BioSystems 128 (2015) 9–18

Contents lists available at ScienceDirect

BioSystems

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

On models of the genetic code generated by binary dichotomic algorithms

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

[a] Mannheim University of Applied Sciences, Institute for Medical Informatics, Paul-Wittsack-Straße 10, D-68163 Mannheim, Germany
[b] Mannheim University of Applied Sciences, Institute for Applied Mathematics, Paul-Wittsack-Straße 10, D-68163 Mannheim, Germany
[c] University of Bologna, Department of Pharmacy and Biotechnology, Via Irnerio 42, 40126 Bologna, Italy

ARTICLE INFO

Article history:
Received 30 September 2014
Received in revised form
24 November 2014
Accepted 16 December 2014
Available online 18 December 2014

Keywords:
Genetic code
Dichotomic partition
Rumer
Scan algorithm
Ribosome

ABSTRACT

In this paper we introduce the concept of a BDA-generated model of the genetic code which is based on binary dichotomic algorithms (BDAs). A BDA-generated model is based on binary dichotomic algorithms (BDAs). Such a BDA partitions the set of 64 codons into two disjoint classes of size 32 each and provides a generalization of known partitions like the Rumer dichotomy. We investigate what partitions can be generated when a set of different BDAs is applied sequentially to the set of codons. The search revealed that these models are able to generate code tables with very different numbers of classes ranging from 2 to 64. We have analyzed whether there are models that map the codons to their amino acids. A perfect matching is not possible. However, we present models that describe the standard genetic code with only few errors. There are also models that map all 64 codons uniquely to 64 classes showing that BDAs can be used to identify codons precisely. This could serve as a basis for further mathematical analysis using coding theory, for example. The hypothesis that BDAs might reflect a molecular mechanism taking place in the decoding center of the ribosome is discussed. The scan demonstrated that binary dichotomic partitions are able to model different aspects of the genetic code very well.

The search was performed with our tool Beady-A. This software is freely available at http://mi.informatik.hs-mannheim.de/beady-a. It requires a JVM version 6 or higher.

1. Introduction

Almost all organisms use the same mapping of codons to amino acids – the universal genetic code (see Fig. 1(b)), which maps 64 codons to only 20 amino acid species (and 3 stop signals). There are 18 alternative genetic code tables annotated which are slightly different in the assignment of the codons (Osawa et al., 1992; Jukes and Osawa, 1993).[1] Nevertheless, these minor variations largely preserve the redundancy in amino acid assignment. Hence the degeneracy of orthologous genetic codes is highly conserved. It has been shown that there are more than 10^{84} different genetic codes with the same degeneracy and it was demonstrated that the genetic code is far from being random (Koonin and Novozhilov, 2009). Besides several hypotheses that explain why the genetic code could have been developed as it is (i. e. the frozen accident theory developed by Crick (Crick, 1968; Sella and Ardell, 2006), the stereo-chemical-theory (originally proposed by Gamow, 1954), the co-evolution theory (Wong, 1975; Giulio, 2008) or the adaptive theory (Haig and Hurst, 1991; Freeland et al., 2003), ongoing research still addresses the question whether its structure reflects informational properties useful for error detection or frame retrieval in translation (Crick et al., 1957; Arquès and Michel, 1996; Seligmann, 2007; Guilloux and Jestin, 2012). Recent mathematical models addressed the structure of the genetic code using dichotomic partitions or dichotomic classes, classifying the codons into two partitions of equal size. The dichotomic classes and their non-random correlations were proposed to contribute to frame retrieval and error detection properties, underpinning a robustness of the code (Giannerini, 2012).

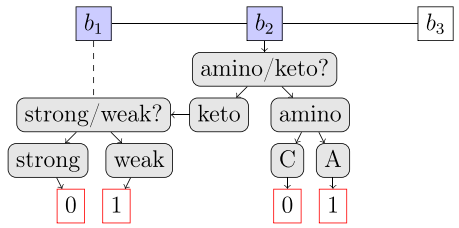
The Russian theoretical physicist Rumer was the first to introduce dichotomic partitions in 1966 (Rumer, 1966). He found that the set of codons can be divided into two subsets of the same cardinality such that (1) for codons belonging to the first subset (represented as digit 0 or class H_0) the first two positions within a codon are sufficient to determine the corresponding amino acid

* Corresponding author. Tel.: +49 621 292 6246.
E-mail addresses: m.gumbel@hs-mannheim.de (M. Gumbel), e.fimmel@hs-mannheim.de (E. Fimmel), alberto.danielli@unibo.it (A. Danielli), l.struengmann@hs-mannheim.de (L. Strüngmann).

[1] See also online at http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c.

http://dx.doi.org/10.1016/j.biosystems.2014.12.001
0303-2647/

a)



b)

	U		C		A		G		
U	Phe	**111**	Ser	001	Tyr	101	Cys	**111**	U
U	Phe	**111**	Ser	001	Tyr	101	Cys	**111**	C
U	Leu	**110**	Ser	010	!	110	!	**110**	A
U	Leu	**100**	Ser	000	!	100	Trp	**100**	G
C	Leu	010	Pro	000	His	**100**	Arg	010	U
C	Leu	010	Pro	000	His	**100**	Arg	010	C
C	Leu	010	Pro	010	Gln	**110**	Arg	010	A
C	Leu	000	Pro	000	Gln	**100**	Arg	000	G
A	Ile	**111**	Thr	001	Asn	101	Ser	**111**	U
A	Ile	**111**	Thr	001	Asn	101	Ser	**111**	C
A	Ile	**110**	Thr	010	Lys	**110**	Arg	**110**	A
A	Met	**100**	Thr	000	Lys	**100**	Arg	**100**	G
G	Val	011	Ala	001	Asp	101	Gly	011	U
G	Val	011	Ala	001	Asp	101	Gly	011	C
G	Val	011	Ala	011	Glu	**111**	Gly	011	A
G	Val	001	Ala	001	Glu	**101**	Gly	001	G

c)

BDA	(i_1, i_2)	Q_1	Q_2
Rumer	(2, 1)	(C, A)≙Am	{C, G}≙ S
Parity	(3, 2)	(G, A)≙ R	{C, A}≙ Am
Complementary	(1, 3)	(C, G)≙ S	{G, A}≙ R

Fig. 1. (a) Visual representation of the Rumer-class algorithm (adapted from Fimmel et al., 2013). (b) Universal genetic code and three overlapping dichotomic partitions generating 8 classes (shown in different colors). A cell contains the dichotomic classes (from left to right) Rumer, Parity and Complementary. 0: H_0, 1: H_1. All digits 1 in the Rumer's class are set in bold. (c) Parameters for the three BDAs. Am: Amino, etc. as explained in the text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

while (2) for codons from the second dichotomic class (digit 1, H_1) the third position is indispensable for it (see Fig. 1(b)). These partitions have also the remarkable feature that a codon of one class can be swapped to the other class and vice versa by a bijective transformation of their bases: $A \leftrightarrow C$ and $U \leftrightarrow G$. Rumer also gave an algorithmic description how the Rumer-class can be determined for a given codon (see Fig. 1(a)).

Gonzalez and co-authors adopted Rumer's algorithm to their model of the genetic code and defined in an analogous way two algorithms for the so called Parity and Hidden classes which matches in a natural way to their model of the genetic code (Gonzalez, 2008). The model of the genetic code developed in Gonzalez (2008) uses a non-power representation of numbers to explain how 64 codons can be mapped to 20 amino acids. Each codon is uniquely represented as a binary string of length 6 (referring to $2^6 = 64$ codons). The Parity class, for instance, is defined as the parity of the associated binary string. It was demonstrated that there are short-range correlations of the Parity and Hidden classes in genes (Giannerini, 2012; Gonzalez et al., 2008).

Interestingly, each of the three algorithms for the calculation of the Rumer-, Parity- and Hidden-class derive their decision from biochemical properties of the bases involved. The algorithms distinguish whether a base is of type purine (denoted as R = {A, G}) or pyrimidine (Y = {C, U}), keto (K = {U, G} or amino (Am = (C, A), or strong (S = {C, G}) or weak (W = {A, U}).

As an example, Fig. 1(a) shows a visual representation of the algorithm producing the Rumer classes. As one can see, the algorithm first determines if the second base of the codon belongs to the amino or keto class and, if not, it considers the first base of the codon and asks whether it is strong or weak.

The algorithmic approach for the definition of dichotomic partitions was recently further investigated and generalized in Fimmel et al. (2013) where the concept of a **Binary Dichotomic Algorithm** (in short: BDA) was introduced and shown to have a tentative biological counterpart in the decoding center of the ribosome. It has been shown that different dichotomic partitions can be linked to the chemical nature of the nucleoside bases which carry the information in the codon (e. g. S/W, Keto/amino; R/Y) (Giannerini (2012)). From an evolutionary point of view this could be of relevance for a decoding process, given that these chemical dichotomies are readily discriminable by biochemical as well as inorganic means, and therefore could provide a simple basis for deciphering the information.

Apparently, partitions and in particular dichotomic partitions like Rumer play an important role in models of the genetic code and it is therefore interesting to study BDAs in general (Hervé Seligmann, 2014; Seligmann, 2014; Demeshkina et al., 2012). The decisions which the algorithms make are binary, i. e. this is a very simple and yet elegant way to make a decision which can be performed on a biological entity like the ribosome or while recognizing a tRNA and its cognate amino acid. This paper investigates the possibility of explaining the structure of the genetic code as an overlapping of dichotomic partitions as generated by a set of BDAs. For instance, Fig. 1(b) shows a code table where the overlapping of the Rumer, Parity and Complementary partition leads to an assignment of the codons into 8 classes represented as a binary string of size 3.

Herein, we propose a new way to create novel models for the genetic code based on overlappings of dichotomic partitions. We will analyze by means of the software Beady-A what kind of code tables of this kind can be generated and whether they are suitable to model aspects of the genetic code. To this aim, different scan algorithms were developed to find solutions suitable to model the genetic code through BDAs.

2. Mathematical background

2.1. Preliminary definitions and notations

In the sequel B= A,C,G,U(T) will denote the set of four nucleotide bases Uracil (Thymine), Cytosine, Adenine, and Guanine, in short $U(T)$, C, A, G. A codon is an element of $\mathcal{B}^3$, e.g. ACU.

In Fimmel et al. (2013) the notion of binary dichotomic algorithms was given taking sequences of nucleotide bases of arbitrary length as input. For our investigations, it suffices to restrict ourselves to codons, i. e. sequences of length three. Let us recall the definition of a BDA in this special case:

Definition 2.1. An ordered pair (H_0, H_1) of subsets $H_0, H_1 \subseteq \mathcal{B}^3$ is called a **dichotomic partition** of $\mathcal{B}^3$ if $H_0 \cap H_1 = \emptyset$, $H_0 \cup H_1 = \mathcal{B}^3$ and $|H_0| = |H_1|$.

In other words, the set of 64 codons is divided into two disjoint subsets of equal size as shown in Fig. 1(b). The next definition shows the algorithmic way to obtain dichotomic partitions as discussed in Fimmel et al. (2013).

Definition 2.2. Let (H_0, H_1) be a dichotomic partition of $\mathcal{B}^3$. We call an algorithm $\mathcal{A}$ a **binary dichotomic algorithm (BDA) with dichotomic partition** (H_0, H_1) if it follows the following scheme:

$\mathcal{A}$ chooses two indices $i_1, i_2 \in \{1, 2, 3\}$ with $i_1 \neq i_2$, an ordered pair of different nucleotide bases $Q_1 = (B_1, B_2)$ and a subset $Q_2 \subset \mathcal{B}$ with $|Q_2| = 2$. Now $\mathcal{A}$ classifies $c = (b_1, b_2, b_3) \in \mathcal{B}^3$ as follows: