Journal of Theoretical Biology 389 (2016) 206–213

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Maximal dinucleotide comma-free codes

Elena Fimmel *, Lutz Strüngmann

Institute of Mathematical Biology, Faculty of Computer Sciences, Mannheim University of Applied Sciences, 68163 Mannheim, Germany

HIGHLIGHTS

- Complete classification of all maximal dinucleotide comma-free codes.
- Any max. dinucleotide comma-free code is induced by a max. dinucleotide circular code.
- Any max. dinucleotide circular code contains exactly 3 max. dinucleotide comma-free codes.
- Comparison with the situation for trinucleotide codes.
- The results discussed with respect to the Crick's hypothesis on comma-free codes.

ARTICLE INFO

Article history:
Received 22 May 2015
Received in revised form
16 October 2015
Accepted 19 October 2015
Available online 10 November 2015

Keywords:
Genetic code
Circular code
Comma-free code
Dinucleotides

ABSTRACT

The problem of retrieval and maintenance of the correct reading frame plays a significant role in RNA transcription. Circular codes, and especially comma-free codes, can help to understand the underlying mechanisms of error-detection in this process. In recent years much attention has been paid to the investigation of trinucleotide circular codes (see, for instance, Fimmel et al., 2014; Fimmel and Strüngmann, 2015a; Michel and Pirillo, 2012; Michel et al., 2012, 2008), while dinucleotide codes had been touched on only marginally, even though dinucleotides are associated to important biological functions. Recently, all maximal dinucleotide circular codes were classified (Fimmel et al., 2015; Michel and Pirillo, 2013). The present paper studies maximal dinucleotide comma-free codes and their close connection to maximal dinucleotide circular codes. We give a construction principle for such codes and provide a graphical representation that allows them to be visualized geometrically. Moreover, we compare the results for dinucleotide codes with the corresponding situation for trinucleotide maximal self-complementary C^3-codes. Finally, the results obtained are discussed with respect to Crick's hypothesis about frame-shift-detecting codes without commas.

1. Introduction

In 1957, Crick et al. (1957) suggested a new idea for the explanation of frame-retrieval in the process of protein synthesis and introduced 'codes without commas' or, ambiguously, comma-free codes. The strong property of such codes is the immediate recognition of a wrong reading frame in RNA transcription. The retrieval and maintenance of a correct reading frame is essential in protein synthesis, since a frame-shift usually leads to abortion of the reading process. Comma-free trinucleotide codes are sets of codons (sequences of three nucleotide bases) which are meaningful only in the right frame, i.e. any out-of-frame sequence of codons from such a code consists solely of codons outside of the code. Thus, in genetic information written only with words from a comma-free code, a window of three nucleotide bases (i.e. one codon) suffices to detect the correct reading frame. Unfortunately, shortly after the discovery of the standard genetic code (Nirenberg and Matthaei, 1961), it was shown (see, for instance, Bollenbach et al., 2007) that the theory by Crick et al. is not valid in the form it was postulated since, among other things, it is impossible to encode all 20 amino acids using comma-free codes (see e.g. Fimmel and Strüngmann, 2015a for a theoretical discussion of this fact). Nevertheless, comma-free codes were mathematically well studied in the next decades (Eastman, 1965; Golomb et al., 1958; Lam, 2003; Pearson, 2003) and became a subject of investigation within coding theory. However, until quite recently, they were essentially discarded from research in theoretical biology for the reasons given above, but they have become an object of intensive investigations again (Michel et al., 2012, 2008).

The motivation for this resurrection came from the discovery of so-called *circular codes*, which form a weaker version of comma-free codes. While comma-free codes are self-synchronising in

* Corresponding author. Tel.:+49 6212926243; fax: +49 6212926237.
E-mail addresses: e.fimmel@hs-mannheim.de (E. Fimmel), l.struengmann@hs-mannheim.de (L. Strüngmann).

http://dx.doi.org/10.1016/j.jtbi.2015.10.022
0022-5193/

some sense, trinucleotide circular codes may need a longer window (more than one codon) to discover a frame-shift error. In fact, any message in the correct reading frame and consisting of words from a circular code does not allow a second decomposition in a different frame when arranged circularly, i.e. the last letter of the message is connected to the first one like on a circle. The first *natural* circular code was discovered empirically by Arquès and Michel (1996) in coding sequences of both eukaryotes and prokaryotes. The so-called X_0-code had, in addition to its own error-detecting property, the nice feature that it is maximal (consists of 20 codons), self-complementary (i.e. the anticodon of every codon from X_0 is also in the code), and C^3, this latter meaning that the sets of circularly shifted codons again form maximal circular codes. Hence the discovered X_0-code is not only error-detecting in the normal reading frame, but also in frames 1 and 2. A computer calculation showed that there are exactly 216 such maximal, self-complementary C^3-codes, but so far no construction algorithm for them is known. Due to their appearance in natural coding sequences, circular codes have been studied intensively recently by several authors, in order to get a better insight into error-detecting mechanisms in RNA-translation process (see Fimmel et al., 2014; Michel and Pirillo, 2013, 2012; Michel et al., 2012; Michel, 2012). However, although these codes possess a lot of symmetries, they are highly complex and difficult in their structure.

The desire to understand nature's use of circular codes for possibly solving the frame-shift problem and their connection to Crick's comma-free codes, put the focus of most studies on trinucleotide codes. However, dinucleotide codes are important as well since on the one hand, they can be seen as a pre-step to trinucleotides, and on the other hand they play a considerable role in, for instance, attempts to explain evolution (Patel, 2005; Shepherd, 1981) or the degeneracy of the genetic code (Rumer, 1969) or the building of dichotomic classes of codons (Fimmel et al., 2013; Giannerini et al., 2012). For example, Y.B. Rumer figured out in Rumer (1969) (see also Fimmel and Strüngmann, 2015c) that dinucleotides built from the first two bases of a codon uniquely determine whether or not the third base is still necessary for detecting the corresponding amino acid. He called these dinucleotides *roots*. Moreover, recently, two variants of the genetic codes from which the standard genetic code could have originated were found (José et al., 2007). Both codes are based on dinucleotides of the form RY, where R stands for purine, Y for pyrimidine. For these reasons, maximal circular dinucleotide codes were completely classified in Michel and Pirillo (2013) and later on, but using different arguments and techniques, in Fimmel et al. (2015). In contrast to the trinucleotide case, an explicit description and construction principle can be given in this case. Moreover, a geometric approach developed in Fimmel et al. (2015) can be used to highlight their symmetry properties.

In the present paper we investigate maximal dinucleotide comma-free codes and their connection to maximal dinucleotide circular codes. A complete classification of all maximal dinucleotide comma-free codes is given using two approaches: it will be shown that any maximal dinucleotide comma-free code must be of length 5 and can be obtained by removing a single dinucleotide from a maximal dinucleotide circular code. In fact, any maximal dinucleotide circular code contains exactly three maximal dinucleotide comma-free codes and dually, any maximal dinucleotide comma-free code can be included in exactly two maximal dinucleotide circular codes. Moreover, a graph representation will be used to visualize this connection geometrically.

The results obtained might be seen as another hint to an evolutionary process conjectured in Fimmel and Strüngmann (2015b). The authors explained in Fimmel and Strüngmann (2015b) that there are several theories about ancient genetic code tables that are predecessors of the current standard genetic code table but were coding for less amino acids than today and were using less codons than the 64 possible ones, e.g. the primeval code, the RNY-code, SNS-code, and the NNS-code, that have been discussed in Di Giulio (2008, 2001), José et al. (2007), and Jolivet and Rothen (2001). As was pointed out in Fimmel and Strüngmann (2015b), in all these theories the full genetic code contained a large comma-free subcode that encoded all except for one of the present amino acids. This was discussed in Fimmel and Strüngmann (2015b) as a hint for the hypothesis that in ancient codes Crick's (Crick et al., 1957) hypothesis might have been true in a weaker form and as soon as the genetic code became more complex and difficult, nature passed from comma-free codes to circular codes paying the price of longer error detecting windows. Our results indicate that in the dinucleotide world this process could have been easily achieved by combining comma-free codes to circular ones while in the trinucleotide world the situation is less obvious. However, there certainly is a relation between trinucleotide comma-free codes and trinucleotide circular codes that still has to be explored.

2. Comma-free codes in the dinucleotide case

As usual we will denote by $\mathcal{B}=\{A,C,G,T\}$ the set of *nucleotide bases* and by $S_{\mathcal{B}}$ the *group of all permutations* of $\mathcal{B}$ with composition as operation. The set of dinucleotides is given by

$$\mathcal{B}^2=\{AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT\}$$

and has 16 elements.

The *reversing* (*mirroring*) *permutation* $\overleftarrow{\ }:\mathcal{B}^2\to\mathcal{B}^2$ of the dinucleotide bases acts as

$$\overleftarrow{N_1N_2}=N_2N_1 \quad \text{for } N_1,N_2\in\mathcal{B}.$$

We will call $N_1N_2, N_1,N_2\in\mathcal{B}$ and $\overleftarrow{N_1N_2}=N_2N_1$ *conjugated* (compare Michel and Pirillo, 2013) dinucleotides.

Definition 2.1. A subset D of the set of dinucleotides $\mathcal{B}^2$ is called a *comma-free* code if given any two dinucleotides $d_1=N_1N_2, d_2=N_3N_4\in D, N_i\in\mathcal{B}$ the dinucleotide N_2N_3 does not belong to D.

The intention of the definition above is to obtain codes that detect a frame-shift in sequences of dinucleotides immediately after reading the first dinucleotide. Clearly, the only possible frame-shift that can be detected is a shift of 1 basis. Moreover, it is obvious that any comma-free code is also circular. Recall, that a subset D of the set of dinucleotides $\mathcal{B}^2$ is called *circular* if any word over D has exactly one decomposition over D when read on a circle (see, for instance, Fimmel et al., 2015).

Remark 2.2. Obviously, a comma-free dinucleotide code D cannot contain the periodic dinucleotides AA,CC,GG,TT and cannot contain at the same time two conjugated dinucleotides N_1N_2 and $N_2N_1, N_1,N_2\in\mathcal{B}$ according to the definition above. A code $D\subseteq\mathcal{B}^2$ having this property is called *1-circular* (see Fimmel et al., 2015). Thus, a *dinucleotide comma-free code* can contain at most $(16-4)/2=6$ dinucleotides.

Example 2.3 (*Fimmel et al. (2015)*). Consider the maximal dinucleotide circular code

$$D=\{AC,AG,AT,CG,CT,GT\}.$$

The code is not comma-free since, for the concatenation

$$ACGT,$$