(397, 412, 16) = TG, GAG, GAA, GAA, GCC, CT (red and blue in bold) and $m_{16SrRNA\text{-}7}$(1347, 1361, 15) = GC, GGT, GAA, TAC, GTT, C (red and brown in bold).

```
1     TTGGAGAGTTTGATCCTGGCTCAGGGTGAACGCTGGCGGCGTGCCTAAGACATGCAAGTCGTGCGGGCCGCGGGGTTTTA
81    CTCCGTGGTCAGCGGCGGACGGGTGAGTAACGCGTGGGTGACCTACCCGGAAGAGGGGGACAACCCGGGGAAACTCGGGC
161   TAATCCCCCATGTGGACCCGCCCCTTGGGGTGTGTCCAAAGGGCTTTGCCCGCTTCCGGATGGGCCCGCGTCCCATCAGC
241   TAGTTGGTGGGGTAATGGCCCACCAAGGCGACGACGGGTAGCCGGTCTGAGAGGATGGCCGGCCACAGGGGCACTGAGAC
321   ACGGGCCCCACTCCTACGGGAGGCAGCAGTTAGGAATCTTCCGCAATGGGCGCAAGCCTGACGGAGCGACGCCGCTTGGA
401   GGAAGAAGCCCTTCGGGGTGTAAACTCCTGAACCCGGGACGAAACCCCCGACGAGGGGACTGACGGTACCGGGGTAATAG
481   CGCCGGCCAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGCGCGAGCGTTACCCGGATTCACTGGGCGTAAAGGGCGT
561   GTAGGCGGCCTGGGGCGTCCCATGTGAAAGACCACGGCTCAACCGTGGGGGAGCGTGGGATACGCTCAGGCTAGACGGTG
641   GGAGAGGGTGGTGGAATTCCCGGAGTAGCGGTGAAATGCGCAGATACCGGGAGGAACGCCGATGGCGAAGGCAGCCACCT
721   GGTCCACCCGTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAACCGGATTAGATACCCGGGTAGTCCACGCCCTAAACGAT
801   GCGCGCTAGGTCTCTGGGTCTCCTGGGGGCCGAAGCTAACGCGTTAAGCGCGCCGCCTGGGGAGTACGGCCGCAAGGCTG
881   AAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACCA
961   GGCCTTGACATGCTAGGGAACCCGGGTGAAAGCCTGGGGTGCCCCGCGAGGGGAGCCCTAGCACAGGTGCTGCATGGCCG
1041  TCGTCAGCTCGTGCCGTGAGGTGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCCGCCGTTAGTTGCCAGCGGTTCGGCC
1121  GGGCACTCTAACGGGACTGCCCGCGAAAGCGGGAGGAAGGAGGGGACGACGTCTGGTCAGCATGGCCCTTACGGCCTGGG
1201  CGACACACGTGCTACAATGCCCACTACAAAGCGATGCCACCCGGCAACGGGGAGCTAATCGCAAAAAGGTGGGCCCAGTT
1281  CGGATTGGGGTCTGCAACCCGACCCCATGAAGCCGGAATCGCTAGTAATCGCGGATCAGCCATGCCGCGGTGAATACGTT
1361  CCCGGGCCTTGTACACACCGCCCGTCACGCCATGGGAGCGGGCTCTACCCGAAGTCGCCGGGAGCCTACGGGCAGGCGCC
1441  GAGGGTAGGGCCCGTGACTGGGGCGAAGTCGTAACAAGGTAGCTGTACCGGAAGGTGCGGCTGGATCACCTCCTTT 1516
```

if Remark 12 is considered) of $m_{tRNA\text{-}Phe}$ retrieve the reading frame with only three or four nucleotides (Table 4b).

3.4.2. Identification of X circular code motifs in 16S rRNA

Seven X circular code motifs of length greater or equal to 15 nucleotides are identified in 16S rRNA of 3I8G (Table 5a). Table 5b gives the reading frame retrieval properties of their trinucleotides. The longest X motif $m_{16SrRNA\text{-}1}$(694, 713, 20) has 20 nucleotides and all its trinucleotides retrieve the reading frame with four nucleotides. The X motifs $m_{16SrRNA\text{-}2}$(1189, 1206, 18), $m_{16SrRNA\text{-}3}$(559, 574, 16) and $m_{16SrRNA\text{-}4}$(813, 827, 15) each have a trinucleotide belonging to $\tilde{X}$ (CTG, CTG and CTC, respectively) retrieving the reading frame with three nucleotides and a preceding "flexible" trinucleotide (GGC, GGC and GGT, respectively)

Table 5b

Nucleotide window length of the reading frame retrieval and its associated non-ambiguous word (from Table 1) of each trinucleotide of the seven X motifs $m_{16SrRNA\text{-}1}$(694, 713, 20), $m_{16SrRNA\text{-}2}$(1189, 1206, 18), $m_{16SrRNA\text{-}3}$(559, 574, 16), $m_{16SrRNA\text{-}4}$(813, 827, 15), $m_{16SrRNA\text{-}5}$(1461, 1475, 15), $m_{16SrRNA\text{-}6}$(397, 412, 16) and $m_{16SrRNA\text{-}7}$(1347, 1361, 15) (Table 5a and with the same color convention) identified in 16S rRNA of the crystal structure 3I8G.

Trinucleotides of the X motif $m_{16SrRNA\text{-}1}(694,713,20)$	AAC	GCC	GAT	GGC	GAA	GGC
Nucleotide window length of reading frame retrieval	4	4	4	4	4	4
Non-ambiguous word of the circular code X	AACG	GCCG	GATG	GGCG	GAAG	GGCA
Trinucleotides of the X motif $m_{16SrRNA\text{-}2}(1189,1206,18)$	TAC	GGC	CTG	GGC	GAC	
Nucleotide window length of reading frame retrieval	4	6	3	4	4	
Non-ambiguous word of the circular code X	TACG	GGCCTG	CTG $\in \tilde{X}$	GGCG	GACA	
Trinucleotides of the X motif $m_{16SrRNA\text{-}3}(559,574,16)$	GTA	GGC	GGC	CTG		
Nucleotide window length of reading frame retrieval	4	4	6	3		
Non-ambiguous word of the circular code X	GTAG	GGCG	GGCCTG	CTG $\in \tilde{X}$		
Trinucleotides of the X motif $m_{16SrRNA\text{-}4}(813,827,15)$	CTG	GGT	CTC	CTG		
Nucleotide window length of reading frame retrieval	3	6	3	3		
Non-ambiguous word of the circular code X	CTG $\in \tilde{X}$	GGTCTC	CTC $\in \tilde{X}$	CTG $\in \tilde{X}$		
Trinucleotides of the X motif $m_{16SrRNA\text{-}5}(1461,1475,15)$	GGC	GAA	GTC	GTA		
Nucleotide window length of reading frame retrieval	4	4	4	>5		
Non-ambiguous word of the circular code X	GGCG	GAAG	GTCG	GTAAC...		
Trinucleotides of the X motif $m_{16SrRNA\text{-}6}(397,412,16)$	GAG	GAA	GAA	GCC		
Nucleotide window length of reading frame retrieval	3	4	4	4		
Non-ambiguous word of the circular code X	GAG $\in \tilde{X}$	GAAG	GAAG	GCCC		
Trinucleotides of the X motif $m_{16SrRNA\text{-}7}(1347,1361,15)$	GGT	GAA	TAC	GTT		
Nucleotide window length of reading frame retrieval	4	7	4	>4		
Non-ambiguous word of the circular code X	GGTG	GAATACG	TACG	GTTC...		

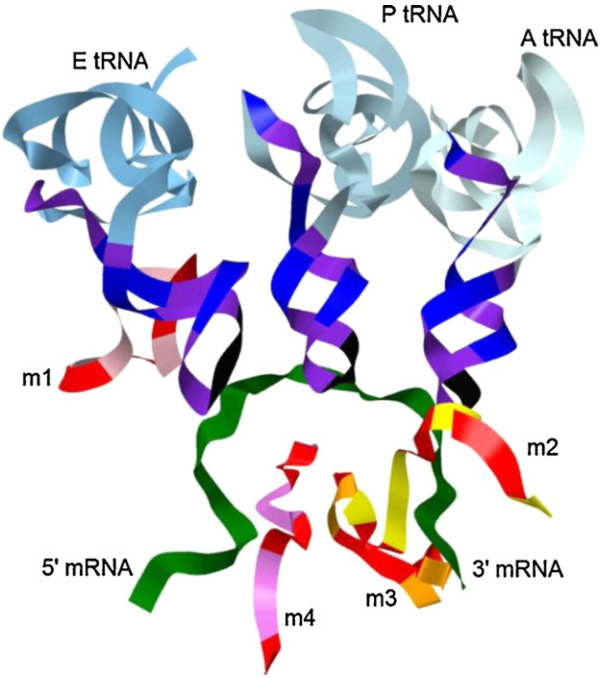

Fig. 6. Overview of the X circular code motifs in the messenger, transfer and 16S ribosomal RNAs of the crystal structure 3I8G. X circular code motifs in mRNA (green). X circular code motif in A-tRNA (lightcyan), P-tRNA (lightblue) and E-tRNA (lightskyblue): $m_{\text{tRNA-Phe}}(18, 43, 26)$ (blue and blueviolet with the anticodon in black) (Table 4a). X circular code motifs in 16S rRNA: $m_{\text{16SrRNA-1}}(694, 713, 20)$ (shortly m1 in red and pink), $m_{\text{16SrRNA-2}}(1189, 1206, 18)$ (shortly m2 in red and yellow), $m_{\text{16SrRNA-3}}(559, 574, 16)$ (shortly m3 in red and orange) and $m_{\text{16SrRNA-4}}(813, 827, 15)$ (shortly m4 in red and violet) (Table 5a). The remaining ribosomal RNA and proteins are cleared.

retrieving the reading frame with six nucleotides. The X motifs $m_{\text{16SrRNA-5}}(1461, 1475, 15)$ and $m_{\text{16SrRNA-6}}(397, 412, 16)$ each have three trinucleotides retrieving the reading frame with four nucleotides. The X motif $m_{\text{16SrRNA-7}}(1347, 1361, 15)$ has a very flexible trinucleotide GAA with a reading frame retrieval of seven nucleotides.

These seven X circular code motifs do not show particular relation in the primary structure of 16S rRNA. Four 16S rRNA X motifs, $m_{\text{16SrRNA-1}}$, $m_{\text{16SrRNA-2}}$, $m_{\text{16SrRNA-3}}$ and $m_{\text{16SrRNA-4}}$, have interesting spatial configurations with mRNA and/or tRNA in 3I8G (Section 3.5). Visualization of the three other 16S rRNA X motifs in 3I8G does not reveal spatial properties in regards to the translation concept developed here. However, two nucleotides of the X motif $m_{\text{16SrRNA-5}}$ are known to be very important in the decoding center (Section 3.6). The 16S rRNA X motifs $m_{\text{16SrRNA-6}}$ and $m_{\text{16SrRNA-7}}$ may have translation properties identified in other works.

3.5. Spatial visualization of the tRNA X circular code motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ and the four 16S rRNA X circular code motifs $m_{\text{16SrRNA-1}}(694, 713, 20)$, $m_{\text{16SrRNA-2}}(1189, 1206, 18)$, $m_{\text{16SrRNA-3}}(559, 574, 16)$ and $m_{\text{16SrRNA-4}}(813, 827, 15)$ in the crystal structure 3I8G

The crystal structure 3I8G recently obtained (Jenner et al., 2010) allows to visualize the 3D relations of the identified tRNA and 16S rRNA X motifs with the classical mRNA X motifs (Arquès and Michel, 1996). These identified X motifs allow, by complementary pairing (complementarity property of X) in various spatial configurations mRNA-tRNA-16SrRNA, to retrieve the reading frame (according to the properties described in Tables 4b and 5b) and the two shifted frames (C^3 property of X).

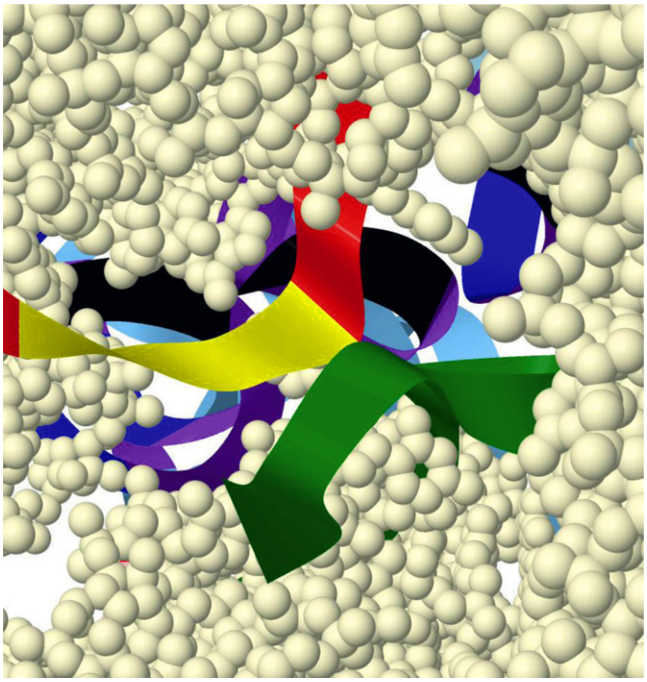

Fig. 7. Spatial relation of the mRNA X motifs (green), the A-tRNA (lightcyan) X motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ (blue and blueviolet with the anticodon in black) and the rRNA X motif $m_{\text{16SrRNA-2}}(1189, 1206, 18)$ (red and yellow). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs.

Fig. 6 gives a 3D overview of these mRNA, tRNA and 16S rRNA X motifs in the crystal structure 3I8G. There is a triple spatial relation between the mRNA X motifs, the A-tRNA X motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ and the 16S rRNA $m_{\text{16SrRNA-2}}(1189, 1206, 18)$ (Fig. 7). There is another triple spatial relation between the mRNA X motifs, the E-tRNA X motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ and the 16S rRNA X motif

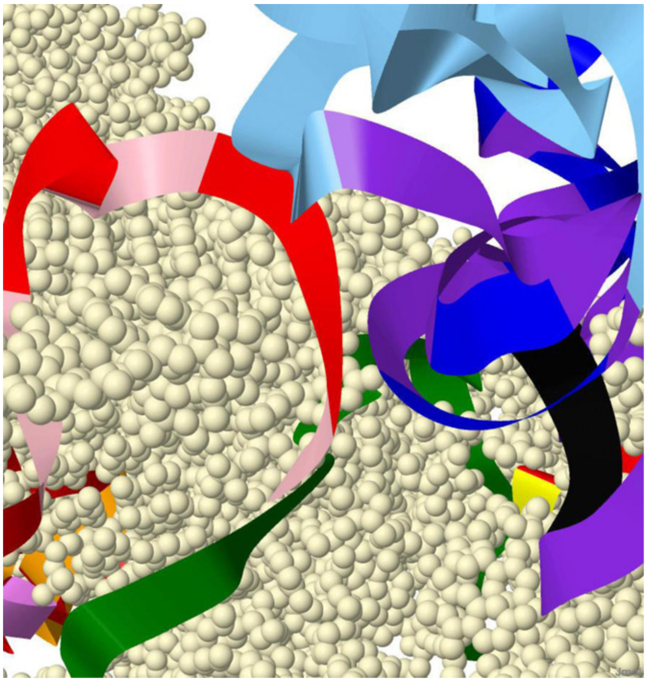

Fig. 8. Spatial relation of the mRNA X motifs (green), the E-tRNA (lightskyblue) X motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ (blue and blueviolet with the anticodon in black) and the rRNA X motif $m_{\text{16SrRNA-1}}(694, 713, 20)$ (red and pink). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs.

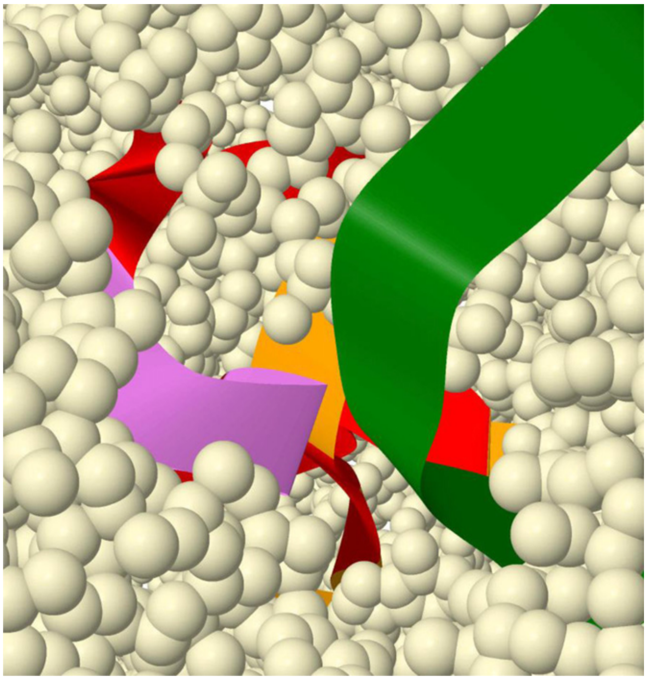

Fig. 9. Spatial relation of the mRNA X motifs (green) and the two rRNA X motifs $m_{16SrRNA-3}(559, 574, 16)$ (red and orange) and $m_{16SrRNA-4}(813, 827, 15)$ (red and violet). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs.

$m_{16SrRNA-1}(694, 713, 20)$ (Fig. 8). The mRNA X motifs also have a direct spatial relation, i.e. without tRNA X motifs, with two 16S rRNA X motifs, $m_{16SrRNA-3}(559, 574, 16)$ and $m_{16SrRNA-4}(813, 827, 15)$ (Fig. 9). All these X motifs may have chemical interactions as the remaining rRNA is outside their spatial neighborhood.

3.6. The 16S rRNA X circular code motif $m_{16SrRNA-5}(1461, 1475, 15)$

The X motif $m_{16SrRNA-5}(1461, 1475, 15)$ = G, GGC, GAA, GTC, GTA, AC belongs to the helix 44 of 16S rRNA. It has two successive adenines A1466 and A1467 which are well known to be involved in the decoding center (associated to the numbering A1492 and A1493 in *E. coli* 16S rRNA; review in Zaher and Green, 2009a,b). These two adenines recognize the codon–anticodon helix in A-tRNA and their base substitutions lead to a ribosome defective in AA-tRNA selection.

4. Discussion

The first part of this paper concerns a complete analysis of ambiguous words of the common circular code X of eukaryotes and prokaryotes. The type and the multiplicity of ambiguous words of X in the shifted frames 1 and 2 are identified. The circular code X has ambiguous words mainly in frame 2. This X property could be an explanation of the degeneracy of the genetic code at the 3rd codon site. Four trinucleotides of X, i.e. the subset $\tilde{X} = \{\text{CAG, CTC, CTG, GAG}\}$, are not ambiguous. Furthermore, $\tilde{X}$ is a C^3 self-complementary trinucleotide comma-free code. Maximal ambiguous words of X are also identified. They all begin with the nucleotide G. These results allow to define some probabilities of reading frame retrieval of X. The common trinucleotide circular code X retrieves the reading frame in genes with a probability of about 90% with a window length of 6 nucleotides, and a probability of 99.9% with a window length of 9 nucleotides (100% with a window length of 12 nucleotides, by definition of the circular code X). Thus, already two trinucleotides of X retrieve in average the reading frame with a good confidence.

In the second part, we identify X circular code motifs in transfer RNA and 16S ribosomal RNA. A tRNA X motif $m_{tRNA-Phe}(18, 43, 26)$ of 26 nucleotides is found in the anticodon stem-loop with trinucleotides "in frame" with the anticodon. Seven X motifs of length greater or equal to 15 nucleotides are identified in 16S rRNA. The window lengths of reading frame retrieval are also determined for each trinucleotide of these X motifs.

The crystal structure 3I8G allows the identified X motifs in the 3D ribosomal complex to be visualized. Four 16S rRNA X motifs have spatial relations with the mRNA X motifs and the tRNA X motif. Two spatial configurations mRNA-tRNA-16SrRNA are observed: mRNA X motifs, A-tRNA X motif $m_{tRNA-Phe}(18, 43, 26)$ and 16S rRNA X motif $m_{16SrRNA-2}(1189, 1206, 18)$, and mRNA X motifs, E-tRNA X motif $m_{tRNA-Phe}(18, 43, 26)$ and 16S rRNA X motif $m_{16SrRNA-1}(694, 713, 20)$. Another spatial configuration mRNA-16SrRNA-16SrRNA is also visualized: mRNA X motifs with the two 16S rRNA X motifs $m_{16SrRNA-3}(559, 574, 16)$ and $m_{16SrRNA-4}(813, 827, 15)$. Finally, another 16S rRNA X motif $m_{16SrRNA-5}(1461, 1475, 15)$ is involved in the decoding center by recognizing the codon–anticodon helix in A-tRNA. Thus, the path of mRNA X motifs in the ribosome may interact at several regions with tRNAs and rRNA X motifs.

From a code theory point of view, the identified X circular code motifs and their mathematical properties may constitute a translation code involved in retrieval, maintenance and synchronization of the reading frame in genes. Such a translation code based on X motifs can be confirmed or invalidated by experimental and crystallographic studies, e.g. by mutating some trinucleotides of X motifs into trinucleotides which do not belong to X, i.e. into trinucleotides of X_1, X_2 or $T_{id} = \{\text{AAA, CCC, GGG, TTT}\}$. Furthermore, if such a translation code based on series of trinucleotides is confirmed, the dynamic of its spatial molecular process, i.e. as a function of time, also remains to be discovered. Finally, from an evolutionary aspect, such a translation code sets the problem of the origin of the genetic code, i.e. the code for amino acids. In particular, are the genetic code and its degeneracy property a consequence of a translation code?

Acknowledgments

I thank Denise Besch, Svetlana Gorchkova, Elisabeth Michel and Jean-Marc Vassards for their support.

References

Ahmed, A., Michel, C.J., 2008. Plant microRNA detection using the circular code information. Comput. Biol. Chem. 32, 400–405.

Ahmed, A., Frey, G., Michel, C.J., 2010. Essential molecular functions associated with the circular code evolution. J. Theor. Biol. 264, 613–622.

Ahmed, A., Michel, C.J., 2011. Circular code signal in frameshift genes. J. Comput. Sci. Syst. Biol. 4, 7–15.

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

Brodersen, D.E., Clemons Jr., W.M., Carter, A.P., Wimberly, B.T., Ramakrishnan, V., 2002. Crystal structure of the 30 S ribosomal subunit from *Thermus thermophilus*: structure of the proteins and their interactions with 16 S RNA. J. Mol. Biol. 316, 725–768.

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

Devaraj, A., Shoji, S., Holbrook, E.D., Fredrick, K., 2009. A role for the 30S subunit E site in maintenance of the translational reading frame. RNA 15, 255–265.

Gonzalez, D.L., Giannerini, S., Rosa, R., 2011. Circular codes revisited: a statistical approach. J. Theor. Biol. 275, 21–28.

Gorini, L., Kataja, E., 1964. Phenotypic repair by streptomycin of defective genotypes in *E. coli*. Proc. Natl. Acad. Sci. U.S.A. 51, 487–493.

Gustilo, E.M., Vendeix, F.A., Agris, P.F., 2008. tRNA's modifications bring order to gene expression. Curr. Opin. Microbiol. 11, 134–140.

Jenner, L.B., Demeshkina, N., Yusupova, G., Yusupov, M., 2010. Structural aspects of messenger RNA reading frame maintenance by the ribosome. Nat. Struct. Mol. Biol. 17, 555–560.

Jorgensen, F., Kurland, C.G., 1990. Processivity errors of gene expression in *Escherichia coli*. J. Mol. Biol. 215, 511–521.

Kurland, C.G., Hughes, D., Ehrenberg, M., 1996. Limitation of translation accuracy in *Escherichia coli* and *Salmonella*. In: Neidhardt, F.C. (Ed.), Cellular and Molecular Biology. American Society for Microbiology, Washington, DC, pp. 979–1004.

Lipsett, M.N., Heppel, L.A., Bradley, D.F., 1960. Complex formation between adenine oligonucleotides and polyuridylic acid. Biochim. Biophys. Acta 41, 175–177.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids Res. 25, 955–964.

Márquez, V., Wilson, D.N., Tate, W.P., Triana-Alonso, F., Nierhaus, K.H., 2004. Maintaining the ribosomal reading frame. Cell 118, 45–55.

McLaughlin, C.S., Dondon, J., Grunberg-Manago, M., Michelson, A.M., Saunders, G., 1966. Stability of the messenger RNA-sRNA-ribosome complex. Cold Spring Harb. Symp. Quant. Biol. 31, 601.

Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.

Moazed, D., Noller, H.F., 1990. Binding of tRNA to the ribosomal A and P sites protects two distinct sets of nucleotides in 16 S rRNA. J. Mol. Biol. 211, 135–145.

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

Ogle, J.M., Brodersen, D.E., Clemons Jr., W.M., Tarry, M.J., Carter, A.P., Ramakrishnan, V., 2001. Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science 292, 897–902.

Rozenski, J., Crain, P.F., McCloskey, J.A., 1999. The RNA modification database: 1999 update. Nucleic Acids Res. 27, 196–197.

Sundararajan, A., Michaud, W.A., Qian, Q., Stahl, G., Farabaugh, P.J., 1999. Near-cognate peptidyl-tRNAs promote +1 programmed translational frameshifting in yeast. Mol. Cell 4, 1005–1015.

Zaher, H.S., Green, R., 2009a. Quality control by the ribosome following peptide bond formation. Nature 457, 161–166.

Zaher, H.S., Green, R., 2009b. Fidelity at the molecular level: lessons from protein synthesis. Fidelity at the molecular level: lessons from protein synthesis. Cell 136, 746–762.