Computational Biology and Chemistry 45 (2013) 17–29

Contents lists available at SciVerse ScienceDirect

Computational Biology and Chemistry

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

Circular code motifs in transfer RNAs

Christian J. Michel*

Equipe de Bioinformatique Théorique, BFO, ICube, Université de Strasbourg, CNRS, 300 Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Article history:
Received 14 January 2013
Accepted 28 February 2013

Keywords:
Circular code motif
Search algorithm
Statistical distribution
Transfer RNA
5′ and 3′ regions of transfer RNA

ABSTRACT

In 1996, a trinucleotide circular code X is identified in genes of prokaryotes and eukaryotes (Arquès and Michel, 1996). In 2012, X motifs are identified in the transfer RNA (tRNA) Phe and 16S ribosomal RNA (Michel, 2012). A statistical analysis of X motifs in all available tRNAs of prokaryotes and eukaryotes in the genomic tRNA database (September 2012, http://lowelab.ucsc.edu/GtRNAdb/, Lowe and Eddy, 1997) is carried out here. For this purpose, a search algorithm of X motifs in a DNA sequence is developed. Two definitions allow to determine the occurrence probabilities of X motifs and the circular codes X, $X_1 = \mathcal{P}(X)$ and $X_2 = \mathcal{P}^2(X)$ ($\mathcal{P}$ being a circular permutation map applied on X) in a population of tRNAs. This approach identifies X motifs in the 5′ and/or 3′ regions of 16 isoaccepting tRNAs (except for the tRNAs Arg, His, Ser and Trp). The statistical analyses are performed on different and large tRNA populations according to the taxonomy (prokaryotes and eukaryotes), tRNA length and tRNA score. Finally, a circular code property observed in genes of prokaryotes and eukaryotes is identified in the 3′ regions of 19 isoaccepting tRNAs of prokaryotes and eukaryotes (except for the tRNA Leu). The identification of X motifs and a gene circular code property in tRNAs strengthens the concept proposed in Michel (2012) of a possible translation (framing) code based on a circular code.

1. Introduction

The structure and the function of transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), molecules which are involved for translating the genetic information into proteins, are detailed, e.g., in a review of Zaher and Green (2009). 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 messenger RNA (mRNA). It must decode only successive codons and not trinucleotides in shifted frames. However, an mRNA lacks punctuation (or comma) which could be used by the transfer and ribosomal RNAs to identify the trinucleotides in reading frame. There are multiple genetic codes, moreover, multiple translational codes, while one of them is the well-known triplet genetic code, and another one could be a circular code.

The maintenance of the correct reading frame in 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 are framing codes, i.e. with the property of reading frame retrieval, synchronization and maintenance. Furthermore, there are trinucleotide 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 trinucleotide 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 $\mathcal{C}(Y_2) = Y_1$.

However, the search for a framing code was a very difficult task. 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). However, no trinucleotide comma-free code was identified in genes, theoretically or statistically. The main reason that I propose to explain the successive failures for many years, is related to the fact that almost all developed statistical analyses during several years to understand the genetic code, were based on the reading frame (frame 0). For example, the codon usage, a classical method based on the frequencies of codons (i.e. trinucleotides in frame 0), which is widely used by biostatisticians, bioinformaticians and biologists, can (only) explain the variations of the genetic code, e.g. the differences between the codons coding the same amino acid, the codon differences between species, etc. But it cannot explain the "origin" of the genetic code. The genetic code is associated to the reading frame "space". To understand the origin of the genetic code, a more general space should be considered, e.g. the three frames of genes. The difficulty of methods for identifying such codes is not concerned, only the choice of the space. 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.

* Tel.: +33 368854462.
E-mail address: c.michel@unistra.fr

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.02.004

In 1996, a statistical analysis of occurrence frequencies of the 64 trinucleotides $A_4^3 = \{AAA, \ldots, TTT\}$ 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. By excluding the four trinucleotides with identical nucleotides {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, simultaneously of two large gene populations (protein coding regions): eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686 sequences, 4,709,758 trinucleotides) (Arquès and Michel, 1996). The set X contains the 20 following trinucleotides

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

The two sets X_1 and X_2, of 20 trinucleotides each, in the frames 1 and 2 of genes can be deduced from X by circular permutation (see below). These three trinucleotide sets present several strong mathematical properties, particularly the fact that they are trinucleotide circular codes (Arquès and Michel, 1996). We have also proved that a comma-free code, a code proposed by Crick et al. (1957), is a particular circular code. Precisely, a hierarchy of circular codes is closed by the strongest ones which are comma-free and the weakest ones which have a large "necklace" (Michel et al., 2008).

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). Thus, the common trinucleotide circular code $X(1)$ of 20 trinucleotides allows the reading frames in genes to be retrieved locally, i.e. anywhere in genes and in particular without start codon.

We briefly recall a few properties of the common trinucleotide circular code $X(1)$ which are involved in this paper.

Notation 1. The letters (or nucleotides or bases) define the genetic alphabet $A_4 = \{A, C, G, T\}$. The set of non-empty words (words resp.) over A_4 is denoted by A_4^+ (A_4^* resp.). The set of the 16 words of length 2 (dinucleotides or diletters) on A_4 is denoted by $A_4^2 = \{AA, \ldots, TT\}$. The set of the 64 words of length 3 (trinucleotides or triletters) on A_4 is denoted by $A_4^3 = \{AAA, \ldots, TTT\}$. Let $x_1 \cdots x_n$ be the concatenation of the words x_i for $i = 1, \ldots, n$. We use here the definitions in coding theory.

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 $x_1 \cdots x_n = y_1 \cdots y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \ldots, n$.

Remark 1. The set A_4^3 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 $sx_2 \cdots x_n r = y_1 \cdots y_m$ and $x_1 = rs$ imply $n = m$, $r = \varepsilon$ (empty word) and $x_i = y_i$ for $i = 1, \ldots, n$.

Remark 2. A set containing a trinucleotide with identical nucleotides, e.g. AAA, cannot be a circular code (details, e.g., in Michel, 2008). Thus, the set A_4^3 is not a trinucleotide circular code.

Remark 3. A set containing two trinucleotides related to circular permutation $\mathcal{P}$, e.g. AAC and $\mathcal{P}(AAC) = ACA$, cannot be a circular code (details, e.g., in Michel, 2008). Thus, the set $A_4^3 \setminus \{AAA, CCC, GGG, TTT\}$ 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.

Example 1. Let the set Y be composed of the six following words: $Y = \{AAT, ATG, CCT, CTA, GCC, GGC\}$ and the word w, be a series of the nine following letters: $w = ATGGCCCTA$. 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. Therefore, Y is not a circular code. In contrast, if the set Z obtained by replacing the word GGC of Y by GTC is considered, i.e. $Z = \{AAT, ATG, CCT, CTA, GCC, GTC\}$, then there never exists an ambiguous word with Z, such as w for Y, and then Z is a circular code (Michel, 2012, Fig. 2 for details).

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 (shortly 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 classical window length defined for biinfinite words of a circular code Y is described in Michel (2012, Fig. 3 for details).

In order to analyze 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 have introduced 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 (Michel, 2012).

Property 1. *The window length l to retrieve the construction frame for any right infinite word of the common circular code X is the letter length of the longest ambiguous word plus one letter and is equal to $l = 12$ nucleotides. In other words, there is no ambiguous word of the common circular code X when $l \geq 12$ nucleotides.*

Definition 4. Complementarity map $\mathcal{C}$: The complementarity map $\mathcal{C} : A_4^+ \to A_4^+$ is defined by $\mathcal{C}(A) = T, \mathcal{C}(C) = G, \mathcal{C}(G) = C, \mathcal{C}(T) = A$ 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. $\mathcal{C}(AAC) = GTT$. This map on words is naturally extended to word sets: a complementary trinucleotide set is obtained by applying the complementarity map $\mathcal{C}$ to all its trinucleotides.

Definition 5. Circular permutation map $\mathcal{P}$: The circular permutation map $\mathcal{P} : A_4^3 \to A_4^3$ permutes circularly each trinucleotide $l_0 l_1 l_2$ as follows $\mathcal{P}(w_0) = l_1 l_2 l_0$, e.g. $\mathcal{P}(AAC) = ACA$. The k th iterate of $\mathcal{P}$ is denoted $\mathcal{P}^k$, e.g. $\mathcal{P}^2(AAC) = CAA$. 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 k th 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. Permuted trinucleotide set: A trinucleotide set $Y_1 = \mathcal{P}(Y)$ of a trinucleotide circular code Y is permuted if, for each $y \in Y$, $\mathcal{P}(y) \in \mathcal{P}(Y)$. The permuted trinucleotide set $Y_2 = \mathcal{P}^2(Y)$ is defined similarly.

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 (1) coding the reading frames (frames 0) in eukaryotic and prokaryotic genes is a C^3 self-complementary trinucleotide circular code. The circular code $X_1 = \mathcal{P}(X)$ contains the 20 following trinucleotides.

$$X_1 = \{\text{AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG, GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG}\} \quad (2)$$

and the circular code $X_2 = \mathcal{P}^2(X)$, the 20 following trinucleotides

$$X_2 = \{\text{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT, CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT}\}. \quad (3)$$

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

Result 2. (Gonzalez et al., 2011). A new definition of a statistical function analysing the covering capability of a circular code has recently showed on a 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 (Arquès and Michel, 1996; Michel et al., 2008). Furthermore, permutation tests of bases in the codon sites of X also suggest a reading frame synchronization property of X.

Result 3. (Michel, 2012). In 2012, X motifs (motifs of the common trinucleotide circular code X) are identified in transfer and 16S ribosomal RNAs: an almost perfect tRNA–Phe X motif of 26 nucleotides including the anticodon stem-loop and seven 16S rRNA X motifs of length greater or equal to 15 nucleotides. A 3D visualization of X motifs in the ribosome (crystal structure 3I8G, Jenner et al., 2010) 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 motifs and their mathematical properties may constitute a framing code. The aim of this paper is to identify X motifs in transfer RNAs in order to strengthen the concept of a translation code based on a circular code, a concept recently proposed in Michel (2012).

The Section 2 presents a search algorithm of X motifs in a DNA sequence. Such an algorithm was never proposed since 1996. Two definitions allow to determine the occurrence probabilities of X motifs and the circular codes X, X_1 and X_2 in a population of tRNAs. The tRNA data are described and the populations of the 20 isoaccepting tRNAs based on their taxonomy (prokaryotes, eukaryotes), length and score which are analyzed statistically, are characterized. The Section 3 presents the distribution of X motifs in random tRNAs. Then, X motifs are identified in the 5′ and 3′ regions of tRNAs. These results are confirmed by several statistical analyses of different and large tRNA populations. Finally, a circular code property observed in genes of prokaryotes and eukaryotes is identified in the 3′ regions of 19 isoaccepting tRNAs of prokaryotes and eukaryotes.

2. Methods

2.1. Search algorithm of X motifs in a DNA sequence

We propose here a search algorithm of X motifs in a DNA sequence. Even if its principle is simple, such an algorithm was never proposed since 1996. It will allow bioinformaticians to implement it in various biological contexts: search of X motifs in DNA databases, sequence alignment based on X motifs, etc. We will apply it here to the analysis of the statistical distribution of X motifs in the 20 isoaccepting tRNAs of prokaryotes and eukaryotes.

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

$$\text{Pref}_{let}(X) = \{\text{A, C, G, T}\}, \quad (4)$$

$$\text{Pref}_{dilet}(X) = \{\text{AA, AC, AT, CA, CT, GA, GC, GG, GT, TA, TT}\}. \quad (5)$$

Remark 5. Card (Pref_{let}(X)) = 4 and Card (Pref_{dilet}(X)) = 11 (among 16 dinucleotides).

The description of 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 $x[i]x[i+1]\cdots x[j]$ of length $j-i+1$.

The Algorithm_Search_X_motif proposed here searches in a DNA sequence the X motif of greatest length having at least a given number of nucleotides. Its input parameters are the DNA sequence, named seq, and the minimum number of nucleotides of the X motif, named lgMinX. Its output parameter is either a list containing the X motif, its length and its start and end positions in the DNA sequence seq, or an empty list. It uses the function Xmotif which searches a X motif at a given position startX in seq (input parameter).

```
Xmotif[startX]
1.  endX = startX // endX: end position of Xmotif in seq
2.  iX = 1 // index on X
3.  testX = true
4.  while testX
5.     if iX = 1[3] then // Case 1 modulo 3: Pref_let
6.        if endX ≤ |seq| and {seq[endX] ∩ Pref_let} ≠ {} then
7.           iX++
8.           endX++
9.        else testX = false
10.    if iX = 2[3] then // Case 2 modulo 3: Pref_dilet
11.       if endX ≤ |seq| and {seq[endX-1..endX] ∩ Pref_dilet } ≠ {} then
12.          iX++
13.          endX++
14.       else testX = false
15.    if iX = 0[3] then // Case 0 modulo 3: X
16.       if endX ≤ |seq| and {seq[endX-2..endX] ∩ X} ≠ {} then
17.          iX++
18.          endX++
19.       else testX = false
20. return endX--
```

```
Algorithm_Search_Xmotif[seq,lgMinX]
1. for start ← 1 to |seq| step +1 do
2.    end ← Xmotif[start] // start: start position of Xmotif in seq
3.    lg ← end-start+1
4.    if lg ≥ lgMinX then listXMotif ← {seq[start..end],lg,start,end}
5. listSortXMotif ← sort[listXMotif,#1[[2]] > #2[[2]] &]
6. if listSortXMotif ≠ {} then return take[listSortXMotif,1]
7. else return {}
```

Instruction 4 of Algorithm_Search_Xmotif stores all the *X* motifs in listXMotif (a list) of lengths greater or equal to the constant lgMinX. Instruction 5 sorts the *X* motifs in listXMotif by descending order of their lengths (the length is the 2nd element in listXMotif) and ranges them in listSortXMotif (a different list in order to perform eventually additional treatment in listXMotif). Instruction 6 returns the *X* motif of greatest length having at least lgMinX nucleotides in the DNA sequence seq if listSortXMotif is not empty.

The *X* motif returned by the Algorithm_Search_Xmotif beginning at position *b* and ending at position *e* in a sequence *s* is defined for Section 2.2 by the set *X*[*b*,*e*,*s*] containing the following nucleotides

$$X[b, e, s] = \{s[b], s[b+1], \ldots, s[e]\}. \quad (6)$$

2.2. Definition of an occurrence probability of X motifs in a tRNA population

Let $\mathcal{F}$ be a tRNA population with $N(\mathcal{F})$ sequences *s*. Let $n_{i(s)}$ be the nucleotide $n \in A_4$ in a position *i* of a tRNA *s* of $\mathcal{F}$. By convention, the position $i = 0$ in a tRNA refers to the anticodon (symbolized by the three nucleotides $n_0n_0n_0$, Fig. 1a). All tRNAs *s* of $\mathcal{F}$ are centered according to their anticodon position $i = 0$. The statistical analysis of *X* motifs will exclude the anticodons whose nucleotides depend on the isoaccepting tRNAs. By convention, all nucleotides before the anticodon (5′ region) have a negative position $i < 0$ and all nucleotides after the anticodon (3′ region) have a positive position $i > 0$. It should be stressed that both regions are analyzed in the same direction, i.e. the 5′–3′ direction. As the tRNAs have different lengths, the occurrence probability of *X* motifs is defined for a nucleotide range [min(*s*), max(*s*)] varying for each tRNA *s* of $\mathcal{F}$, min(*s*) being the minimum position (negative value) and max(*s*), the maximum position. Thus, for each tRNA *s*, two regions are studied: the 5′ region in the nucleotide range [min(*s*), −1], i.e. max(*s*) has the particular value −1, and the 3′ region in the nucleotide range [1, max(*s*)], i.e. min(*s*) has the particular value 1. As some positions *i* may not exist with some nucleotide ranges, let $N(i, \mathcal{F})$ be the number of tRNAs *s* of $\mathcal{F}$ having a position *i*. Obviously, at the anticodon position $i = 0$, the number of sequences is maximal and $N(0, \mathcal{F}) = N(\mathcal{F})$.

The function $\delta(i, s)$ indicates whether or not the nucleotide $n_{i(s)}$ in a position *i* of a tRNA *s* belongs or not to the *X* motif $X[b, e, s] = \{s[b], s[b+1], \ldots, s[e]\}$ obtained by the Algorithm_Search_Xmotif

$$\delta(i, s) = \begin{cases} 1 & \text{if } n_{i(s)} \in X[b, e, s] \\ 0 & \text{otherwise} \end{cases}. \quad (7)$$

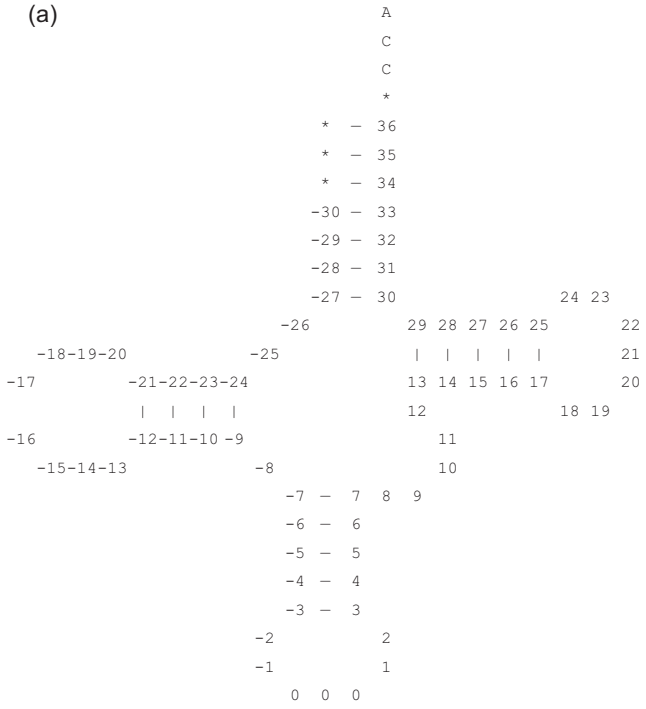

Fig. 1. (a) Nucleotide numbering in a common tRNA. The three sites of the anticodon are numbered 000, the nucleotides in the 5′ region of a tRNA are numbered negatively and the nucleotides in the 3′ region of a tRNA are numbered positively. (b) Trinucleotide numbering in a common tRNA. A trinucleotide numbered *iii* is in position *i*. The anticodon is in position 0, the trinucleotides in 5′ the region of a tRNA are numbered negatively and the trinucleotides in the 3′ region of a tRNA are numbered positively.

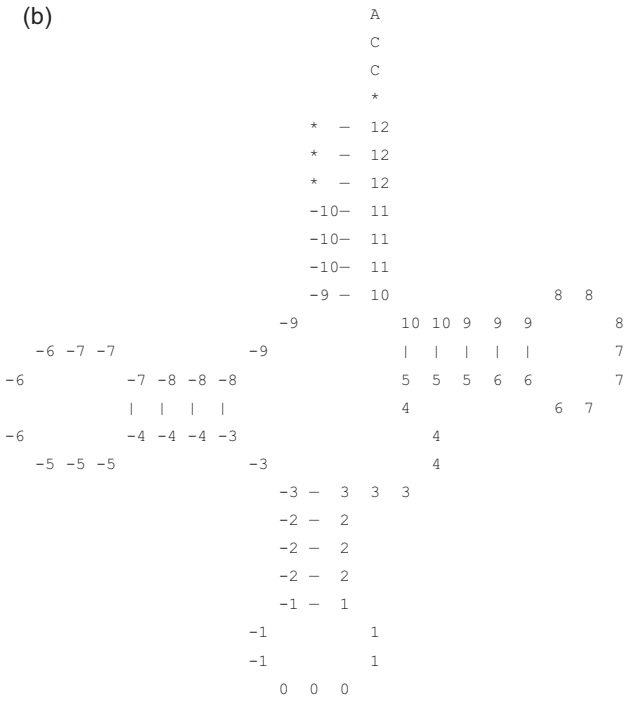

Fig. 1. (*continued*).

Then, the function $Pr(i, \mathcal{F})$ computes the occurrence probability of X motifs in the nucleotide range [min(s), max(s)] in each tRNA s of the population $\mathcal{F}$ as follows

$$Pr(i, \mathcal{F}) = \frac{1}{N(i, \mathcal{F})} \sum_{s \in \mathcal{F} | i \in [\min(s), \max(s)]} \delta(i, s). \quad (8)$$

Property 2. *If each sequence s of $\mathcal{F}$ has a X motif of length greater or equal to lgMinX then $Pr(i, \mathcal{F}) = 1$.*

Property 3. *If all sequences s of $\mathcal{F}$ have no X motif of length greater or equal to lgMinX then $Pr(i, \mathcal{F}) = 0$.*

2.3. Definition of an occurrence probability of the circular codes X, X_1 and X_2 in a tRNA population

This definition extends the previous definition on nucleotides (Section 2.2) to trinucleotide circular codes (sets of trinucleotides). Let $\mathcal{F}$ be a tRNA population with $N(\mathcal{F})$ sequences s. Let $t_{i(s)}$ be the trinucleotide $t \in A_4^3$ in a position i in a trinucleotide range [min(s), max(s)] (Fig. 1b) of a tRNA s of $\mathcal{F}$. Let the circular codes $X = X_0$ (1), X_1 (2) and X_2 (3) be defined by X_j with $j \in \{0,1,2\}$.

The function $\delta(X_j,i,s)$ indicates whether the trinucleotide $t_{i(s)}$ in a trinucleotide position i of a tRNA s belongs or not to a circular code X_j

$$\delta(X_j, i, s) = \begin{cases} 1 & \text{if } t_{i(s)} \in X_j \\ 0 & \text{otherwise} \end{cases}. \quad (9)$$

Then, the function $N(X_j,s)$ computes the occurrence number of a circular code X_j in a trinucleotide range [min(s), max(s)] of a tRNA s as follows

$$N(X_j, s) = \sum_{i \in [\min(s), \max(s)]} \delta(X_j, i, s). \quad (10)$$

Finally, the function $Pr(i, \mathcal{F})$ computes the occurrence probability of a circular code X_j in a tRNA population $\mathcal{F}$ as follows

$$Pr(X_j, \mathcal{F}) = \frac{1}{\sum_{s \in \mathcal{F}} \sum_{j=1}^{3} N(X_j, s)} \sum_{s \in \mathcal{F}} N(X_j, s). \quad (11)$$

Remark 6. $\sum_{j=1}^{3} Pr(X_j, \mathcal{F}) = 1$.

2.4. tRNA data

The tRNAs are extracted from the genomic tRNA database (September 2012, http://lowelab.ucsc.edu/GtRNAdb/, Lowe and Eddy, 1997). Usual preliminary tests exclude tRNAs with nucleotides different from A_4 and with inconsistent lengths, anticodon positions and anticodon types. We recall that isoaccepting tRNAs code the same amino acid.

2.4.1. tRNAs of prokaryotes

Fig. 2 gives the histograms representing the numbers of isoaccepting tRNAs of prokaryotes PRO as a function of their lengths. Its distribution greater than 10% of data shows:

(i) eight tRNAs have two preferential lengths: Ala {76,73}, Arg {77,74}, Asn {76,75}, Asp {77,76}, Ile {77,74}, Lys {76,73}, Phe {76,73} and Pro {77,74} (the tRNA lengths in each list being given by descending order of their numbers);
(ii) ten tRNAs have three preferential lengths: Cys {74,71,75}, Gln {75,72,74}, Glu {76,75,72}, Gly {76,74,75}, His {76,77,73}, Leu {87,85,86}, Met {77,74,76}, Ser {90,91,88}, Thr {76,73,75} and Tyr {85,86,84};
(iii) two tRNAs have four preferential lengths: Trp {76,73,74,77} and Val {76,75,77,73}.

Thus, the preferential lengths of the 20 isoaccepting tRNAs of prokaryotes are:

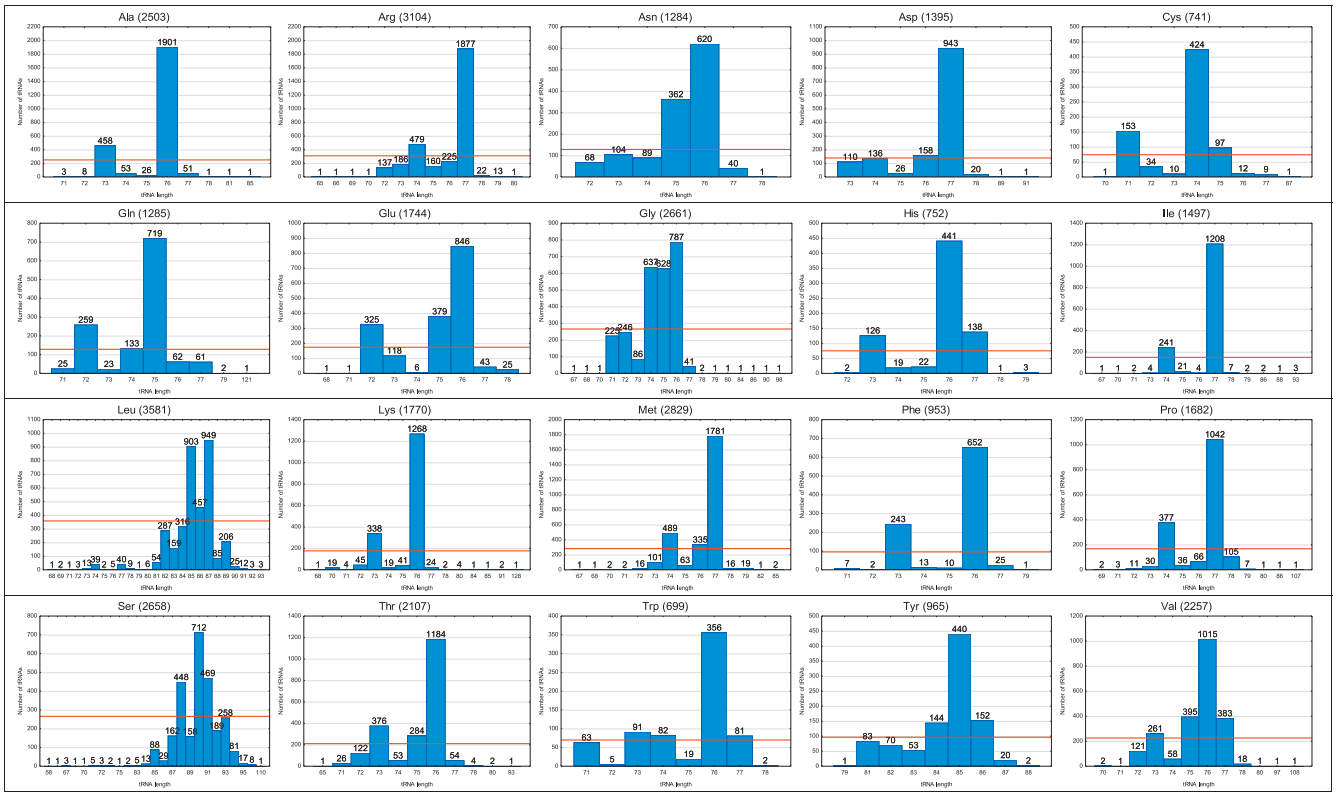

Fig. 2. Histograms representing the number of isoaccepting tRNAs as a function of their lengths in prokaryotes PRO. The title of each histogram gives the isoaccepting tRNA and its total number (in parenthesis). The horizontal line in red represents 10% of data.

(i) 76 nucleotides for 10 tRNAs: Ala, Asn, Glu, Gly, His Lys, Phe, Thr, Trp and Val;
(ii) 77 nucleotides for five tRNAs: Arg, Asp, Ile, Pro and Met;
(iii) 74 and 75 nucleotides for two tRNAs: Cys and Gln, respectively;
(iv) 85 and 87 nucleotides for two tRNAs: Tyr and Leu, respectively;
(v) 90 nucleotides for one tRNA: Ser.

Some modulo 3 differences between the isoaccepting tRNA lengths, e.g. 76 and 73 for Ala, 77 and 74 for Arg etc., could be explained by the presence or absence of the terminal trinucleotide CCA which can be added during processing and therefore, which does not appear in the tRNA genes. Other length differences, e.g. 76 and 75 for Asn, 77 and 76 for Asp etc., could be related to some initial or terminal missing nucleotides.

2.4.2. tRNAs of eukaryotes

A similar data analysis is now carrying out with the isoaccepting tRNAs of eukaryotes. Fig. 3 gives the histograms representing the numbers of isoaccepting tRNAs of eukaryotes EUK as a function of their lengths. Its distribution greater than 10% of data shows:

(i) nine tRNAs have a unique preferential length: Asn {74}, Asp {72}, Cys {72}, His {72}, Lys {73}, Phe {73}, Pro {72}, Ser {82} and Val {73};
(ii) eight tRNAs have two preferential lengths: Ala {72,73}, Arg {73,85}, Gln {72,85}, Glu {72,73}, Ile {74,94}, Leu {82,83}, Met {73,72} and Trp {72,73};
(iii) three tRNAs have three preferential lengths: Gly {71,73,72}, Thr {74,72,73} and Tyr {87,84,89}.

Thus, the preferential lengths of the 20 isoaccepting tRNAs of eukaryotes are:

(i) 72 nucleotides for eight tRNAs: Ala, Asp, Cys, Gln, Glu, His, Pro and Trp;
(ii) 73 nucleotides for five tRNAs: Arg, Lys, Met, Phe and Val;
(iii) 74 nucleotides for three tRNAs: Asn, Ile and Thr;
(iv) 82 nucleotides for two tRNAs: Leu and Ser;
(v) 71 and 87 nucleotides for two tRNAs: Gly and Tyr, respectively.

In eukaryotes, three tRNAs Arg {73,85}, Gln {72,85} and Ile {74,94} may be constituted of two length classes having different primary, secondary and tertiary structures.

2.4.3. tRNA variability

These tRNA data based on their lengths already demonstrate the well-known great variety of tRNAs of prokaryotes and eukaryotes. This tRNA diversity is also revealed by other parameters, such as: (i) 81 modified nucleotides reported so far (see the RNA Modification Database at http://rna-mdb.cas.albany.edu/RNAmods/, Rozenski et al., 1999) which is the largest number among RNA molecules (rRNA, mRNA, snRNA): D arm with dihydrouridine, T arm with pseudouridine, first anticodon base (wobble base) sometimes modified to inosine, pseudouridine or lysidine; (ii) variability in the secondary structures (stems and loops); (iii) complexity in the tertiary structure with non-Watson–Crick base pairs, etc. This tRNA variability explains the great difficulty of statistical analyses of tRNAs in order to reveal general patterns. This is the reason why several populations of tRNAs based on their lengths and scores will be analysed statistically in order to identify *X* motifs in tRNAs significantly.

2.4.4. Statistical analysis 1: tRNA populations of prokaryotes $\mathcal{F}^1_{PRO}$ and eukaryotes $\mathcal{F}^1_{EUK}$

The statistical analysis 1 is based on the tRNA populations of prokaryotes $\mathcal{F}^1_{\mathrm{PRO}}$ and eukaryotes $\mathcal{F}^1_{\mathrm{EUK}}$ constituted of the 20

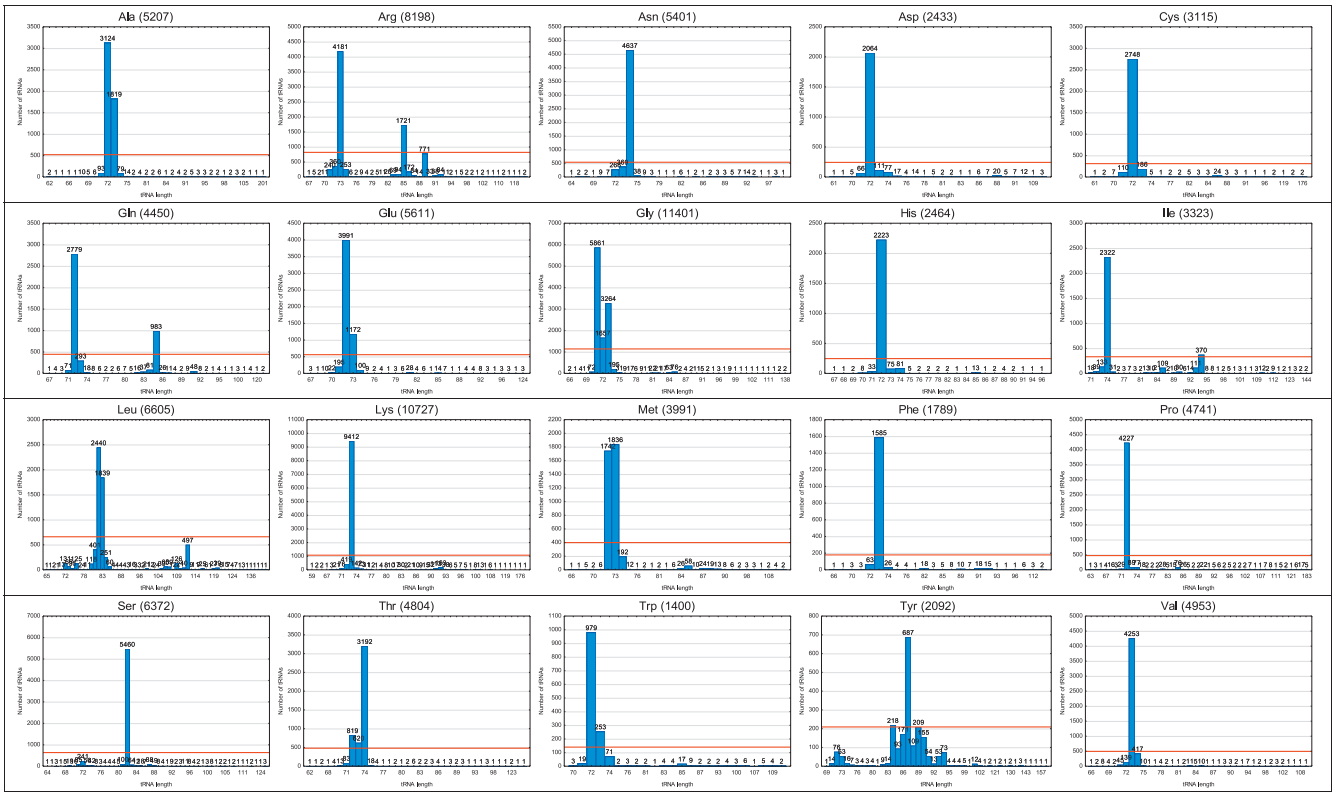

Fig. 3. Histograms representing the number of isoaccepting tRNAs as a function of their lengths in eukaryotes EUK. The title of each histogram gives the isoaccepting tRNA and its total number (in parenthesis). The horizontal line in red represents 10% of data.

isoaccepting tRNAs having a unique length according to the maximum number of data. It is the most filtered (Figs. 2 and 3):

$\mathcal{F}^1_{\mathrm{PRO}}$: Ala {76} (1901), Arg {77} (1877), Asn {76} (620), Asp {77} (943), Cys {74} (424), Gln {75} (719), Glu {76} (846), Gly {76} (787), His {76} (441), Ile {77} (1208), Leu {87} (949), Lys {76} (1268), Met {77} (1781), Phe {76} (652), Pro {77} (1042), Ser {90} (712), Thr {76} (1184), Trp {76} (356), Tyr {85} (440) and Val {76} (1015), where the lengths of isoaccepting tRNAs are in bracket and the numbers of isoaccepting tRNAs are in parenthesis;

$\mathcal{F}^1_{\mathrm{EUK}}$: Ala {72} (3124), Arg {73} (4181), Asn {74} (4637), Asp {72} (2064), Cys {72} (2748), Gln {72} (2779), Glu {72} (3991), Gly {71} (5861), His {72} (2223), Ile {74} (2322), Leu {82} (2440), Lys {73} (9412), Met {73} (1836), Phe {73} (1585), Pro {72} (4227), Ser {82} (5460), Thr {74} (3192), Trp {72} (979), Tyr {87} (687) and Val {73} (4253).

2.4.5. Statistical analysis 2: tRNA populations of prokaryotes $\mathcal{F}^2_{PRO}$ and eukaryotes $\mathcal{F}^2_{EUK}$

The statistical analysis 2 is based on the tRNA populations of prokaryotes $\mathcal{F}^2_{\mathrm{PRO}}$ and eukaryotes $\mathcal{F}^2_{\mathrm{EUK}}$ constituted of the 20 isoaccepting tRNAs having several lengths according to the distribution of data greater than 10%. It contains almost all data of the genomic tRNA database (deduced from Figs. 2 and 3):

$\mathcal{F}^2_{\mathrm{PRO}}$: Ala (2359), Arg (2356), Asn (982), Asp (1101), Cys (674), Gln (1111), Glu (1550), Gly (2052), His (705), Ile (1449), Leu (2309), Lys (1606), Met (2605), Phe (895), Pro (1419), Ser (1629), Thr (1844), Trp (610), Tyr (736) and Val (2054), where the numbers of isoaccepting tRNAs are in parenthesis, their preferential lengths being described above in Section 2.4.1;

$\mathcal{F}^2_{\mathrm{EUK}}$: Ala (4943), Arg (5902), Asn (4637), Asp (2064), Cys (2748), Gln (3762), Glu (5163), Gly (10,782), His (2223), Ile (2692), Leu (4279), Lys (9412), Met (3578), Phe (1585), Pro (4227), Ser (5460), Thr (4631), Trp (1232), Tyr (1114) and Val (4253), their preferential lengths being described above in Section 2.4.2.

Note: As nine eukaryotic tRNAs Asn, Asp, Cys, His, Lys, Phe, Pro, Ser and Val have a unique preferential length, their numbers of isoaccepting tRNAs are identical in $\mathcal{F}^1_{\mathrm{EUK}}$ and $\mathcal{F}^2_{\mathrm{EUK}}$.

2.4.6. Statistical analysis 3: tRNA populations of prokaryotes $\mathcal{F}^3_{PRO}$ and eukaryotes $\mathcal{F}^3_{EUK}$

The statistical analysis 3 is based on the tRNA populations of prokaryotes $\mathcal{F}^3_{\mathrm{PRO}}$ and eukaryotes $\mathcal{F}^3_{\mathrm{EUK}}$ constituted of the 20 isoaccepting tRNAs having several lengths according to the distribution of data greater than 10% with an added filter based on a minimum score of 70, the score being a parameter of the genomic tRNA database. Thus, the tRNA populations $\mathcal{F}^3_{\mathrm{PRO}}$ and $\mathcal{F}^3_{\mathrm{EUK}}$ are the tRNA populations $\mathcal{F}^2_{\mathrm{PRO}}$ and $\mathcal{F}^2_{\mathrm{EUK}}$ in which a score filter is added. The description of $\mathcal{F}^3_{\mathrm{PRO}}$ and $\mathcal{F}^3_{\mathrm{EUK}}$ are given in Tables 1 and 2, respectively.

2.4.7. Choice of the nucleotide range [min(s), max(s)] in the 5′ and 3′ regions of tRNAs and the length of X motifs

A preliminary statistical analysis shows a significant drop of the number of isoaccepting tRNAs for the nucleotide positions $i < -30$ and $i > 36$. Thus, the two regions of tRNAs analyzed with the probability $Pr(i, \mathcal{F})$ (Eq. (8)) use the following nucleotide range [min(s), max(s)]: the 5′ regions in [−30, −1] and the 3′ regions in [1,36] (recall that the anticodon in nucleotide position 0 is excluded). Note that "detailed" statistical analyses outside these nucleotide ranges, i.e. for $i < -30$ and $i > 36$, may also identify X motifs at the beginning of the 5′ regions and at the end of the 3′ regions of tRNAs. However, this study is not investigated as the purpose here is to perform statistical analyses on tRNA populations, i.e. "large" statistical analyses.

Table 1
Identification of *X* motifs in the 20 isoaccepting tRNAs of prokaryotes PRO. The tRNA population of prokaryotes $\mathcal{F}^1_{PRO}$ is constituted of the 20 isoaccepting tRNAs having a unique length according to the maximum number of data (associated to Fig. 5). The tRNA population of prokaryotes $\mathcal{F}^2_{PRO}$ is constituted of the 20 isoaccepting tRNAs having several lengths according to the distribution of data greater than 10% (associated to Fig. 7). The tRNA population of prokaryotes $\mathcal{F}^3_{PRO}$ is constituted of the 20 isoaccepting tRNAs having several lengths according to the distribution of data greater than 10% with a minimum score of 70 (associated to Fig. 9). The *X* motifs underlined in the populations $\mathcal{F}^2_{PRO}$ and $\mathcal{F}^3_{PRO}$ have a probability curve shape modified from the population $\mathcal{F}^1_{PRO}$ taken as reference.

tRNA	Initial data	Statistical analysis 1: $\mathcal{F}^1_{PRO}$		Statistical analysis 2: $\mathcal{F}^2_{PRO}$		Statistical analysis 3: $\mathcal{F}^3_{PRO}$	
	Nb seq.	Nb seq.	Distribution of *X* (Fig. 5)	Nb seq.	Distribution of *X* (Fig. 7)	Nb seq.	Distribution of *X*
Ala	2503	1901	3′ *X* motifs	2359	3′ *X* motifs	2358	3′ *X* motifs
Arg	3104	1877	Random	2356	Random	2147	Random
Asn	1284	620	5′ *X* motifs	982	5′ *X* motifs	976	5′ *X* motifs
Asp	1395	943	5′ *X* motifs	1101	5′ *X* motifs	1082	5′ *X* motifs
Cys	741	424	3′ *X* motifs	674	3′ *X* motifs	180	3′ *X* motifs (Fig. 9)
Gln	1285	719	5′ *X* motifs	1111	5′ *X* motifs	530	5′ and 3′ *X* motifs (Fig. 9)
Glu	1744	846	5′ and 3′ *X* motifs	1550	5′ and 3′ *X* motifs	320	5′ and 3′ *X* motifs (Fig. 9)
Gly	2661	787	3′ *X* motifs	2052	3′ *X* motifs	1778	3′ *X* motifs
His	752	441	Random	705	Random	562	Random
Ile	1497	1208	3′ *X* motifs	1449	3′ *X* motifs	1448	3′ *X* motifs
Leu	3581	949	5′ *X* motifs	2309	5′ *X* motifs	1209	5′ *X* motifs (Fig. 9)
Lys	1770	1268	5′ and 3′ *X* motifs	1606	5′ and 3′ *X* motifs	1584	5′ and 3′ *X* motifs
Met	2829	1781	3′ *X* motifs	2605	3′ *X* motifs	2577	3′ *X* motifs
Phe	953	652	5′ and 3′ *X* motifs	895	5′ and 3′ *X* motifs	888	5′ and 3′ *X* motifs
Pro	1682	1042	Random	1419	Random	1391	Random
Ser	2658	712	Random	1629	Random	718	Random
Thr	2107	1184	3′ *X* motifs	1844	3′ *X* motifs	1817	3′ *X* motifs
Trp	699	356	Random	610	Random	523	Random
Tyr	965	440	3′ *X* motifs	736	3′ *X* motifs	246	3′ *X* motifs
Val	2257	1015	3′ *X* motifs	2054	3′ *X* motifs	2023	3′ *X* motifs
Tot	36,467	19,165		30,046		24,357	

The *X* motifs analyzed with the three statistical analyses in the 5′ and 3′ regions of the 20 isoaccepting tRNAs of prokaryotes and and eukaryotes are the *X* motifs of greatest lengths having at least 9 nucleotides (3 trinucleotides). Indeed, we have recently proved that *X* motifs of 9 nucleotides retrieve the reading frame with a probability of 99.9% (Table 3 and Fig. 4 in Michel, 2012). In fact, as $\mathrm{Pref}_{let}(X) = \{A,C,G,T\}$ (4), the *X* motifs have at least 10 nucleotides. Thus, the statistical approach developed here may identify the tRNAs with a strong capacity of reading frame retrieval. Such tRNAs could be involved in a translation code (details in Michel, 2012). In the following and for simplification, "distribution of *X* motifs" means distribution of *X* motifs of greatest lengths having at least 9 nucleotides in the 5′ and 3′ regions of isoaccepting tRNAs.

3. Results

3.1. Distribution of X motifs in a randomized tRNA population

As the probability $Pr(i, \mathcal{F})$ (Eq. (8)) depends on the length of *X* motifs (having at least 9 nucleotides), the nucleotide range [min(*s*), max(*s*)] and the nucleotide composition of tRNAs, the determination of an analytical probability of *X* motifs in a random tRNA

Table 2
Identification of *X* motifs in the 20 isoaccepting tRNAs of eukaryotes EUK. The tRNA population of eukaryotes $\mathcal{F}^1_{EUK}$ is constituted of the 20 isoaccepting tRNAs having a unique length according to the maximum number of data (associated to Fig. 6). The *X* motifs in bold in the tRNA population of eukaryotes $\mathcal{F}^1_{EUK}$ differ from the *X* motifs in the tRNA population of prokaryotes $\mathcal{F}^1_{PRO}$ (Table 1). The tRNA population of eukaryotes $\mathcal{F}^2_{EUK}$ is constituted of the 20 isoaccepting tRNAs having several lengths according to the distribution of data greater than 10% (associated to Fig. 8). The tRNA population of eukaryotes $\mathcal{F}^3_{EUK}$ is constituted of the 20 isoaccepting tRNAs having several lengths according to the distribution of data greater than 10% with a minimum score of 70 (associated to Fig. 10). The *X* motifs underlined in the populations $\mathcal{F}^2_{EUK}$ and $\mathcal{F}^3_{EUK}$ have a probability curve shape modified from the population $\mathcal{F}^1_{EUK}$ taken as reference.

tRNA	Initial data	Statistical analysis 1: $\mathcal{F}^1_{EUK}$		Statistical analysis 2: $\mathcal{F}^2_{EUK}$		Statistical analysis 3: $\mathcal{F}^3_{EUK}$	
	Nb seq.	Nb seq.	Distribution of *X* (Fig. 6)	Nb seq.	Distribution of *X* (Fig. 8)	Nb seq.	Distribution of *X*
Ala	5207	3124	3′ *X* motifs	4943	3′ *X* motifs	2016	3′ *X* motifs
Arg	8198	4181	Random	5902	Random	1470	3′ *X* motifs (Fig. 10)
Asn	5401	4637	**3′ *X* motifs**	4637	3′ *X* motifs	2073	3′ *X* motifs
Asp	2433	2064	**Random**	2064	Random	921	Random
Cys	3115	2748	**Random**	2748	Random	1553	Random
Gln	4450	2779	**3′ *X* motifs**	3762	3′ *X* motifs	1412	3′ *X* motifs
Glu	5611	3991	**Random**	5163	Random	2518	Random
Gly	11,401	5861	**5′ *X* motifs**	10,782	5′ *X* motifs	2765	5′ *X* motifs
His	2464	2223	Random	2223	Random	171	3′ *X* motifs (Fig. 10)
Ile	3323	2322	**5′ *X* motifs**	2692	5′ *X* motifs	1796	5′ *X* motifs
Leu	6605	2440	**Random**	4279	5′ *X* motifs	1477	5′ *X* motifs
Lys	10,727	9412	**Random**	9412	Random	4509	3′ *X* motifs (Fig. 10)
Met	3991	1836	**5′** and 3′ *X* motifs	3578	3′ *X* motifs	1247	5′ and 3′ *X* motifs
Phe	1789	1585	**3′ *X* motifs**	1585	3′ *X* motifs	1124	3′ *X* motifs
Pro	4741	4227	**5′ *X* motifs**	4227	5′ *X* motifs	2466	5′ *X* motifs
Ser	6372	5460	Random	5460	Random	4691	Random
Thr	4804	3192	3′ *X* motifs	4631	3′ *X* motifs	3104	3′ *X* motifs
Trp	1400	979	Random	1232	Random	694	Random
Tyr	2092	687	3′ *X* motifs	1114	3′ *X* motifs	798	3′ *X* motifs
Val	4953	4253	3′ *X* motifs	4253	3′ *X* motifs	3246	3′ *X* motifs
Tot	99,077	68,001		84,687		40,051	

Table 3
Probability $Pr(X_j, \mathcal{F})$ (Eq. (11)) (in %) of the circular codes X, X_1 and X_2 in the 3′ regions of the tRNA populations of prokaryotes $\mathcal{F}^2_{3'PRO}$ and eukaryotes $\mathcal{F}^2_{3'EUK}$ which are constituted of the 20 isoaccepting tRNAs having several lengths associated to the distribution of data greater than 10%. The property $Pr(X, \mathcal{F}_{3'tRNA}) > Pr(X_1, \mathcal{F}_{3'tRNA}) > Pr(X_2, \mathcal{F}_{3'tRNA})$ identified in the 3′ regions of 19 isoaccepting tRNA populations $\mathcal{F}_{3'tRNA}$ of prokaryotes and eukaryotes (except for the tRNA Leu) is also observed in genes of prokaryotes and eukaryotes (Arquès et al., 1999; Bahi and Michel, 2008). The numbers underlined do not verify the property identified.

tRNA	$Pr(X_j, \mathcal{F}^2_{3'PRO})$ (in %)			$Pr(X_j, \mathcal{F}^2_{3'EUK})$ (in %)		
	X	X_1	X_2	X	X_1	X_2
Ala	47.6	42.6	9.8	43.8	38.0	18.2
Arg	42.1	31.9	26.1	48.2	33.6	18.2
Asn	36.9	28.3	34.9	44.7	42.8	12.5
Asp	43.6	42.6	13.8	51.3	25.3	23.4
Cys	30.6	35.9	33.5	63.4	19.8	16.8
Gln	43.6	37.9	18.5	39.8	33.7	26.5
Glu	35.7	44.9	19.4	53.2	23.3	23.5
Gly	38.4	37.4	24.3	49.6	20.1	30.3
His	50.6	30.3	19.2	44.0	18.4	37.6
Ile	43.2	30.8	26.0	50.8	35.2	14.0
Leu	36.4	40.7	23.0	35.7	37.8	26.5
Lys	44.8	25.9	29.3	45.7	44.5	9.8
Met	49.3	28.6	22.1	42.1	39.6	18.3
Phe	50.4	24.2	25.4	56.2	23.4	20.4
Pro	53.0	27.7	19.3	45.3	28.6	26.1
Ser	40.1	32.0	27.8	32.0	42.6	25.3
Thr	55.4	30.0	14.6	60.6	26.7	12.7
Trp	37.4	34.4	28.2	51.5	28.4	20.1
Tyr	45.4	30.5	24.0	38.5	34.7	26.8
Val	43.6	46.3	10.1	40.8	38.5	20.7
Tot	43.8	34.7	21.5	46.4	32.5	21.1

population is impossible (analytical formulas for simple cases could be investigated in future). However, approximated probability of X motifs in a random tRNA population $\mathcal{F}_{RAN}$ can be performed by computer simulation which randomizes the tRNAs of $\mathcal{F}_{RAN}$ by considering the length and nucleotide composition of each tRNA region (5′ and 3′). The probability $Pr(i, \mathcal{F}_{RAN})$ of X motifs in the 20 randomized isoaccepting tRNAs of prokaryotes and eukaryotes have almost identical curves (data not shown). Fig. 4 gives an example of a distribution of X motifs in a randomized tRNA population. The random curve has a "bicorn" shape and a steady state of probability less than 0.25 in the 5′ and 3′ regions of $\mathcal{F}_{RAN}$.

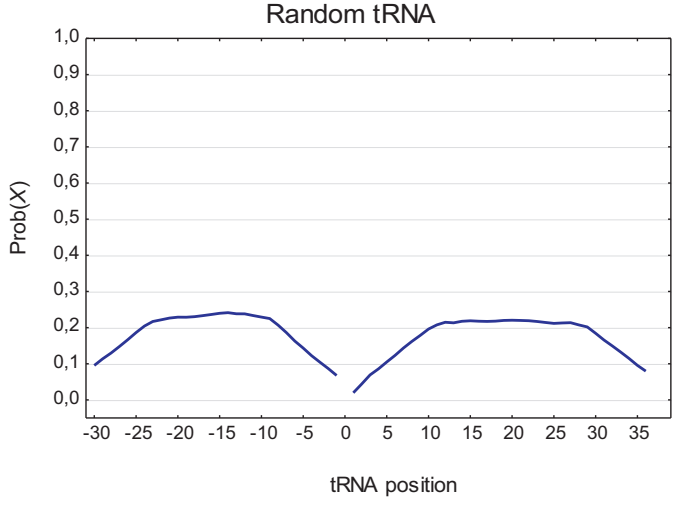

Fig. 4. Probability $Pr(i, \mathcal{F}_{RAN})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of a randomized tRNA population $\mathcal{F}_{RAN}$.

3.2. Distribution of X motifs in the tRNA populations of prokaryotes $\mathcal{F}^1_{PRO}$ and eukaryotes $\mathcal{F}^1_{EUK}$

The tRNA populations of prokaryotes $\mathcal{F}^1_{PRO}$ and eukaryotes $\mathcal{F}^1_{EUK}$ are constituted of the 20 isoaccepting tRNAs having a unique length associated to the maximum number of data (statistical analysis 1).

Fig. 5 shows the probability $Pr(i, \mathcal{F}^1_{PRO})$ (Eq. (8)) of X motifs in the tRNA population of prokaryotes $\mathcal{F}^1_{PRO}$. Five tRNAs Arg, His, Pro, Ser and Trp have random curves, i.e. a random (or a very low significance) occurrence probability of X motifs in their 5′ and 3′ regions. Eight tRNAs Ala, Cys, Gly, Ile, Met, Thr, Tyr and Val have 3′ X motifs, i.e. X motifs with unexpected high occurrence probabilities (compared to the random case in Fig. 4) in their 3′ regions. For example, the maximum of the probability $Pr(i, \mathcal{F}^1_{Ala})$ in the 3′ regions of the Ala tRNA population $\mathcal{F}^1_{Ala}$ obtained at 0.74 for the positions $10 \le i \le 13$ (Fig. 5) means that 74% of tRNAs in $\mathcal{F}^1_{Ala}$, i.e. 74% of 1901 tRNAs (Table 1), have the positions $10 \le i \le 13$ in their 3′ regions occupied by X motifs. Recall that the random curve shape in Fig. 4 has probabilities always less than 0.25. Four tRNAs Asn, Asp, Gln and Leu have 5′ X motifs. Three tRNAs Glu, Lys and Phe have 5′ and 3′ X motifs. The properties of X motifs in $\mathcal{F}^1_{PRO}$ are summarized in Table 1.

Fig. 6 shows the probability $Pr(i, \mathcal{F}^1_{EUK})$ (Eq. (8)) of X motifs in the tRNA population of eukaryotes $\mathcal{F}^1_{EUK}$. Nine tRNAs Arg, Asp, Cys, Glu, His, Leu, Lys, Ser and Trp have random curves. Seven tRNAs Ala, Asn, Gln, Phe, Thr, Tyr and Val have 3′ X motifs. Three tRNAs Gly, Ile and Pro have 5′ X motifs. One tRNA Met has 5′ and 3′ X motifs. The properties of X motifs in $\mathcal{F}^1_{EUK}$ are summarized in Table 2.

3.3. Distribution of X motifs in the tRNA populations of prokaryotes $\mathcal{F}^2_{PRO}$ and eukaryotes $\mathcal{F}^2_{EUK}$

The tRNA populations of prokaryotes $\mathcal{F}^2_{PRO}$ and eukaryotes $\mathcal{F}^2_{EUK}$ are constituted of the 20 isoaccepting tRNAs having several lengths according to the distribution of data greater than 10% (statistical analysis 2). They contain almost all data of the genomic tRNA database.

Fig. 7 shows the probability $Pr(i, \mathcal{F}^2_{PRO})$ (Eq. (8)) of X motifs in the tRNA population of prokaryotes $\mathcal{F}^2_{PRO}$. The number of tRNAs in $\mathcal{F}^2_{PRO}$ (30,046 tRNAs) differs very significantly from the number of tRNAs in $\mathcal{F}^1_{PRO}$ (19,165 tRNAs), i.e. a sequence increase of 56.8% (Table 1). However, the 20 probability curves $Pr(i, \mathcal{F}^2_{PRO})$ are very similar to the 20 probability curves $Pr(i, \mathcal{F}^1_{PRO})$ (Figs. 5 and 7 and Table 1). There are two significant curve shape modifications: the tRNA Gly with a maximum which decreases from 0.7 to 0.4 in its 3′ regions but still with a significant probability (greater than 0.25) and the tRNA Val with a different curve shape in its 3′ regions.

Nine tRNAs Asn, Asp, Cys, His, Lys, Phe, Pro, Ser and Val have the same numbers of sequences in the two populations $\mathcal{F}^1_{EUK}$ and $\mathcal{F}^2_{EUK}$. Thus, only 11 tRNAs Ala, Arg, Gln, Glu, Gly, Ile, Leu, Met, Thr, Trp and Tyr are considered in the statistical analysis 2. Fig. 8 shows their probability $Pr(i, \mathcal{F}^2_{EUK})$ (Eq. (8)) of X motifs. The number of sequences of these 11 tRNAs in $\mathcal{F}^2_{EUK}$ (48,078 tRNAs) differs very significantly from the one in $\mathcal{F}^1_{EUK}$ (31,392 tRNAs), i.e. a sequence increase of 53.1% (deduced from Table 2). The 11 probability curves $Pr(i, \mathcal{F}^2_{EUK})$ are, again, very similar to the 11 probability curves $Pr(i, \mathcal{F}^1_{EUK})$ (Figs. 6 and 8 and Table 2). There are six curve shape modifications: the tRNA Ala with a peak in the steady state in its 3′ regions, the tRNA Gly with a maximum which decreases from 0.75 to 0.45 in its 5′ regions but still with a significant probability (greater than 0.25), the tRNA Leu with appearance of 5′ X motifs, the tRNA Met with a disappearance of 5′ X motifs and the tRNAs Thr and Tyr with a steady state which decreases from 0.9 to 0.7 in their 3′ regions.

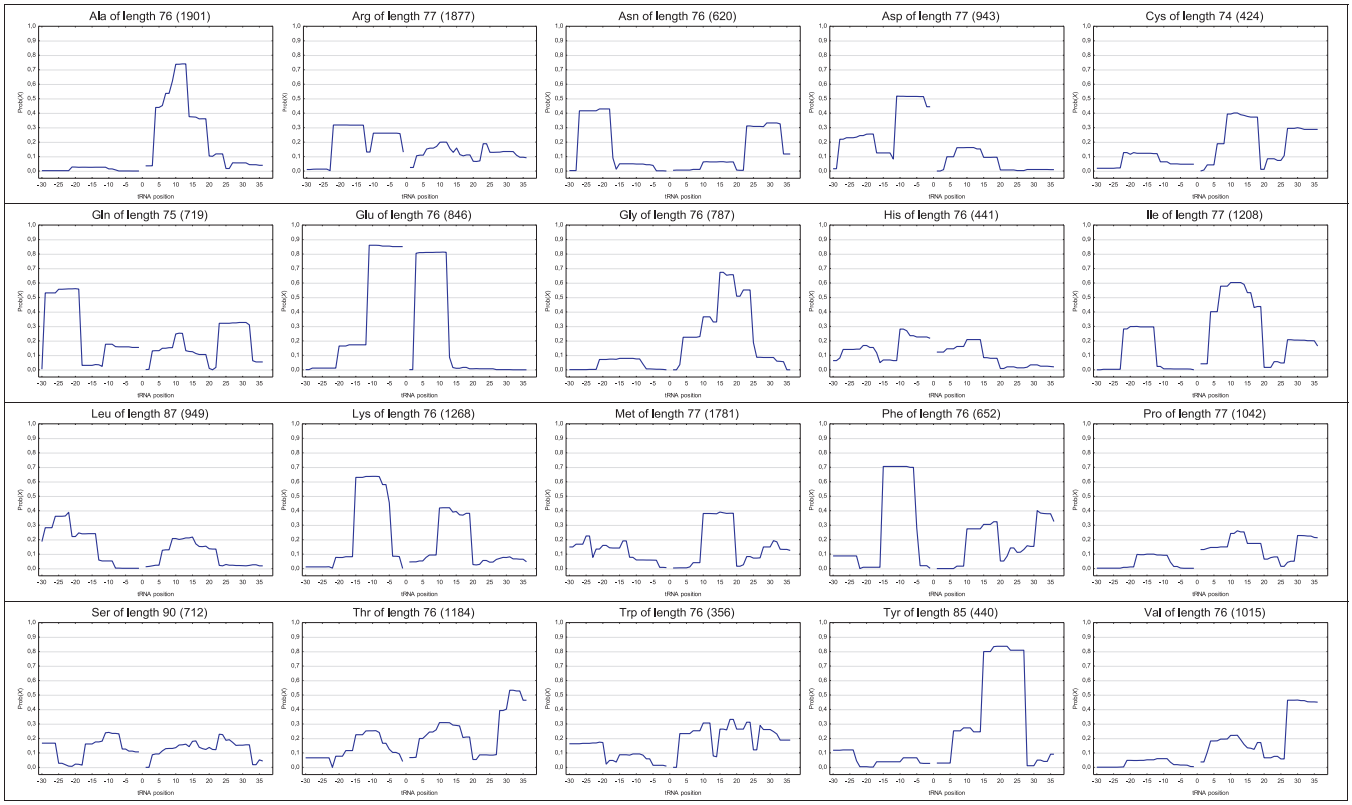

Fig. 5. Probability $Pr(i, \mathcal{F}^1_{\mathrm{PRO}})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of the tRNA population of prokaryotes $\mathcal{F}^1_{\mathrm{PRO}}$ constituted of the 20 isoaccepting tRNAs having a unique length according to the maximum number of data. The number of isoaccepting tRNAs is in parenthesis.

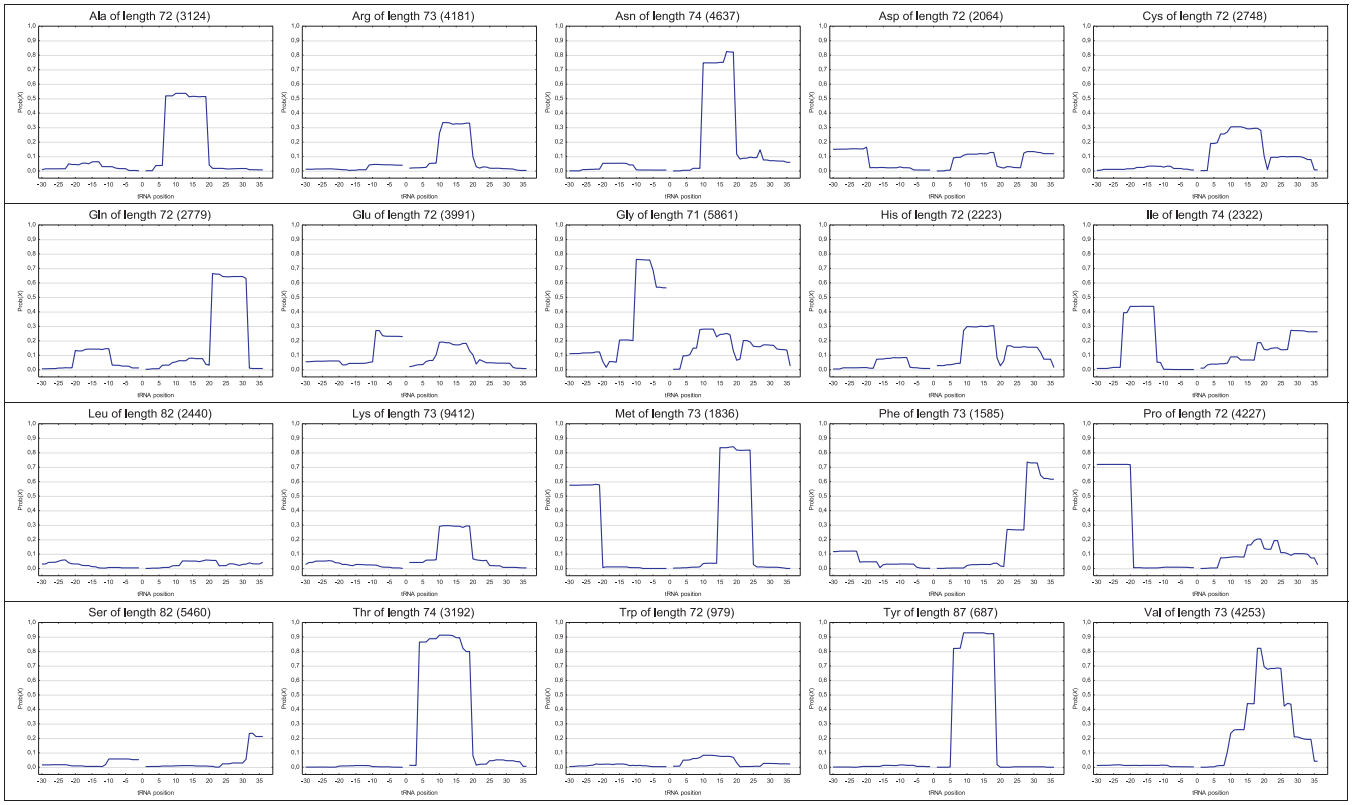

Fig. 6. Probability $Pr(i, \mathcal{F}^1_{\mathrm{EUK}})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of the tRNA population of eukaryotes $\mathcal{F}^1_{\mathrm{EUK}}$ constituted of the 20 isoaccepting tRNAs having a unique length according to the maximum number of data. The number of isoaccepting tRNAs is in parenthesis.

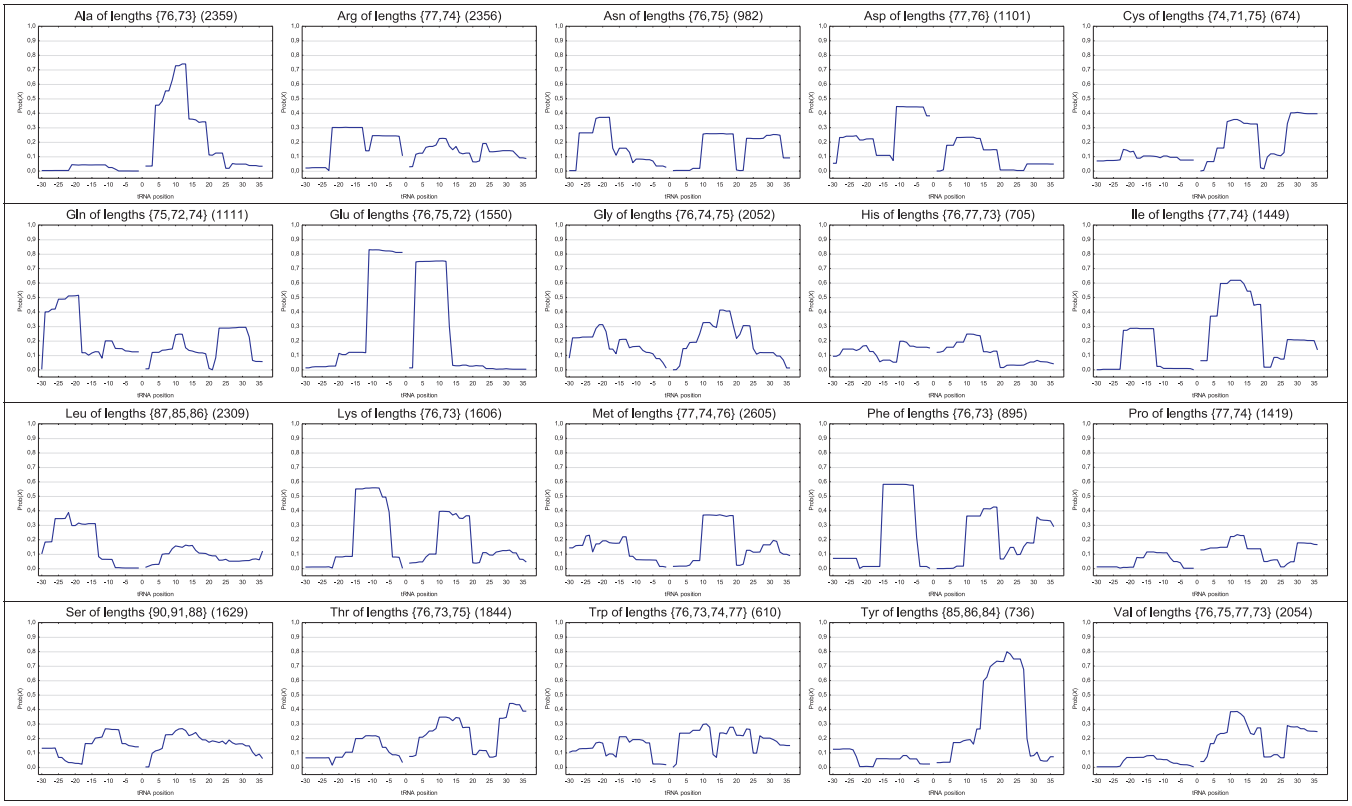

Fig. 7. Probability $Pr(i, \mathcal{F}^2_{\mathrm{PRO}})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of the tRNA population of prokaryotes $\mathcal{F}^2_{\mathrm{PRO}}$ constituted of the 20 isoaccepting tRNAs having several lengths according to the distribution of data greater than 10%. The tRNA lengths are in bracket by descending order of their numbers and the number of isoaccepting tRNAs is in parenthesis.

3.4. Distribution of X motifs in the tRNA populations of prokaryotes $\mathcal{F}^3_{PRO}$ and eukaryotes $\mathcal{F}^3_{EUK}$

The tRNA populations of prokaryotes $\mathcal{F}^3_{\mathrm{PRO}}$ and eukaryotes $\mathcal{F}^3_{\mathrm{EUK}}$ are constituted of the 20 isoaccepting tRNAs having several lengths associated to the distribution of observations greater than 10% with a minimum score of 70 (statistical analysis 3).

The 20 probability curves $Pr(i, \mathcal{F}^3_{\mathrm{PRO}})$ are very similar to the 20 probability curves $Pr(i, \mathcal{F}^1_{\mathrm{PRO}})$ and $Pr(i, \mathcal{F}^2_{\mathrm{PRO}})$ with the same identification of X motifs (Table 1) except for the tRNA Gln with the

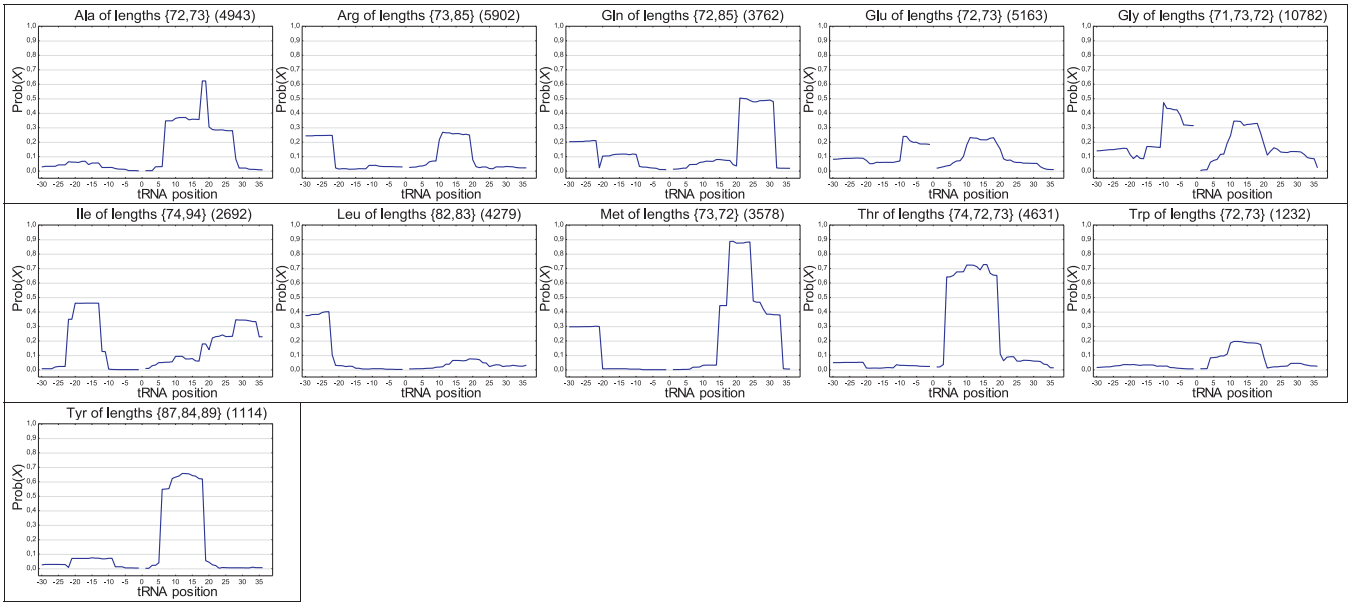

Fig. 8. Probability $Pr(i, \mathcal{F}^2_{\mathrm{EUK}})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of the tRNA population of eukaryotes $\mathcal{F}^2_{\mathrm{EUK}}$ given for the 11 isoaccepting tRNA populations having several lengths according to the distribution of data greater than 10% and whose sequence numbers differ from those of $\mathcal{F}^1_{\mathrm{EUK}}$ (see Table 2). The tRNA lengths are in bracket by descending order of their numbers and the number of isoaccepting tRNAs is in parenthesis.

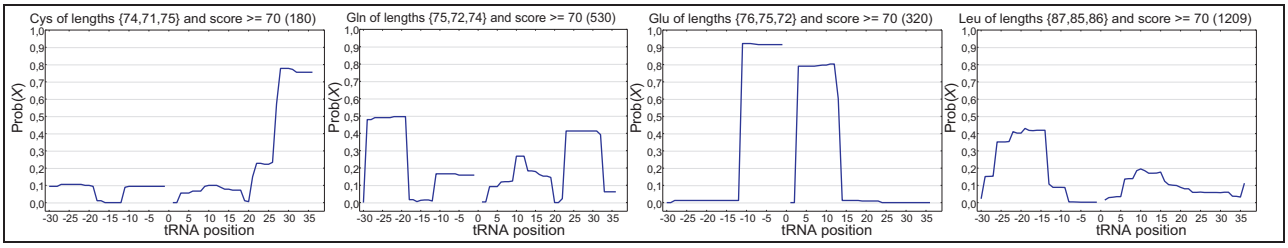

Fig. 9. Probability $Pr(i, \mathcal{F}^3_{PRO})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of the tRNA population of prokaryotes $\mathcal{F}^3_{PRO}$ given for the four isoaccepting tRNAs Cys, Gln, Glu and Leu having several lengths according to the distribution of data greater than 10% with a minimum score of 70. The tRNA lengths are in bracket by descending order of their numbers and the number of isoaccepting tRNAs is in parenthesis.

appearance of 3′ X motifs in addition to 5′ X motifs (Fig. 9). Otherwise, the tRNA Cys has a different curve shape in its 3′ regions and the tRNA Leu has an occurrence of X motifs in its 5′ regions with a level close to the statistical significance (Fig. 9). Also interestingly from a sample perspective, the tRNA Glu with only 320 sequences in $\mathcal{F}^3_{PRO}$ has a curve shape very similar to the curve shapes in $\mathcal{F}^1_{PRO}$ with 846 sequences and $\mathcal{F}^2_{PRO}$ with 1550 sequences (Fig. 9).

The 20 probability curves $Pr(i, \mathcal{F}^3_{EUK})$ are very similar to the 20 probability curves $Pr(i, \mathcal{F}^1_{EUK})$ and $Pr(i, \mathcal{F}^2_{EUK})$ with the same identification of X motifs (Table 2) except for three tRNAs Arg, His and Lys with the appearance of 3′ X motifs instead of random curves (Fig. 10). Note that the number of sequences for the tRNA His drops significantly, from 2223 sequences in $\mathcal{F}^2_{EUK}$ to 171 sequences in $\mathcal{F}^3_{EUK}$ (Table 2).

3.5. Identification of a gene circular code property in the 3′ regions of the tRNA populations of prokaryotes $\mathcal{F}^2_{3'PRO}$ and eukaryotes $\mathcal{F}^2_{3'EUK}$

The probability $Pr(X_j, \mathcal{F})$ (Eq. (11)) of the circular codes $X = X_0$, X_1 and X_2 are computed in the 3′ regions of the tRNA populations of prokaryotes $\mathcal{F}^2_{3'PRO}$ and eukaryotes $\mathcal{F}^2_{3'EUK}$ which are constituted of the 20 isoaccepting tRNAs having several lengths associated to the distribution of data greater than 10%. Recall that $\mathcal{F}^2_{3'PRO}$ and $\mathcal{F}^2_{3'EUK}$ contain almost all data of the genomic tRNA database. Very unexpectedly, for 19 isoaccepting tRNA populations $\mathcal{F}_{3'tRNA}$ (except for the tRNA Leu), the probabilities of the circular codes X, X_1 and X_2 verify, with a very few exceptions, the inequality (Table 3)

$$1 > Pr(X, \mathcal{F}_{3'tRNA}) > Pr(X_1, \mathcal{F}_{3'tRNA}) > Pr(X_2, \mathcal{F}_{3'tRNA}) > 0. \quad (12)$$

The average values (in %) (Table 3) of this probability inequality (12) are for the 3′ regions of the tRNA population of prokaryotes $\mathcal{F}^2_{3'PRO}$ (30,046 tRNAs, Table 1),

$$Pr(X, \mathcal{F}^2_{3'PRO}) = 43.8 > Pr(X_1, \mathcal{F}^2_{3'PRO}) = 34.7 > Pr(X_2, \mathcal{F}^2_{3'PRO}) = 21.5 \quad (13)$$

and for the 3′ regions of the tRNA population of eukaryotes $\mathcal{F}^2_{3'EUK}$ (84,687 tRNAs, Table 2),

$$Pr(X, \mathcal{F}^2_{3'EUK}) = 46.4 > Pr(X_1, \mathcal{F}^2_{3'EUK}) = 32.5 > Pr(X_2, \mathcal{F}^2_{3'EUK}) = 21.1. \quad (14)$$

From a theoretical point of view, as the circular code X is C^3 self-complementary (Result 1), the probabilities of the circular codes X, X_1 and X_2 verify the following inequalities

(i) in primitive (pure) sequences $\mathcal{F}_X$ based on the circular code X,

$$Pr(X, \mathcal{F}_X) = 1 > Pr(X_1, \mathcal{F}_X) = Pr(X_2, \mathcal{F}_X) = 0; \quad (15)$$

(ii) in primitive sequences $\mathcal{F}_X$ subjected to random mutations leading to mutated sequences $\mathcal{F}_M$,

$$1 > Pr(X, \mathcal{F}_M) > Pr(X_1, \mathcal{F}_M) = Pr(X_2, \mathcal{F}_M) > 0; \quad (16)$$

(iii) in random sequences $\mathcal{F}_{RAN}$,

$$Pr(X, \mathcal{F}_{RAN}) = Pr(X_1, \mathcal{F}_{RAN}) = Pr(X_2, \mathcal{F}_{RAN}) = 1/3. \quad (17)$$

As $Pr(X, \mathcal{F}_{3'tRNA}) < 1$ (12–14), the circular code X was subjected to random mutations in tRNAs. As $Pr(X, \mathcal{F}_{3'tRNA}) > 1/3$ (12–14), the circular code X has a non-random occurrence in tRNAs. Thus, from a theoretical point of view, the circular code X in tRNAs is associated to the probability inequality (16). According to (16), the equality $Pr(X_1, \mathcal{F}_M) = Pr(X_2, \mathcal{F}_M)$ should also be satisfied. However, an expected asymmetry between the circular codes X_1 and X_2 with $Pr(X_1, \mathcal{F}_{3'tRNA}) > Pr(X_2, \mathcal{F}_{3'tRNA})$ is observed in tRNAs with the probability inequalities (12–14).

Very surprisingly, this probability inequality (12) is a circular code property which is also observed in gene populations of both prokaryotes and eukaryotes. Indeed, a computer calculus in 1999 on a gene population $\mathcal{F}_{G(PRMV)}$ of primates, rodents, mammals and vertebrates (17,072 genes) showed the inequality (12) with the

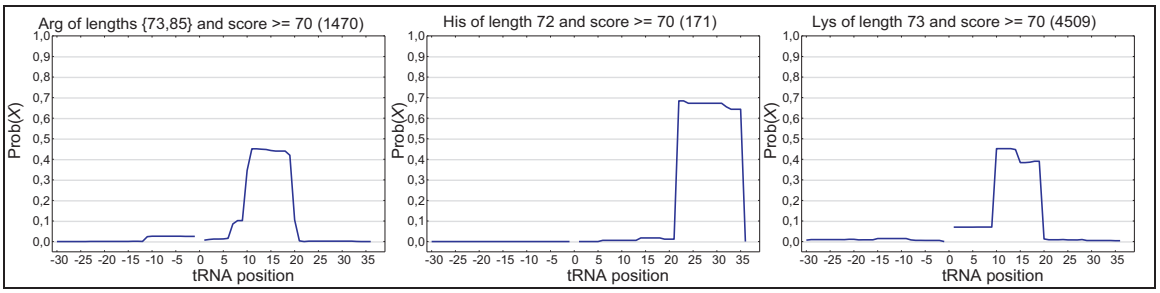

Fig. 10. Probability $Pr(i, \mathcal{F}^3_{EUK})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of the tRNA population of eukaryotes $\mathcal{F}^3_{EUK}$ given for the three isoaccepting tRNAs Arg, His and Lys having several lengths according to the distribution of data greater than 10% with a minimum score of 70. The tRNA lengths are in bracket by descending order of their numbers and the number of isoaccepting tRNAs is in parenthesis.

following values (in %) (Table 2 in Arquès et al., 1999)

$$Pr(X, \mathcal{F}_{G(\mathrm{PRMV})}) = 48.5 > Pr(X_1, \mathcal{F}_{G(\mathrm{PRMV})}) = 29.0 > Pr(X_2, \mathcal{F}_{G(\mathrm{PRMV})}) = 22.5. \quad (18)$$

Furthermore, this inequality (in %) was also retrieved with a computer calculus in 2008 on a gene population $\mathcal{F}_{G(\mathrm{PRO})}$ of 175 complete genomes of prokaryotes (487,758 genes, 454 megabases) (inequality (3.1) in Bahi and Michel, 2008):

$$Pr(X, \mathcal{F}_{G(\mathrm{PRO})}) = 48.8 > Pr(X_1, \mathcal{F}_{G(\mathrm{PRO})}) = 28.0 > Pr(X_2, \mathcal{F}_{G(\mathrm{PRO})}) = 23.2. \quad (19)$$

This inequality (12) has been studied from two aspects. In combinatorics, two propositions were recently proved on circular codes which are C^3 and/or self-complementary:

(i) A self-complementary trinucleotide circular code X has two permuted sets X_1 and X_2 which are either both circular codes or both not circular codes (Bussoli et al., 2011).
(ii) A trinucleotide circular code X is self-complementary if and only if the two permuted sets X_1 and X_2 are complementary (Bussoli et al., 2012).

In a stochastic approach, we have developed substitution models of gene evolution allowing to explain an asymmetry between the circular codes X_1 and X_2. Indeed, random substitutions in "primitive" genes based on the circular code X with different rates in the three sites of the 20 trinucleotides of X, rates which can be constant (Arquès et al., 1999) or chaotic (Bahi and Michel, 2008) during time, can generate the codes X_1 and X_2 according to an unbalanced way and retrieve the probability inequality (12). Nevertheless, the biological foundations of this asymmetry between the circular codes X_1 and X_2 in genes and now in the 3′ regions of tRNAs remain mysterious and unexplained so far.

In contrast to the 3′ regions, the probability inequality (12) is not observed in the 5′ regions of the tRNA populations of prokaryotes $\mathcal{F}^2_{\mathrm{PRO}}$ and eukaryotes $\mathcal{F}^2_{\mathrm{EUK}}$ (data not shown).

4. Conclusion

X motifs (circular code motifs) are identified in the 5′ and/or 3′ regions of several isoaccepting tRNAs. For the tRNAs of prokaryotes and with the three statistical analyses (Table 1), eight isoaccepting tRNAs have 3′ X motifs: Ala, Cys, Gly, Ile, Met, Thr, Tyr and Val, four isoaccepting tRNAs have 5′ X motifs: Asn, Asp, Gln and Leu (Gln has also 3′ X motifs in $\mathcal{F}^3_{\mathrm{PRO}}$), and three isoaccepting tRNAs have 5′ and 3′ X motifs: Glu, Lys and Phe. Note that 5′ and 3′ X motifs were already observed in a particular tRNA Phe (Fig. 5 in Michel, 2012). The statistical approach developed does not reveal preferential occurrence of X motifs in five prokaryotic isoaccepting tRNAs: Arg, His, Pro, Ser and Trp. For the tRNAs of eukaryotes and with the three statistical analyses (Table 1), seven isoaccepting tRNAs have 3′ X motifs: Ala, Asn, Gln, Phe, Thr, Tyr and Val, four isoaccepting tRNAs have 5′ X motifs: Gly, Ile, Leu and Pro (Leu has no 5′ X motif in $\mathcal{F}^1_{\mathrm{EUK}}$), and one isoaccepting tRNA has 5′ and 3′ X motifs: Met (Met has no 5′ X motif in $\mathcal{F}^2_{\mathrm{EUK}}$). No preferential occurrence of X motifs is observed in eight eukaryotic isoaccepting tRNAs: Arg, Asp, Cys, Glu, His, Lys, Ser and Trp (Arg, His and Lys have 3′ X motifs in $\mathcal{F}^3_{\mathrm{EUK}}$). Among the 20 isoaccepting tRNAs, four tRNAs have no identified X motifs both in prokaryotes and eukaryotes: Arg, His, Ser and Trp. Interestingly, the four attached amino acids Arg, His, Ser and Trp are not coded by the trinucleotides of the common trinucleotide circular code X (1) (see also Table 4(a) in Arquès and Michel, 1996). The circular code X (1) codes the 12 amino acids Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr and Val (Arquès and Michel, 1996) which are attached for the prokaryotes to tRNAs containing all X motifs, and for the eukaryotes to tRNAs containing X motifs except for the two tRNAs Asp and Glu (random curves observed with this statistical approach).

A circular code property (probability inequality (12)) observed in genes of prokaryotes and eukaryotes (Arquès et al., 1999; Bahi and Michel, 2008) is identified in the 3′ regions of 19 isoaccepting tRNAs of prokaryotes and eukaryotes (except for the tRNA Leu). It suggests that the 3′ regions of tRNAs have a primary structure similar to the primary structure of genes with a "frame" starting from the anticodon: trinucleotides numbered 111, 222, etc. (Fig. 1b). The 5′ regions of tRNAs do not have this gene circular code property which remains unexplained so far.

The identification, using a statistical analysis of large populations of tRNAs, of X motifs of greatest lengths having at least 9 nucleotides which allow to retrieve the reading frame with a probability of 99.9% (Table 3 and Fig. 4 in Michel, 2012) and also of a gene circular code property in tRNAs strengthens the concept proposed in Michel (2012) of a possible translation (framing) code based on a circular code. Short X motifs, i.e. of lengths less than 9 nucleotides, or series of short X motifs may also have a capacity of reading frame retrieval in tRNAs. Indeed, X motifs of 6 nucleotide length retrieve the reading frame with a probability of about 90% (Table 3 and Fig. 4 in Michel, 2012). The statistical analysis of short X motifs in tRNAs should also be investigated in future.

Acknowledgments

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

References

Arquès, D.G., Fallot, J.-P., Marsan, L., Michel, C.J., 1999. An evolutionary analytical model of a complementary circular code. Biosystems 49, 83–103.
Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. Journal of Theoretical Biology 182, 45–58.
Bahi, J.M., Michel, C.J., 2008. A stochastic model of gene evolution with chaotic mutations. Journal of Theoretical Biology 255, 53–63.
Bussoli, L., Michel, C.J., Pirillo, G., 2011. On some forbidden configurations for self-complementary trinucleotide circular codes. Journal for Algebra and Number Theory Academia 2, 223–232.
Bussoli, L., Michel, C.J., Pirillo, G., 2012. On conjugation partitions of sets of trinucleotides. Applied Mathematics 3, 107–112.
Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proceedings of the National Academy of Sciences of the United States of America 43, 416–421.
Gonzalez, D.L., Giannerini, S., Rosa, R., 2011. Circular codes revisited: a statistical approach. Journal of Theoretical Biology 275, 21–28.
Jenner, L.B., Demeshkina, N., Yusupova, G., Yusupov, M., 2010. Structural aspects of messenger RNA reading frame maintenance by the ribosome. Nature Structural and Molecular Biology 17, 555–560.
Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research 25, 955–964.
Michel, C.J., 2008. A 2006 review of circular codes in genes. Computers and Mathematics with Applications 55, 984–988.
Michel, C.J., 2012. Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Computational Biology and Chemistry 37, 24–37.
Michel, C.J., Pirillo, G., Pirillo, M.A., 2008. A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theoretical Computer Science 401, 17–26.
Rozenski, J., Crain, P.F., McCloskey, J.A., 1999. The RNA modification database: 1999 update. Nucleic Acids Research 27, 196–197.
Zaher, H.S., Green, R., 2009. Fidelity at the molecular level: lessons from protein synthesis. Cell 136, 746–762.