Journal of Theoretical Biology 355 (2014) 83–94

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A genetic scale of reading frame coding

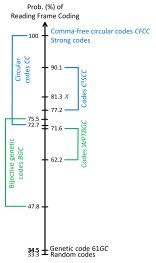
Christian J. Michel

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

HIGHLIGHTS

- Determination of a genetic scale of reading frame coding.
- Trinucleotide circular codes.
- Bijective genetic codes.
- Trinucleotide codes of amino acids.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 6 November 2013
Received in revised form
18 March 2014
Accepted 18 March 2014
Available online 31 March 2014

Keywords:
Comma free-code
Circular code
Bijective genetic code
Random code

ABSTRACT

The reading frame coding (RFC) of codes (sets) of trinucleotides is a genetic concept which has been largely ignored during the last 50 years. A first objective is the definition of a new and simple statistical parameter PrRFC for analysing the probability (efficiency) of reading frame coding (RFC) of any trinucleotide code. A second objective is to reveal different classes and subclasses of trinucleotide codes involved in reading frame coding: the circular codes of 20 trinucleotides and the bijective genetic codes of 20 trinucleotides coding the 20 amino acids. This approach allows us to propose a genetic scale of reading frame coding which ranges from 1/3 with the random codes (RFC probability identical in the three frames) to 1 with the comma-free circular codes (RFC probability maximal in the reading frame and null in the two shifted frames). This genetic scale shows, in particular, the reading frame coding probabilities of the 12,964,440 circular codes (PrRFC = 83.2% in average), the 216 C^3 self-complementary circular codes (PrRFC = 84.1% in average) including the code *X* identified in eukaryotic and prokaryotic genes (PrRFC = 81.3%) and the 339,738,624 bijective genetic codes (PrRFC = 61.5% in average) including the 52 codes without permuted trinucleotides (PrRFC = 66.0% in average). Otherwise, the reading frame coding probabilities of each trinucleotide code coding an amino acid with the universal genetic code are also determined. The four amino acids Gly, Lys, Phe and Pro are coded by codes (not circular) with RFC probabilities equal to 2/3, 1/2, 1/2 and 2/3, respectively. The amino acid Leu is coded by a circular code (not comma-free) with a RFC probability equal to 18/19. The 15 other amino acids are coded by comma-free circular codes, i.e. with RFC probabilities equal to 1. The identification of coding properties in some classes of trinucleotide codes studied here may bring new insights in the origin and evolution of the genetic code.

1. Introduction

The reading frame coding (RFC) of a code (set) of trinucleotides, e.g. the genetic code, is a fascinating and open problem. It is also an old problem. Almost 60 years ago (in 1957), before the discovery of

E-mail address: c.michel@unistra.fr
URL: http://dpt-info.u-strasbg.fr/~c.michel/

http://dx.doi.org/10.1016/j.jtbi.2014.03.029
0022-5193/

the genetic code, a class of trinucleotide 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 could code amino acids. The two questions of interest were: why are there more trinucleotides than amino acids and, how does one choose the reading frame? Crick et al. (1957) proposed that only 20 trinucleotides among 64 code the 20 amino acids. Such a bijective code implies that the coding trinucleotides are found only in one frame. The determination of a set of 20 trinucleotides forming a comma-free code has several constraints:

(i) A periodic permuted trinucleotide, i.e. a trinucleotide with identical nucleotides, must be excluded from such a code. Indeed, the concatenation of AAA with itself, for instance, does not allow the (original) reading frame to be retrieved as there are three possible decompositions: …AAA · AAA · AAA… (original frame), …A · AAA · AAA · AA…, and …AA · AAA · AAA · A…, the concatenation operator " · " showing the adopted decomposition.
(ii) Two non-periodic permuted trinucleotides, i.e. two trinucleotides related to the circular permutation map, e.g. ACG and CGA, must also be excluded from such a code. Indeed, the concatenation of ACG with itself, for instance, does not allow the (original) reading frame to be retrieved as there are two possible decompositions: …ACG · ACG · ACG… (original frame) and …A · CGA · CGA · CG…

Therefore, by excluding the four periodic permuted trinucleotides {AAA, CCC, GGG, TTT} and by gathering the 60 remaining trinucleotides in 20 classes of three trinucleotides such that, in each class, three trinucleotides are deduced from each other by the circular permutation map, e.g. ACG, CGA and GAC, we see that a comma-free code has only one trinucleotide per class and therefore contains at most 20 trinucleotides. This trinucleotide number is identical to the amino acid number, thus leading to a code assigning one trinucleotide per amino acid without ambiguity. A few combinatorial results on trinucleotide comma-free codes were obtained by Golomb et al. (1958a, 1958b). However, no trinucleotide comma-free code was identified in genes statistically. Furthermore, in the late 1950s, the discovery that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes phenylalanine (Nirenberg and Matthaei, 1961), led to the abandonment of the concept of a comma-free code over the alphabet {A, C, G, T}. For several biological reasons, in particular the interaction between mRNA and tRNA, this concept was again taken up later over the purine/pyrimidine alphabet {R, Y} (R = {A, G}, Y = {C, T}) with two trinucleotide comma-free codes for primitive genes: RRY (Crick et al., 1976) and RNY (N = {R, Y}) (Eigen and Schuster, 1978).

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, 4,709,758 trinucleotides) (Arquès and Michel, 1996). This set X contains the 20 following trinucleotides:

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

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 present several strong mathematical properties, particularly the fact that X is a 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 a few nucleotides, whose nucleotide length depends on the circular code, positioned anywhere in a sequence generated with the circular code always retrieves the reading frame (Lassez, 1976; Berstel and Perrin, 1985). For crossing the largest ambiguous words of the circular code X (words, not necessarily unique, in two or three frames), this window needs a length of 13 nucleotides with X (Fig. 3 in Michel, 2012). A window of 13 nucleotide length is the largest window of X, i.e. it allows to retrieve the reading frame for all the ambiguous words of X. Gonzalez et al. (2011), by defining a statistical function analysing the covering capability of a circular code, have recently showed on a gene data set from 13 classes of proteins that the 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; list given in Tables 4a, 5a and 6a in Michel et al., 2008a). A review of this code X gives some additional properties (Michel, 2008). Recently, X motifs, i.e. motifs generated with the circular code X, are identified in the 5′ and/or 3′ regions of 16 isoaccepting tRNAs of prokaryotes and eukaryotes (Michel, 2013). Seven X motifs of length greater or equal to 15 nucleotides are also found in 16S rRNAs, in particular in the decoding center which recognizes the codon–anticodon helix in A-tRNA (Michel, 2012). 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. These results led to the concept of a possible translation (framing) code based on circular code (Michel, 2012).

Comma-free and circular codes have two different definitions in combinatorics (Definitions 5 and 10 below). In fact, these two classes of codes should not be considered as different. Indeed, it was proved recently that a comma-free code is a particular circular code (Proposition 3 in Michel et al., 2008a). Precisely, a hierarchy of circular codes is closed by the strongest ones which are comma-free and the weakest ones which are circular with large "necklaces" (Proposition 4 and Remark 4 in Michel et al., 2008a). Furthermore, it exists as circular codes even stronger than the comma-free codes and called strong circular codes (Michel and Pirillo, 2011). There are 12,964,440 (maximal, i.e. of 20 trinucleotide length) trinucleotide circular codes (Table 2(d) in Arquès and Michel, 1996; growth function in Table 1 in Michel and Pirillo, 2010) which include the 408 comma-free codes (growth function in Table 2a and list in Table 2b in Michel et al., 2008b).

There are two mathematical approaches for proving that a trinucleotide code is circular or not: a classical proof based on the flower automaton (Lassez, 1976; Berstel and Perrin, 1985) and a modern proof, more refined, using the necklaces *5LDCN* (Pirillo, 2003) and *nLDCCN* (Michel and Pirillo, 2010). Indeed, the necklace proof allows not only to decide if a trinucleotide code is circular or not, but also to classify the circular codes. Comma-free codes have their trinucleotides only in reading frame, thus short necklaces, while circular codes (not comma-free) have their trinucleotides in reading frame but also in the two shifted frames 1 and 2, thus large necklaces (see the most general hierarchy given in Proposition 4.1 in Michel and Pirillo, 2011).

The first objective of this study is to define a new and simple statistical parameter PrRFC for analysing the probability (efficiency) of reading frame coding (RFC) of any trinucleotide code C.

The second objective is to reveal different classes of trinucleotide codes C involved in reading frame coding, almost all of them