Journal of Theoretical Biology 336 (2013) 221–230

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

On dichotomic classes and bijections of the genetic code

Elena Fimmel [a,*], Alberto Danielli [b], Lutz Strüngmann [a]

[a] Institute of Applied Mathematics, Faculty of Computer Sciences, Mannheim University of Applied Sciences, 68163 Mannheim, Germany
[b] Department of Pharmacy and BioTechnology, University of Bologna, 40126 Bologna, Italy

HIGHLIGHTS

- A general algorithm underlying significant dichotomic partitions of the code is presented.
- The complementarity to codons arises as a coherent dichotomy of this algorithm.
- The algorithm mirrors operations having a real counterpart in the decoding center of the ribosome.
- Dichotomic classes are very symmetric with respect to the bijective transformations in the code.

ARTICLE INFO

Article history:
Received 20 January 2013
Received in revised form
5 June 2013
Accepted 25 July 2013
Available online 26 August 2013

Keywords:
Dichotomic class
Rumer class
Transformations of the nucleotide bases
Genetic code

ABSTRACT

Dichotomic classes arising from a recent mathematical model of the genetic code allow to uncover many symmetry properties of the code, and although theoretically derived, they permitted to build statistical classifiers able to retrieve the correct translational frame of coding sequences. Herein we formalize the mathematical properties of these classes, first focusing on all the possible decompositions of the 64 codons of the genetic code into two equally sized dichotomic subsets. Then the global framework of bijective transformations of the nucleotide bases is discussed and we clarify when dichotomic partitions can be generated. In addition, we show that the parity dichotomic classes of the mathematical model and complementarity dichotomic classes obtained in the present article can be formalized in the same algorithmic way the dichotomic Rumer's degeneracy classes. Interestingly, we find that the algorithm underlying dichotomic class definition mirrors biochemical features occurring at discrete base positions in the decoding center of the ribosome.

1. Introduction

The genetic code is a set of 64 lemmas, called codons, encompassing all the combinations of four letters, or nucleotide bases (G, A, U, C), in groups of three (the number of bases per codon). The genetic code instructs the translational machinery to correctly incorporate one of the 20 naturally occurring amino acids into a growing polypeptide chain, using the information encoded in the mRNA nucleotide sequence as template. In this decoding process tRNAs are used as adaptor molecules, charged on one end with the pertinent amino acid, while providing on another site a specific anticodon sequence, capable to base pair with the codon.

The genetic code was cracked in the 1960s by a series of enlightening experiments, proving the absence of commas or gaps (i.e. each base in the coding sequence of the mRNA is part of a codon), as well as the non-overlapping nature of the code (i.e. each base in the coding sequence belongs to only one codon). The mapping of 64 codons to 20 amino acids (and 3 stop signals) implicates that the code is highly degenerate (Crick, 1968; Woese, 1965). Part of the degeneracy of the code is provided by isoaccepting tRNAs that are charged with the same amino acid but recognize different (synonymous) codons. The remaining degeneracy derives by wobble pairing, in which the third base of a codon is allowed to form a shifty non-Watson-Crick base pair with the anticodon (Agris et al., 2007).

The study of the degeneracy of the genetic code and its biological implications have been subject of intensive research over the years, and it is still debated whether the natural code is a frozen accident or a honed product of evolution, or perhaps both. However, several studies provided evidence of inherent advantages of the natural code over other possible ones, suggesting a non-randomness of its degeneracy (Vogel, 1998; Itzkovitz and Alon, 2007; Freeland et al., 2000; Di Giulio, 2005).

One of the first to address these questions from a theoretical point of view was the Russian physicist Rumer (1969). He showed that codons can be divided into two classes of 32 codons each: the first Rumer class identifies amino acids with degeneracy 4 (for which first two bases of the triplet are sufficient to define

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

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.07.027

unambiguously the amino acid), while the second one specifies amino acids with degeneracy non-4 (i.e. 1, 2 or 3). One fundamental aspect behind the duality of Rumer's classes was the identification of a global bijective transformation of bases

$$r : \{A, C, G, U\} \to \{A, C, G, U\}$$

$$A \mapsto C$$

$$C \mapsto A$$

$$G \mapsto U$$

$$U \mapsto G$$

able to transform respectively degeneracy 4 class codons into degeneracy non-4 class ones, and viceversa. We will abbreviate this transformation, which is called *Rumer's transformation*, by

$$r : U, C, A, G \to G, A, C, U.$$

Later Jestin and Soulé noticed that r is not the unique bijective transformation of the set of codons which converts one Rumer's class into the other (see Jestin and Soulé, 2007; Jestin, 2006). However the bijective transformation they found acts different in different bases of the codon, hence it is not induced by a transformation on the set of bases.

Recently, the degeneracy distribution of the code has been addressed by an interesting mathematical model based on non-power representation of integer numbers (Gonzalez et al., 2008). In this model, a 6 digit binary string is assigned to each of the 64 codons together with an integer number from 0 to 23, identifying the corresponding amino acid or stop codon. It has been demonstrated that the unique 6 digit set of non-power bases 1, 1, 2, 4, 7, 8 can exactly represent the degeneracy distribution of the genetic code (Gonzalez, 2008). The mathematical properties of these length-6 binary strings led to the definition of dichotomic classes, i.e. non-linear functions of the information of two adjacent bases, that are intimately linked with the chemical properties of the codon bases, as well as with the duality of Rumer's codon classes. In particular, it was shown that the parity of a codon, defined as the parity of the associated binary string, can be obtained from the chemical class (weak/strong, keto/amino, purine/pyrimidine) of the last two bases of the codon. The transformation of bases exchanging A and G as well as C and U converts one parity class into another one:

$$p : U, C, A, G \to C, U, G, A.$$

Furthermore it has been shown that the same algorithmic rules defining the parity of a codon, also hold for the determination of the Rumer's degeneracy class, when the algorithm is applied to the first and the second base of the codon ($b_2 \to b_1$), instead of the second and the third one ($b_3 \to b_2$). Consequently, it has been shown that a third dichotomic class, the hidden class, can be computed by shifting the dinucleotide window further upstream, with the algorithm rules operating on the first base of one codon and the last base of the previous one (Gonzalez et al., 2008; Giannerini et al., 2012).

Statistically significant short-range correlations between specific combinations of dichotomic classes have been observed in coding sequences (Gonzalez et al., 2008). Strikingly, such correlations appear universal, as they have been detected in a set of both prokaryotic and eukaryotic sequences, irrespectively of their base composition or GC content . In addition, it has been shown that the three dichotomic classes (parity, hidden and Rumer) are linked to the Klein V group through the set of global transformations (c, p, r) acting on a codon, highlighting almost periodic structures related to the short-range organization of coding sequences, in analogy with the properties of quasi-crystals (Gonzalez et al., 2008; Giannerini et al., 2012).

Thus, the mathematical model opened unexpected perspectives to investigate the degeneracy and encoding of the genetic code through the lens of non-linear classifiers represented by the dichotomic classes, both from the theoretical and from the biological point of view. In fact, the short-range correlations between dichotomic classes have been successfully exploited to build statistical classifiers able to retrieve the correct translational frame of coding sequences, using the information contained in only nine codons (Giannerini et al., 2012). This is a strong indication that, together with the renowned Rumer dichotomy, the parity and the hidden classes, although theoretically derived, reflect a local informational structure of the reading frame that may be of great interest to molecular biologists and bioinformaticians.

In the present paper we contribute to the formalization of the model with a detailed analysis of the mathematical properties arising when the dichotomic classes are defined circularly on the codon. We will show in Section 2 that the codon complementarity (corresponding to the anticodon sequence read in $3' > 5'$ direction) can be obtained analogously to the Rumer and parity dichotomic classes discussed above, if we order the dinucleotide bases of a codon circularly ($b_3 \to b_2 \to b_1 \to b_3$).

A general algorithm underlying all three dichotomies will be presented and illustrated in Section 3. This algorithm, derived from that of Gonzalez and collaborators (Gonzalez et al., 2008), classifies codons into dichotomic classes based on two questions that are successively asked about two out of the three nucleotide bases in a codon. There are three kinds of such questions: the first one relates to energy (number of hydrogen bonds) formed between complementary codon and anticodon bases (weak/strong; $W/S; \{A, U\}/\{G, C\}$), the second one asks for the presence of an electron donor or acceptor group in the nitrogen atom of the base (keto/amino groups; $K/Am; \{A, C\}/\{U, G\}$), the third one asks for a space configuration of the base (purine/pyrimidine; $R/Y; \{A, G\}/\{U, C\}$). We will show that the algorithm enquires for these different chemical/physical questions at discrete base positions of the codon, and show that these theoretical questions have a biological counterpart in the way the ribosome interacts with the codon–anticodon duplex in the decoding center.

Finally, it will be shown in Section 4 which transformations of the nucleotide bases can induce dichotomic partitions of sequences of nucleotide bases (not only triplets). We will determine the number of such possible partitions and it is shown that the same partition may be induced by several transformations.

2. Dichotomic classes: complementarity dichotomy

Let us denote the ribonucleotide bases alphabet as

$$\mathcal{B} := \{U, C, A, G\}.$$

Easy combinatorics show that there are exactly $C(4;2)/2 = 3$ possibilities to partition $\mathcal{B}$ into two disjoint parts of equal size:

$$\mathcal{B} = \{C, G\} \cup \{A, U\}; \quad \mathcal{B} = \{C, A\} \cup \{U, G\}; \quad \mathcal{B} = \{C, U\} \cup \{A, G\}$$

where $C(4;2)$ is the binomial coefficient. Remarkably, each of this partitions has a biological or chemical meaning: The first partition divides the set into ‘strong’ and ‘weak’ bases, the second divides the set into amino and keto nucleotide bases and the third partition into pyrimidines and purines. In Gonzalez et al. (2008) it was shown that the parity and Rumer's dichotomic classes arise the same algorithmic way. To define the parity classes we consider the last two bases of a codon and classify it asking if the last base is