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 $i = 1, \ldots, n$.