Computational Biology and Chemistry 59 (2015) 158–176

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Circular code motifs near the ribosome decoding center

Karim El Soufi, Christian J. Michel*

Theoretical Bioinformatics, ICube, University of Strasbourg, CNRS, 300 Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Article history:
Received 12 June 2015
Accepted 13 July 2015
Available online 14 September 2015

Keywords:
Circular code motif
Ribosomal RNA
Transfer RNA
Translation code

ABSTRACT

A maximal C^3 self-complementary trinucleotide circular code X is identified in genes of bacteria, eukaryotes, plasmids and viruses (Michel, 2015; Arquès and Michel, 1996). A translation (framing) code based on the circular code was proposed in Michel (2012) with the identification of several X circular code motifs (X motifs shortly) in both ribosomal RNAs (rRNAs) and their decoding center, and transfer RNAs (tRNAs). We extended these results in two ways. First, three universal X motifs were determined in the ribosome decoding center: the X motif m_{AA} containing the conserved nucleotides A1492 and A1493, the X motif m_G containing the conserved nucleotide G530 and the X motif m with unknown biological function (El Soufi and Michel, 2014). Secondly, statistical analysis of X motifs of greatest lengths performed on different and large tRNA populations according to taxonomy, tRNA length and tRNA score showed that these X motifs have occurrence probabilities in the 5′ and/or 3′ regions of 16 isoaccepting tRNAs of prokaryotes and eukaryotes greater than the random case (Michel, 2013). We continue here the previous works with the identification of X motifs in rRNAs of prokaryotes and eukaryotes near the ribosome decoding center. Seven X motifs *PrRNAXm* conserved in 16S rRNAs of prokaryotes P and four X motifs *ErRNAXm* conserved in 18S rRNAs of eukaryotes E are identified near the ribosome decoding center. Furthermore, four very large X motifs of length greater than or equal to 20 nucleotides, 14 large X motifs of length between 16 and 19 nucleotides and several X motifs of length greater or equal to 9 nucleotides are found in tRNAs of prokaryotes. Some properties of these X motifs in tRNAs are described. These new results strengthen the concept of a translation code based on the circular code (Michel, 2012).

1. Introduction

The ribosome is a complex ribonucleoprotein particle responsible for the synthesis of the cell protein by translating messenger RNA (mRNA). Ribosomes are composed of two subunits, a large subunit and a small subunit. Each subunit is formed by ribosomal RNAs (rRNAs) and proteins. A ribosome contains three transfer RNA (tRNA) binding sites: A-site (aminoacyl), P-site (peptidyl) and E-site (exit). During translation, the aminoacyl tRNA binds to the A-site where the decoding center containing the universally conserved dinucleotide AA (A1492 and A1493) distinguishes cognate from non-cognate tRNAs by anticodon-codon interactions (Wilson, 2014). The transfer of the amino acid from the P-site to the A-site results in the peptide-bond forming between the carboxyl group at the P-site and the newly arrived amino acid at the A-site. As the ribosome progresses by three nucleotides, the peptidyl tRNA moves from the A-site to the P-site. Finally the unloaded tRNA moves from the P-site to the E-site.

In 1996, a statistical analysis of occurrence frequencies of the 64 trinucleotides {AAA, . . . , TTT} in the three frames 0, 1 and 2 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 frame 0 is the reading frame in a gene 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 periodic permuted trinucleotides {AAA, CCC, GGG, TTT} and by assigning each trinucleotide to a preferential frame (frame of its highest occurrence frequency), three subsets $X = X_0$, X_1 and X_2 of 20 trinucleotides are found in the frames 0, 1 and 2, respectively, simultaneously of two large gene populations (protein coding regions): eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686 sequences, 4709,758 trinucleotides) (Arquès and Michel, 1996). This set X contains the 20 following trinucleotides:

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

The two sets X_1 and X_2, of 20 trinucleotides each, in the shifted frames 1 and 2 of genes can be deduced from X by the circular permutation map (see below). These three trinucleotide sets

* Corresponding author.
E-mail addresses: kelsoufi@unistra.fr (K. El Soufi), c.michel@unistra.fr (C.J. Michel).

http://dx.doi.org/10.1016/j.compbiolchem.2015.07.015
1476-9271/

present several strong mathematical properties, particularly the fact that X is a maximal C^3 self-complementary trinucleotide circular code (Arquès and Michel, 1996). A trinucleotide circular code has the fundamental property to always retrieve the reading frame in any position of any sequence generated with the circular code. In particular, initiation and stop trinucleotides as well as any frame signals are not necessary to define the reading frame. Indeed, a window of 12 nucleotide length positioned anywhere in a sequence generated with the circular code X always retrieves the reading frame (Tables 2 and 3 in Michel, 2012).

A translation (framing) code based on the circular code was proposed in Michel (2012) with the identification of X circular code motifs (X motifs shortly) in the bacterial ribosomal RNA of *Thermus thermophilus*, in particular in the ribosome decoding center which recognizes the codon-anticodon helix in A-tRNA, and its tRNA of phenylalanine. A 3D visualization of X motifs in the ribosome shows several spatial configurations involving X motifs of mRNA, tRNA and rRNA. These results were extended in two ways. Firstly, three universal X motifs were identified in the ribosome decoding center of all the studied rRNAs from different kingdoms: bacteria *Escherichia coli* and *T. thermophilus*, archaea *Pyrococcus furiosus*, nuclear eukaryotes *Saccharomyces cerevisiae*, *Triticum aestivum* and *Homo sapiens*, and chloroplast *Spinacia oleracea*. These three X motifs are m_{AA} containing the conserved nucleotides A1492 and A1493, m_G containing the conserved nucleotide G530 and m whose biological function is unknown (El Soufi and Michel, 2014). Secondly, a statistical analysis of X motifs of greatest lengths performed on different and large tRNA populations according to the taxonomy, tRNA length and tRNA score showed that these X motifs have occurrence probabilities in the 5′ and/or 3′ regions of 16 isoaccepting tRNAs of prokaryotes and eukaryotes greater than the random case (Michel, 2013). By developing a search algorithm of X motifs in a DNA global multiple sequence alignment, we extend here the previous works to the identification of X motifs in rRNAs of prokaryotes and eukaryotes near the ribosome decoding center. Furthermore, in contrast to the statistical analysis of the distribution of X motifs of greatest lengths (Michel, 2013), a detailed analysis of X motifs is performed in the 20 isoaccepting tRNAs of bacteria. Several properties of X motifs are described according to (i) their type; (ii) their length: very large with a length greater than 20 nucleotides, large with a length between 16 and 19 nucleotides, otherwise with a length between 9 and 15 nucleotides (X motifs of lengths equal to 9 nucleotides retrieve the reading frame with a probability of 99.9%, Table 3 and Fig. 4 in Michel, 2012); (iii) their position in the tRNAs: 5′ regions, anticodon regions, 3′ regions; and (iv) their relations: X motifs shifted in frame by +1 or +2 nucleotides from other X motifs or the anticodons of tRNAs.

2. Method

2.1. Recall

The definitions of code, trinucleotide code, trinucleotide circular code, self-complementary trinucleotide circular code, C^3 trinucleotide circular code and C^3 self-complementary trinucleotide circular code related to the X motifs, i.e. motifs from the circular code X (Eq. (1)), are given in Michel (2012, 2013) and El Soufi and Michel (2014).

The trinucleotide set X (Eq. (1)) coding the reading frames in eukaryotic and prokaryotic genes is a maximal (20 trinucleotides) C^3 self-complementary trinucleotide circular code with a window length equal to 12 nucleotides for retrieving the reading frame. The fundamental property of a circular code is the ability to retrieve the reading (original or constructed) frame of any sequence generated with this circular code. A circular code is a set of words over an alphabet such that any sequence written on a circle (the next letter after the last letter of the sequence being the first letter) has a unique decomposition (factorization) into words of the circular code (Fig. 1 in Michel, 2012; for a graphical representation of the circular code definition and Fig. 2 in Michel, 2012; for an example). The reading frame in a sequence (a gene) is retrieved after the reading of a certain number of letters (nucleotides), called the window of the circular code. The length of this window for retrieving the reading frame is the letter length of the longest ambiguous words which can be read in at least two frames, plus one letter (Fig. 3 in Michel, 2012; for an example). For the circular code X, this window needs a length of 12 nucleotides as the two longest ambiguous words GGTAATTACCA and GGTAATTACCT of X have 11 nucleotides (Tables 2 and 3 in Michel, 2012).

In this paper, we study X circular code motifs (X motifs shortly) near the ribosome decoding center. It is important to remind the reader that there are two concepts: (i) the circular code X, which is a set of 20 trinucleotides (Eq. (1)); and (ii) X motifs, which are motifs (words) obtained with the circular code X. We give here a few examples of X motifs: AAC,AAT (a concatenation of the 1st and 2nd trinucleotides of X, the commas showing the adopted decomposition), TTC,TAC,AAC (a concatenation of the 20th, 19th and 1st trinucleotides of X), AG,AAC,AAT (a concatenation of the suffix AG of the 6th or 11th trinucleotides of X, and the 1st and 2nd trinucleotides of X), AG,AAC,AAT,AC (a concatenation of the suffix of the 6th or 11th trinucleotides of X, the 1st and 2nd trinucleotides of X, and the prefix AC of the 3rd trinucleotide of X), etc. The motifs, for example, AAC,AAT,AAG, CA,AAC,AAT and AAC,AAT,AG are not X motifs.

2.2. Search algorithm of X motifs in a DNA global multiple sequence alignment

We present here a search algorithm of X motifs of lengths greater than a given number of nucleotides in a DNA global multiple sequence alignment (global MSA with the program ClustalW2). It will identify common X motifs in multiple aligned RNA sequences. The algorithm is presented with DNA sequences, i.e. on the 4-letter alphabet $A_4 = \{A, C, G, T\}$, its extension on RNA sequences, i.e. on the 4-letter alphabet {A, C, G, U}, being trivial.

Let a trinucleotide t of the circular code X defined in Eq. (1) be the three letters $t = l_1 l_2 l_3 \in A_4^3 = \{AAA, \ldots, TTT\}$. Let $\mathrm{Pref}_{let}(X)$ be the set containing the letters $l_1 \in A_4$ of X and $\mathrm{Pref}_{dilet}(X)$ be the set containing the diletters $l_1 l_2 \in A_4^2 = \{AA, \ldots, TT\}$ of X. Then, by inspection of X, we have:

$$\mathrm{Pref}_{let}(X) = \{A, C, G, T\} = A_4, \tag{2}$$

$$\mathrm{Pref}_{dilet}(X) = \{AA, AC, AT, CA, CT, GA, GC, GG, GT, TA, TT\}. \tag{3}$$

Remark 1. $\mathrm{Card}(\mathrm{Pref}_{let}(X)) = 4$ and $\mathrm{Card}(\mathrm{Pref}_{dilet}(X)) = 11$ (among 16 dinucleotides).

The algorithm uses the following classical notions of language theory. Let x be a word (sequence) on A_4 of length $|x|$. $x[i]$ denotes the letter at index i of x and $x[i..j]$ denotes the factor of x defined by $x[i]x[i+1]..x[j]$ of length $j - i + 1$.

The function Xmotif searches an X motif at a given position startX (input parameter in integer) in a DNA sequence seq of length |seq| on A_4 or $A_4 \bigcup \{-\}$ (an aligned sequence with gaps) and returns its end position endX (output parameter in integer).

The function Search_Xmotif_seq searches all the X motifs in a DNA sequence seq (input parameter in string) of length |seq| on A_4 or $A_4 \bigcup \{-\}$ which are greater or equal to a minimum number of nucleotides in the X motif, named lgMinX (input parameter in integer), and returns a list listXMotif (output parameter) of X