ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: José MV, Zamudio GS, Morgado ER. 2017 A unified model of the standard genetic code. *R. Soc. open sci.* **4**: 160908. http://dx.doi.org/10.1098/rsos.160908

Received: 14 November 2016
Accepted: 30 January 2017

Subject Category:
Genetics

Subject Areas:
bioinformatics/evolution/theoretical biology

Keywords:
standard genetic code, symmetry groups, aminoacyl-tRNA synthetases, group actions, automorphisms, polar requirement

Author for correspondence:
Marco V. José
e-mail: marcojose@biomedicas.unam.mx

THE ROYAL SOCIETY
PUBLISHING

A unified model of the standard genetic code

Marco V. José[1], Gabriel S. Zamudio[1] and Eberto R. Morgado[2]

[1]Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Mexico D.F. 04510, Mexico
[2]Facultad de Matemática, Física y Computación, Universidad Central 'Marta Abreu' de Las Villas, Santa Clara, Cuba

MVJ, 0000-0001-8497-6681; GSZ, 0000-0003-4486-9843

The Rodin–Ohno (RO) and the Delarue models divide the table of the genetic code into two classes of aminoacyl-tRNA synthetases (aaRSs I and II) with recognition from the minor or major groove sides of the tRNA acceptor stem, respectively. These models are asymmetric but they are biologically meaningful. On the other hand, the standard genetic code (SGC) can be derived from the primeval RNY code (R stands for purines, Y for pyrimidines and N any of them). In this work, the RO-model is derived by means of group actions, namely, symmetries represented by automorphisms, assuming that the SGC originated from a primeval RNY code. It turns out that the RO-model is symmetric in a six-dimensional (6D) hypercube. Conversely, using the same automorphisms, we show that the RO-model can lead to the SGC. In addition, the asymmetric Delarue model becomes symmetric by means of quotient group operations. We formulate isometric functions that convert the class aaRS I into the class aaRS II and vice versa. We show that the four polar requirement categories display a symmetrical arrangement in our 6D hypercube. Altogether these results cannot be attained, neither in two nor in three dimensions. We discuss the present unified 6D algebraic model, which is compatible with both the SGC (based upon the primeval RNY code) and the RO-model.

1. Introduction

The insight that all organisms on Earth are related by common descent [1] is a remarkable scientific achievement. Indeed, the Last Universal Common Ancestor seemed to obey already the standard genetic code (SGC), which is *nearly universal*. The problem of the origin and evolution of the SGC is a fundamental challenge in biology. After the decipherment of the SGC [2], there have been several proposals that account for both the origin and evolution of the genetic code [3–11]. There seems to be a consensus that the SGC conserves vestiges of earlier codes, to wit, the operational [12,13] and anticodon codes [14,15]. The amino acid specific aminoacylation of tRNAs (operational code) is localized in the acceptor stem of the tRNAs and is recognized by the corresponding aminoacyl-tRNA synthetases (aaRSs) [12,13]. Indeed, most living organisms still contain relics of these primeval codes, which are a palimpsest over which the evolving codes were later additions in order to arrive at the frozen SGC [16,17]. In fact, the primeval RNY code was already frozen [18].

The SGC is written in an alphabet of four letters (C, A, U, G), grouped into words three letters long, called triplets or codons. In general, and in most textbooks, the genetic code is represented in a two-dimensional (2D) table arranged in such a way that it is possible to readily find any amino acid from the three letters, written in the $5'$ to $3'$ direction of the codon [4,19,20]. Each of the 64 codons specifies one of the 20 amino acids or else serves as a punctuation mark signalling the end of a message. The standard table of codon assignments derives from the obvious representation of the triplet code as a $4 \times 4 \times 4$ cube. Three-dimensional (3D) algebraic models using a Galois Field of four elements GF(4) [21,22] or Lie algebras [22] have also been formulated. More revealing representations have been attained using the six-dimensional (6D) hypercube [23,24] of the 64 codons of the SGC. Observing that 64 is equal not only to 4^3 but also to 2^6, the codon table can be organized as a 6D hypercube or 6D vector space $(\mathbb{Z}_2)^6$ over the binary field $\mathbb{Z}_2 = \{0, 1\}$ [24]. The phenotypic graphs of amino acids have been obtained from the topology of the SGC [15]. Additionally, circular representations of the SGC have been proposed [25–27]. Given 64 codons and 20 amino acids plus a punctuation mark, there are $21^{64} \approx 4 \times 10^{84}$ possible genetic codes. The result that only one in every million random alternative codes is more efficient than the SGC [28] implies that there could be approximately 4×10^{78} genetic codes as efficient as the SGC. This calculation does not offer deeper insights concerning the origin and structure of the SGC, particularly the frozen accident. Francis Crick [4] argued that the SGC need not be special at all; it could be nothing more than a ‘frozen accident’. Yet as we show in this article, there are indeed several features that are special about the SGC: firstly, it can be partitioned *exactly* into two classes of aaRRs in six dimensions; secondly, it displays symmetry groups when the polar requirement (PR) is used; and thirdly, the SGC can be broken down into a product of simpler groups reflecting the pattern of degeneracy observed, and the salient fact that evolution did not erase its own evolutionary footsteps.

The search for symmetries in the SGC has been made by examining the tRNA [29,30] and aaRSs [3,6–8], using phylogenetic methods [31,32]. Less popular have been algebraic models seeking to unveil hidden symmetries of the SGC [33,34]. For example, the SGC has been theoretically derived from a primeval RNY (R means purines, Y pyrimidines, and N any of them) genetic code [9] under a model of sequential symmetry breakings [16,21,35]. Universal vestiges of these evolutionary steps were found in current genomes of both Eubacteria and Archaea [35]. The SGC is implemented via the tRNAs that bind each codon with its anticodon. These molecules define the genetic code, by linking the specific amino acids and tRNAs with the corresponding anticodons [7]. The tRNA molecule itself displays two codes, the operational code and the anticodon code. Typically, two genetic codes are considered, to wit, the ‘classic’ code represented in tRNA by an anticodon for reading codons in mRNA, and the other is the ‘second’ [12] operational RNA code [13,36] mapped mainly to the acceptor for appropriate aminoacylation at its $3'$ terminus. In addition, there are also two separate codes, embedded in the tRNA anticodon and acceptor-stem bases that correspond, respectively, to amino acid size and hydrophobicity [37,38]. These coding elements evolved separately and independently [38]. The earlier appearance of an acceptor-stem code, before the emergence of the universal genetic code [13] is supported experimentally by (i) the reciprocal biochemistry of minihelix acylation by full-length synthetases [39] and (ii) the acylation of full-length tRNAs by truncated synthetases called Urzymes [40].

PR is an abiotic feature of free amino acids in solution. PR is a physico-chemical property of each amino acid, defined by their migration in paper chromatographic experiments in aqueous solutions of nucleobases [41]. PR is directly related to the organization of the codon table and its amino acids [42]. In addition, PR is related to the partition of amino acid in a polar–non-polar interface [43]. The SGC is also robust to errors of single base mutations and this is reflected when PR is used as a metric of amino acid similarity [28,44,45]. Moreover, the phenotypic graphs of amino acids exhibit disjoint clusters of amino acids when their PR values are used [15]. The genetic code became optimized with respect to PR. By observing the microscopic environments of the amino acids in binary solution, it is apparent that the PR is related to how an amino acid partitions across a polar–non-polar interface. Several theoretical studies have found a high degree of error tolerance in the genetic code when PR is used as a measure of amino acid similarity [28,45–47]. Polar–non-polar interfaces may have played a role in the establishment or development of the early genetic code. It is highly improbable that the genetic code became optimized with respect to PR purely by chance.

As far as translation is concerned, it does not make sense to consider one code without the other. The present-day operational code is intricately carved in the structure of tRNA acceptors and cognate aaRSs, whereas the anticodon code is reduced to codon–anticodon interactions. The catalytic proteins required to accelerate this binding are divided between two very ancient enzyme superfamilies, the class I and class II aaRSs, each activating 10 of the 20 canonical amino acids [8]. The present correspondence of the two codes is provided by 20 specific aaRSs divided into two strikingly dissimilar classes of 10 members each. There are only 20 aaRSs, one for each amino acid (and, respectively, for isoacceptor tRNAs); hence, the operational code is non-degenerate [12]. Such a non-degeneracy, inherent only to the acceptor code, may indicate the historically subsidiary role of anticodons in aminoacylation. Otherwise, more than 20 aaRSs could exist, one for each anticodon rather than one for each amino acid. The two aaRSs recognize the acceptor helix from opposite sides: class I aaRS approaches the helix from the side of its minor groove and attaches the amino acid to the 2′OH group of the terminal adenosine ribose, while class II aaRS approaches from the side of major groove and attaches the amino acid to the 3′OH group [8]. The aaRSs are divided into two classes distinguished by their structures [8]. The term ‘class’ is used to distinguish both the enzymes and the amino acids that they activate [8]. Polarity and size are used to distinguish between the two classes of amino acids [37,39]. Class II amino acids occur significantly more frequently at the surfaces of proteins, whereas class I amino acids occur more frequently in their cores [39]. Notably, the two synthetases classes seem to have descended from ancestors coded by opposite strands of the same gene [48]. There is no need for the aaRS to recognize the anticodon in order to properly aminoacylate the tRNA. This means that the two codes coevolved right at the origin of translation. This encoding system seems now lost in the dimness of the past. Rodin & Ohno [49] found that the two families of aaRSs exhibit significant sequence similarity, but only when their coding sequences are compared in the opposite direction. This finding prompted Rodin & Ohno [49] to suggest that the two synthetases families originated as two-protein coding genes located on the complementary strands of the same primordial double-stranded RNA. Assuming that the partition into two mechanisms of tRNA-aminoacylation is a relic that dates back to the primordial genetic code in the RNA world, Delarue [3] proposed a simple model based upon successive binary choices for the assignment of codons to amino acids. Both Delarue’s [3] and Rodin & Rodin’s [7] models reorganize the codon table to reflect these contrasting molecular recognition modes by the two aaRS classes. These authors propose that this dual complementarity is frozen from an earlier stage in the code’s development, at which triplet reading frames had been established, but only the middle bases of the anticodons had been fixed, perhaps coinciding with the second step of Delarue’s differentiation genealogy [7]. They concluded that new codons were recruited in pairs, because translation of both sense and antisense strands would require that meaning be attached to both codons and their anticodons. We chose these models in order to prove the power of algebraic methods to understand each model and because our approach facilitates the comparison of the predictions among different models. In particular, the RO-model has a sound experimental background [37–40,48].

Herein the RO [7,49,50] and Delarue (D) [3] models for the origin of the genetic code are analysed in terms of its symmetrical properties. The RO- and D-models are asymmetrical. In this work, we assume a primeval RNY code [9], and make the same assumption of the RO-model, i.e. that the SGC can be divided according to the two classes of aaRSs I and II. We formulate isometries with which we arrive precisely to our symmetrical algebraic model [15,21,35].

The article is organized as follows. We start with some basic definitions of group theory and we provide the definition of the group action over the set of nucleotides. Then, we analyse the Rodin–Rodin model [50] of dividing the table of the genetic code according to the two classes of aaRSs. This table is symmetric but it is biologically incorrect. Then, we formulate simple isometric transformations that allow us to transform the RO-model which is asymmetric but biologically correct, into the SGC model based on the primeval RNY code and vice versa. We define an automorphism that converts the class aaRS I into the class aaRS II. We also model the asymmetric D-model into a symmetrical one by means of quotient groups. As a direct application of the 6D model of the SGC, we used the four scales of PR of each amino acid [41] and it neatly divides the SGC into four symmetrical groups. Finally, we discuss the results in terms of our model, which is compatible with the RO- and D-models and the primeval RNY code [9]. In other words, we have a unique 6D model, which is consistent with the RNY primeval genetic code and with the distribution of the two classes of aaRS.

2. Mathematical background

Group theory is a branch of abstract algebra that deals with the notion of symmetry of a geometrical object, making the set of symmetries of an object a group structure.

Table 1. The multiplication table of the Four-Klein group (K_4,○).

○	*e*	*a*	*b*	*ab*
e	*e*	*a*	*b*	*ab*
a	*a*	*e*	*ab*	*b*
b	*b*	*ab*	*e*	*a*
ab	*ab*	*b*	*a*	*e*

2.1. Definition of a group

A group is a set G with a binary operation ○ that combines any two elements of G and returns an element in G. This ordered pair is denoted as (G, ○) which satisfies the following properties:

1. Closure: For all *a*,*b* in G, the resulting element is also in G.
2. Associativity: For all *a*,*b*,*c* in G, the next equality holds: $(a \circ b) \circ c = a \circ (b \circ c)$.
3. Identity element: There exists an element *e* in G such that $a \circ e = e \circ a = a$ for all *a* in G.
4. Inverse element: For all *a* in G, there exists an element a' such that $a \circ a' = a' \circ a = e$, where *e* is the identity element.

2.2. Definition of a group action

If G is a group and X is a set then a group action is a function $\mathrm{f} : \mathrm{G} \times \mathrm{X} \to \mathrm{X}$, $(a, \mathrm{x}) \to a * \mathrm{x}$ that satisfies the following axioms:

1. Compatibility: For all *a*,*b* in G and all x in X the equality $(a \circ b) * \mathrm{x} = a * (b * \mathrm{x})$ holds.
2. Identity: For all x in X, $e * \mathrm{x} = \mathrm{x}$, where *e* is the identity element of G.

Then, it is said that G acts on X and X is a G− set.

A group action is the description of symmetries of an object using an external group. The essential elements of the object are described in a set and the operating group is known as the group of symmetries and its members correspond to some of the one-to-one transformations of the set. When considering a point $\mathrm{x} \in \mathrm{X}$ and the group G operating over X, the set $\mathrm{Gx} = \{\mathrm{g} * x | \mathrm{g} \in \mathrm{G}\}$ is called the orbit of the point X under the action of G. The set of orbits from a set X under the action of a group G is a partition of the set X, and it is known as the quotient set of the action, denoted by X/G.

2.3. Four-Klein group

Herein, we develop a novel and logically equivalent approach, where fewer algebraic properties are required, to that followed in our previous works [16,21,24] in which a group structure in the set N = {C, U, G, A} of the four nucleotides was defined. Herein, the ordering of the nucleotides and their arbitrary binary assignments are no longer necessary. A group is naturally constructed with the two types of mutations, transversions and transitions, represented by *a* and *b*, respectively. These two types of transformations are used like generators of the group with the property that the composition (denoted by ○) of a mutation with itself is equal to the identical mutation. The new approach starts with the symmetry group that corresponds to an abstract rectangle, which in group theory is known as the Four-Klein group, here symbolized as $(K_4, \circ)$, where $K_4 = \{e, a, b, ab = ba\}$ is the set, and ○ is the group operation (table 1). The Four-Klein group is identified as an abelian group in the direct product $\mathbb{Z}_2 \times \mathbb{Z}_2$, where $\mathbb{Z}_2 = \{0, 1\}$ represents the cyclic group of two elements. The set $\mathbb{Z}_2 \times \mathbb{Z}_2$ is regarded as the set of the four duplets of zeros and ones.

2.4. Group action in the set of nucleotides

Herein, the set of nucleotides N and its mutations will be considered. The Four-Klein group that will act over the set N, making it mutate, just as a rectangle is transformed in itself through its symmetries. This is represented as the Cayley graph of the group with the nucleotides as vertices. As an example consider the following: $a * (\mathrm{A}) = \mathrm{U}$, $a * (\mathrm{G}) = \mathrm{C}$, $b * (\mathrm{A}) = \mathrm{G}$, $b * (\mathrm{U}) = \mathrm{C}$, while $(a \circ b) * (\mathrm{A}) = (b \circ a) * (\mathrm{A}) = \mathrm{C}$, and $(a \circ b) * (\mathrm{U}) = (b \circ a) * (\mathrm{U}) = \mathrm{G}$. For the sake of simplicity, the symbols ○, ∗ and the parentheses will be here and further omitted where no misinterpretation can be made, so that $(a \circ b) * (\mathrm{A}) = ab\mathrm{A}$.

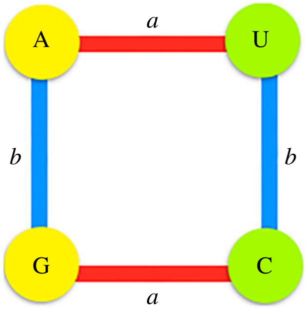

Figure 1. Representation of the action of the generators of the group over the set of nucleotides, where *a* represents transversions and *b* transitions. Purines are coloured in yellow and pyrimidines in green.

Now we extend our nucleotide level group action to the set of 64 triplets, $N \times N \times N = N^3$ as follows: $f: K_4^3 \times N^3 \rightarrow N^3, ((a_1, a_2, a_3), (x_1x_2x_3)) \rightarrow (a_1x_1, a_2x_2, a_3x_3)$, where we have used the vector notation and f is well defined because the mapping is component-wise.

A common classification of the nucleotides can be done through their chemical properties [24]. Herein, we consider purines and pyrimidines represented as R and Y, respectively, where R = {A, G} and Y = {C, U}. Next, we will deal with codons, which in set notation are the sets: RNY, YNR, YNY and RNR.

2.5. Defining a metric or distance in N^3

For a given choice of generators, one has to define a metric, i.e. the natural distance on the Cayley graph. Here, we have the group K_4 and its two generators a and b. The metric is defined in the following manner for x_1, x_2 in N, for single nucleotides:

1. $d(x_1, x_2) = 0$. If and only if $x_1 = ex_2$.
2. $d(x_1, x_2) = 1$. If and only if $x_1 = ax_2$ or $x_1 = bx_2$.
3. $d(x_1, x_2) = 2$. If and only if $x_1 = abx_2 = bax_2$.

This is a discrete metric that is similar to the one known as Hamming distance, but here the distance is given by the minimum number of generators of the group that are used to take one nucleotide and mutate it into another one. An extension in the definition of distance is natural for triplets so that it will be the sum of the distances of the nucleotides that conform the triplet. Formally, for two triplets $x_1y_1z_1$ and $x_2y_2z_2$, the distance is: $d(x_1y_1z_1, x_2y_2z_2) = d(x_1, x_2) + d(y_1, y_2) + d(z_1, z_2)$.

The genetic code is then represented as a 6D hypercube. This geometric figure can also be interpreted as a graph $G = (V, E)$ of vertices, representing the codons, and edges, joining the codons at distance one, making it possible to analyse its symmetries through the group of automorphisms of the graph. This group consists of all the bijective functions of the graph G, $f: (V, E) \rightarrow (V, E)$ that preserve its adjacencies. With the metric defined above, these automorphisms comprise all the isometric transformations of the cube. It is worth mentioning that there are, in essence, only three different Cayley graphs that determine the action of the group over the nucleotides. The pairs of opposite edges of the graph chosen here (figure 1) represent the generators of the group (transversions and transitions), which is in agreement with a common evolutionary interpretation [51]. In our previous approach [16,21,24], the distance of a codon and its anticodon in the 6D hypercube is at the maximum distance of 6. It is worth remarking that, if the Cayley graph associated with our previous works is used, the interchange of the action a for ab, and ab for a, applied as described above, will result in the same conclusions. Hence, the two approaches do not contradict each other, neither in biological aspects nor in mathematical ones, owing to the fact that with the present approach the ordering of nucleotides and arbitrary binary assignments are not required. In fact, the four nucleotides A,C,G,U can be situated at the vertices of a given rectangle in 4! = 24 ways. Interestingly, the assumption that a and b represent transversion and transition, respectively, being a the transversion that converts each nucleotide into its complementary, reduces all the possible graphs to only three.

3. The Rodin–Rodin model

In the original proposal made by Rodin & Ohno [49], the table of the genetic code is arranged in such a manner, that complementary codons appear vis-à-vis each other. Each of the 20 different aaRSs

Table 2. Symmetric table of the SGC that is biologically incorrect.

	U			A			G			C	
U	Phe	U	A	Arg	A	U	Cys	U	A	Thr	A
U	Phe	C	G	Glu	A	U	Cys	C	G	Ala	A
U	Leu	A	U	Stop	A	U	Stop	A	U	Ser	A
U	Leu	G	C	Gln	A	U	Trp	G	C	Pro	A
C	Leu	U	A	Arg	G	C	Arg	U	A	Thr	G
C	Leu	C	G	Glu	G	C	Arg	C	G	Ala	G
C	Leu	A	U	Stop	G	C	Arg	A	U	Ser	G
C	Leu	G	C	Gln	G	C	Arg	G	C	Pro	G
A	Ile	U	A	Asn	U	A	Ser	U	A	Thr	U
A	Ile	C	G	Asp	U	A	Ser	C	G	Ala	U
A	Ile	A	U	Tyr	U	A	Lys	A	U	Ser	U
A	Met	G	C	His	U	A	Lys	G	C	Pro	U
G	Val	U	A	Asn	C	G	Gly	U	A	Thr	C
G	Val	C	G	Asp	C	G	Gly	C	G	Ala	C
G	Val	A	U	Tyr	C	G	Gly	A	U	Ser	C
G	Val	G	C	His	C	G	Gly	G	C	Pro	C

Table 3. Biologically correct table of the SGC that is not symmetric. Phe and Tyr are ambiguous and they are marked with an asterisk.

	U			A			G			C	
U	Phe*	U	A	Lys	A	U	Cys	U	A	Thr	A
U	Phe*	C	G	Glu	A	U	Cys	C	G	Ala	A
U	Leu	A	U	Stop	A	U	Stop	A	U	Ser	A
U	Leu	G	C	Gln	A	U	Trp	G	C	Pro	A
C	Leu	U	A	Lys	G	C	Arg	U	A	Thr	G
C	Leu	C	G	Glu	G	C	Arg	C	G	Ala	G
C	Leu	A	U	Stop	G	C	Arg	A	U	Ser	G
C	Leu	G	C	Gln	G	C	Arg	G	C	Pro	G
A	Ile	U	A	Asn	U	A	Ser	U	A	Thr	U
A	Ile	C	G	Asp	U	A	Ser	C	G	Ala	U
A	Ile	A	U	Tyr*	U	A	Arg	A	U	Ser	U
A	Met	G	C	His	U	A	Arg	G	C	Pro	U
G	Val	U	A	Asn	C	G	Gly	U	A	Thr	C
G	Val	C	G	Asp	C	G	Gly	C	G	Ala	C
G	Val	A	U	Tyr*	C	G	Gly	A	U	Ser	C
G	Val	G	C	His	C	G	Gly	G	C	Pro	C

recognizes the cognate amino acid, and then attaches it to isoacceptor tRNAs with the corresponding anticodons. The operational code provides virtually errorless aminoacylation of tRNAs [6,12,13]. The 20 aaRSs are divided into two 10-member non-overlapping classes, I and II, that have virtually nothing in common with each other as far as the primary sequence, secondary elements and 3D structures are concerned [8].

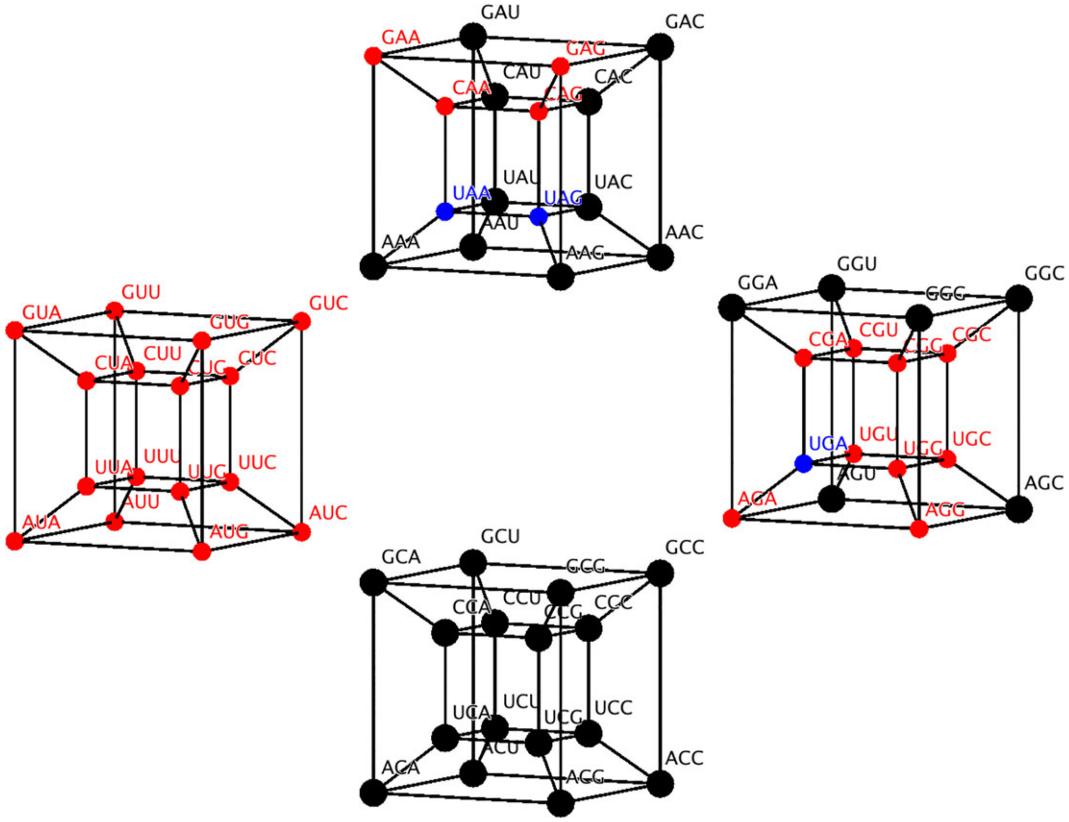

Figure 2. The six-dimensional cube of the genetic code coloured according to the aaRS class, class I is red and class II is black and bold. Stop codons (UUA, UAG and UGA) are in blue although the known cases of their 'capture' by amino acids are mostly from class I [52]. The edges joining the four-dimensional cube are not shown for better appreciation.

Table 4. The automorphisms used in each subcode of the SGC to interchange the aaRS classes.

T_1	T_2
RNY $\underset{(a,b,a)}{\leftrightarrow}$ YNR	RNY $\underset{(a,b,ab)}{\leftrightarrow}$ YNR
RNR $\underset{(e,b,e)}{\leftrightarrow}$ RNR	RNR $\underset{(e,b,b)}{\leftrightarrow}$ RNR
YNY $\underset{(e,b,e)}{\leftrightarrow}$ YNY	YNY $\underset{(e,b,b)}{\leftrightarrow}$ YNY

In their table, amino acids of class aaRS I are coloured in red, while those of class aaRS II are coloured in black (table 2). The amino acids from the first column of the code table tend to belong to class I (Phe being the only exception), whereas the amino acids from the second column all belong to class II.

3.1. A remarkable observation: a flaw in table 2

In table 2, there is a flaw, which conspires against the symmetries. Lysine and arginine are incorrectly placed. In arginine, two (AGA and AGG) out of its six coding triplets are incorrectly assigned to lysine, whereas the two triplets of lysine, AAA and AAG are assigned to arginine. Rodin & Rodin [50] and Rodin & Ohno [52,53] corrected table 2 [7,49], which is biologically correct but it is not symmetric (see table 3).

3.2. An automorphism that converts the class aaRS I into the class aaRS II and vice versa

The RO corrected table of codons associated to each class of aaRS lost symmetry, but in the 6D model this symmetry is recovered. Symmetries are represented with automorphisms of the cube that interchange the

Table 5. Automorphisms to convert the Rodin–Ohno model partitions of the genetic code into the RNR, RNY, YNR, YNY partitions.

F	F
RAR $\underset{(e,e,e)}{\leftrightarrow}$ RAR	RGR $\underset{(a,b,e)}{\leftrightarrow}$ YAR
YGR $\underset{(e,e,e)}{\leftrightarrow}$ YGR	YAR $\underset{(a,b,e)}{\leftrightarrow}$ RGR
RUY $\underset{(e,e,e)}{\leftrightarrow}$ RUY	RCY $\underset{(a,b,e)}{\leftrightarrow}$ YUY
YCY $\underset{(e,e,e)}{\leftrightarrow}$ YCY	YUY $\underset{(a,b,e)}{\leftrightarrow}$ RCY
RUR $\underset{(e,a,a)}{\leftrightarrow}$ RAY	RCR $\underset{(a,ab,a)}{\leftrightarrow}$ YAY
YCR $\underset{(e,a,a)}{\leftrightarrow}$ YGY	YUR $\underset{(a,ab,a)}{\leftrightarrow}$ RGY
RAY $\underset{(e,a,a)}{\leftrightarrow}$ RUR	RGY $\underset{(a,ab,a)}{\leftrightarrow}$ YUR
YGY $\underset{(e,a,a)}{\leftrightarrow}$ YCR	YAY $\underset{(a,ab,a)}{\leftrightarrow}$ RCR

codons of class I with class II and vice versa. In fact, there are two such functions, T_1 and T_2 defined piece-wise (table 4). These automorphisms form a subgroup that under composition yields a class invariant transformation $(T_1 \circ T_2) = T_3 = (e, e, b)$, which is a transition in the wobble position. In figure 2, the codons in the 6D model are coloured according to the aaRS class as in table 2 and table 3 but black is replaced by blue. Each isolated cube is actually a four-dimensional (4D) cube and the union of all of them with their respective extra edges forms the complete 6D cube. The edges joining each 4D cube are omitted for a better appreciation of the complete figure.

3.3. From the RO-model to the standard genetic code

According to the RO-model [49], the table of the genetic code can be divided into the sub-codes NAN, NGN, NUN, NCN. There exists an automorphism F of the cube defined also piece-wise, which transforms that division into the sub-codes RNR, YNR, RNY, YNY, respectively (table 5), which is precisely our algebraic model [16,21,35]. As an example, consider the codon AGC in the RO-model. AGC is an element of the RGY subcode, so the action required to transform it to our 6D model is (a,ab,a) as described in table 5. From the definition of the group action, this codon will be transformed to the triplet UUG. Note also that, owing to the order of the elements of the group, the same action over UUG on the 6D model will send it back to AGC in the RO-model.

4. The polar requirement in the six-dimensional SGC

PR was scaled into four categories [41]. We assign a particular colour (red, yellow, blue and green) to each scale. When such categories are set on the 6D genetic code, new symmetries emerge (figure 3). Now the SGC in six dimensions can be symmetrically divided into four colours according to the PR. Each category, or colour, comprises 16 codons that are arranged in 4D hypercubes, whose symmetry is given by the wreath product $S_2\mathrm{Wr}S_4$, where S_n is a permutation group of n elements [54]. Such group can be represented by the group of orthogonal matrices of 4×4 whose entries are all integers [54]. To interchange whole categories, it is sufficient to use the symmetries of a square Dih_4 (figure 3). Hence, the 6D representation of the SGC can reflect this property using its automorphisms as a biological classifier.

4.1. Delarue's model

Delarue [3] argues that the partition of codons according to the aaRS class distinction facilitated a hierarchical process by which additions to the code reduced codon ambiguity to produce the extant table with just five binary choices. The code started with undifferentiated and nonsense triplets, NNN. Codons were given meaning beginning with the second base and ending with the third. The NYN triplets could interact with a synthetase, whereas the NRN could not and remained stop codons. At each step, the ambiguous codon family differentiated to give descendants with opposite groove recognition, while descent of the stop codon family generated a new ambiguous family and retained a stop codon,

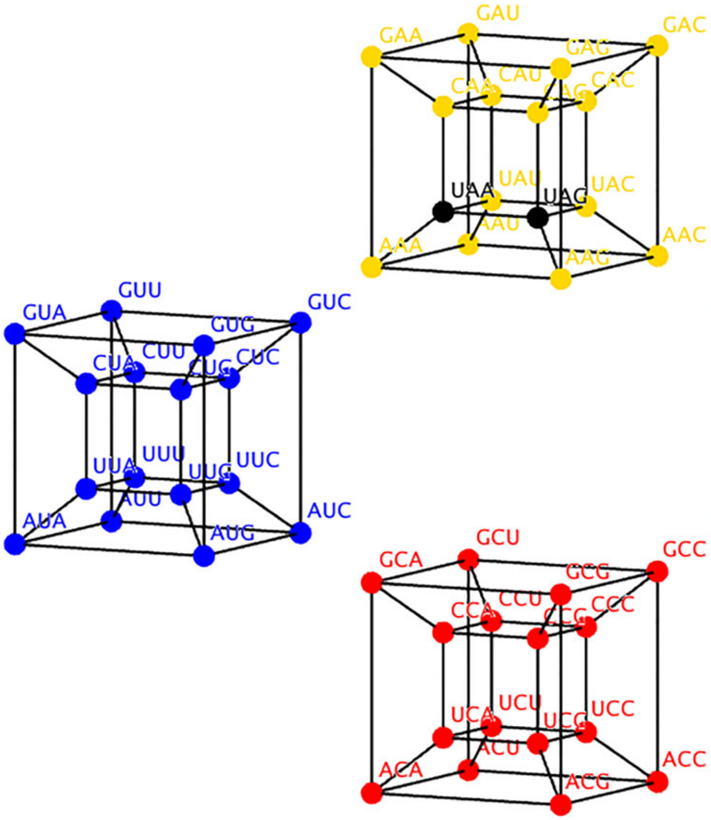

Figure 3. Six-dimensional hypercube of the SGC coloured by amino acid polar requirement values [41]. The four-dimensional hypercubes are yellow (upper); blue (left); red (lower); green (right); Stop codon are in black (UUA, UAG and UGA).

which was always present in the code. These asymmetric division rules provide a unique differentiation order, rendering the exhaustive exploration of the initial assignment of codons plausible, and suggesting that the appearance of the code conferred meaning successively from redundancy by a deterministic elimination of the most frequent errors. Notably, tRNAs with complementary anticodons also have statistically significant complementarity in their acceptor-stem operational codes [3].

With the concept of group action in mind, it is possible to analyse the D-model and elaborate an algebraic model. As the order 2 subgroup T generated by b is the group of transitions of the set N, $T = \{e, b\}$ is isomorphic to the cyclic group $\mathbb{Z}_2$. The quotient N/T represents precisely the partitions {R, Y} of the set of nucleotides. Considering the quotient N/e, where e is the trivial group, we obtain the nucleotides separated in different sets: $\mathrm{N}/e = \{\{A\}, \{\mathrm{G}\}, \{\mathrm{U}\}, \{\mathrm{C}\}\}$. Finally, the quotient with the entire group is a trivial operation, with only one class, as $\mathrm{N}/\mathrm{K}_4 = \mathrm{N}$. In order to analyse triplets, a component-wise operation naturally arises from these definitions resulting in: $\mathrm{NNN}/GGG = \mathrm{N}/G \times \mathrm{N}/G \times \mathrm{N}/G$, where G is any subgroup of K_4, i.e. analysing the quotients component by component and then relating them with the cartesian products of sets. Now the Delarue's model given by six binary choices can be algebraically analysed. The quotient $\mathrm{NNN}/\mathrm{K}_4 T \mathrm{K}_4 = \{\mathrm{NRN}, \mathrm{NYN}\}$, where $T = \{e,b\}$ is the subgroup of transitions, yields the first binary choice and for the next steps doing $\mathrm{NRN}/\mathrm{K}_4 e \mathrm{K}_4$ and $\mathrm{NYN}/\mathrm{K}_4 e \mathrm{K}_4$, respectively, and for the rest what is only needed is to use as quotients the products: $Te\mathrm{K}_4$, $ee\mathrm{K}_4$, eeT and eee in that order. We have just replaced the six binary decisions (including the wobbling assignments) in the D-model by six algebraic well-defined mathematical representations. The value of the latter is that we can follow the groups of symmetries in each step. Furthermore, we can make the model parsimonious and simpler, if we now make quaternary decisions so that the nucleotides in each position of the codon are determined at each step, by the use of only three group products, $\mathrm{K}_4 e \mathrm{K}_4$, $ee\mathrm{K}_4$ and eee.

5. Discussion

In this work, we have been able to formulate algebraic expressions for two well-known models of the origin and evolution of the genetic code, to wit, the RO-model and Delarue's model. Both models are consistent with the RNY code [9], as partitioning of aaRSs in two classes could have been encoded in a strand-symmetric RNA world [7,50]. We have shown that by assuming both a primeval RNY code and that the code can be divided into two classes of the aaRSs, we arrive at a symmetrical representation of the genetic code in a 6D algebraic model. We have also shown that PR displays a symmetrical pattern in this 6D model. PR is an empirical scale unrelated to either of the two transfer equilibria that best represent the partitioning of amino acids between pure phases, rather than between a pure phase and cellulose. PR seems to be also unrelated to other measures such as hydropathy. Further experimental work is needed to clarify these issues.

The aaRSs are a prime example of horizontal gene transfer [55,56]. Evolutionary replacements of aaRSs accompanied the evolution of the genetic code [31]. The assignments seemed to minimize errors in a primitive translation mechanism that was highly inaccurate [57,58]. The evolutionary phylogenies of synthetases do not obey the basic division of all life into the three primary groupings Bacteria, Archaea and Eukaryotes [56]. The two aaRS classes are presumably the oldest protein superfamilies. The RO hypothesis [52] implies that they arose at nearly the same instant in geological time because, at the nucleic acid level, the information necessary for function of each class is indistinguishable from that necessary for function of the other [40]. Complementarity means that one strand implies the existence of the other. Sense/antisense coding thus projects back past the genetic coding nexus to chemistry. The sense/antisense ancestry of the aaRS appears to be solidly established [40,59]. The authors, Rodin & Ohno, observed that their model is *almost perfectly symmetric* [49,52,53]. But in front of this unusual assertion we argue that something that is almost perfectly symmetric is not symmetric at all. Interestingly, the automorphisms T_1 and T_2 show the so-called symmetry that only exists in our 6D model, and the function F converts the partitions of the RO-model {NUN, NAN, NGN, NCN} into the partition {RNR, RNY, RNY, YNR}, which corresponds to our symmetric model. As the functions presented are isometric, the RO-model may be considered as equivalent to this one and it only takes a different point of view of the same model to reach one's conclusions from the other. The D-model is a phenomenological model of progressive differentiation-like reduction of codon ambiguity [60]. Indeed, it has been suggested that the primitive ribosome worked to synthesize peptides randomly, without the need of a code [61]. This elegant model is also based on the pattern of tRNA aminoacylation by class I and II aaRSs. However, in contrast with our complementarity-based model, Delarue's asymmetric model consists of a binary decision tree, like in a longitudinal differentiation process [3]. The whole SGC is derived from binary decisions but it remains unclear why the minor or major groove side is preferred in each particular step. We propose an algebraic model that accounts for the simultaneous selection of pairs of complementary triplets following the RO-model, and a set of six algebraic well-defined algebraic operations that account for the six binary decisions of the D-model. We have shown that the D-model can be built from simple operations of action groups. The preservation of symmetries is noteworthy. With only two transformations, we can derive, from a single codon, the 32 triplets forming the RNY and YNR subsets, as well as the 32 triplets comprising the sets RNR and YNY. All the transformations required for the construction are subgroups of K_4^3 which is the general group acting on the codon space, therefore making impossible the creation of new codons without a symmetry breaking which is the action of a new subset of operators.

Until now, participation of two aaRS classes in genetic coding has been rationalized as a result of successive binary choices [3] or as a means of avoiding coding ambiguity [60]. It has been shown that this distinction appears to be related to the complementary roles of class I and II amino acids in protein folding. Members of subclass IA (Leu, Ile, Val and Met) have aliphatic side-chains and are found in hydrophobic cores. Members of subclass IIA (Ser, Thr and His) are small amino acids with water-favouring side-chains. Subclasses B (with carboxyl, amide, primary amine side-chains) and C (aromatic) in both classes contain similar amino acids. Class I amino acids tend to be buried; those in class II remain largely on the surface. Class I amino acids allowed formation of non-polar cores and class II amino acids populated the surfaces of globular proteins. The linkage between classes arising from their sense/antisense ancestry [38,62] would be expected to simplify the search for reduced amino acid alphabets that may have been used during early protein evolution, leading to the universal genetic code. The order in which predictors emerge in the stepwise regressions discussed above is similar, but not identical to, the series of decisions by which Delarue suggested that genetic coding actually became fixed [3]. Although tRNA identity elements have probably been confounded by horizontal gene transfer [32], ancestral tRNA sequence reconstruction may clarify further how identity elements and the synthetase class recognition evolved.

With our approach, we have shown that the whole SGC can be derived starting from a pair of reverse complementary codons with just six steps or just three if we follow quaternary decisions. The present algebraic approach is general and abstract enough as it deals with the algebra from outside of the genetic code making it possible to build bridges among different models. This approach permits the direct comparison of different genetic models that otherwise would be difficult to perform. For example, the self-referential (SR) model for the formation of the SGC [14] is appealing because it considers a self-modifying genetic code that alters its own instructions while it is evolving. Consequently, the instruction path length is reduced and improves its performance and maintenance through the mechanism of natural selection. It is called SR because it is centred on the integration of self-feeding ribonucleoprotein structures where the protein and RNA activities are mutually stimulatory, after having been formed on top of the basic tRNA dimers. It assumes that during early stages of the formation of the SGC, protein synthesis was directed by tRNA dimers. The SR-model lacks experimental support but it is compatible with the appearance of the metabolic pathways [63]. The proposed dimer-directed transferase activity should be experimentally tested, either utilizing present-day tRNAs or the various kinds of mini-tRNAs that have been used as acceptors for the aaRS function or for spontaneous aminoacylation. The genetic eukaryotic anticode comprises 46 anticodons as there are not anticodons ending with adenine $(3' \to 5')$ direction. The group actions required to describe the symmetries of this model are given by the direct product $K_4 \times K_4 \times \mathbb{Z}_3$, where the last set is the cyclic group of three elements that corresponds to the rotations of a triangle. The cyclic groups are generated with one element so the biological interpretation of this action is ambiguous, in contrast with the generators of K_4 representing transitions and transversions. Another difference is that this model can only be fully described in five dimensions. These differences in the mathematical properties of the SR-model with our 6D model show that they are non-equivalent and that there is no smooth way to mathematically complete the SGC. Essentially, the problem lies in the fact that the group K_4 cannot be obtained from $\mathbb{Z}_3$. The SR-model lacks an explanation of how the dinucleotides formed the codons. Did they appear gradually? Or did codons appear simultaneously from a given set of principal dinucleotides? The chronology of appearance of codons is absent.

The partition of the table of the genetic code into the two classes of aaRSs is entirely consistent with the complementary symmetry of the RNA world in general, and the hypothesis of its initial double-strand coding in particular. It has been shown that the elimination of any amino acid encoded by the primeval RNY code would be strongly selected against and therefore at this stage the RNY code was already frozen [18]. The very existence of the ying-yang (formerly dubbed 'ying-yang-*like*' [7]) pattern of aminoacylation that certainly has little if anything at all to do with the present-day protein aaRSs, points to the 'anticodon first' scenario of the genetic code origin [64,65]. The anticodon is indeed essential for 17 of the 20 *Escherichia coli* isoaccepting groups [66]. The second operational code does not make sense without the anticodon code. However, the early relevance of the acceptor mini-helix in evolutionary development of the tRNA molecule cannot be understated [13,36,52,59,67,68]. Consistent with the hypothesis that the acceptor double-stranded stem is older than the anticodon loop, the GC-biased codon–anticodon-like triplet pairs located just next to the 73rd base-determinator of the acceptor stem may better reflect the very initial shaping stage of the genetic code than the single-stranded anticodon [50,53].

We have developed mathematical models for the RO-hypothesis and the D-genealogy. We highlight that these mathematical models are different despite the fact that they share the fundamental fact that the SGC can be divided by the two classes of aaRSs. We emphasize that our 6D model is completely equivalent to the mathematical model of the RO-hypothesis. The mathematical model of the SR-hypothesis underscores the differences with the other three models. The 6D symmetrical model has been enriched by the RO-model and the RO-model has acquired a sound mathematical structure. All presented models deal with the same biological aspects of the SGC, but differently. The 6D structure has been exploited not only for comparing different models but more importantly to give a step forward to unify models and reinforce (or weaken) models' hypotheses.

In conclusion, the most adequate model for the SGC can be represented in a 6D hypercube. Each dimension describes a type of mutation, transition and transversion as given by the Cayley graph, acting on each of three bases of any codon. Consequently, we obtain the six dimensions. When considering the hydropathy scale of amino acids [69], there are no symmetries that would interchange the four categories. However, if the codon UGA were associated with an amino acid that falls into the category of 'moderately hydrophobic', then the transformation (e, e, b) would be invariant to the hydropathy classes. In the same manner, when considering the polarity of amino acids [37–39], it would be needed that UGA were a non-polar amino acid, the transformation (e, e, b) would be invariant to polarity. If, in addition, the other stop codons are assigned to polar amino acids, the transformation (e, e, a) would be another invariant symmetry, as well as their composition. This means that a biological classification can also be interpreted as symmetries that would maintain the classification.

Undoubtedly, the 6D description of the genetic code as the hypercube $(\mathbb{Z}_2)^6$, becomes essential for a better understanding of the evolution of the code. The SGC, as derived from the primeval genetic code, and the RO-model are one and the same. We have shown that these different models of the genetic code are mathematically equivalent. Hence, the 6D algebraic model presented here unifies different models of the genetic code.

Data accessibility. The values of polar requirement were taken from [41]; the hydropathy scale of amino acids were taken from [69]. The rest of the work is essentially theoretical.

Authors' contributions. M.V.J. conceived the whole work, performed the analysis, interpretation, coordinated the research and wrote the manuscript; G.S.Z. and E.R.M. contributed to ideas, design, calculations, and wrote drafts of the manuscript. All authors approved the final version to be published.

Competing interests. The authors declare no competing interests.

Funding. M.V.J. was financially supported by PAPIIT-IN224015, UNAM, México. G.S.Z. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship number: 737920 from CONACYT.

Acknowledgements. We thank Juan R. Bobadilla for the technical computer support.

References

1. Darwin C. 1859 *The origin of species: by means of natural selection, or the preservation of favoured races in the struggle for life*. Cambridge, UK: Cambridge University Press.
2. Nirenberg MW, Matthaei JH. 1961 The dependance of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl Acad. Sci. USA* **47**, 1588–1602. (doi:10.1073/pnas.47.10.1588)
3. Delarue M. 2008 An asymmetric underlying rule in the assignment of codons: possible clue to a quick early evolution of the genetic code via successive binary choices. *RNA* **13**, 161–169. (doi:10.1261/rna.257607)
4. Crick FHC. 1968 The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379. (doi:10.1016/0022-2836(68)90392-6)
5. Crick FHC, Brenner S, Klug A, Pieczenik GA. 1976 Speculation on the origin of protein synthesis. *Orig. Life* **7**, 389–397. (doi:10.1007/BF00927934)
6. Ribas de Pouplana L, Schimmel P. 2001 Aminoacyl-tRNA synthetases: potential markers of genetic code development. *Trends Biochem. Sci.* **26**, 591–596. (doi:10.1016/S0968-0004(01)01932-6)
7. Rodin SN, Rodin SA. 2008 On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. *Heredity* **100**, 341–355. (doi:10.1038/sj.hdy.6801086)
8. Eriani G, Delarue M, Poch O, Gangloff J, Moras D. 1990 Partition of aminoacyl-tRNA synthetases into two classes based on mutually exclusive sets of conserved motifs. *Nature* **347**, 203–206. (doi:10.1038/347203a0)
9. Eigen M, Schuster P. 1978 The hypercycle: a principle of natural selection. *Naturwissenschaften* **65**, 341–369. (doi:10.1007/BF00439699)
10. Eigen M, Winkler-Oswatitsch R. 1981 Transfer-RNA: the early adaptor. *Naturwissenschaften* **68**, 217–228. (doi:10.1007/BF01047323)
11. Eigen M, Winkler-Oswatitsch R. 1981 Transfer-RNA, an early gene? *Naturwissenschaften* **68**, 282–292. (doi:10.1007/BF01047470)
12. De Duve C. 1988 The second genetic code. *Nature* **333**, 117–118. (doi:10.1038/333117a0)
13. Schimmel P, Giégé R, Moras D, Yokoyama S. 1993 An operational RNA code for amino acids and possible relationship to genetic code. *Proc. Natl Acad. Sci. USA* **90**, 8763–8768. (doi:10.1073/pnas.90.19.8763)
14. Guimarães RC, Costa Moreira CH, Farias ST. 2008 AP self-referential model for the formation of the genetic code. *Theory Biosci.* **127**, 249–270. (doi:10.1007/s12064-008-0043-y)
15. José MV, Morgado ER, Guimarães RC, Zamudio GS, Farías ST, Bobadilla JR, Sosa D. 2014 Three-dimensional algebraic models of the tRNA code and the 12 graphs for representing the amino acids. *Life* **4**, 341–373. (doi:10.3390/life4030341)
16. José MV, Morgado ER, Govezensky T. 2007 An extended RNA code and its relationship to the standard genetic code: an algebraic and geometrical approach. *Bull. Math. Biol.* **69**, 215–243. (doi:10.1007/s11538-006-9119-3)
17. Novozhilov AS, Wolf YI, Koonin E. 2007 Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *BM Centr. Biol. Dir.* **2**, 1–24. (doi:10.1186/1745-6150-2-1)
18. José MV, Zamudio GS, Palacios-Pérez M, Bobadilla JR, Farías ST. 2015 Symmetrical and thermodynamic properties of phenotypic graphs of amino acids encoded by the primeval RNY code. *Orig. Life Evol. Biosph.* **45**, 77–83. (doi:10.1007/s11084-015-9427-4)
19. Lewin B. 2000 *Genes* (vol. VII). New York, NY: Oxford University Press.
20. Crick FHC. 1966 Genetic code: yesterday, today and tomorrow. *Cold Spring Harb. Symp. Quant. Biol.* **31**, 1–5. (doi:10.1101/SQB.1966.031.01.006)
21. José MV, Morgado ER, Govezensky T. 2011 Genetic hotels for the standard genetic code: evolutionary analysis based upon novel three-dimensional algebraic models. *Bull. Math. Biol.* **73**, 1443–1476. (doi:10.1007/s11538-010-9571-y)
22. Sánchez R, Grau R, Morgado E. 2006 A novel Lie algebra of the genetic code over the Galois field of four DNA bases. *Math. Biosci.* **202**, 156–174. (doi:10.1016/j.mbs.2006.03.017)
23. Jiménez-Montaño MA, de la Mora-Basañez CR, Pöschel T. 1996 The hypercube structure of the genetic code explains conservative and non-conservative amino acid substitutions *in vivo* and *in vitro*. *Biosystems* **39**, 117–125. (doi:10.1016/0303-2647(96)01605-X)
24. José MV, Morgado ER, Sánchez R, Govezensky T. 2012 The 24 possible algebraic representations of the standard genetic code in six and three dimensions. *Adv. Stud. Biol.* **4**, 119–152.
25. Arquès DG, Michel CJ. 1996 A complementary circular code in the protein coding genes. *J. Theor. Biol.* **182**, 45–58. (doi:10.1006/jtbi.1996.0142)
26. Michel CJ, Pirillo G, Pirillo MA. 2008 A relation between trinucleotide comma-free codes and trinucleotide circular codes. *J. Theor. Biol.* **401**, 17–26. (doi:10.1016/j.tcs.2008.02.049)
27. Pohlmeyer R. 2008 The genetic code revisited. *J. Theor. Biol.* **253**, 623–624. (doi:10.1016/j.jtbi.2008.04.028)
28. Freeland SJ, Hurst LD. 1998 The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248. (doi:10.1007/PL00006381)
29. Eigen M *et al.* 1989 How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **244**, 673–679. (doi:10.1126/science.2497522)
30. Nicholas HB, McClain WH. 1995 Searching tRNA sequences for relatedness to aminoacyl-tRNA synthetase families. *J. Mol. Evol.* **40**, 482–486. (doi:10.1007/BF00166616)
31. Nagel GM, Doolittle RF. 1995 Phylogenetic analysis of aminoacyl-tRNA synthetases. *J. Mol. Evol.* **40**, 487–498. (doi:10.1007/BF00166617)
32. Woese CR, Olsen GJ, Ibba M, Söll D. 2000 Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* **64**, 202–236. (doi:10.1128/MMBR.64.1.202-236.2000)
33. Hornos JEM, Hornos YMM. 1993 Algebraic model for the evolution of the genetic code. *Phys. Rev. Lett.* **71**, 4401–4404. (doi:10.1103/PhysRevLett.71.4401)
34. Sánchez R, Morgado ER, Grau R. 2005 A genetic code boolean structure, I: the meaning of Boolean deductions. *Bull. Math. Biol.* **67**, 1–14. (doi:10.1016/j.bulm.2004.05.005)
35. José MV, Govezensky T, García JA, Bobadilla JR. 2009 On the evolution of the standard genetic code: vestiges of scale invariance from the RNA World in current prokaryote genomes. *PLoS ONE* **4**, e4340. (doi:10.1371/journal.pone.0004340)
36. Schimmel P. 1995 An operational RNA code for amino acids and variations in critical nucleotide

sequences in evolution. *J. Mol. Evol.* **40**, 531–536. (doi:10.1007/BF00166621)

37. Wolfenden R, Lewis CA, Yuan Y, Carter Jr CW. 2015 Temperature dependence of amino acid hydrophobicities. *Proc. Natl Acad. Sci. USA* **112**, 7484–7488. (doi:10.1073/pnas.1507565112)
38. Carter Jr CW, Wolfenden R. 2015 tRNA acceptor stem and anticodon bases form independent codes related to protein folding. *Proc. Natl Acad. Sci. USA* **112**, 7489–7494. (doi:10.1073/pnas.1507569112)
39. Carter Jr CW, Wolfenden R. 2016 tRNA acceptor-stem and anticodon bases embed separate features of amino acid chemistry. *RNA Biol.* **13**, 145–151. (doi:10.1080/15476286.2015.1112488)
40. Carter Jr CW *et al.* 2014 The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed. *Biol. Direct* **9**, 11. (doi:10.1186/1745-6150-9-11)
41. Woese CR, Dugre DH, Saxinger WC, Dugre SA. 1966 The molecular basis for the genetic code. *Proc. Natl Acad. Sci. USA* **55**, 966–974. (doi:10.1073/pnas.55.4.966)
42. Alff-Steinberger C. 1969 The genetic code and error transmission. *Proc. Natl Acad Sci. USA* **64**, 584–591. (doi:10.1073/pnas.64.2.584)
43. Mathew DC, Luthey-Schulten Z. 2008 On the physical basis of the amino acid polar requirement. *J. Mol. Evol.* **66**, 519–528. (doi:10.1007/s00239-008-9073-9)
44. Freeland SJ, Knight RD, Landwebber LF, Hurst LD. 2000 Early fixation of an optimal genetic code. *Mol. Biol. Evol.* **17**, 511–518. (doi:10.1093/oxfordjournals.molbev.a026331)
45. Haig D, Hurst LD. 1991 A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **33**, 412–417. (doi:10.1007/BF02103132)
46. Caporaso JG, Yarus M, Knight R. 2005 Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. *J. Mol. Evol.* **61**, 597–607. (doi:10.1007/s00239-004-0314-2)
47. Di Giulio M. 1989 The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J. Mol. Evol.* **29**, 288–293. (doi:10.1007/BF02103616)
48. Martinez-Rodriguez L *et al.* 2015 Functional class I and II amino acid-activating enzymes can be coded by opposite strands of the same gene. *J. Biol. Chem.* **290**, 19 710–19 725. (doi:10.1074/jbc.M115.642876)
49. Rodin SN, Ohno S. 1995 Two types of aminoacyl-tRNA synthetases originally encoded by complementary strands of the same nucleic acid. *Orig. Life Evol. Biosph.* **25**, 565–589. (doi:10.1007/BF01582025)
50. Rodin SN, Rodin SA. 2006 Partitioning of aminoacyl-tRNA synthetases in two classes could have been encoded in a strand-symmetric RNA world. *DNA Cell Biol.* **25**, 617–626. (doi:10.1089/dna.2006.25.617)
51. Kimura M. 1981 Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl Acad. Sci. USA* **78**, 454–458. (doi:10.1073/pnas.78.1.454)
52. Rodin S, Rodin A, Ohno S. 1996 The presence of codon-anticodon pairs in the acceptor stem of tRNAs. *Proc. Natl Acad. Sci. USA* **93**, 4537–4542. (doi:10.1073/pnas.93.10.4537)
53. Rodin SN, Ohno S. 1997 Four primordial modes of tRNA synthetase recognition, determined by the (G,C) operational code. *Proc. Natl Acad. Sci. USA* **94**, 5183–5188. (doi:10.1073/pnas.94.10.5183)
54. Young A. 1930 On quantitative substitutional analysis 5. *Proc. Lond. Math. Soc. Second Ser.* **31**, 273–288. (doi:10.1112/plms/s2-31.1.273)
55. Wolf YI, Aravind L, Grishin NV, Koonin EV. 1999 Evolution of aminoacyl-tRNA synthetases: analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfers. *Genet. Res.* **9**, 689–710.
56. Woese CR, Kandler O, Wheelis ML. 1990 Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA* **87**, 4576–4579. (doi:10.1073/pnas.87.12.4576)
57. Woese CR. 1965 On the evolution of the genetic code. *Proc. Natl Acad. Sci. USA* **54**, 1546–1552. (doi:10.1073/pnas.54.6.1546)
58. Woese CR. 1973 Evolution of the genetic code. *Naturwissenschaften* **60**, 447–459. (doi:10.1007/BF00592854)
59. Rodin AS, Rodin SN, Carter Jr CW. 2009 On primordial sense–antisense coding. *J. Mol Evol.* **69**, 555–567. (doi:10.1007/s00239-009-9288-4)
60. Carter Jr CW. 2008 Thawing the ‘Frozen Accident’. *Heredity* **100**, 339–340. (doi:10.1038/hdy.2008.7)
61. Belousoff MJ, Davidovich C, Bashan A, Yonath A. 2010 On the development towards the modern world: a plausible role of uncoded peptides in the RNA world. In *Origins of life and evolution of biospheres* (eds K Ruiz-Mirazo, PL Luisi), pp. 415–419. Berlin, Germany: Springer.
62. Chandrasekaran SN, Yardimici GG, Erdogan O, Roach J, Carter Jr CW. 2013 Statistical evaluation of the Rodin-Ohno hypothesis: sense/antisense coding of ancestral class I and II aminoacylt-tRNA synthetases. *Mol. Biol. Evol.* **30**, 1588–1604. (doi:10.1093/molbev/mst070)
63. Guimarães RC. 2011 Metabolic basis for the self-referential genetic code. *Orig. Life Evol. Biosph.* **41**, 357–371. (doi:10.1007/s11084-010-9226-x)
64. Rodin AS, Szathmáry E, Rodin SN. 2011 On origin of genetic code and tRNA before translation. *Biol. Direct* **6**, 14. (doi:10.1186/1745-6150-6-14)
65. Szathmáry E. 1991 Codon swapping as a possible evolutionary mechanism. *J. Mol. Evol.* **32**, 178–182. (doi:10.1007/BF02515390)
66. Saks MS, Sampson JR, Abelson JN. 1994 The transfer RNA identity problem: a search for rules. *Science* **263**, 191–197. (doi:10.1126/science.7506844)
67. Fox GE, Naik AK. 2004 The evolutionary history of the translation machinery. In *The genetic code and the origin of life* (ed. LR de Pouplana), pp. 92–105. New York, NY: Landes Bioscience.
68. Maizels N, Weiner AM. 1994 Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc. Natl Acad. Sci. USA* **91**, 6729–6734. (doi:10.1073/pnas.91.15.6729)
69. Farías ST, Costa Moreira CH, Guimarães RC. 2007 Structure of the genetic code suggested by the hydropathy correlation between anticodons and amino acid residues. *Orig. Life Evol. Biosph.* **37**, 83–103. (doi:10.1007/s11084-006-9008-7)