Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computational Biology and Chemistry

Computational Biology and Chemistry 30 (2006) 87–101

www.elsevier.com/locate/compbiolchem

Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes

Gabriel Frey, 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

Received 30 September 2005; received in revised form 7 November 2005; accepted 7 November 2005

Abstract

We developed a statistical method that allows each trinucleotide to be associated with a unique frame among the three possible ones in a (protein coding) gene. An extensive gene study in 175 complete bacterial genomes based on this statistical approach resulted in identification of 72 new circular codes. Finding a circular code enables an immediate retrieval of the reading frame locally anywhere in a gene. No knowledge of location of the start codon is required and a short window of only a few nucleotides is sufficient for automatic retrieval. We have therefore developed a factorization method (that explores previously found circular codes) for retrieving the reading frames of bacterial genes. Its principle is new and easy to understand. Neither complex treatment nor specific information on the nucleotide sequences is necessary. Moreover, the method can be used for short regions in nucleotide sequences (less than 25 nucleotides in protein coding genes). Selected additional properties of circular codes and their possible biological consequences are also discussed.

Keywords: Bacterial genome; Circular code; Frame

1. Introduction

Each bacterial genome has its own trinucleotide distribution (Grantham et al., 1980). Indeed, the synonymous codons (codons coding for the same amino acid) do not occur with the same frequencies in bacterial genes. This synonymous codon usage is biased: a restricted subset of codons is preferred in genes. Codon usage is generally correlated with gene expressivity (Grantham et al., 1981; Ikemura, 1985; Sharp and Matassi, 1994) even if its strength varies among bacterial species (Sharp et al., 2005). A proposed explanation is that codon usage reflects the variation in the concentration of tRNAs. Major codons encoded by more abundant tRNAs should increase translational efficacy (Bulmer, 1991; Akashi and Eyre-Walker, 1998). Nevertheless, tRNA abundance could also have evolved for matching codon pattern in a genome (Fedorov et al., 2002) and then would rather be a consequence of the synonymous codon bias.

Several other processes may influence codon usage (Llopart and Aguade, 2000; Smith and Eyre-Walker, 2001; Konu and Li, 2002; Krakauer and Jansen, 2002; Rogozin et al., 2005). In particular, codon choice may depend on its context, i.e. the surrounding nucleotides (Yarus and Folley, 1984; Shpaer, 1986; Berg and Silva, 1997). These pressures might be frame independent (Antezana and Kreitman, 1999). In this line of research, we have studied the trinucleotide occurrences in the three frames of genes by computing their $3 \times 64 = 192$ frequencies. This approach has led to the identification of particular codes in genes called circular codes.

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, three subsets of 20 trinucleotides per frame have been identified in the gene populations of both eukaryotes EUK and prokaryotes PRO (Arquès and Michel, 1996). These three 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

* Corresponding author. Tel.: +33 3 90 24 44 62.
E-mail addresses: frey@dpt-info.u-strasbg.fr (G. Frey), michel@dpt-info.u-strasbg.fr (C.J. Michel).

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.11.001

briefly pointed out without mathematical notations after a short historical presentation of an another class of code which has been searched but not found in genes (over the alphabet {A,C,G,T}).

A code in genes has been proposed by Crick et al. (1957) in order to explain how the reading of a series of nucleotides could code for the amino acids constituting the proteins. The two problems stressed were: why are there more trinucleotides than amino acids and how to choose the reading frame? Crick et al. (1957) have then proposed that only 20 among 64 trinucleotides code for the 20 amino acids. Furthermore, 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. 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 three possible decompositions: ...AAA,AAA,AAA,..., ...A,AAA,AAA,AA... and ...AA,AAA,AAA,A...

(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 two 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 three trinucleotides such that, in each class, three 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.

The determination of comma-free codes and their properties are unrealizable without computer as there are $3^{20} \approx 3.5$ billions potential codes. A comma-free code search algorithm demonstrates in particular that there are only 408 comma-free codes of 20 trinucleotides. None of them is complementary as the maximal complementary comma-free codes contain only 16 trinucleotides (results not shown). Furthermore, in the late 1950s, the two discoveries that the trinucleotide TTT, an excluded trinucleotide in a comma-free code, codes for phenylalanine (Nirenberg and Matthaei, 1961) and that genes are placed in reading frames with a particular start trinucleotide, have led to give up the concept of comma-free code over the alphabet {A,C,G,T}. For several biological reasons, in particular the interaction between mRNA and tRNA, this concept is taken up again later over the alphabet {R,Y} (R = purine = A or G, Y = pyrimidine = C or T) with two comma-free codes for primitive genes: RRY (Crick et al., 1976) and RNY (N = R or Y) (Eigen and Schuster, 1978).

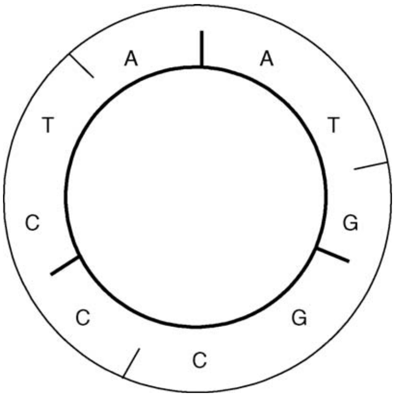

Fig. 1. The set X = {AAT,ATG,CCT,CTA,GCC,GGC} is not a circular code as the word w = ATGGCCCTA, written on a circle, can be factorized into words of X according to two different ways: ATG, GCC, CTA (thick line) and AAT, GGC, CCT (thin line).

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. As an example, let the set X be composed of the six following words: X = {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 X according to two different ways: ATG, GCC, CTA and AAT, GGC, CCT (Fig. 1). Therefore, X is not a circular code. In contrast, if the set $\tilde{X}$ obtained by replacing the word GGC of X by GTC is considered, i.e. $\tilde{X}$ = {AAT, ATG, CCT, CTA, GCC, GTC}, then there never exists an ambiguous word with $\tilde{X}$, in particular w is not ambiguous, and $\tilde{X}$ 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 W 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 two shifted frames 1 and 2 (see below). On the other hand, the lengths of the windows W of a comma free code and a circular code are less than or equal to 4 and 13 nucleotides respectively (Section 2.2.4).

Definition of the trinucleotide (left circular) permutation: the (left circular) permutation 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 $\mathsf{P}(w_0) = w_1 = l_1l_2l_0$, e.g. P(AAC) = ACA, and $\mathsf{P}(\mathsf{P}(w_0)) = \mathsf{P}(w_1) = w_2 = l_2l_0l_1$, e.g. $\mathsf{P}(\mathsf{P}$(AAC)) = CAA. This definition is naturally extended to the trinucleotide set permutation: the permutation P of a set of trinucleotides is the permuted trinucleotide set obtained by the permutation P of all its trinucleotides.

The first identified circular code is the set X_0(EUK_PRO) = {AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC,GAG,