Available online at www.sciencedirect.com
ScienceDirect
Computers and Mathematics with Applications 55 (2008) 984–988
An International Journal
computers & mathematics
with applications
www.elsevier.com/locate/camwa

A 2006 review of circular codes in genes

Christian J. Michel*

Equipe de Bioinformatique Théorique, LSIIT (UMR CNRS-ULP 7005), Université Louis Pasteur de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France

Abstract

We present a review of circular codes in genes 10 years after they were discovered. After an introduction of the research context, codes are replaced in their historical aspect with the comma-free codes which have been searched but not found in genes over the alphabet {A, C, G, T}. Then, the circular codes identified in genes and genomes are presented according to 3 axes without mathematical notation: identification, evolution and possible function.

Keywords: Circular code; Genome; Gene; Evolution; Analytical model

1. Research context

Each genome has its own trinucleotide distribution [1]. Indeed, the synonymous codons (codons coding for the same amino acid) do not occur with the same frequencies in genes. This synonymous codon usage is biased: a restricted subset of codons is preferred in genes. Codon usage is generally correlated with gene expressivity [2–4] even if its strength varies among bacterial species [5]. Codon choice may depend on its context, i.e. the surrounding nucleotides [6–8]. These pressures might be frame independent [9]. In order to understand all these properties (genetic code, variant genetic codes, codon usage, frame independence), we have studied the trinucleotide occurrences in the 3 frames of genes by different new statistical methods. This approach has led to the identification of particular codes in genes called circular codes [10–13].

By convention, the reading frame established by a start codon (ATG, GTG and TTG) is the frame 0, and the frames 1 and 2 are the reading frame shifted by 1 and 2 nucleotides in the $5'$–$3'$ direction respectively. After excluding the trinucleotides with identical nucleotides (AAA, CCC, GGG and TTT) and by assigning each trinucleotide to a preferential frame, 3 subsets of 20 trinucleotides per frame have been identified in the gene populations of both eukaryotes EUK and prokaryotes PRO [10]. These 3 sets X_0 (EUK_PRO), X_1 (EUK_PRO) and X_2 (EUK_PRO) associated with the frames 0, 1 and 2 respectively have several strong properties, in particular the property of circular code. The circular code concept will be briefly pointed out without mathematical notations after a short historical presentation of an another class of code, i.e. the comma-free code, which has been searched but not found in genes (over the alphabet {A,C,G,T}).

* Tel.: +33 3 90 24 44 62.
E-mail address: michel@dpt-info.u-strasbg.fr.

0898-1221/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2006.12.090

2. History: Comma-free code

A code in genes has been proposed by Crick et al. in 1957 [14] in order to explain how the reading of a series of nucleotides could code for the amino acids constituting the proteins. The 2 problems stressed were: why are there more trinucleotides than amino acids and how to choose the reading frame? Crick et al. [14] have then proposed that only 20 among 64 trinucleotides code for the 20 amino acids. Such a bijective code implies that the coding trinucleotides are found only in one frame. Such a particular code is called a comma-free code or a code without commas [15–18]. However, the determination of a set of 20 trinucleotides forming a comma-free code has several constraints:

(i) A trinucleotide with identical nucleotides must be excluded from such a code. Indeed, the concatenation of AAA with itself, for example, does not allow the reading (original) frame to be retrieved as there are 3 possible decompositions: ... AAA, AAA, AAA, ..., ... A, AAA, AAA, AA ... and ... AA, AAA, AAA, A ..., the commas showing the way of construction (decomposition).

(ii) Two trinucleotides related to circular permutation, for example AAC and ACA, must be also excluded from such a code. Indeed, the concatenation of AAC with itself, for example, also does not allow the reading frame to be retrieved as there are 2 possible decompositions: ... AAC, AAC, AAC, ... and ... A, ACA, ACA, AC ...

Therefore, by excluding AAA, CCC, GGG and TTT, and by gathering the 60 remaining trinucleotides in 20 classes of 3 trinucleotides such that, in each class, 3 trinucleotides are deduced from each other by circular permutations, e.g. AAC, ACA and CAA, 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 one, thus leading to a comma-free code assigning one trinucleotide per amino acid without ambiguity.

Some basic results with comma-free codes have been obtained by Golomb et al. [15,16]. However, the determination of comma-free codes of different lengths and with different properties are unrealizable without a computer. Indeed, there are $\binom{20}{i} \times 3^i$ potential codes of length $i \in \{1, 20\}$. For example, there are 3.5 billions potential codes of length 20. A comma-free code search algorithm demonstrates in particular that among the 408 comma-free codes of 20 trinucleotides, none of them is complementary and none of them is C^3 (see the definition below). Indeed, there are 4 maximal complementary comma-free codes with 16 trinucleotides and 18 maximal C^3 comma-free codes with 16 trinucleotides (unpublished results). Furthermore, in the late fifties, the 2 discoveries that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes for phenylalanine [19] and that genes are placed in reading frames with a particular start trinucleotide, have led to the concept of comma-free code over the alphabet {A,C,G,T} being given up. For several biological reasons, in particular the interaction between mRNA and tRNA, this concept is taken again later over the alphabet {R,Y} (R = purine = A or G, Y = pyrimidine = C or T) with 2 comma-free codes for primitive genes: RRY [20] and RNY (N = R or Y) [21].

3. Circular code

A circular code also allows the reading frames of genes to be retrieved but with weaker conditions compared to a comma-free code. 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 at most one decomposition into words of the circular code [22,17,18,10,23–25]. As an example, let the set X be composed of the 6 following words: $X = \{\text{AAT, ATG, CCT, CTA, GCC, GGC}\}$ and the word w, be a series of the 9 following letters: $w = \text{ATGGCCCTA}$. The word w, written on a circle, can be factorized into words of X according to 2 different ways: ATG, GCC, CTA and AAT, GGC, CCT. Therefore, X is not a circular code. In contrast, if the set Y obtained by replacing the word GGC of X by GTC is considered, i.e. $Y = \{\text{AAT, ATG, CCT, CTA, GCC, GTC}\}$, then there never exists an ambiguous word with Y, in particular w is not ambiguous, and Y is a circular code. The construction frame of a word generated by any concatenation of words of a circular code can be retrieved after the reading, anywhere in the generated word, of a certain number of nucleotides depending on the code. This series of nucleotides is called the window of the circular code.

A comma-free code has conditions stronger than a circular code. Indeed, the 20 trinucleotides of a comma-free code are found only in one frame, i.e. in the reading frame, while some trinucleotides of a circular code can be found in the 2 shifted frames 1 and 2 (see below). On the other hand, the lengths of the windows of a comma-free code and a circular code are less than or equal to 4 and 13 nucleotides respectively.

Definition 1. The (left circular) permutation $\mathcal{P}$ of a trinucleotide $w_0 = l_0l_1l_2, l_0, l_1, l_2 \in \{A, C, G, T\}$, is the permuted trinucleotide $\mathcal{P}(w_0) = w_1 = l_1l_2l_0$, e.g. $\mathcal{P}(AAC) = ACA$, and $\mathcal{P}(\mathcal{P}(w_0)) = \mathcal{P}(w_1) = w_2 = l_2l_0l_1$, e.g. $\mathcal{P}(\mathcal{P}(AAC)) = CAA$. This definition is naturally extended to the trinucleotide set permutation: the permutation $\mathcal{P}$ of a set of trinucleotides is the permuted trinucleotide set obtained by the permutation $\mathcal{P}$ of all its trinucleotides.

The first identified circular code is the set $X(EUK\_PRO) = X_0(EUK\_PRO) = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$ in the frame 0 (reading frame) of genes of eukaryotes EUK and prokaryotes PRO [10]. It has several important properties:

(i) Maximality: $X_0(EUK\_PRO)$ is a maximal circular code (20 trinucleotides) as it is not contained in a larger circular code, i.e. in a circular code with more words.

(ii) Permutation: $X_0(EUK\_PRO)$ generates $X_1(EUK\_PRO)$ by one permutation and $X_2(EUK\_PRO)$ by another permutation, i.e. $\mathcal{P}(X_0(EUK\_PRO)) = X_1(EUK\_PRO)$ and $\mathcal{P}(\mathcal{P}(X_0(EUK\_PRO))) = X_2(EUK\_PRO)$.

(iii) Complementarity: $X_0(EUK\_PRO)$ is self-complementary (10 trinucleotides of $X_0(EUK\_PRO)$ are complementary to 10 other trinucleotides of $X_0(EUK\_PRO)$) and, $X_1(EUK\_PRO)$ and $X_2(EUK\_PRO)$ are complementary to each other (the 20 trinucleotides of $X_1(EUK\_PRO)$ are complementary to the 20 trinucleotides of $X_2(EUK\_PRO)$).

(iv) C^3 code: $X_1(EUK\_PRO)$ and $X_2(EUK\_PRO)$ obtained by permutation of $X_0(EUK\_PRO)$ (property (ii)) are maximal circular codes. It is important to stress that a circular code X_0 does not necessarily imply that X_1 and X_2 obtained by permutation, are also circular codes. Therefore, if $X_0(EUK\_PRO)$, $X_1(EUK\_PRO)$ and $X_2(EUK\_PRO)$ are circular codes, then $X_0(EUK\_PRO)$, $X_1(EUK\_PRO)$ and $X_2(EUK\_PRO)$ are C^3 codes. As the circular code $X_0(EUK\_PRO)$ is coding for the reading frame (frame 0) in genes, i.e. the most important frame, and as it is self-complementary (property (iii)), it is considered as the main C^3 code and noted $X(EUK\_PRO)$ simply.

(v) Rarity: the occurrence probability of the C^3 code $X(EUK\_PRO)$ is equal to $216/3^{20} \approx 6 \times 10^{-8}$, i.e. the computed number of complementary C^3 codes (216) divided by the number of potential codes ($3^{20} = 3486\,784\,401$).

(vi) Flexibility:

(via) The lengths of the minimal windows of $X_0(EUK\_PRO)$, $X_1(EUK\_PRO)$ and $X_2(EUK\_PRO)$ for retrieving automatically the frames 0, 1 and 2 respectively, are all equal to 13 nucleotides and represent the largest window length among the 216 C^3 codes.

(vib) The frequencies of "misplaced" trinucleotides in the shifted frames 1 and 2 are both equal to 24.6%. If the trinucleotides of $X_0(EUK\_PRO)$ are randomly concatenated, for example as follows:

... GAA, GAG, GTA, GTA, ACC, AAT, GTA, CTC, TAC, TTC, ACC, ATC ...

then, the trinucleotides in frame 1:

... G, AAG, AGG, TAG, TAA, CCA, ATG, TAC, TCT, ACT, TCA, CCA, TC ...

and the trinucleotides in frame 2:

... GA, AGA, GGT, AGT, AAC, CAA, TGT, ACT, CTA, CTT, CAC, CAT, C ...

mainly belong to $X_1(EUK\_PRO)$ and $X_2(EUK\_PRO)$ respectively. A few trinucleotides are misplaced in the shifted frames. With this example, in frame 1, 9 trinucleotides belong to $X_1(EUK\_PRO)$ but 1 trinucleotide TAC to $X_0(EUK\_PRO)$ and 1 trinucleotide TAA to $X_2(EUK\_PRO)$ ($\mathcal{P}(TAA) = AAT \in X_0(EUK\_PRO)$). In frame 2, 8 trinucleotides belong to $X_2(EUK\_PRO)$ but 2 trinucleotides GGT and AAC to $X_0(EUK\_PRO)$ and 1 trinucleotide ACT to $X_1(EUK\_PRO)$ ($\mathcal{P}(\mathcal{P}(ACT)) = TAC \in X_0(EUK\_PRO)$). By computing exactly, the frequencies of misplaced trinucleotides in frame 1 are 11.9% for $X_0(EUK\_PRO)$ and 12.7% for $X_2(EUK\_PRO)$. In frame 2, the frequencies of misplaced trinucleotides are 11.9% for $X_0(EUK\_PRO)$ and 12.7% for $X_1(EUK\_PRO)$. The complementarity property (iii) explains on the one hand, the identical frequencies of $X_0(EUK\_PRO)$ in frames 1 and 2 (such words are impossible with a comma-free code), and on the other hand, the identical frequencies of $X_2(EUK\_PRO)$ in frame 1 and $X_1(EUK\_PRO)$ in frame 2 (such words are also impossible with a comma-free code). Then, the frequency sum of misplaced trinucleotides in frame 1 ($X_0(EUK\_PRO)$ and $X_2(EUK\_PRO)$) is equal to the one of misplaced trinucleotides in frame 2 ($X_0(EUK\_PRO)$ and $X_1(EUK\_PRO)$) and is equal to 24.6%. This value is close to the highest frequency (27.9%) of misplaced trinucleotides among the 216 C^3 codes.

(vic) The 4 types of nucleotides occur in the 3 trinucleotide sites of $X_0(EUK\_PRO)$, and also obviously by the permutation property (ii) of $X_1(EUK\_PRO)$ and $X_2(EUK\_PRO)$. It is important to stress that C^3 codes can have missing nucleotides in their trinucleotide sites.

Similarly to the existence of variant genetic codes (compared to the universal one) and different codon usage, several circular codes have been found in genes: 1 code $X\,(MIT)$ in mitochondria [11], 15 codes $X\,(G_{\text{archaea}})$ in archaeal genomes [12] and 72 codes $X\,(G_{\text{bacteria}})$ in 175 complete bacterial genomes [13], several bacterial genomes having the same codes, by using a sensitive statistical method [12].

4. Models of gene evolution

We have developed new analytical models of gene evolution. They can be applied to various problems, in particular to the study of circular code evolution. The most recent analytical evolutionary model is based on a trinucleotide mutation matrix 64×64 with 9 substitution parameters associated with the 3 types of substitutions in the 3 trinucleotide sites [26]. It generalizes the previous models based on the nucleotide mutation matrices 4×4, in particular at 1 substitution parameter [27], 2 parameters (transitions and tranversions) [28], 3 parameters [29], 4 parameters [30] and 6 parameters [29], and based on the trinucleotide mutation matrix 64×64 with 3 and 6 substitution parameters [31, 32]. It determines at some evolutionary time t the exact occurrence probabilities of trinucleotides mutating randomly according to these 9 substitution parameters.

One application of these analytical models has allowed an evolutionary study of the common circular code X (EUK_PRO) and the 15 archaeal circular codes $X\,(G_{\text{archaea}})$ [32]. Very unexpectedly, it demonstrates that the archaeal circular codes can derive from the common circular code subjected to random substitutions with particular values for the substitutions parameters. It has a strong correlation with the statistical observations of 3 archaeal codes in actual archaeal genes. Furthermore, the properties of these substitution rates allow proposal of an evolutionary classification of the 15 archaeal codes into 3 main classes according to this evolutionary model. In almost all the cases, they agree with the actual degeneracy of the genetic code with substitutions more frequent in the third trinucleotide site and with transitions more frequent that transversions in any trinucleotide site.

We have also developed a new class of gene evolution models in which the mutations are time dependent [33]. These models study nonlinear gene evolution by accelerating or decelerating the substitutions rates at different evolutionary times. They generalize the previous analytical models which are based on constant substitution rates.

An application of this nonlinear model allows evolution of the common circular code X (EUK_PRO) to be analysed. Different classes of functions for the substitution parameters have been studied. One among 12 models retrieves after mutation the statistical properties of the common circular code in the 3 frames of actual genes. In this model, the mutation rate in the third trinucleotide site increases exponentially during gene evolution while the mutation rates in the first and second sites decrease exponentially. This property also agrees with the actual degeneracy of the genetic code.

5. Conclusion

The main property of a circular code is the retrieval of the reading frames in genes, both locally, i.e. anywhere in genes and in particular without a start codon, and automatically with a window of a few nucleotides. Furthermore, a C^3 code can retrieve any frame in genes, both locally and automatically. Circular codes could be issued from primitive genes. However, it is still not known to date which biological apparatus could have used these circular codes and which function could have their words in actual genes. Our results lead to the conclusion that there are 2 types of codes in genes: genetic codes for coding the amino acids, the most important one being the universal one, and circular codes for retrieving the reading frames in genes. The different genetic codes have been determined experimentally. The existence of several circular codes in genes is shown by our recent statistical observations in genomes [12,13]. Until now, no experimental proof has been searched for to confirm or reject the concept of circular codes in genes.

References

[1] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pave, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1980) r49–r62.
[2] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, R. Mercier, Codon catalog usage is a genome strategy modulated for gene expressivity, Nucleic Acids Res. 9 (1981) r43–r74.
[3] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 12–34.
[4] P.M. Sharp, G. Matassi, Codon usage and genome evolution, Curr. Opin. Genet. Dev. 4 (1994) 851–860.

[5] P.M. Sharp, E. Bailes, R.J. Grocock, J.F. Peden, R.E. Sockett, Variation in the strength of selected codon usage bias among bacteria, Nucleic Acids Res. 33 (2005) 1141–1153.
[6] M. Yarus, L.S. Folley, Sense codons are found in specific contexts, J. Mol. Biol. 182 (1984) 529–540.
[7] E.G. Shpaer, Constraints on codon context in Escherichia coli genes: Their possible role in modulating the efficiency of translation, J. Mol. Biol. 188 (1986) 555–564.
[8] O.G. Berg, P.J.N. Silva, Codon bias in Escherichia coli: The influence of codon context on mutation and selection, Nucleic Acids Res. 25 (1997) 1397–1404.
[9] M.A. Antezana, M. Kreitman, The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences, J. Mol. Evol. 49 (1999) 36–43.
[10] D.G. Arquès, C.J. Michel, A complementary circular code in the protein coding genes, J. Theoret. Biol. 182 (1996) 45–58.
[11] D.G. Arquès, C.J. Michel, A circular code in the protein coding genes of mitochondria, J. Theoret. Biol. 189 (1997) 273–290.
[12] G. Frey, C.J. Michel, Circular codes in archaeal genomes, J. Theoret. Biol. 223 (2003) 413–431.
[13] G. Frey, C.J. Michel, Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes, J. Comput. Biol. Chem. 30 (2006) 87–101.
[14] F.H.C. Crick, J.S. Griffith, L.E. Orgel, Codes without commas, Proc. Natl. Acad. Sci. USA 43 (1957) 416–421.
[15] S.W. Golomb, B. Gordon, L.R. Welch, Comma-free codes, Canad. J. Math. 10 (1958) 202–209.
[16] S.W. Golomb, L.R. Welch, M. Delbrück, Construction and properties of comma-free codes, Biol. Medd. Dan. Vid. Selsk. 23 (1958).
[17] J. Berstel, D. Perrin, Theory of Codes, Academic Press, New York, 1985.
[18] M.-P. Béal, Codage Symbolique, Masson, Paris, 1993.
[19] M.W. Nirenberg, J.H. Matthaei, The dependence of cell-free protein synthesis in E. Coli upon naturally occurring or synthetic polyribonucleotides, Proc. Natl. Acad. Sci. USA 47 (1961) 1588–1602.
[20] F.H.C. Crick, S. Brenner, A. Klug, G. Pieczenik, A speculation on the origin of protein synthesis, Origins of Life 7 (1976) 389–397.
[21] M. Eigen, P. Schuster, The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle, Naturwissenschaften 65 (1978) 341–369.
[22] J.-L. Lassez, Circular codes and synchronization, Int. J. Comput. System Sci. 5 (1976) 201–208.
[23] J. Lacan, C.J. Michel, Analysis of a circular code model, J. Theoret. Biol. 213 (2001) 159–170.
[24] G. Pirillo, Remarks on the Arques-Michel code, Rivista di Biologia (Biology Forum) 94 (2001) 327–330.
[25] G. Pirillo, M.A. Pirillo, Growth function of self-complementary circular codes, Rivista di Biologia (Biology Forum) 98 (2005) 97–110.
[26] C.J. Michel, An analytical model of gene evolution with 9 mutation parameters: An application to the amino acids coded by the common circular code, Bull. Math. Biol. 69 (2007) 677–698.
[27] T.H. Jukes, C.R. Cantor, Evolution of protein molecules, in: H.N. Munro (Ed.), Mammalian Protein Metabolism, Academic Press, New York, 1969, pp. 21–132.
[28] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, J. Mol. Evol. 16 (1980) 111–120.
[29] M. Kimura, Estimation of evolutionary distances between homologous nucleotide sequences, Proc. Natl. Acad. Sci. USA 78 (1981) 454–458.
[30] N. Takahata, M. Kimura, A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes, Genetics 98 (1981) 641–657.
[31] D.G. Arquès, J.-P. Fallot, C.J. Michel, An evolutionary analytical model of a complementary circular code simulating the protein coding genes, the 5' and 3' regions, Bull. Math. Biol. 60 (1998) 163–194.
[32] G. Frey, C.J. Michel, An analytical model of gene evolution with 6 mutation parameters: An application to archaeal circular codes, J. Comput. Biol. Chem. 30 (2006) 1–11.
[33] J.M. Bahi, C.J. Michel, A stochastic gene evolution model with time dependent mutations, Bull. Math. Biol. 66 (2004) 763–778.