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

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.

© 2015 Elsevier Ireland Ltd. All rights reserved.

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′ → 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. 1 a 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).

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

First codon base

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* → *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

four, three or two codons and includes, in addition to the quaternions assigned to the four bases, an extra number of quaternions, related with the splitting of the "atomic levels" due to the symmetry breaking during the evolution, which, in principle, are indeterminate. These extra quaternions are determined by demanding that the degeneration for amino acids encoded by more than four (concretely six) codons be also verified. In order that this scheme works in practice we need to explicitly give the four quaternions for the bases. Of the infinitely many options the one we choose clearly has a Pythagorean flavor: we consider a subset of four quaternions from the complete set of eight prime integer quaternions with norm 7. The subset we take does not contain pairs of conjugate quaternions, four being the maximum cardinality for a subset with this property (Davidoff et al., 2003). Once a quaternion of this subset has been assigned to each of the four nucleotide bases, the quaternion corresponding to each amino acid is directly determined by the above mentioned function. This way the quaternionic description of the genetic code is completed. In order to remark the potentiality of the quaternionic representation of amino acids for opening new perspectives beyond the description of the genetic code degeneration, we appeal to another fundamental question: the protein folding problem, say the establishment of the native tertiary structure of the protein from the knowledge of its amino acids sequence (primary structure) (Creighton, 1992; Ben Naim, 2013). The protein folding problem is *per se* a phenomenal task that in some sense can be considered as experimentally solved through X-ray diffraction, Nuclear Magnetic Resonance and other techniques. However, theoretically the problem remains unsolved and a lot of work has been done by many researchers since the middle of the past century in order to develop a computational procedure that allows predicting the tertiary structure of a given protein from its amino acids sequence. Here we avoid to mention the lot of methods proposed to attack the question and simply give our own, maybe rather heuristic, approach as to show the advantages of associating amino acids with quaternions. This will be done in Section 4 were we show the procedure that we have designed in order to go from the amino acids quaternions to the coordinates of the backbone alpha-carbon atoms of a protein whose spatial structure we assume is the native one for the given amino acids sequence. These coordinates can be compared with those experimentally obtained as given in the Protein Data Bank (PDB). The procedure involves a set of real quaternions associated with the order of the amino-acids in the chain so that each amino-acid in a protein is represented by an integer quaternion (type quaternion) and a real quaternion (order quaternion). If this quaternions are the same ones for all the proteins, then the protein folding problem would be solved. In this work we limit ourselves to show how the type and order quaternions can be used to transform the primary structure of a given protein into its spatial configuration. The problem of obtaining the set of order quaternions which is adequate to all proteins (if it exists), say the possibility of solving the protein folding problem, is left for future work.

Two Appendices, one with the one and three letters convention for identifying the 20 standard amino acids and another one with the main properties of the quaternions are finally given for completeness.

2. A diagram for the evolution of the genetic code

In Fig. 2 we show the diagram that we propose to take account of the evolution of the genetic code. It is mainly inspired in pioneering ideas by Crick (1968) and also in the physical concept of broken symmetry, first applied in relation with the genetic code by Hornos and Hornos (1993).

According to Crick if the genetic code is at present time a triplet code, in the sense that the reading mechanism moves along three

bases at each step, then it must always have been a triplet code since otherwise a loss of Darwinian fitness can occur. Thus we assume that the codons were always formed by three bases of the set {A, C, G, U}. We must mention that Crick also have analyzed the plausibility of primitive nucleic acids constituted by just two bases. However even if this were the case, since the passage to a four bases system had to occur in some moment of the evolution without to substantially alter the message carried by the old two bases chain (Principle of continuity), we can take the four bases alphabet as being always available since a given moment at the origins of the code. Therefore we accept that since the beginning codons are triplets of bases chosen from the set {A, C, G, U}. Moreover, we consider that, in the first evolution steps, only the second base of the codon was effective in codifying amino acids. Accordingly only four amino acids could be codified, each one by one of the four bases C, G, U and A independently of which the first and third bases are. In the diagram this fact is denoted with a rectangle containing the four letters. This is consistent with Crick's suggestion that only a few amino acids were coded at the beginning. According with the diagram, C would codify alanine (A); G, glycine (G); U, valine (V) and A aspartic acid (D) whatever the first and third bases are. It is worth noting here that the four amino acids that we assume were the first ones to be codified are the first four in the Trifonov (2000) consensus temporal order scale for the appearance of the amino acids (column of natural numbers in Fig. 2). The four amino acids A, G, V and D were also the first four that appeared under simulation of the primitive earth conditions in Miller experiments (Miller, 1953).

As the left part of diagram shows, our version of the primitive code is highly degenerate: in principle each of the four amino acids, A, G, V and D, could be encoded by $4^2 = 16$ codons. Physically the idea of degeneration is closely related with the concept of symmetry and a very illustrative form to think about these concepts is by doing an analogy with the energy levels of an atom. In our case we would have four levels indexed each one with the letter corresponding to the second codon base, say C, G, U and A (main quantum number). We thus assume that, as the code evolves, the symmetry that causes that the amino acid codification be independent of the first base of the codon, disappears for some reason. The reason could be that with time the recognition mechanism becomes more precise as to differentiate between two codons with distinct first base. Because of this symmetry breaking, a part of the degeneration also disappears. In the diagram each of the four initial levels splits into four new levels, one for each of the possible bases (C, G, U and A) at the first place of the codon (secondary quantum number). Now we have a total of 16 levels indexed each one by two letters (the first and second bases of the codon). Each level is fourfold degenerate in the codons third base. One of the new levels follows codifying the same amino acid as before that the level splits whereas the other three codify a new amino acid each. We indicate with an arrow the four groups of codons that conserve the amino acid and with a simple line those that substitute the amino acid by a new one. Note that the codons that follow codifying the same amino acid are those whose first base is guanine (G). This is consistent with the above mentioned temporal order of appearance and with the present time correct assignation of amino acids in the case of fourfold degeneration as is shown in Fig. 1. This way 9 new amino acids (that with the old four sum 13) and the stop signal are coded. Note also that we assume that the amino acids serine (S) and leucine (L) at that moment were codified by two groups of codons: S by UC (*third base arbitrary*) and AG (*third base arbitrary*), whereas L by CU (*third base arbitrary*) and UU (*third base arbitrary*).

As the code follows evolving it suffers new breaking of symmetry so that the third base of some codons bring into use or, in the atomic analogy, some of the fourfold degenerate levels split into two levels each one twofold degenerate. Those levels pointed out with an arrow follow codifying the same amino acid whereas the

other levels substitute it for a new one. Eventually, in subsequent steps, a few of the twofold degenerate levels split once more given two non-degenerate levels each. This is the case of codons that codify methionine (M), tryptophan (W) and (again) the stop signal. The case of isoleucine (I) is a particular one since the split level coincides with the twofold one which represents the two codons that follow codifying the same amino acid. This way, isoleucine is the only amino acid which is coded by three codons. The stop signal is also threefold degenerate since it is coded by two groups of codons one twofold degenerate and the other one non-degenerate. At this step of the evolution the code frozen to give its present form. It is worth mentioning that the code evolution gives as a particular result that the amino acids serine (S), arginine (R) and leucine (L) are at present coded by two groups of codons each one. In the three cases one of the groups is fourfold degenerate and the other one is twofold degenerate, so that these amino acids are the only three which are sixfold degenerates. We point out this property in the diagram with a broken line linking the two groups of codons. The two groups of codons that codify the stop signal are also linked by a broken line.

3. Mathematical representation of the genetic code

We proceed now to describe the genetic code by using quaternions. Define the sets:

$$\mathcal{B} = \{C, G, U, A\}, \tag{1}$$

$$\mathcal{A} = \{P, A, S, T, R, G, C, W, L, V, F, I, M, H, Q, D, E, Y, N, K, Stop\} \tag{2}$$

and

$$\mathbf{H}_{7,\,red.}(\mathbb{Z}) = \{(2, 1, 1, 1), (2, -1, 1, 1), (2, 1, -1, 1), (2, 1, 1, -1)\}. \tag{3}$$

We propose a quaternionic representation of the genetic code according to the following scheme:

$$\begin{array}{ccc} \mathcal{B}^3 & \longrightarrow & \mathcal{A} \\ \downarrow & & \downarrow \\ \mathbf{H}_{7,\,red.}^3(\mathbb{Z}) & \longrightarrow & \mathbf{H}(\mathbb{Z}) \end{array} \tag{4}$$

where $\mathbf{H}(\mathbb{Z})$ denotes the set of integer quaternions (see Appendix B). $\mathcal{B}^3$ is the set of the 64 codons and we assume that the correspondence $\mathcal{B}^3 \to \mathcal{A}$ is the present day standard genetic code as described by Fig. 1, whereas the function $\mathcal{B}^3 \to \mathbf{H}_{7,\,red.}^3(\mathbb{Z})$ assigns to each codon a triplet of quaternions of the set $\mathbf{H}_{7,\,red.}(\mathbb{Z})$ (the subindex red. is for reduced). This set is a maximum cardinality subset of

$$\mathbf{H}_7(\mathbb{Z}) = \{(a_0, a_1, a_2, a_3) : a_0, a_1, a_2, a_3 \in \mathbb{Z}; a_0^2 + a_1^2 + a_2^2 + a_3^2 = 7,$$

$$a_0 > 0 \text{ and even}\}$$

with the property that it does not contain pairs of conjugate quaternions. The set $\mathbf{H}_7(\mathbb{Z})$ has 7 + 1 = 8 elements (Davidoff et al., 2003) and so $\mathbf{H}_{7,\,red.}(\mathbb{Z})$ has 4 quaternions as it should be. It is worth-noting that all the integer quaternions in $\mathbf{H}_7(\mathbb{Z})$ are prime quaternions in the sense that they can not be expressed as the product of two integer quaternions if neither of them can be the unit quaternion (1, 0, 0, 0). This is consistent with the fact that an integer quaternion is prime if and only if its norm is a prime number (Davidoff et al., 2003). Note that taking the nucleotide bases as prime quaternions gives them a certain character of *elemental* molecules. Apart from this, the election of $\mathbf{H}_{7,\,red.}(\mathbb{Z})$ may seem rather arbitrary. However we are just looking for a quaternionic representation of the genetic code so that, whatever the set of quaternions that we assign to the

codons is, the important issue is that the function $\mathbf{H}_{7,\,red.}^3(\mathbb{Z}) \to \mathbf{H}(\mathbb{Z})$ preserves the essential properties of the correspondence $\mathcal{B}^3 \to \mathcal{A}$.

In what follows, in order to simplify the notation, we assign natural numbers to identify the bases and the amino acids: $C \to 1$, $G \to 2$, $U \to 3$, $A \to 4$ and $P \to 1$, $A \to 2$, $S \to 3$, $T \to 4$, $R \to 5$, $G \to 6$, $C \to 7$, $W \to 8$, $L \to 9$, $V \to 10$, $F \to 11$, $I \to 12$, $M \to 13$, $H \to 14$, $Q \to 15$, $D \to 16$, $E \to 17$, $Y \to 18$, $N \to 19$, $K \to 20$, Stop $\to 21$.

Inspired by the diagram of Fig. 2 we define the quaternionic function

$$F : \mathbf{H}_{7,\,red.}^3(\mathbb{Z}) \to \mathbf{H}(\mathbb{Z})$$

$$(q_\beta, q_\gamma, q_\delta) \to \alpha_i = F[(q_\beta, q_\gamma, q_\delta)] \tag{5}$$

by (see Appendix B for the operations between quaternions):

$$P \to \alpha_1 = q_1 q_1 \quad (\beta = 1, \gamma = 1, \delta = 1, 2, 3, 4)$$

$$A \to \alpha_2 = q_2 q_1 \quad (\beta = 2, \gamma = 1, \delta = 1, 2, 3, 4)$$

$$S \to \alpha_3 = q_3 q_1 = q_4 q_2 + \gamma_{2;13} \quad (\beta = 3, \gamma = 1, \delta = 1, 2, 3, 4 \text{ or}$$
$$\beta = 4, \gamma = 2, \delta = 1, 3)$$

$$T \to \alpha_4 = q_4 q_1 \quad (\beta = 4, \gamma = 1, \delta = 1, 2, 3, 4)$$

$$R \to \alpha_5 = q_1 q_2 = q_4 q_2 + \gamma_{2;24} \quad (\beta = 1, \gamma = 2, \delta = 1, 2, 3, 4 \text{ or}$$
$$\beta = 4, \gamma = 2, \delta = 2, 4)$$

$$G \to \alpha_6 = q_2 q_2 \quad (\beta = 2, \gamma = 2, \delta = 1, 2, 3, 4)$$

$$C \to \alpha_7 = q_3 q_2 + \gamma_{2;13} \quad (\beta = 3, \gamma = 2, \delta = 1, 3)$$

$$W \to \alpha_8 = q_3 q_2 + \gamma_{2;24} + \delta_{2;2} \quad (\beta = 3, \gamma = 2, \delta = 2)$$

$$L \to \alpha_9 = q_1 q_3 = q_3 q_3 + \gamma_{3;24} \quad (\beta = 1, \gamma = 3, \delta = 1, 2, 3, 4 \text{ or}$$
$$\beta = 3, \gamma = 3, \delta = 2, 4)$$

$$V \to \alpha_{10} = q_2 q_3 \quad (\beta = 2, \gamma = 3, \delta = 1, 2, 3, 4)$$

$$F \to \alpha_{11} = q_3 q_3 + \gamma_{3;13} \quad (\beta = 3, \gamma = 3, \delta = 1, 3)$$

$$I \to \alpha_{12} = q_4 q_3 + \gamma_{3;13} = q_4 q_3 + \gamma_{3;24} + \delta_{3;4}$$
$$(\beta = 4, \gamma = 3, \delta = 1, 3, 4)$$

$$M \to \alpha_{13} = q_4 q_3 + \gamma_{3;24} + \delta_{3;2} \quad (\beta = 4, \gamma = 3, \delta = 2))$$

$$H \to \alpha_{14} = q_1 q_4 + \gamma_{4;13} \quad (\beta = 1, \gamma = 4, \delta = 1, 3)$$

$$Q \to \alpha_{15} = q_1 q_4 + \gamma_{4;24} \quad (\beta = 1, \gamma = 4, \delta = 2, 4)$$

$$D \to \alpha_{16} = q_2 q_4 + \gamma_{4;13} \quad (\beta = 2, \gamma = 4, \delta = 1, 3)$$

$$E \to \alpha_{17} = q_2 q_4 + \gamma_{4;24} \quad (\beta = 2, \gamma = 4, \delta = 2, 4)$$

$$Y \to \alpha_{18} = q_3 q_4 + \gamma_{4;13} \quad (\beta = 3, \gamma = 4, \delta = 1, 3)$$

$$N \to \alpha_{19} = q_4 q_4 + \gamma_{4;13} \quad (\beta = 4, \gamma = 4, \delta = 1, 3))$$

$$K \to \alpha_{20} = q_4 q_4 + \gamma_{4;24} \quad (\beta = 4, \gamma = 4, \delta = 2, 4))$$

$$\text{Stop} \to \alpha_{21} = q_3 q_2 + \gamma_{2;24} + \delta_{2;4} = q_3 q_4 + \gamma_{4;24}$$

$$(\beta = 3, \gamma = 2, \delta = 4 \text{ or } \gamma = 4, \delta = 2, 4) \tag{6}$$

From these expressions we can appreciate the importance of working with objects that obey a non-commutative algebra. In fact, if the quaternions product where commutative then amino acids A and R would have associated the same quaternion and the same would occur with S and L.

In Eq. (6), the quaternions $\gamma_{i;jk}$ accounts for the level splitting when the second base of codon is i and the third base is $jk = 13$ (CU) or 24 (GA). Analogously, the quaternion $\delta_{i;j}$ accounts for the level splitting when the second base of the codon is i and the third base is $j = 2$ (G) or 4 (A). Thus, in principle we have as unknown quaternions $\gamma_{2;13}$, $\gamma_{2;24}$, $\gamma_{3;13}$, $\gamma_{3;24}$, $\gamma_{4;13}$, $\gamma_{4;24}$ and $\delta_{2;2}$, $\delta_{2;4}$, $\delta_{3;2}$ and $\delta_{3;4}$.

Of the 10 unknown quaternions we can find 5, say $\gamma_{2;13}$, $\gamma_{2;24}$, $\gamma_{3;13}$, $\gamma_{3;24}$, $\gamma_{4;24}$, by requiring that those amino acids which are coded by two different groups of codons (case of codons sixfold degenerates or codons that codify the stop signal) have associated an unique quaternion and also that the two ways to reach isoleucine (I) give the same quaternion (see Fig. 2), so we must solve the system

$$\begin{cases} q_3 q_1 = q_4 q_2 + \gamma_{2;13} & (\alpha_3) \\ q_1 q_2 = q_4 q_2 + \gamma_{2;24} & (\alpha_5) \\ q_1 q_3 = q_3 q_3 + \gamma_{3;24} & (\alpha_9) \\ q_4 q_3 + \gamma_{3;13} = q_4 q_3 + \gamma_{3;24} + \delta_{3;4} & (\alpha_{12}) \\ q_3 q_2 + \gamma_{2;24} + \delta_{2;4} = q_3 q_4 + \gamma_{4;24} & (\alpha_{21}). \end{cases} \tag{7}$$

The solution is:

$$\gamma_{2;13} = q_3 q_1 - q_4 q_2$$

$$\gamma_{2;24} = q_1 q_2 - q_4 q_2$$

$$\gamma_{3;13} = q_1 q_3 - q_3 q_3 + \delta_{3;4} \tag{8}$$

$$\gamma_{3;24} = q_1 q_3 - q_3 q_3$$

$$\gamma_{4;24} = q_3 q_2 + q_1 q_2 - q_4 q_2 - q_3 q_4 + \delta_{2;4}.$$

To obtain the quaternions $\delta_{2;2}$, $\delta_{2;4}$, $\delta_{3;2}$ and $\delta_{3;4}$ we assign to those levels that can not split more (non degenerate levels) the product of the quaternions associated with each of the corresponding bases: $\alpha_8 = q_3 q_2 q_2$; $\alpha_{13} = q_4 q_3 q_2$; $\alpha_{21} = q_3 q_2 q_4$; $\alpha_{12} = q_4 q_3 q_4$. This way we have

$$\delta_{2;2} = q_3 q_2 q_2 - q_3 q_2 - \gamma_{2;24}$$

$$\delta_{3;2} = q_4 q_3 q_2 - q_4 q_3 - \gamma_{3;24}$$

$$\delta_{2;4} = q_3 q_2 q_4 - q_3 q_2 - \gamma_{2;24} \tag{9}$$

$$\delta_{3;4} = q_4 q_3 q_4 - q_4 q_3 - \gamma_{3;24}.$$

Finally for the remaining unknown quaternion $\gamma_{4;13}$ we propose:

$$\gamma_{4;13} = -\gamma_{4;24}. \tag{10}$$

Eqs. (6), (8), (9) and (10) solve completely the problem of assigning quaternions to the amino acids in such a way that the pattern of redundancy of the genetic code is verified. Taking: $q_1 = (2, 1, 1, 1)$, $q_2 = (2, -1, 1, 1)$, $q_3 = (2, 1, -1, 1)$ and $q_4 = (2, 1, 1, -1)$, we explicitly obtain

$\alpha_1 = (1, 4, 4, 4)$	$\alpha_8 = (6, -15, -1, 9)$	$\alpha_{15} = (16, -3, 7, 1)$
$\alpha_2 = (3, 0, 6, 2)$	$\alpha_9 = (3, 6, 0, 2)$	$\alpha_{16} = (-8, 3, 3, -3)$
$\alpha_3 = (3, 2, 0, 6)$	$\alpha_{10} = (5, 2, 2, 4)$	$\alpha_{17} = (18, -7, 5, -1)$
$\alpha_4 = (3, 6, 2, 0)$	$\alpha_{11} = (2, 17, 1, 3)$	$\alpha_{18} = (-8, 9, 1, 1)$
$\alpha_5 = (3, 0, 2, 6)$	$\alpha_{12} = (6, 17, 3, -3)$	$\alpha_{19} = (-12, 9, 3, -5)$
$\alpha_6 = (1, -4, 4, 4)$	$\alpha_{13} = (18, 3, -1, 3)$	$\alpha_{20} = (14, -1, 5, -3)$
$\alpha_7 = (3, -2, -6, 8)$	$\alpha_{14} = (-10, 7, 5, -1)$	$\alpha_{21} = (18, -1, 3, 3).$

$$\tag{11}$$

We will denote $\mathbf{H}_\alpha(\mathbb{Z})$ the set of quaternions assigned to the amino acids as given by Eq. (11).

At first sight this set of quaternions could seem to say nothing special by itself, however when we watch it more carefully we start to discover some patterns of regularity or symmetries. The first thing that we observe is that the norm of all these quaternions is odd: $N(\alpha_i) = a_0^2 + a_1^2 + a_2^2 + a_3^2 \equiv 1 \bmod(2)$ ($i = 1, 2, \ldots, 21$) and can roughly be taken as a measure of the information needed to codify the corresponding amino acid in the sense that the larger
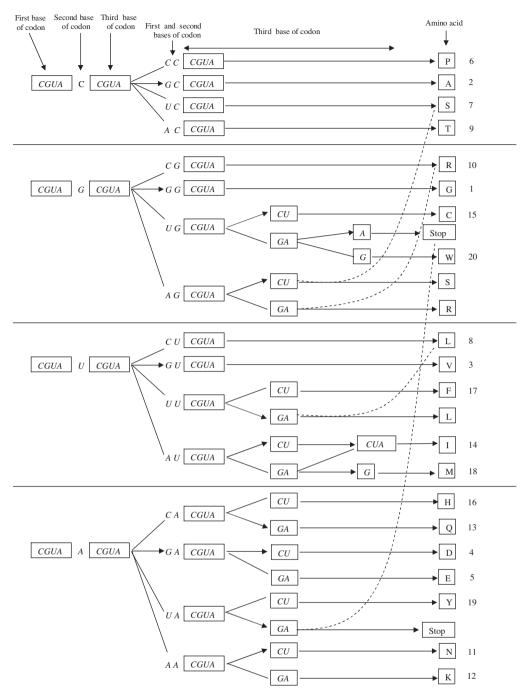
the norm the larger the necessary information. In fact, taking into account the multiplicative property of the quaternions norm we can easily see from Eq. (6) that those quaternions associated with amino acids which need just the first and second codon bases to be recognized, say α_1, α_2, α_3, α_4, α_5, α_6, α_9 and α_{10}, have as norm $N(\alpha_i) = N(q_\beta q_\gamma) = N(q_\beta)N(q_\gamma) = 49$ whereas those which need of the three bases to that effect, say the quaternions α_8 and α_{13} corresponding to the amino acids methionine (M), tryptophan (W) and also α_{12} associated with the amino acid isoleucine (I) and α_{21} with the stop signal, both coming (in one of two possible ways) from a non-degenerate level (see Fig. 2 and also Eq. (6)), have $N(\alpha_i) = N(q_\beta q_\gamma q_\delta) = N(q_\beta)N(q_\gamma)N(q_\delta) = 343$. Here we have used the fact that the norms of the quaternions that represent the nucleotide bases are $N(q_\beta) = 7$ ($\beta = 1, 2, 3, 4$). If the information about what amino acids will be added during the protein synthesis is encoded in the quaternions triplets $(q_\beta, q_\gamma, q_\delta)$ then for amino acids which are determined by quaternions of the type $\alpha_i = q_\beta q_\gamma$ the lack of information is compensated with the degeneration in the third base whereas for amino acids specified by quaternions of the form $\alpha_i = q_\beta q_\gamma q_\delta$ there is no lack of information and redundancy would be, in principle, not necessary. The amino acids which are twofold degenerate have norms which lie, with just one exception (α_{17}), in between these extreme values.

We can also use the norm to divide the set $\mathbf{H}_\alpha(\mathbb{Z})$ into classes: the norm of the quaternions corresponding to four or sixfold degenerate levels verifies $N(\alpha_i) \equiv 1 \bmod(4)$ whereas all the remaining quaternions, say α_8, α_{11}, α_{12}, α_{13}, α_{14}, α_{15}, α_{16}, α_{17}, α_{18}, α_{19}, α_{20} and α_{21} that come from levels with lower degeneration, have norm that fulfills $N(\alpha_i) \equiv 3 \bmod(4)$. The exception is the quaternion corresponding to the amino-acid cysteine which is coded by two codons but verifies $N(\alpha_7) = 113 \equiv 1 \bmod(4)$. At the respect we can say that in the *euplotid nuclear* variant of the genetic code the codon *UGA* codifies the amino acid C instead of the stop signal. If we consider this variant then α_7 would play in some sense the role of α_{21} and vice versa and the exception would be the stop signal which could be eliminated of the discussion that mainly concerns with amino acids. However since we are actually interested into the standard code we simply take the quaternion α_7 as the exception to the rule and momentarily ignore it in our discussion here. The class of quaternions that verifies $N(\alpha_i) \equiv 3 \bmod(4)$ can still be split into a couple of groups: one (α_{15}, α_{16}, α_{18}, α_{19}) with $N(\alpha_i) \equiv 3 \bmod(8)$ and the other one (α_8, α_{11}, α_{12}, α_{13}, α_{14}, α_{17}, α_{20}, α_{21}) with $N(\alpha_i) \equiv 7 \bmod(8)$. Although we have not clear the actual meaning of this separation we suspect that it has to do with symmetries involved in the translation process at molecular level. Any way we think that these simple observations are enough as to give a preliminary idea about the potential usefulness of quaternions to discover hidden patterns of symmetry inside the genetic code.

4. Amino acids as quaternions and the folding of proteins

As we have seen in the previous section, our quaternionic representation of the genetic code reproduces its structure, particularly the code redundancy and allows to make evident some regularity patterns. However the point that we wish to remark here is the special richness that gives to the description the close relationship between quaternions and rotations (see Appendix B). Because of the advantages of using quaternions to describe spatial rotations, the association of amino acids with quaternions opens new horizons beyond the genetic code representation. In this context, we consider the suitability of this association to take account of the folding of the proteins that the amino acids form.

The primary structure of a protein of N amino acids is a sequence $A_1, A_2, \ldots, A_N$ with $A_i \in \mathcal{A}$. The protein folding problem consists in obtaining from this sequence the spatial coordinates of each one of the atoms of all the amino acids that constitute the protein
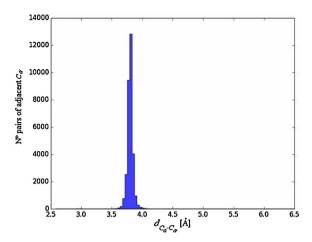
Fig. 2. Authors proposal for the genetic code evolution. The one letter convention for amino acids is used (see Appendix A). The direction of the temporal evolution is from left to right. Rectangles with two or more bases implies degeneration with respect to those ones. The broken lines link different sets of codons that encode the same amino acid in the case of sixfold degeneration. Arrows and common lines indicate what codons follow codifying the same amino acid and what will start to codify a new one, respectively, after the symmetry is broken (see text). The natural numbers at the right side of the diagram give the temporal order of the amino acids in the Trifonov consensus scale (Trifonov, 2000).
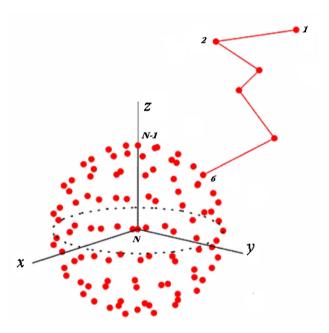
when this one is in the native -or functional- state (tertiary structure). As such we consider the one corresponding to the protein in physiological solution whose coordinates can be obtained, after crystallization, by application of, for example, X-ray diffraction methods. That is the case of most of the proteins whose coordinates are stored at the PDB. In principle we restrict ourselves to determine the coordinates for just the alpha-carbon atoms of the chain which is not a severe restriction since it is known that there exist very efficient algorithms for going from this trace representation to the full atoms one (Rotkiewicz and Skolnick, 2008). We also take into account that, in our quaternionic representation, the

amino acids sequence is expressed as a sequence of quaternions $p_1, p_2, \ldots, p_N$ with $p_i \in \mathbf{H}_\alpha(\mathbb{Z})$. Under these conditions we proceed now to present an algorithm to determine the spatial coordinates of the alpha-carbon atoms of the protein.

First we observe that although adjacent alpha-carbon atoms are not covalently bonded their distance is notably stable and take very similar values for all the pairs within a given protein and also for those belonging to different proteins, as the histogram of Fig. 3 shows. So in our calculations we assume that all these distances are equal to a unique value $d_{C_\alpha - C_\alpha} = 3.80$ Å. Thus we determine on the unit sphere with center at the origin a point for each of the

Fig. 3. Histogram for the distance $d_{C_\alpha - C_\alpha}$ between adjacent alpha-carbon atoms. The distances were calculated from the alpha-carbon atoms coordinates corresponding to a sample of 110 proteins of different length stored at the PDB (31,332 pairs of adjacent alpha-carbon atoms). The mean value and the standard deviation are $\langle d_{C_\alpha - C_\alpha} \rangle = 3.801$ Å and $\sigma_{C_\alpha - C_\alpha} = [\langle d^2_{C_\alpha - C_\alpha} \rangle - \langle d_{C_\alpha - C_\alpha} \rangle^2]^{1/2} = 0.061$ Å, respectively.

Fig. 4. Development of the alpha-carbon atoms backbone of a hypothetical protein of length N from its position on the sphere surface into its spatial configuration (schematic). The last two alpha-carbon atoms, as well as some of the first ones, are labeled by their order number inside the sequence.

amino acids (alpha-carbon atoms) in the protein sequence. To the last one we assign directly the origin, the preceding one is located at the intersection between the axis z and the sphere surface (versor $\widehat{e}_z$). To each of the remaining alpha-carbon atoms we assign a point on the sphere surface that results of rotating the versor $\widehat{e}_z$ by a quaternion (see Appendix B). For the jth alpha-carbon atom in the sequence, the quaternion responsible of the rotation is denoted $\widehat{\beta}_j (j = 1, 2, \ldots, N - 2)$. We then expand the chain of alpha-carbon atoms from their location on the sphere into the back-bone protein three dimensional configuration (see Fig. 4) by means of the following iterative procedure, where initially the $\vec{r}_j$'s are on the sphere surface:

```
do i = 1, N − 2
    δr⃗ = r⃗ᵢ₊₁
    do j = 1, i
        r⃗ⱼ = r⃗ⱼ + δr⃗
    end do
end do
```

According to the algorithm, the distance between adjacent alpha-carbon atoms is the unit so, to establish the correct distance, we must multiply the final calculated coordinates by $d_{C_\alpha - C_\alpha}$.

It remains to determine how to calculate the quaternions $\widehat{\beta}_j (j = 1, 2, \ldots, N - 2)$. We do this in a somewhat heuristic way. We take into account that the jth amino acid interacts in some way with the $j - 1$ previous amino acids in the sequence and also with the $N - j$ subsequent ones. Of course that in these interactions the effect of the medium should be incorporated in some form, for example in the form of effective interactions between amino acids. Actually we are trying for a sort of decodification and so we are not directly interested into the detailed form of the interactions, but we recognize that in any codification of information that involves those interactions, some trace of their general form should be. In general it is reasonable to think that the global interaction includes two body, three body, ..., until N body (effective) interactions so by analogy we choose with generality for $\widehat{\beta}_j$ the normalized version of the quaternion

$$\beta_j = (S^<_{j,1} + S^<_{j,2} + \cdots + S^<_{j,j-1})p_j + p_j(S^>_{j,1} + S^>_{j,2} + \cdots + S^>_{j,N-j}) \qquad (12)$$

with

$$S^<_{j,1} = \sum_{1 \leq r \leq j-1} c_r p_r, \quad S^<_{j,2} = \sum_{1 \leq r < s \leq j-1} c_r p_r c_s p_s, \ldots, S^<_{j,j-1}$$
$$= c_1 p_1 c_2 p_2 \cdots c_{j-1} p_{j-1} \qquad (13)$$

and

$$S^>_{j,1} = \sum_{j+1 \leq r \leq N} c_r p_r, \quad S^>_{j,2} = \sum_{j+1 \leq r < s \leq N} c_r p_r c_s p_s, \ldots, S^>_{j,N-j}$$
$$= c_{j+1} p_{j+1} c_{j+2} p_{j+2} \cdots c_N p_N, \qquad (14)$$

where $c_r \in \mathbf{H}(\mathbb{R})$ ($r = 1, 2, \ldots, N$) are in principle unknown real quaternions to be determined. It is worth mentioning that in our election of the form of Eq. (12) we have taken into account the non-commutativity of quaternions too.
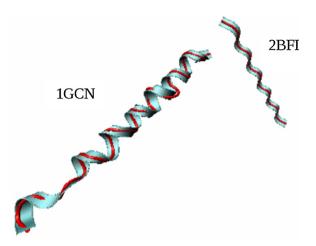
Even for proteins of length N relatively small, the memory and computation time required for evaluating the unknown quaternions $c_1, c_2, \ldots, c_N$ using the complete expression given by Eq. (12) for the β_j's are too large, at least for our computational facilities. Thus in the calculations here we use the simplest version:

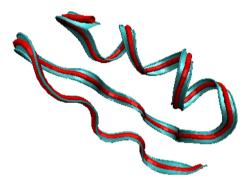$$\beta_j = S^<_{j,1} p_j + p_j S^>_{j,1}, \qquad (15)$$

that, in our analogy, corresponds to consider just pair interactions in the protein total potential energy.

Here we adjust the unknown quaternions by means of an optimization technique. As such we use the particle swarm optimization (PSO) procedure of Kennedy and Eberhart (1995) taking as function of fitness the difference between the coordinates of the alpha-carbon atoms calculated following the previous procedure and the corresponding experimental ones as read from the PDB. We take the rmsd (root-mean-square deviation) as a measure of this difference, using to that effect Bosco K. Ho's implementation of Kabsch algorithm (Kabsch, 1976). This way we assign to each amino-acid in the primary structure of the protein, two quaternions: an integer quaternion belonging to the set $\mathbf{H}_\alpha(\mathbb{Z})$ (type quaternion) and a real one (order quaternion) according to its position inside the protein chain.

In Figs. 5–8 we show the result of the application of our procedure to five small peptides and proteins: in Fig. 5 the synthetic peptide amyloid fibril (PDB ID: 2BFI – length: 12 amino acids)
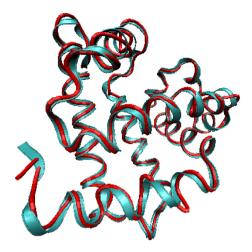
Fig. 5. Trace representation of the alpha-carbon atoms backbone for the small proteins 2BFI and 1GCN. Red (dark grey) tube: from the coordinates obtained using our procedure. Cyan (light grey) ribbon: from the coordinates stored at PDB. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 6. Trace representation of the alpha-carbon atoms backbone for the protein 2CK5. Red (dark grey) tube: from the coordinates obtained using our procedure. Cyan (light grey) ribbon: from the coordinates stored at PDB. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
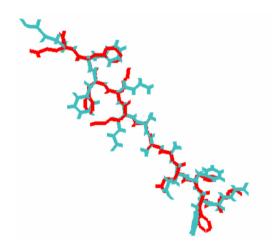

Fig. 7. Trace representation of the alpha-carbon atoms backbone for the protein 1HG7. Red (dark grey) tube: from the coordinates obtained using our procedure. Cyan (light grey) ribbon: from the coordinates stored at PDB. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
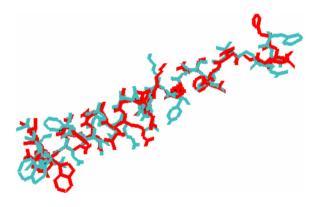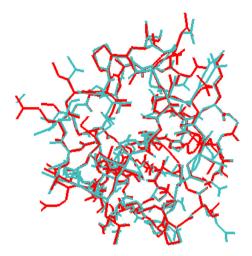

Fig. 8. Trace representation of the alpha-carbon atoms backbone for the protein 1MBN. Red (dark grey) tube: from the coordinates obtained using our procedure. Cyan (light grey) ribbon: from the coordinates stored at PDB. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 9. Full atom line representation of the peptide 2BFI. Red (dark grey): reconstruction from the alpha-carbon atoms backbone coordinates (obtained with our procedure) using the method of Rotkiewicz and Skolnick (2008). Cyan (light grey): from the coordinates stored at PDB. In the rebuilt protein the hydrogen atoms do not appear. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 10. Full atom line representation of the protein 1GCN. Red (dark grey): reconstruction from the alpha-carbon atoms backbone coordinates (obtained with our procedure) using the method of Rotkiewicz and Skolnick (2008). Cyan (light grey): from the coordinates stored at PDB. In the rebuilt protein the hydrogen atoms do not appear. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Conclusions

In this work we have presented a mathematical representation of the standard genetic code. Starting from a set of four prime integer quaternions (one for each of the nucleotide bases that form the mRNA molecules) and guided by a heuristic diagram that we propose for the evolution of the code, we introduce a function that assigns an integer quaternion (type quaternion) to each codon (represented by a triplet of the prime integer quaternions) and preserves the main properties of the genetic code. The diagram we introduce for describing the evolution of the genetic code is based in pioneering ideas by Crick and incorporates, in a way that resembles the energy levels of an atom, the physical notion of broken symmetry. The objects that we use for performing the mathematical representation of the code, the Hamilton quaternions, have as remarkable properties the fact that they verify a non commutative algebra and their capability for describing spatial rotations. In particular, this last property gives a special character to the representation in the sense that it allows to develop a procedure for going from the primary to the tertiary structure of proteins. To this effect we introduce a set of real quaternions (order quaternions) that, together with the integer type quaternions, univocally identify each amino acid of the proteins. Given an amino acids sequence we present an algorithm that determines the coordinates of the alpha-carbon atoms of the corresponding protein using the type and order quaternions. However here we simply adjust the order quaternions in order to reproduce the experimental coordinates stored at PDB. As already was commented above, we postpone for future studies the question of searching for a set of order quaternions which be common to all the proteins, say the possibility of approaching the protein folding problem by using our procedure. In our criterion this possibility distinguishes the above quaternionic representation of the genetic code among the diverse reported mathematical representations.

Acknowledgments

Appendix A. One and three letters convention for the 20 standard amino acids

Amino acid	Three letter	One letter
Alanine	ala	A
Arginine	arg	R
Asparagine	asn	N
Aspartic acid	asp	D
Cysteine	cys	C
Glutamic acid	glu	E
Glutamine	gln	Q
Glycine	gly	G
Histidine	his	H
Isoleucine	ile	I
Leucine	leu	L
Lysine	lys	K
Methionine	met	M
Phenylalanine	phe	F
Proline	pro	P
Serine	ser	S
Threonine	thr	T
Tryptophan	trp	W
Tyrosine	tyr	Y
Valine	val	V

Fig. 11. Full atom line representation of the protein 1HG7. Red (dark grey): reconstruction from the alpha-carbon atoms backbone coordinates (obtained with our procedure) using the method of Rotkiewicz and Skolnick (2008). Cyan (light grey): from the coordinates stored at PDB. In the rebuilt protein the hydrogen atoms do not appear. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

together with the hormone glucagon (PDB ID: 1GCN – length: 29 amino acids); in Fig. 6 the ion channel inhibitor osk1 toxin (PDB ID: 2CK5 – length: 31 amino acids); in Fig. 7 a type III antifreeze protein (PDB ID:1HG7 – length: 66 amino acids); in Fig. 8 the "hydrogen atom" of proteins, say myoglobin (PDB ID: 1MBN – length: 153 amino acids). The two proteins of Fig. 5 were adjusted simultaneously so that the order quaternions for 2BFI are the same ones as the first 12 of 1GCN. The 29 order quaternions we have obtained for 1GCN differ of the first 29 ones of the remainder proteins instead. With respect to this last fact we must mention that, at least within the error (rmsd) considered here, the set of quaternions we found for a given protein by fitting the alpha-carbon atoms coordinates is not unique. This is an important point since otherwise the possibility of finding a common set of order quaternions valid for all the proteins would be definitively closed. In the figures we compare the chains of alpha-carbon atoms calculated with our algorithm with the corresponding ones obtained from the coordinates stored at PDB. The resultant rmsd's are: 0.06 Å for 2BFI; 0.26 Å for 1GCN, 0.14 Å for 2CK5, 0.29 Å for 1HG7 and 0.79 Å for 1MBN.

For 2BFI, 1GCN and 1HG7 we have reconstructed the full-atom protein models from their alpha-carbon atoms representations using Rotkiewicz and Skolnick algorithm (PULCHRA) (Rotkiewicz and Skolnick, 2008). The results are shown in Figs. 9–11 were we also display the corresponding proteins as obtained from the PDB coordinates.

It must be remarked again that in this work we simply have shown a way to pass from the primary to the tertiary structure of the proteins assuming as known the corresponding order quaternions. These quaternions were obtained by fitting the coordinates of the alpha-carbon atoms obtained following our algorithm with the corresponding ones stored at PDB. The problem of using the procedure here described in order *to predict* the tertiary structure of proteins just from their amino acids sequences, which implies to know *a priori* a unique set of order quaternions that be adequate for all proteins, is left for future studies. Despite this important question, we believe that the results we have obtained until now already give a good idea about the usefulness of associating amino acids with quaternions, this being the main objective of this section.

Appendix B. Hamilton quaternions

Quaternions were invented by mathematician William Rowlan Hamilton (Hamilton, 1843, 1866) in 1843 as a generalization of the complex numbers with the aim of describing rotations in the space in the same sense as complex numbers describe rotations in the plane. Here we give for completeness some of the main properties of quaternions. We concentrate ourselves into their definition, the algebra they fulfill and their relation with rotations in the space (Kuipers, 1999).

B.1. Definition

A quaternion q is an ordered list of four numbers: $q = (a_0, a_1, a_2, a_3)$ with $a_0, a_1, a_2, a_3 \in \mathbb{R}$. In the particular case in that the four numbers are integers we talk of integer quaternions (Lipschitz integers). Alternatively we can introduce the placeholders $\mathbf{i}, \mathbf{j}, \mathbf{k}$ and represent the same quaternion as $q = a_0 + a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$. The placeholders $\mathbf{i}, \mathbf{j}, \mathbf{k}$ verify the product rules

$$\mathbf{ii} = -1 \quad \mathbf{jj} = -1 \quad \mathbf{kk} = -1$$
$$\mathbf{ij} = \mathbf{k} \quad \mathbf{jk} = \mathbf{i} \quad \mathbf{ki} = \mathbf{j}$$
$$\mathbf{ji} = -\mathbf{k} \quad \mathbf{kj} = -\mathbf{i} \quad \mathbf{ik} = -\mathbf{j}$$

Note that the placeholders play for quaternions a role in some sense similar to that of the imaginary unit $i = \sqrt{-1}$ for the complex numbers. In this context the triplet (a_1, a_2, a_3) would be the "imaginary" part of the quaternion. Defining the (real and imaginary) quaternions $q_R = (a_0, 0, 0, 0)$ and $q_I = (0, a_1, a_2, a_3)$, we can write: $q = q_R + q_I$.

B.2. Algebra

Let s be a real number and $q = (a_0, a_1, a_2, a_3)$, $p = (b_0, b_1, b_2, b_3)$ and $r = (c_0, c_1, c_2, c_3)$ quaternions, we give here the definition of a few operations:

– Conjugation: $\widetilde{q} = (a_0, -a_1, -a_2, -a_3)$.
– Scalar multiplication: $sq = (sa_0, sa_1, sa_2, sa_3)$.
– Addition of quaternions: $q + p = (a_0 + b_0, a_1 + b_1, a_2 + b_2, a_3 + b_3)$.
– Multiplication of quaternions: $qp = r$ where

$$c_0 = a_0 b_0 - a_1 b_1 - a_2 b_2 - a_3 b_3$$
$$c_1 = a_0 b_1 + a_1 b_0 + a_2 b_3 - a_3 b_2$$
$$c_2 = a_0 b_2 - a_1 b_3 + a_2 b_0 + a_3 b_1$$
$$c_3 = a_0 b_3 + a_1 b_2 - a_2 b_1 + a_3 b_0$$

Note that this product is not commutative say, in general, $qp \neq pq$.
– Norm: $N(q) = q\widetilde{q} = \widetilde{q}q = a_0^2 + a_1^2 + a_2^2 + a_3^2$.
  A quaternion q with $N(q) = 1$ is called a unit quaternion.
  An important property of the norm is that it is multiplicative: $N(pq) = N(p)N(q)$.
– Inverse: $q^{-1} = \widetilde{q}/N(q) (q \neq (0,0,0,0))$.

B.3. Quaternions and 3D rotations

If $N(q) = 1$ then the matrix

$$R_q = \begin{pmatrix} a_0^2 + a_1^2 + a_2^2 + a_3^2 & 0 & 0 & 0 \\ 0 & a_0^2 + a_1^2 - a_2^2 - a_3^2 & 2a_1a_2 - 2a_0a_3 & 2a_1a_3 + 2a_0a_2 \\ 0 & 2a_1a_2 + 2a_0a_3 & a_0^2 - a_1^2 + a_2^2 - a_3^2 & 2a_2a_3 - 2a_0a_1 \\ 0 & 2a_1a_3 - 2a_0a_2 & 2a_2a_3 + 2a_0a_1 & a_0^2 - a_1^2 - a_2^2 + a_3^2 \end{pmatrix}$$

is a rotation matrix. The oriented axis of rotation $\vec{e}$ is given by

$$\vec{e} = \frac{\vec{q}}{|\vec{q}|},$$

with $\vec{q} = a_1\widehat{e}_x + a_2\widehat{e}_y + a_3\widehat{e}_y$ where $\widehat{e}_x, \widehat{e}_y$ and $\widehat{e}_z$ are versors along the three Cartesian axis. The angle θ that determines the rotation around the axis $\vec{e}$ satisfies the following equation:

$$\tan(\theta/2) = \frac{\sqrt{a_1^2 + a_2^2 + a_3^2}}{a_0}.$$

Moreover, if we denote with R_3 the 3×3 matrix that results when in matrix R_q the first row and the first column are deleted, then we can see that the quaternion q transforms by rotation a vector $\vec{r}_0 = x_0\widehat{e}_x + y_0\widehat{e}_y + z_0\widehat{e}_z$ into the vector $\vec{r}_1 = x_1\widehat{e}_x + y_1\widehat{e}_y + z_1\widehat{e}_z$ according with

$$\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = R_3 \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}.$$

References

Ben Naim, A., 2013. The Protein Folding Problem and its Solutions. World Scientific Publishing Co., Singapore.
Creighton, T.E. (Ed.), 1992. Protein Folding. W.H. Freeman and Company, New York.
Crick, F.H.C., Barnett, L., Brenner, S., Watts-Tobin, R.J., 1961. General nature of the genetic code for proteins. Nature 192, 1227–1232.
Crick, F.H.C., 1958. On protein synthesis. Symp. Soc. Exp. Biol. 12, 138–163.
Crick, F.H.C., 1968. The origin of the genetic code. J. Mol. Biol. 38, 367–379.
Davidoff, G., Sarnak, P., Valette, A., 2003. Elementary Number Theory, Group Theory and Ramanujan Graphs. Cambridge University Press, Cambridge, UK.
Gonzalez, D.L., 2004. Can the genetic code be mathematically described? Med. Sci. Monit. 10, HY11–17.
Hamilton, W.R., 1843. On quaternions; or on a new system of imaginaries in algebra (letter to John T. Graves dated October 17, 1843).
Hamilton, W.R., 1866. In: Hamilton Longmans, W.E. (Ed.), Elements of Quaternions. Green & Co., London.
Hartman, H., 1995. Speculations on the origin of the genetic code. J. Mol. Evol. 40, 541–544.
Hornos, J.E., Hornos, Y.M.M., 1993. Algebraic model for the evolution of the genetic code. Phys. Rev. Lett. 71, 4401–4404.
Ibba, M., Söll, D., 1999. Quality control mechanisms during translation. Science 286, 1893–1897.
Ibba, M., Becker, H.D., Stathopoulos, C., Tumbula, D.L., Söll, D., 2000. The adaptor hypothesis revisited. Trends Biochem. Sci. 25, 311–316.
Jiménez-Sánchez, A., 1995. On the origin and evolution of the genetic code. J. Mol. Evol. 41, 712–716.
Jukes, T.H., 1973. Possibilities for the evolution of the genetic code from a preceding form. Nature 246, 22–26.
Kabsch, W., 1976. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A 32, 922–923.
Kennedy, J., Eberhart, R., 1995. R. particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks IV, pp. 1942–1948.
Kuipers, J.B., 1999. Quaternions and Rotation Sequences. A Primer with Applications to Orbits, Aerospace, and Virtual Reality. Princeton University Press, Princeton, NJ.
Maddox, J., 1994. The genetic code by numbers. Nature 367, 111.
Miller, S.L., 1953. A production of amino acids under possible primitive earth conditions. Science 117, 528–529.
Osawa, S., Jukes, T.H., Watanabe, K., Muto, A., 1992. Recent evidence for evolution of the genetic code. Microbiol. Rev. 56, 229–264.
Petoukhov, S.V., 2006. Bioinformatics: matrix genetics, algebras of the genetic code and biological harmony. Symmetry Cult. Sci. 17, 251–290.
Protein Data Bank. http://www.rcsb.org/pdb/.
Rotkiewicz, P., Skolnick, J., 2008. Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comput. Chem. 29, 1460–1465.
Sciarrino, A., 2003. A mathematical model accounting for the organization in multiplets of genetic code. BioSystems 69, 1–13.
Stewart, I., 1994. Broken symmetry in the genetic code? New Scientist 1915, 16.
Trifonov, E.N., 2000. Consensus temporal order of amino acids and evolution of the triplet code. Gene 261, 139–151.
Wong, J.T.-F., 1988. Evolution of the genetic code. Microbiol. Sci. 5, 174–181.