BioSystems 141 (2016) 10–19

Contents lists available at ScienceDirect

BioSystems

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

Quaternionic representation of the genetic code

C. Manuel Carlevaro[a,b], Ramiro M. Irastorza[a,c], Fernando Vericat[a,d,*]

[a] Instituto de Física de Líquidos y Sistemas Biológicos, 59 Nro. 789, 1900 La Plata, Argentina
[b] Universidad Tecnológica Nacional, Facultad Regional Buenos Aires, Mozart Nro. 2300, C14071VT Buenos Aires, Argentina
[c] Instituto de Ingeniería y Agronomía, Universidad Nacional Arturo Jauretche, 1888 Florencio Varela, Buenos Aires, Argentina
[d] Grupo de Aplicaciones Matemáticas y Estadísticas de la Facultad de Ingeniería (GAMEFI), Universidad Nacional de La Plata, Calle 115 y 48, 1900 La Plata, Argentina

ARTICLE INFO

Article history:
Received 11 September 2015
Received in revised form 4 December 2015
Accepted 28 December 2015
Available online 2 January 2016

Keywords:
Genetic code evolution
Broken symmetry
Hamilton quaternions
Protein folding

ABSTRACT

A heuristic diagram of the evolution of the standard genetic code is presented. It incorporates, in a way that resembles the energy levels of an atom, the physical notion of broken symmetry and it is consistent with original ideas by Crick on the origin and evolution of the code as well as with the chronological order of appearance of the amino acids along the evolution as inferred from work that mixtures known experimental results with theoretical speculations. Suggested by the diagram we propose a Hamilton quaternions based mathematical representation of the code as it stands now-a-days. The central object in the description is a codon function that assigns to each amino acid an integer quaternion in such a way that the observed code degeneration is preserved. We emphasize the advantages of a quaternionic representation of amino acids taking as an example the folding of proteins. With this aim we propose an algorithm to go from the quaternions sequence to the protein three dimensional structure which can be compared with the corresponding experimental one stored at the Protein Data Bank. In our criterion the mathematical representation of the genetic code in terms of quaternions merits to be taken into account because it describes not only most of the known properties of the genetic code but also opens new perspectives that are mainly derived from the close relationship between quaternions and rotations.

1. Introduction

The standard genetic code (Crick et al., 1961), say the correspondence between the sequence of nucleotide bases of mRNA molecules and the sequence of amino acids in the ribosomal protein synthesis as occurring at the cells of most of the animals and plants, is now-a-days fairly well known. The mRNA bases belong to the set $\{A, C, G, U\}$ where A stands for adenine, C for cytosine, G for guanine and U for uracil. Non-overlapping triplets of consecutive bases (codons) encode just one of the 20 standard amino acids (see Appendix A) or a stop signal each one. In principle, there is no any kind of separation between adjacent codons in the sequence. Of the $4^3 = 64$ possible different codons, 61 translate into amino acids and the remaining three determine a stop signal. We are then speaking about a code of four letters that can form 64 words three letters each. The words translate into amino acids or the stop signal.

The mechanism that performs this translation involves a very sophisticated molecular machinery which is no completely known yet. However, Crick's adaptor hypothesis (Crick, 1958) and further refinements (Ibba and Söll, 1999; Ibba et al., 2000) are, in general, widely accepted as accurate enough as to describe, at molecular level, the complex translation procedure in most of the cases. The image currently accepted is that tRNA molecules act as intermediaries (adaptors) between the template (mRNA) and the amino acids that will form the protein. The amino acid to be incorporated into the protein chain is covalently bonded to the tRNA 3′ extreme (forming an aminoacyl–tRNA complex) at the time that, in another part of the tRNA chain, a triplet of nucleotide bases (anticodon) specifically interacts with the codon of the mRNA template that codifies the amino acid in question. The bases of the anticodon are just the complementary ones of the corresponding codon bases (read in the direction $5' \to 3'$) and the interactions manifest as hydrogen bonds between complementary bases.

Skipping over for the moment the molecular details of the translation and restricting ourselves to the correspondence *codons* → *amino acids* in itself, we reproduce in Fig. 1a classical

* Corresponding author at: Instituto de Física de Líquidos y Sistemas Biológicos, 59 Nro. 789, 1900 La Plata, Argentina.
E-mail addresses: manuel@iflysib.unlp.edu.ar (C.M. Carlevaro), rirastorza@iflysib.unlp.edu.ar (R.M. Irastorza), vericat@iflysib.unlp.edu.ar (F. Vericat).

http://dx.doi.org/10.1016/j.biosystems.2015.12.009
0303-2647/

First codon base	Second codon base: U		C		A		G	
U	*UUU*	Phe	*UCU*	Ser	*UAU*	Tyr	*UGU*	Cys
	UUC	Phe	*UCC*	Ser	*UAC*	Tyr	*UGC*	Cys
	UUA	Leu	*UCA*	Ser	*UAA*	Stop	*UGA*	Stop
	UUG	Leu	*UCG*	Ser	*UAG*	Stop	*UGG*	Trp
C	*CUU*	Leu	*CCU*	Pro	*CAU*	His	*CGU*	Arg
	CUC	Leu	*CCC*	Pro	*CAC*	His	*CGC*	Arg
	CUA	Leu	*CCA*	Pro	*CAA*	Gln	*CGA*	Arg
	CUG	Leu	*CCG*	Pro	*CAG*	Gln	*CGG*	Arg
A	*AUU*	Ile	*ACU*	Thr	*AAU*	Asn	*AGU*	Ser
	AUC	Ile	*ACC*	Thr	*AAC*	Asn	*AGC*	Ser
	AUA	Ile	*ACA*	Thr	*AAA*	Lys	*AGA*	Arg
	AUG	Met	*ACG*	Thr	*AAG*	Lys	*AGG*	Arg
G	*GUU*	Val	*GCU*	Ala	*GAU*	Asp	*GGU*	Gly
	GUC	Val	*GCC*	Ala	*GAC*	Asp	*GGC*	Gly
	GUA	Val	*GCA*	Ala	*GAA*	Glu	*GGA*	Gly
	GUG	Val	*GCG*	Ala	*GAG*	Glu	*GGG*	Gly

Fig. 1. Text book picture of the standard genetic code. The three letters convention for the amino acids is used (see Appendix A) and the third base in the codons is remarked in bold. The order of the codons is in the direction $5' \rightarrow 3'$. The codon *AUG* besides to codify the amino acid methionine (met) also determines the starting point within the mRNA sequence for the protein synthesis.

presentation of the standard genetic code. The structure of the code is evident. Each codon codifies just one amino acid or (in the case of the codons *UAA*, *UAG* and *UGA*) the stop signal. The code is degenerate in the sense that, except for the amino acids methionine (met) and tryptophan (trp) that are codified by a single codon each one, all the other amino acids are codified by two or more codons.

One interesting related question that has received some attention is the origin and evolution of the genetic code. The proposals in this direction are obviously rather speculative (Jukes, 1973; Wong, 1988; Osawa et al., 1992; Hartman, 1995; Jiménez-Sánchez, 1995). However, Crick's scenario (Crick, 1968) according to which originally only a few amino acids were coded by most of the possible three bases codons and that, in subsequent steps, some of those codons were substituting the amino acid they coded by a new one until eventually the code became frozen in its present form, seems reasonable and very attractive. In particular, the idea of an increasing number of amino acids to be coded, can be correlated with the studies on the evolution of the amino acids abundance (Miller, 1953; Trifonov, 2000).

A step further in relation with the genetic code includes several efforts done in order to give mathematical models for describing the present structure of the code and how it has evolved in order to reach this state (Gonzalez, 2004; Hornos and Hornos, 1993; Sciarrino, 2003). The main mathematical tools are tensor algebras and group theory. In particular, in Hornos and Hornos (1993) the authors use the physical concept of broken symmetry to find a mathematical group with a 16-dimensional representation (the highly degenerate primitive code) which can be written as the product of simpler groups that describe the pattern of redundancies observed in Fig. 1. The approach gives a very elegant physical explanation of the code degeneration. However, perhaps because it concerns the application of a relatively complicated mathematical tool to a subject dominated by researchers with main formation in disciplines other than Mathematics and Physics, the work has been taken just as a valuable exercise in classification (Maddox, 1994; Stewart, 1994).

In this work we propose a mathematical description of the genetic code too, but it is based on a tool that, in our judgement, is very friendly and, at the same time, very powerful as to open new perspectives beyond of simply giving a representation of the code structure. We are talking about the Hamilton quaternions (Hamilton, 1843, 1866). These mathematical objects are a sort of generalization of the complex numbers and obey an algebra in many aspects similar to theirs but with the very important (for our purposes) property that the product is, in general, non-commutative (see Appendix B). In addition, the quaternions are ideal for representing rotations with important advantages over the classical matrix representation. This fact has of course already been recognized by bioinformaticians in writing routines involving the tertiary structure of proteins. We must mention that Petoukhov has also applied quaternions to describe the genetic code but from a very different point of view (Petoukhov, 2006).

Our journey starts by presenting in the next section a diagram for the evolution of the genetic code that incorporates the concept of broken symmetry in a way that resembles the energy levels of an atom. Actually, our interest is in the present form of the code, however the evolution diagram gives a picture of the correspondence *bases triplets* $\rightarrow$ *amino acids* that will help us with the mathematical representation of this correspondence by means of quaternions. Moreover, despite the high degree of speculation that exists in any model for the origin and evolution of the genetic code, we can give to our diagram an interpretation which is consistent with the above mentioned ideas by Crick on the subject (Crick, 1968). Thus, inspired by this diagram, in Section 3 we proceed to represent the relationship between the codons and amino acids by using quaternions. First we assign an integer quaternion (Lipschitz integer) to each one of the four nucleotide bases and then, suggested by the diagram structure, we consider a codons function that gives as result the assignation of a quaternion to each one of the amino acids. The explicit form of this function involves simple quaternionic operations (products and sums) that automatically accounts for the degeneration of amino acids encoded by