MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **79** (2018) 527-560

ISSN 0340 - 6253

Symmetric Group of the Genetic–Code Cubes. Effect of the Genetic–Code Architecture on the Evolutionary Process

Robersy Sanchez*

Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68588-0660, USA

rus547@psu.edu

(Received July 13, 2017)

Abstract

The current evidence supports that the genetic code architecture is optimized to minimize the transcriptional and translational errors and to preserve amino-acid hydrophobicity during mutational events. The genetic code is mathematically equivalent to a cube inserted in the ordinary three-dimensional (3D) space, which leads to consistent phylogenetic analyses of DNA protein-coding regions. Herein, the symmetric group $(GC,\circ)$ of the genetic-code cubes is formally developed. Next, it is shown that principal component (PC) scales of amino-acid derived from subsets of the genetic-code cubes are highly correlated with hydrophobicity and other physicochemical amino-acid properties. The effect of this architecture on the evolutionary process was modelled by a Weibull probability distribution to fit the evolutionary mutational cost estimated using amino acid PC-scales optimized on a set of homologous proteins. The application of Weibull model permits the identification of mutational events with high and low probabilities of fixation in gene populations. It is illustrated how this approach conveys a valuable information for *de novo* vaccine design.

* Corresponding address: 360 North Frear, Department of Biology. Eberly College of Science. University Park. Penn State University. PA 16802. USA.

1 Introduction

The genetic code is the biochemical system used to establish the rules by which the DNA nucleotide sequence is transcribed into mRNA codon sequences, and ultimately translated into amino acid protein sequences. This code is an extension of the four-letter alphabet of the DNA bases: adenine, guanine, cytosine and thymine (denoted A, G, C, T), with uracil (U) for thymine in RNA. It has been shown that the genetic code is mathematically equivalent to a cube inserted in the three-dimensional (3D) space R^3 [1,2]. An introductory summary to the subject is provided in Appendix A.

The genetic-code architecture has been studied in the framework of the genetic-code algebraic structures [1–5]. The standard genetic-code cube was introduced in reference [1] as a geometrical model of the standard genetic code presented in Table 1. The standard genetic-code cube is also a 3D vector space over the Galois field *GF*(4) defined on the set of four DNA bases [1].

Table 1. The standard genetic code table.

First base position	Second base position: U		C		A		G		Third base position
	U		C		A		G		
U	UUU	[1]P	UCU	S	UAU	Y	UGU	C	U
	UUC		UCC		UAC		UGC		C
	UUA	L	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG		UGG	W	G
C	CUU	L	CCU	P	CAU	H	CGU	R	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Q	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	I	ACU	T	AAU	N	AGU	S	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	K	AGA	R	A
	AUG	M	ACG		AAG		AGG		G
G	GUU	V	GCU	A	GAU	D	GGU	G	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	E	GGA		A
	GUG		GCG		*Gag*		GGG		G

[1]The one letter symbol of amino acids.

The 3D genetic-code vector space on the Galois field *GF*(4) was derived from the quantitative relationship of the Watson-Crick DNA base-pairing, initially described in [3]), and codon order according to the evolutionary importance of their bases: from the less (base *Z* in codons *XYZ*) to the most important base, the second codon position *Y* (Table 1) [4,6]. Though Table 1 was initially built *ad hoc* based on empirical observations [6], it has been shown that the corresponding columns are mathematically determined in the standard genetic code 3D vector space [1,2]. Indeed, these columns are mathematically derived as quotient subspaces of the standard genetic-code cube, with strong associations with the amino acid physicochemical properties [1,2,4]. In more recent work, it was shown that the 24 possible ways to order the set of bases leads to 24 possible cubes of the standard genetic code [5].

The classification of the 24 possible cubes representations of the genetic code was based on IUPAC criteria [5,7], as given in Appendix B. In section 2 of the current manuscript, it will be shown that this classification leads to 24 algebraic representations of the extended genetic-code cubes over the Galois field *GF*(5), i.e., over the set $\mathbb{Z}_5$ of integers module 5 [2]. However, since all these cubes are isomorphic to the cube $\mathbb{Z}_5^3 = \mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z}_5$, the genetic-code cube is unique up to isomorphism. Moreover, the standard genetic-code cube universality is based on its architecture that depends on the physicochemical properties of the DNA bases rather on a specific encoding system. In contrast, the genetic code is not universal as there are small variations for the amino acids encoded for in archaebacteria, bacteria, chloroplasts, and mitochondria.

To guide the reader across the manuscript, a graphical summary is given in Fig. 1. The formal derivation of the symmetric group of the genetic-code cubes is given in the next section 2. Applications of the theory are provided in section 3 and 4. A concrete example application of the theory to the estimation of immunoescape variants fixation probabilities is given in sections 4.1 with *Env* and *Gag* HIV1 proteins. The application to *de novo* vaccine design (Fig. 1) is discussed in section 5.

A graphic user interface with an interactive didactic introduction to the mathematical biology background is provided in a computable document format (CDF) (free available at a link provided in Appendix C). All the data and tools required to check the claims and results presented in this manuscript are provided in Appendix C. Although all the results presented in this manuscript were derived analytically (and can be derived by readers as well), the graphical-user-interfaces available in Appendix C will support a fast application of the results presented here, as well as, a fast comprehension of the subject for readers not familiar with abstract algebra.

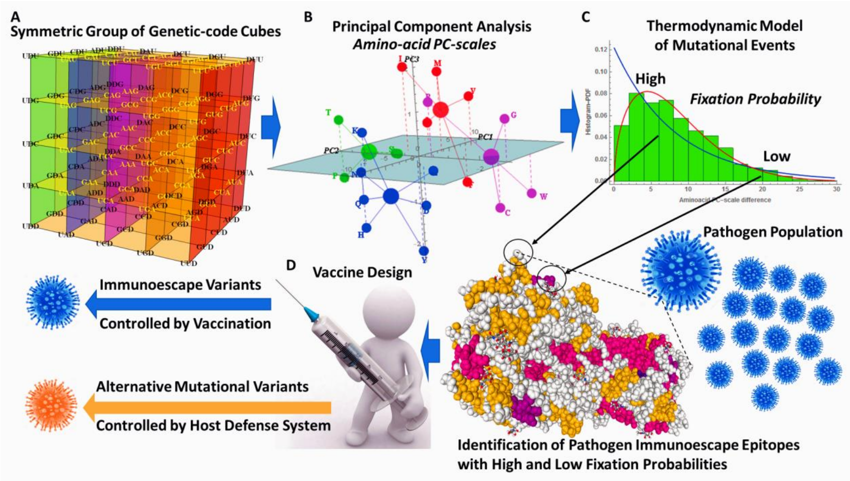


Figure 1. Graphical summary of the subjects covered by this work. **A**, the development of the symmetric group of the genetic-code cubes is presented. **B**, amino-acid PC-scales from codon norms are derived from subsets of the genetic-code cubes and optimized on a set of homologous proteins. It is shown that the amino-acid PC-scales are correlated with the physicochemical indexes reported by studies on protein folding and protein interactions. **C**, a Weibull probability distribution model based on the thermodynamics of the mutational process on gene populations is estimated on experimental datasets of aligned mutational variants of protein sequences. **D**, a feasible application of this result to *de novo* vaccine design is provided.

2 The group of the genetic code cubes $(GC,\circ)$

Twenty-four algebraic representations of the extended genetic code can be defined on the twenty-four sets $B^3 = B \times B \times B$, where B runs over the twenty-four ordered sets of bases {A, C, G, U}. i.e., $B \in \{\{D, A, C, G, U\}, ..., \{D, C, G, A, U\}\}$ (Appendix A). Each cube is named according to the base ordering used to build it. Cubes are classified based on the physicochemical criteria used to ordering the set of codons (Appendix B): number of hydrogen bonds (strong-weak, *SW*), chemical type (purine-pyrimidine, *YR*), and chemical groups (amino versus keto, *MK*).

Since the extended base D remains invariant, there are 24 representations of the extended genetic-code cube (Fig. 2, Table A1 from Appendix A, and Appendix C section 2). The algebraic operations are defined over the Galois field $GF(5)$ as in reference [2] and not over $GF(4)$ as in references [1].

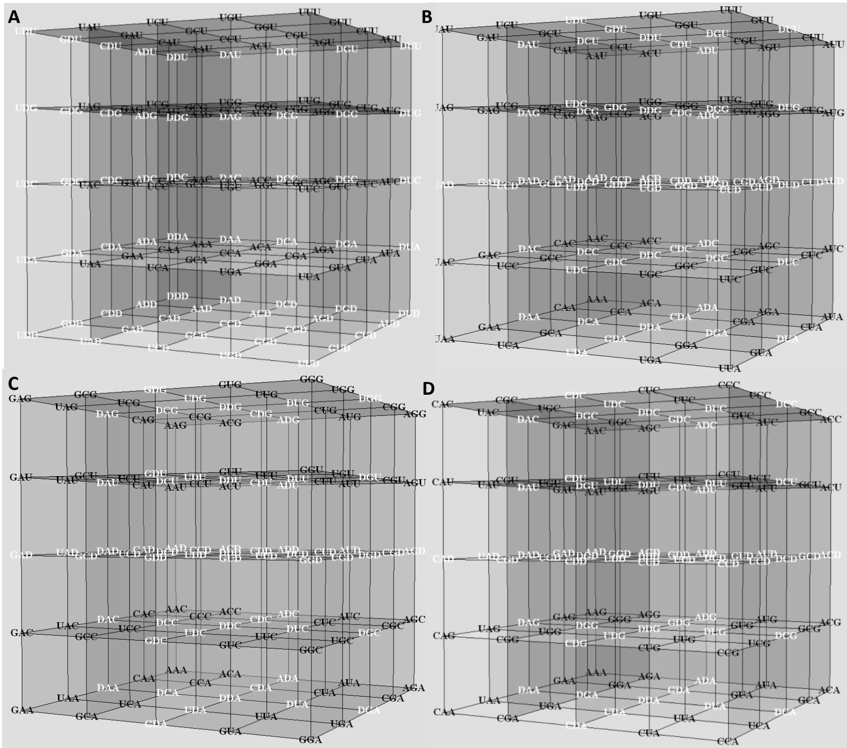


Figure 2. The genetic-code cubes. **A**: cube ACGU centered at codon CCC; **B**, **C** and **D** denote cubes ACGU, ACUG, and AGUC centered at codon DDD, respectively. In **A**, the standard genetic-code cube is inserted (codons in black) in the extended genetic-code cube. In **B**, **C** and **D**, codons of the standard genetic code are in the eight corner cubes, whilst the ancient codons are in the coordinated planes. The codons found in every vertical plane correspond to the main columns in Table 1, and codons found in every vertical line encode for the same amino acid or for an amino acid with similar physicochemical properties. The 24 genetic-code cubes can be visualized using the CDF-1 (section 3) given in Appendix C.

The sum operation is defined (as in [2]), for example, over the ordered set of bases B = {D, A, C, G, U} in such a way that the DNA complementary bases are also complementary algebraic elements (Table 2). That is, for the cube analyzed in [2] (shown in Fig. 2D) and for the eight *SW* cubes, the equalities $\mathrm{A}+\mathrm{U}=\mathrm{D}$ and $\mathrm{C}+\mathrm{G}=\mathrm{D}$ hold (Appendix C, CDF-1, sections 1 and 2.3). The physicochemical criteria listed in Appendix B are the basis to define the sum operations in the rest of the 24 possible algebraic structures of the extended genetic-code cubes. The set of 24 genetic-code cubes shall be denoted *GC*. For each class of *GC* cubes, there are eight ways to define the sum operation over the set of bases, depending on their order.

Table 2. Operation tables of the Galois field (*GF*(5)) on the ordered set of the extended bases alphabet *B*={D, A, C, G, U}, and on $\mathbb{Z}_5$.

Sum						Product					
$+_B$	**D**	**A**	**C**	**G**	**U**	**.**	**D**	**A**	**C**	**G**	**U**
D	D	A	C	G	U	**D**	D	D	D	D	D
A	A	C	G	U	D	**A**	D	A	C	G	U
C	C	G	U	D	A	**C**	D	C	U	A	G
G	G	U	D	A	C	**G**	D	G	A	U	C
U	U	D	A	C	G	**U**	D	U	G	C	A
+	**0**	**1**	**2**	**3**	**4**	**•**	**0**	**1**	**2**	**3**	**4**
0	0	1	2	3	4	**0**	0	0	0	0	0
1	1	2	3	4	0	**1**	0	1	2	3	4
2	2	3	4	0	1	**2**	0	2	4	1	3
3	3	4	0	1	2	**3**	0	3	1	4	2
4	4	0	1	2	3	**4**	0	4	3	2	1

The algebraic complementarity of the elements is preserved in all the cubes from the same class. For example, for all the cubes from class *MK*, the algebraic complementary elements for the sum operation according to the chemical type are: $\underbrace{\mathrm{A}+\mathrm{C}=\mathrm{D}}_{\text{Amino}},\underbrace{\mathrm{G}+\mathrm{U}=\mathrm{D}}_{\text{Keto}}$ (Appendix C, CDF-1, sections 1.2 and 2). As a result, we can define 24 groups $(B,+_i)$, where symbol "$+_i$" denotes the subjacent sum operation defined for the group ($i = 1,\ldots, 24$).

For cubes ACGU and UGCA, the pairwise alignment $\begin{pmatrix}\mathrm{ACGU}\\ \mathrm{UGCA}\end{pmatrix}$ of the ordered bases match in terms of hydrogen bonds and algebraic complementarity. Likewise, the pairwise alignment of the ordered bases from cubes AGUC and CUGC, $\begin{pmatrix}\mathrm{AGUC}\\ \mathrm{CUGA}\end{pmatrix}$, match in terms of chemical types and algebraic complementarity. The definition of a sum operation over the base set B = {D, A, C, G, U} is equivalent to define an order on the set of bases [4,8]. Thus, there is a bijection between the elements of the set of 24 groups $(B,+_i)$ and the elements of the symmetric group of degree four S_4. This is the group of all bijections $\Omega \to \Omega$, where $\Omega = \{1,2,3,4\}$. The elements of the group S_4 (or $S(\Omega)$) are permutations (also called substitutions).

The definition of the symmetric group over the set of 24 permutations of the four DNA bases follows straightforward from the usual definition of S_4. This group shall be denoted $(S_B,\circ)$ where symbol "$\circ$" stands for the product operation. If the set with base order (ACGU) is taken as unit element for the group operation, then we shall denote it as $(S_B^{\mathrm{ACGU}},\circ)$. This means that the base order (ACGU) (which is also the lexicographic order) corresponds to the identity permutation $\begin{pmatrix}\text{ACGU}\\ \text{ACGU}\end{pmatrix}$ and any other set with base order $i_1\, i_2\, i_3\, i_4$ corresponds to the permutation: $\begin{pmatrix}\text{A} & \text{C} & \text{G} & \text{U}\\ i_1 & i_2 & i_3 & i_4\end{pmatrix}$, where $i_k \in \{\text{A}, \text{C}, \text{G}, \text{U}\}$. Next, the multiplication of two permutations from group $(S_B^{\text{ACGU}}, \circ)$ follows the general rule for the composition of two permutations (Appendix C, CDF-1 section 5.1). For example, the multiplication of permutations $\sigma = \begin{pmatrix}\text{ACGU}\\ \text{CGUA}\end{pmatrix}$ and $\tau = \begin{pmatrix}\text{ACGU}\\ \text{UGCA}\end{pmatrix}$ is:

$$\begin{pmatrix}\text{ACGU}\\ \text{CGUA}\end{pmatrix} \circ \begin{pmatrix}\text{ACGU}\\ \text{UGCA}\end{pmatrix} = \begin{matrix}\text{A} & \text{C} & \text{G} & \text{U}\\ \downarrow & \downarrow & \downarrow & \downarrow\\ \text{U} & \text{G} & \text{C} & \text{A}\\ \downarrow & \downarrow & \downarrow & \downarrow\\ \text{A} & \text{U} & \text{G} & \text{C}\end{matrix} = \begin{pmatrix}\text{ACGU}\\ \text{AUGC}\end{pmatrix}$$

Once the unit element for the group operation is set, the base order $i_1i_2i_3i_4$ also specifies the permutation $\begin{pmatrix}\text{A} & \text{C} & \text{G} & \text{U}\\ i_1 & i_2 & i_3 & i_4\end{pmatrix}$. That is, following the usual formal notation, it is not ambivalent to denote permutation $\begin{pmatrix}\text{ACGU}\\ \text{AUGC}\end{pmatrix}$ by the abbreviated expression (AUGC). Thus, for the sake of simplicity, the expression $i_1i_2i_3i_4$ will represent both, permutation and base order. Moreover, since each base order determines a sum operation where the sum of each pair of algebraic complementary bases is base D, each base-order/permutation determines a cube and, consequently, the expression $i_1i_2i_3i_4$ also stands for a genetic-code cube. That is, the symmetric group $(S_B^{\text{ACGU}}, \circ)$ induces a group structure over the set of the 24 cubes representations of the genetic code.

Since each genetic-code cube was derived from a given base order, the multiplication of two cubes is determined by the multiplication of the corresponding permutations. For instance, for the above multiplication of permutations, the product of cubes with base orders CGUA and UGCA is the cube with base order AUGC, which is specified by the permutation $\sigma \circ \tau$ (Appendix C, CDF-1 section 5.1). We shall denote this group as the symmetric group of the genetic code cubes $(GC^{\text{ACGU}}, \circ)$. To build this group, we have chosen cube ACGU as the unit element. The Cayley multiplications table and graph for group $(GC^{\text{ACGU}}, \circ)$ are given in Table 3 and in Fig. 3, respectively (Appendix C, CDF-1 sections 5.1 to 5.2). Notice that by construction, group $(GC^{\text{ACGU}}, \circ)$ is isomorphic to the symmetric group S_4. In consequence, group $(GC^{\text{ACGU}}, \circ)$ is also isomorphic to the tetrahedral group T formed by the set of symmetry operations which all leave at least one of the four vertices of the regular tetrahedron unmoved. This isomorphism is clear after numbering the four vertices of a regular tetrahedron as 1, 2, 3, and 4.

Table 3. Cayley multiplications table for the symmetric groups of DNA base permutations $\left(S_B^{\mathrm{ACGU}},\circ\right)$ and genetic-code cubes $\left(GC^{\mathrm{ACGU}},\circ\right)$.

	SW								*YR*								*MK*							
	ACGU	**AGCU**	**UCGA**	**UGCA**	**CAUG**	**CUAG**	**GAUC**	**GUAC**	**ACUG**	**AUCG**	**GCUA**	**GUCA**	**CAGU**	**CGAU**	**UAGC**	**UGAC**	**AGUC**	**AUGC**	**CGUA**	**CUGA**	**GACU**	**GCAU**	**UACG**	**UCAG**
ACGU	ACGU	AGCU	UCGA	UGCA	CAUG	CUAG	GAUC	GUAC	ACUG	AUCG	GCUA	GUCA	CAGU	CGAU	UAGC	UGAC	AGUC	AUGC	CGUA	CUGA	GACU	GCAU	UACG	UCAG
AGCU	AGCU	ACGU	UGCA	UCGA	GAUC	GUAC	CAUG	CUAG	AGUC	AUGC	CGUA	CUGA	GACU	GCAU	UACG	UCAG	ACUG	AUCG	GCUA	GUCA	CAGU	CGAU	UAGC	UGAC
UCGA	UCGA	UGCA	ACGU	AGCU	CUAG	CAUG	GUAC	GAUC	UCAG	UACG	GCAU	GACU	CUGA	CGUA	AUGC	AGUC	UGAC	UAGC	CGAU	CAGU	GUCA	GCUA	AUCG	ACUG
UGCA	UGCA	UCGA	AGCU	ACGU	GUAC	GAUC	CUAG	CAUG	UGAC	UAGC	CGAU	CAGU	GUCA	GCUA	AUCG	ACUG	UCAG	UACG	GCAU	GACU	CUGA	CGUA	AUGC	AGUC
CAUG	CAUG	CUAG	GAUC	GUAC	ACGU	AGCU	UCGA	UGCA	CAGU	CGAU	UAGC	UGAC	ACUG	AUCG	GCUA	GUCA	CUGA	CGUA	AUGC	AGUC	UCAG	UACG	GCAU	GACU
CUAG	CUAG	CAUG	GUAC	GAUC	UCGA	UGCA	ACGU	AGCU	CUGA	CGUA	AUGC	AGUC	UCAG	UACG	GCAU	GACU	CAGU	CGAU	UAGC	UGAC	ACUG	AUCG	GCUA	GUCA
GAUC	GAUC	GUAC	CAUG	CUAG	AGCU	ACGU	UGCA	UCGA	GACU	GCAU	UACG	UCAG	AGUC	AUGC	CGUA	CUGA	GUCA	GCUA	AUCG	ACUG	UGAC	UAGC	CGAU	CAGU
GUAC	GUAC	GAUC	CUAG	CAUG	UGCA	UCGA	AGCU	ACGU	GUCA	GCUA	AUCG	ACUG	UGAC	UAGC	CGAU	CAGU	GACU	GCAU	UACG	UCAG	AGUC	AUGC	CGUA	CUGA
ACUG	ACUG	AUCG	GCUA	GUCA	CAGU	CGAU	UAGC	UGAC	ACGU	AGCU	UCGA	UGCA	CAUG	CUAG	GAUC	GUAC	AUGC	AGUC	CUGA	CGUA	UACG	UCAG	GACU	GCAU
AUCG	AUCG	ACUG	GUCA	GCUA	UAGC	UGAC	CAGU	CGAU	AUGC	AGUC	CUGA	CGUA	UACG	UCAG	GACU	GCAU	ACGU	AGCU	UCGA	UGCA	CAUG	CUAG	GAUC	GUAC
GCUA	GCUA	GUCA	ACUG	AUCG	CGAU	CAGU	UGAC	UAGC	GCAU	GACU	UCAG	UACG	CGUA	CUGA	AGUC	AUGC	GUAC	GAUC	CUAG	CAUG	UGCA	UCGA	AGCU	ACGU
GUCA	GUCA	GCUA	AUCG	ACUG	UGAC	UAGC	CGAU	CAGU	GUAC	GAUC	CUAG	CAUG	UGCA	UCGA	AGCU	ACGU	GCAU	GACU	UCAG	UACG	CGUA	CUGA	AGUC	AUGC
CAGU	CAGU	CGAU	UAGC	UGAC	ACUG	AUCG	GCUA	GUCA	CAUG	CUAG	GAUC	GUAC	ACGU	AGCU	UCGA	UGCA	CGUA	CUGA	AGUC	AUGC	GCAU	GACU	UCAG	UACG
CGAU	CGAU	CAGU	UGAC	UAGC	GCUA	GUCA	ACUG	AUCG	CGUA	CUGA	AGUC	AUGC	GCAU	GACU	UCAG	UACG	CAUG	CUAG	GAUC	GUAC	ACGU	AGCU	UCGA	UGCA
UAGC	UAGC	UGAC	CAGU	CGAU	AUCG	ACUG	GUCA	GCUA	UACG	UCAG	GACU	GCAU	AUGC	AGUC	CUGA	CGUA	UGCA	UCGA	AGCU	ACGU	GUAC	GAUC	CUAG	CAUG
UGAC	UGAC	UAGC	CGAU	CAGU	GUCA	GCUA	AUCG	ACUG	UGCA	UCGA	AGCU	ACGU	GUAC	GAUC	CUAG	CAUG	UACG	UCAG	GACU	GCAU	AUGC	AGUC	CUGA	CGUA
AGUC	AGUC	AUGC	CGUA	CUGA	GACU	GCAU	UACG	UCAG	AGCU	ACGU	UGCA	UCGA	GAUC	GUAC	CAUG	CUAG	AUCG	ACUG	GUCA	GCUA	UAGC	UGAC	CAGU	CGAU
AUGC	AUGC	AGUC	CUGA	CGUA	UACG	UCAG	GACU	GCAU	AUCG	ACUG	GUCA	GCUA	UAGC	UGAC	CAGU	CGAU	AGCU	ACGU	UGCA	UCGA	GAUC	GUAC	CAUG	CUAG
CGUA	CGUA	CUGA	AGUC	AUGC	GCAU	GACU	UCAG	UACG	CGAU	CAGU	UGAC	UAGC	GCUA	GUCA	ACUG	AUCG	CUAG	CAUG	GUAC	GAUC	UCGA	UGCA	ACGU	AGCU
CUGA	CUGA	CGUA	AUGC	AGUC	UCAG	UACG	GCAU	GACU	CUAG	CAUG	GUAC	GAUC	UCGA	UGCA	ACGU	AGCU	CGAU	CAGU	UGAC	UAGC	GCUA	GUCA	ACUG	AUCG
GACU	GACU	GCAU	UACG	UCAG	AGUC	AUGC	CGUA	CUGA	GAUC	GUAC	CAUG	CUAG	AGCU	ACGU	UGCA	UCGA	GCUA	GUCA	ACUG	AUCG	CGAU	CAGU	UGAC	UAGC
GCAU	GCAU	GACU	UCAG	UACG	CGUA	CUGA	AGUC	AUGC	GCUA	GUCA	ACUG	AUCG	CGAU	CAGU	UGAC	UAGC	GAUC	GUAC	CAUG	CUAG	AGCU	ACGU	UGCA	UCGA
UACG	UACG	UCAG	GACU	GCAU	AUGC	AGUC	CUGA	CGUA	UAGC	UGAC	CAGU	CGAU	AUCG	ACUG	GUCA	GCUA	UCGA	UGCA	ACGU	AGCU	CUAG	CAUG	GUAC	GAUC
UCAG	UCAG	UACG	GCAU	GACU	CUGA	CGUA	AUGC	AGUC	UCGA	UGCA	ACGU	AGCU	CUAG	CAUG	GUAC	GAUC	UAGC	UGAC	CAGU	CGAU	AUCG	ACUG	GUCA	GCUA

[1]This multiplication table created by using the CDF-1 supplied in the supporting information Appendix C. Several isomorphic symmetric groups can be obtained by a different genetic-code cube as unit.

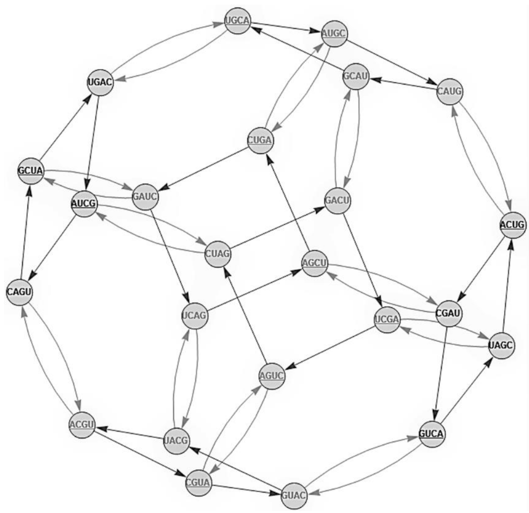


Figure 3. Cayley graph of the symmetric group $\left(GC^{\mathrm{ACGU}},\circ\right)$. This diagram was generated using the set of cubes $G=\{\mathrm{CAGU, CGUA}\}$ as generator. For any cubes $x\in\left(GC^{\mathrm{ACGU}},\circ\right)$ and $g\in G$ the vertices corresponding to the elements x and $x\circ g$ are joined by a directed edge for $g=\mathrm{CGUA}$ and bidirected for $g=\mathrm{CAGU}$.

Every element of the full tetrahedral group permutes the vertices of the regular tetrahedron among themselves. To date there is no biological criteria to favor any cube as unit element. Thus, in principle, we can define 24 groups $\left(GC^{X_1X_2X_3X_4},\circ\right)$ with unit element $X_1X_2X_3X_4$ running over the 24 cubes, and elements integrated by the 24 cube representations of the extended genetic code. These groups are isomorphic between them and isomorphic to group S_4. As result, there is only one symmetric group of the genetic code $(GC,\circ)$ up to isomorphism.

2.1 Sum and product operations between codons from different cubes of $(GC,\circ)$

A sum operation between codons from different cubes is induces by $\left(GC^{\text{ACGU}},\circ\right)$ and can be defined based on the isomorphism between the groups $(\mathbb{Z}_5,+)$ and $(B,+)$. For example, for cube ACGU the DNA base complementarity and the mentioned isomorphism ensure the bijection $\phi_{\text{ACGU}}: \text{D}\leftrightarrow 0, \text{A}\leftrightarrow 1, \text{C}\leftrightarrow 2, \text{G}\leftrightarrow 3, \text{U}\leftrightarrow 4$. Next, cube ACGU can be seen as a function running over the set of codons, i.e., $ACGU(x)$ with $x = x_1x_2x_3 \in B^3_{\text{ACGU}}$, where $B_{\text{ACGU}} = \{\text{A, C, G, U}\}$ $ACGU(x) = (\phi_{\text{ACGU}}(x_1), \phi_{\text{ACGU}}(x_2), \phi_{\text{ACGU}}(x_3))$, and $ACGU\,(x) \in (\mathbb{Z}_5)^3 \subset \mathbb{Z}^3_+ \subset \mathbb{R}^3$ ($\mathbb{Z}_5$ elements are included in the set of positive integers $\mathbb{Z}_+$). The inverse function is given by: $ACGU^{-1}(v) = \left(\phi^{-1}_{\text{ACGU}}(v_1), \phi^{-1}_{\text{ACGU}}(v_2), \phi^{-1}_{\text{ACGU}}(v_3)\right)$, where $v = (v_1, v_2, v_3) \in \mathbb{Z}^3_5 \subset \mathbb{Z}^3_+$, and $ACGU^{-1}(v) \in B^3$ Likewise, we can define the bijection $\phi_{\text{ACUG}}: \text{D}\leftrightarrow 0, \text{A}\leftrightarrow 1, \text{C}\leftrightarrow 2, \text{U}\leftrightarrow 3, \text{G}\leftrightarrow 4$ and $ACUG(x) = (\phi_{\text{ACUG}}(x_1), \phi_{\text{ACUG}}(x_2), \phi_{\text{ACUG}}(x_3))$, $ACUG(x) \in (\mathbb{Z}_5)^3 \subset \mathbb{Z}^3_+ \subset \mathbb{R}^3$ and $ACUG^{-1}(v) \in B^3$.

The composition of functions $X_1X_2X_3X_4(\ldots)$ is defined the same rule as the multiplication of permutation from $(S_B,\circ)$ or cubes from $(GC,\circ)$. Next, if cubes ACGU and ACUG are elements of group $\left(GC^{\text{ACGU}},\circ\right)$, then the sum "$\oplus$" operation between codons $x = x_1x_2x_3 \in \text{ACGU}$ and $y = y_1y_2y_3 \in \text{ACUG}$ can be defined as:

$$x \oplus y = [ACGU \circ ACUG]^{-1}(ACGU\,(x) + ACUG(y))$$

$$\text{Or } x \oplus y = AGUC^{-1}(ACGU\,(x) + ACUG(y))$$

Where $ACGU \circ ACUG = AGUC$ is the composition of functions $ACGU(\ldots)$ and $ACUG(\ldots)$ equivalent to the composition of cubes in $\left(GC^{\text{ACGU}},\circ\right)$, $ACUG^{-1}(\ldots)$ is the inverse of $ACUG(\ldots)$, and $ACGU\,(x) + ACUG(y)$ is the sum on $(\mathbb{Z}_5)^3$ (per coordinate as given in Table 2). In analogous way, we can define a product operation between codons $x = x_1x_2x_3 \in \text{ACGU}$ and $x = y_1y_2y_3 \in \text{ACUG}$ as: $x \otimes y = AGUC^{-1}(ACGU(x) \cdot ACUG(y))$, where symbol "$\cdot$" stands for the product operation in $(\mathbb{Z}_5^3, \cdot)$, which is the multiplicative group of $GF(5)^3$ [2].

2.2 The dihedral subgroups of $(GC, \circ)$

Since groups $(GC^{\text{ACGU}}, \circ)$ and S_4 are isomorphic, each subgroup of group S_4 has an equivalent subgroup in $(GC^{\text{ACGU}}, \circ)$. The subgroups of the symmetric group S_4 are well known. The subset of strong-weak cubes forms a subgroup $(SW^{\text{ACGU}}, \circ)$ of $(GC^{\text{ACGU}}, \circ)$, which is isomorphic to the well-known dihedral group D_4. The Cayley multiplications table on the set of cubes SW^{ACGU} is given in Table 4 (Appendix C, CDF-1 sections 5.1 and 6). Then, the partition of the 24 algebraic representations of the genetic code into the strong-weak, purine-pyrimidine and amino-keto classes is derived from the set of left cosets from the quotient group $(GC^{\text{ACGU}}, \circ)/(SW^{\text{ACGU}}, \circ)$ (or simply, GC/SW^{ACGU}), which is defined as: $GC/SW^{\text{ACGU}} = \{x \circ SW^{\text{ACGU}} \mid x \in GC^{\text{ACGU}}\}$. That is, the mentioned classes are the elements of the set of left cosets GC/SW^{ACGU} (Appendix C, CDF-1 section 6). For example, for any cube $x \in YR^{\text{ACGU}}$, we have $YR^{\text{ACGU}} = x \circ SW^{\text{ACGU}}$.

Group $(SW^{\text{ACGU}}, \circ)$ has two other conjugate subgroups that correspond to dihedral subgroups $(MK^{\text{AGUC}}, \circ)$ and $(YR^{\text{ACUG}}, \circ)$ integrated by *MK* and *YR* cubes, respectively (see below). In principle, we can define 24 dihedral subgroups able to split the set *GC* into the classes *SW*, *MK*, and *YR*. That is, the quotient groups $GC/MK^{\text{ACGU}} = \{x \circ MK^{\text{ACGU}} \mid x \in GC^{\text{ACGU}}\}$ and $GC/YR^{\text{ACGU}} = \{x \circ YR^{\text{ACGU}} \mid x \in GC^{\text{ACGU}}\}$ are well defined and split the set *GC* into the classes *SW*, *MK*, and *YR*.

Table 4. The Cayley multiplications table on set of cubes *SW*[1].

	ACGU	**AGCU**	**UCGA**	**UGCA**	**CAUG**	**CUAG**	**GAUC**	**GUAC**
ACGU	ACGU	AGCU	UCGA	UGCA	CAUG	CUAG	GAUC	GUAC
AGCU	AGCU	ACGU	UGCA	UCGA	GAUC	GUAC	CAUG	CUAG
UCGA	UCGA	UGCA	ACGU	AGCU	CUAG	CAUG	GUAC	GAUC
UGCA	UGCA	UCGA	AGCU	ACGU	GUAC	GAUC	CUAG	CAUG
CAUG	CAUG	CUAG	GAUC	GUAC	ACGU	AGCU	UCGA	UGCA
CUAG	CUAG	CAUG	GUAC	GAUC	UCGA	UGCA	ACGU	AGCU
GAUC	GAUC	GUAC	CAUG	CUAG	AGCU	ACGU	UGCA	UCGA
GUAC	GUAC	GAUC	CUAG	CAUG	UGCA	UCGA	AGCU	ACGU

[1]This multiplication table can be created by using the CDF-1 supplied in the supporting information Appendix C. Several isomorphic symmetric groups can be obtained on different cosets from the symmetric group of the genetic code cubes $(GC^{\text{ACGU}}, \circ)$.

The list of the main sets of dihedral groups for our interest are:

1) Strong-week dihedral group of cubes:

$SW^{\mathrm{ACGU}} = \{\mathrm{ACGU, AGCU, UCGA, UGCA, CAUG, CUAG, GAUC, GUAC}\}$

2) Purine-pyrimidine dihedral group of cubes:

$YR^{\mathrm{ACUG}} = \{\mathrm{ACUG, AUCG, GCUA, GUCA, CAGU, UAGC, UGAC, CGAU}\}$

3) Amino-keto dihedral group of cubes:

$MK^{\mathrm{AGUC}} = \{\mathrm{AGUC, AUGC, CUGA, CGUA, GCAU, GACU, UACG, UCAG}\}$

2.3 Klein four groups of $(GC^{\mathrm{ACGU}},\circ)$

Cubes ACGU, AGCU, UCGA and UGCA forms a Klein four subgroup of $(GC^{\mathrm{ACGU}},\circ)$, which will be denoted as $(SW_K^{\mathrm{ACGU}},\circ)$ (Table 5). The quotient group SW/SW_K^{ACGU} split the strong-weak set of cubes into two subsets (left cosets). While, the quotient group GC/SW_K^{ACGU} split the 24 algebraic representations of the genetic code into six classes (left cosets), each one with four cubes (Appendix C, CDF-1 section 7). As before, we can build 24 different Klein four subgroups of $(GC,\circ)$ by choosing a different cube as unit element at each time, which will originate the same partitions. For example, by taking cubes ACUG and AUCG as the units element, the Klein four groups $(YR_K^{\mathrm{ACUG}},\circ)$ and $(YR_K^{\mathrm{AUCG}},\circ)$ can be defined on the sets $YR_K^{\mathrm{ACUG}} = \{\mathrm{ACUG, AUCG, GUCA, GCUA}\}$ and $YR_K^{\mathrm{AUCG}} = \{\mathrm{AUCG, ACUG, GUCA, GCUA}\}$, respectively, which are isomorphic to $(SW_K^{\mathrm{ACGU}},\circ)$. In other words, Klein four groups can be defined on each one of the six left cosets from the quotient group GC/SW_K^{ACGU}.

Table 5. Klein four group $(SW_K^{\mathrm{ACGU}},\circ)$[1].

	ACGU	**AGCU**	**UCGA**	**UGCA**
ACGU	ACGU	AGCU	UCGA	UGCA
AGCU	AGCU	ACGU	UGCA	UCGA
UCGA	UCGA	UGCA	ACGU	AGCU
UGCA	UGCA	UCGA	AGCU	ACGU

[1]The rest of Klein four groups found in the symmetric groups $(GC,\circ)$ can be visualized by using the CDF-1 supplied in the supporting information Appendix C.

The six Klein four groups and their corresponding subjacent sets of cubes are listed below:

I) Strong-week Klein four groups of cubes:

1) $(SW_K^{\mathrm{ACUG}},\circ)$: subjacent set $SW_K^{\mathrm{ACGU}} = \{\mathrm{ACGU, AGCU, UCGA, UGCA}\}$

2) $\left(SW_{K}^{\text{CAUG}},\circ\right)$: subjacent set $SW_{K}^{\text{CAUG}}=\{\text{CAUG, CUAG, GAUC, GUAC}\}$

II) Purine-pyrimidine Klein four groups of cubes:

3) $\left(YR_{K}^{\text{ACUG}},\circ\right)$: subjacent set $YR_{K}^{\text{ACUG}}=\{\text{ACUG , AUCG , GUCA, GCUA}\}$

4) $\left(YR_{K}^{\text{CAGU}},\circ\right)$: subjacent set $YR_{K}^{\text{CAGU}}=\{\text{CAGU, UAGC, UGAC, CGAU}\}$

III) Amino-keto Klein four groups of cubes:

5) $\left(MK_{K}^{\text{AGUC}},\circ\right)$: subjacent set $MK_{K}^{\text{AGUC}}=\{\text{AGUC, AUGC, CGUA, CUGA}\}$

6) $\left(MK_{K}^{\text{CACU}},\circ\right)$: subjacent set $MK_{K}^{\text{GACU}}=\{\text{GACU, GCAU, UACG, UCAG}\}$

The left cosets of any quotient group of $\left(GC^{\text{ACGU}},\circ\right)$ and a Klein four-group of cubes (for example GC/MK_{K}^{AGUC}) will be integrated by the above subsets. In addition, 24 normal Klein four group can be defined on subsets of cubes. The quotient group of a normal Klein four group with the corresponding dihedral group splits the subjacent dihedral set into two cosets. For example, the quotient group SW/SW_{NK}^{ACGU} split set SW into two subsets, {ACGU, CAUG, GUAC, UGCA} and {AGCU, CUAG, GAUC, UCGA}. The quotient group CG/SW_{NK}^{ACGU} split set GC into six subsets (see Appendix C, CDF-1, section 7). Hence, only six normal Klein four group are defined on six different subsets of cubes:

IV) Strong-week normal Klein four groups of cubes:

3) $\left(SW_{NK}^{\text{ACUG}},\circ\right)$: subjacent set $SW_{NK}^{\text{ACGU}}=\{\text{ACGU, CAUG, GUAC, UGCA}\}$

4) $\left(SW_{NK}^{\text{AGCU}},\circ\right)$: subjacent set $SW_{NK}^{\text{AGCU}}=\{\text{AGCU, CUAG, GAUC, UCGA}\}$

V) Purine-pyrimidine normal Klein four groups of cubes:

7) $\left(YR_{NK}^{\text{ACUG}},\circ\right)$: subjacent set $YR_{NK}^{\text{ACUG}}=\{\text{ACUG, CAGU, GUCA, UGAC}\}$

8) $\left(YR_{NK}^{\text{AUCG}},\circ\right)$: subjacent set $YR_{NK}^{\text{AUCG}}=\{\text{AUCG, CGAU, GCUA, UAGC}\}$

VI) Amino-keto normal Klein four groups of cubes:

9) $\left(MK_{NK}^{\text{AGUC}},\circ\right)$: subjacent set $MK_{NK}^{\text{AGUC}}=\{\text{AGUC, CUGA, GACU, UCAG}\}$

10) $\left(MK_{NK}^{\text{AUGC}},\circ\right)$: subjacent set $MK_{NK}^{\text{AUGC}}=\{\text{AUGC, CGUA, GCAU, UACG}\}$

The above sets are also cosets from the quotient group GC/SW_{NK}^{ACGU}. That is, $GC/SW_{NK}^{\text{ACGU}}=\left\{x\circ SW_{NK}^{\text{ACGU}}\,\middle|\,x\in GC\right\}$.

2.4 Group of duals cubes

Two cubes with complementary base orders shall be called dual subsets of cubes, which is a classification originally given in reference [3] for cubes ACGU and UGCA. That is, the concept of dual cubes is taken borrow from the dual genetic code Boolean lattice of defined in [3]. Following the results presented in [3], twelve pairs of dual Boolean lattices can be defined on the set of 24 genetic-code cubes. For any Boolean lattice ($B_L(B)$, $\vee$, $\wedge$) there exists the "dual Boolean lattice" ($B_L'(B)$, $\wedge$, $\vee$), where the order relation is reversed, the symbols $\vee$ and $\wedge$ are interchanged and the maximum and minimum (1 and 0) are inverted (see [3]). It turns out that these twelve pairs of dual Boolean lattices are in one-to-one correspondence with twelve groups of duals cubes.

Indeed, the group of dual cubes corresponding to the dual Boolean lattices reported in reference [3] is built after setting CAUG cube as unit element of the operation "$\circ$": $\begin{pmatrix}\text{CAUG}\\\text{CAUG}\end{pmatrix}\circ\begin{pmatrix}\text{CAUG}\\\text{GUAC}\end{pmatrix}=\begin{pmatrix}\text{CAUG}\\\text{GUAC}\end{pmatrix}$ and $\begin{pmatrix}\text{CAUG}\\\text{GUAC}\end{pmatrix}\circ\begin{pmatrix}\text{CAUG}\\\text{GUAC}\end{pmatrix}=\begin{pmatrix}\text{CAUG}\\\text{CAUG}\end{pmatrix}$. The 64 codons ordered according to cube CAUG integrates the elements of the *primal lattice* defined in reference [3], while codons ordered according to cube GUAC integrate the elements of the *dual lattice*.

Likewise, group $\left(GC^{\text{ACGU}},\circ\right)$ can be split into subsets of dual cubes. This follows directly from the fact that a group structure can be defined on the set of dual cubes ACGU and UGCA: $\begin{pmatrix}\text{ACGU}\\\text{ACGU}\end{pmatrix}\circ\begin{pmatrix}\text{ACGU}\\\text{UGCA}\end{pmatrix}=\begin{pmatrix}\text{ACGU}\\\text{UGCA}\end{pmatrix}$ and $\begin{pmatrix}\text{ACGU}\\\text{UGCA}\end{pmatrix}\circ\begin{pmatrix}\text{ACGU}\\\text{UGCA}\end{pmatrix}=\begin{pmatrix}\text{ACGU}\\\text{ACGU}\end{pmatrix}$. This group will be denoted as $\left(SW_D^{\text{ACGU}},\circ\right)$. The quotient group GC/SW_D^{ACGU} split the 24 algebraic representations of the genetic-code cube into twelve classes (cosets), each one with two cubes. The elements of the quotient group GC/SW_D^{ACGU} are pairs of dual cubes as well. For example, the twelve groups of dual cubes are listed below:

I) Strong-week groups of dual cubes:

1) $SW_D^{\text{ACGU}}=\{\text{ACGU},\text{UGCA}\}$
2) $SW_D^{\text{AGCU}}=\{\text{AGCU},\text{UCGA}\}$
3) $SW_D^{\text{CAUG}}=\{\text{CAUG},\text{GUAC}\}$
4) $SW_D^{\text{CUAG}}=\{\text{CUAG},\text{GAUC}\}$

II) Purine-pyrimidine groups of dual cubes:

5) $YR_D^{\mathrm{ACUG}} = \{\mathrm{ACUG}, \mathrm{GUCA}\}$

6) $YR_D^{\mathrm{AUCG}} = \{\mathrm{AUCG}, \mathrm{GCUA}\}$

7) $YR_D^{\mathrm{CAGU}} = \{\mathrm{CAGU}, \mathrm{UGAC}\}$

8) $YR_D^{\mathrm{UAGC}} = \{\mathrm{UAGC}, \mathrm{CGAU}\}$

III) Amino-keto groups of dual cubes:

9) $MK_D^{\mathrm{AGUC}} = \{\mathrm{AGUC}, \mathrm{CGUA}\}$

10) $MK_D^{\mathrm{AUGC}} = \{\mathrm{AUGC}, \mathrm{CUGA}\}$

11) $MK_D^{\mathrm{GCAU}} = \{\mathrm{GCAU}, \mathrm{UACG}\}$

12) $MK_D^{\mathrm{GACU}} = \{\mathrm{GACU}, \mathrm{UCAG}\}$

This result leads to a generalization of the results reported in reference [3] and to the clear definition of the symmetric group of genetic-code Boolean lattices, or in terms of the results reported in reference [9], the symmetric group of genetic-code Boolean algebras. The symmetric group $(S_B, \circ)$ induces a group structure over the set of the 24 Boolean lattices, which can be defined following the procedure presented in reference [3]. However, a further development of the symmetric group of genetic-code Boolean lattices goes beyond the limits and purposes of the current manuscript.

2.5 Alternating Group

The subgroups of $(GC, \circ)$ mentioned so far were defined in subsets of cubes that belong to the same class. Alternating group A_4 is the group of even permutations of S_4 and, in accordance with theory exposed above, the alternating group $(A^{\mathrm{ACGU}}, \circ)$ of $(GC^{\mathrm{ACGU}}, \circ)$ is well defined. Cayley graph for alternating group $(A^{\mathrm{ACGU}}, \circ)$ is shown in Fig. 4.

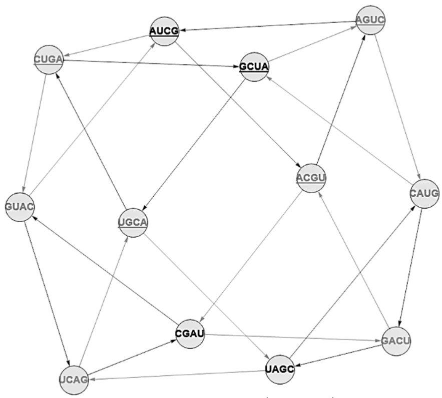


Figure 4. Cayley graph of the symmetric group $x \in \left(A^{\mathrm{ACGU}}, \circ\right)$. This diagram was generated by using the generator set of cubes $G = \{\mathrm{AGUC}, \mathrm{CGAU}\}$. For any cubes $x \in \left(A^{\mathrm{ACGU}}, \circ\right)$ and $g \in G$ the vertices corresponding to the elements x and $x \circ g$ are joined by a directed edge for $g = \mathrm{AGUC}$ and bidirected for $g = \mathrm{CGAU}$

2.6 The norm of codon is preserved in the set of left cosets of SW/SW_K

Since the genetic code cubes are inserted in $\mathbb{R}^3$, giving specific bijections, the norm of codons can be defined for the cubes inserted in $\mathbb{R}^3$. For example, cube ACGU is inserted (with center in codon CCC) in $\mathbb{R}^3$ by the function $ACGU(x) \in (\mathbb{Z}_5)^3 \subset \mathbb{Z}_+^3 \subset \mathbb{R}^3$. Next, the inner product of two codons $x \in B^3$ and $y \in B^3$ can be defined in $\mathbb{R}^3$ as:

$$\langle ACGU(x), ACGU(y)\rangle = x_1 y_1 + x_2 y_2 + x_3 y_3 \tag{1}$$

Then, the norm $\|x\|_{\mathrm{ACGU}}$ of a codon $x \in B^3$ with coordinates $(x_1, x_2, x_3) \in \mathbb{R}^3$ is given by:

$$\|x\|_{\mathrm{ACGU}} = \sqrt{\langle ACGU(x),\ ACGU(x)\rangle} = \sqrt{x_1^2 + x_2^2 + x_3^2} \tag{2}$$

To analyze cube symmetries in $\mathbb{R}^3$ the cubes must be centered on the origin of coordinates. The insertion of *GC* cubes with center in the origin of coordinates is performed by means of the bijection between the sets $\{0,1,2,3,4\}$ and $\{0,-2,-1,1,2\}$, given by $\gamma_{1234}: 0 \leftrightarrow 0,\ 1 \leftrightarrow -2,\ 2 \leftrightarrow -1,\ 3 \leftrightarrow 1, 4 \leftrightarrow 2$. In consequence, the composition of bijections ϕ_{ACGU} and γ_{1234} yields the bijection that maps the set of bases into the set $\{0,-2,-1,1,2\}$, i.e., $\gamma_{1234}(\phi_{\mathrm{ACGU}}) = \gamma_{\mathrm{ACGU}}: \mathrm{D} \leftrightarrow 0, \mathrm{A} \leftrightarrow -2, \mathrm{C} \leftrightarrow -1, \mathrm{G} \leftrightarrow 1, \mathrm{U} \leftrightarrow 2$. Next, we can define function $acgu(x) = (\gamma_{\mathrm{ACGU}}(x_1), \gamma_{\mathrm{ACGU}}(x_2), \gamma_{\mathrm{ACGU}}(x_3))$, where $acgu(x) \in \mathbb{Z}^3 \subset \mathbb{R}^3$ Analogous bijections can be derived for every genetic-code cube. The inner product of two codons $X \in B^3$ and $Y \in B^3$ can be defined in R^3 by means of the bijection ϕ_{ACGU} as:

$$\langle acgu(x), acgu(y) \rangle = x_1 y_1 + x_2 y_2 + x_3 y_3 \tag{3}$$

Then, for cube ACGU, the norm $\|x\|_{\mathrm{ACGU}}$ of a codon $x \in B^3$ with coordinates $(x_1, x_2, x_3) \in \mathsf{R}^3$ is given by:

$$\|x\|_{\mathrm{ACGU}} = \sqrt{\langle acgu(x),\, acgu(x) \rangle} = \sqrt{x_1^2 + x_2^2 + x_3^2} \tag{4}$$

After last definition, we can propose the following:

Theorem 1. Let $X \in B^3$ be a codon of the set of left cosets $GC/K^{X_1^0 X_2^0 X_3^0 X_4^0}$ with cube $X_1^0 X_2^0 X_3^0 X_4^0$ as unit element of the (non-normal) Klein four group $K^{X_1^0 X_2^0 X_3^0 X_4^0}$. For any cube $X_1 X_2 X_3 X_3$ from a coset of the quotient group $GC/K^{X_1^0 X_2^0 X_3^0 X_4^0}$ and $X \in B^3$, the norm $\|x\|_{X_1 X_2 X_3 X_3}$ given by Eq. 4 is preserved.

Proof. The coordinates (x_1, x_2, x_3) of any codon $X \in B^3$ in R^3 will change depending on which cube representation $X_1 X_2 X_3 X_3 \in GC/K^{X_1^0 X_2^0 X_3^0 X_4^0}$ is used. However, the possible changes of codon coordinates between the cube representations from the same coset from the quotient group $GC/K^{X_1^0 X_2^0 X_3^0 X_4^0}$ only involve coordinate changes of one or more bases by its (their) algebraic inverse(s). For example, the coordinates of codon ACG in the cubes ACGU, AGCU, UCGA, and UGCA from coset SW_K^{ACGU} are $(-2,-1,1)$, $(-2,1,-1)$, $(2,-1,1)$ and $(2,1,-1)$, respectively. It is not difficult to see that all these codon coordinates yield the same norm value, as given by Eq. 4 (Appendix C, CDF-1, section 4.2). Therefore, the norm $\|x\|_{X_1 X_2 X_3 X_3}$ (Eq. 4) is preserved in any cube $X_1 X_2 X_3 X_3$ representations from coset $GC/K^{X_1^0 X_2^0 X_3^0 X_4^0}$ □

Readers can explore Theorem 1 in the CDF-1 from Appendix C (section 4.2) and verify that the norm given by Eq. 2 is not preserved in the set of left cosets $GC/SW_K^{X_1^0 X_2^0 X_3^0 X_4^0}$. As was pointed out before, in practice, we can define 24 different but isomorphic groups $(GC, \circ)$, where a different cube is taken as the unit element in each one of these groups and a corresponding Klein four subgroup can be defined. For example, for the Klein four subgroups $(YR_K^{\mathrm{ACUG}}, \circ)$ and

$\left(YR_K^{\mathrm{AUCG}},\circ\right)$ (Appendix C, CDF-1, sections 2.2 and 7), according to Theorem 1, for any codon $X \in B^3$, $\|x\|_{\mathrm{ACUG}} = \|x\|_{\mathrm{AUCG}} \neq \|x\|_{\mathrm{ACGU}}$. In general, the Klein four subgroups $\left(YR_K^{ACUG},\circ\right)$, $\left(MK_K^{AGUC},\circ\right)$, and $\left(SW_K^{ACGU},\circ\right)$ determine the subsets of cubes from GC where the norm of codons is preserved.

In addition to the codon norm definition given by Eqs. 2 and 4, a weighted codon norm can be defined on a cube $X_1X_2X_3X_4$ as:

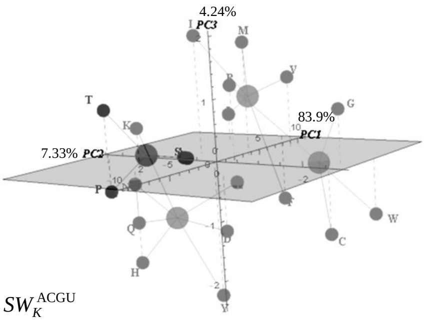
$$\|x\|_{X_1X_2X_3X_4} = \sqrt{w_1x_1^2 + w_2x_2^2 + w_3x_3^2} \tag{5}$$

Where $0 \le w_i \le 1$. Each set of weights w_i will produce a different norm.

3 Principal component analysis of the genetic–code cube scales

Each amino acid can be represented by a statistic of its synonymous codon norms. We can consider the minimum, the maximum, the median and the mean of codon norm. Hence, each single amino acid can be represented as single number or a vector with four coordinates, corresponding to the mentioned statistics of its synonymous codon norms calculated for a given cube. If the 24 cubes are used simultaneously, then each amino acid can be represented as a vector with 24x4 coordinates. In this way, several coordinates can be correlated and the amino representation can be carrying redundant information. This is precisely the scenario to apply principal component analysis (PCA). The application PCA will permit us to reduce dimension and to represent the set of amino acids by new orthogonal (uncorrelated) variables, the principal components (PCs) [10]. Results indicate that for all the cubes and codon subsets mentioned above, the three first PCs carry more than the 80% of sample variance (Fig. 5 and CDF-2, Appendix C).

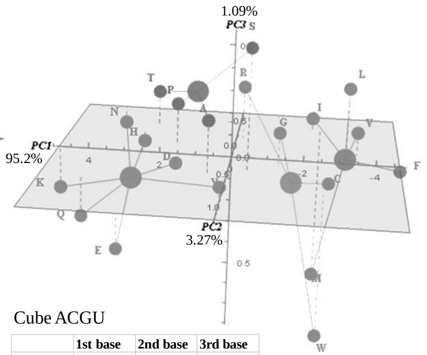
Consequently, for any subset of genetic code cubes, each amino acid can be represented by the sum of its three PCs coordinate values. In other words, an amino acid scale can be derived from any subset of genetic code cubes by applying the above-mentioned procedure. Then, it will be natural to verify whether a genetic-code cube-scale is correlated with some reported amino acid physicochemical property. Studies on protein folding by the end of the 20th century resulted in the development of numerous physicochemical and biochemical indexes to empirically describe the interaction of amino acids in protein 3D structures [11–13]. It turned out that codon norm and the weighted codon norm defined in Eqs. 2, 4, and 5 can be used to defined amino-acid scales correlated with the physicochemical indexes reported by many authors, which are currently available in the AAindex database [12].



SW_K^{ACGU}

	1st base	2nd base	3rd base
Weights 1	1	1	0.8
Weights 2	0.56	0.98	0.48

Physico-Chemical AAindex	Correlation
Linker propensity from small dataset (linker length is less than six residues) (George-Heringa, 2003)	0.880263
Hydropathy index (Kyte-Doolittle, 1982)	0.85735
Modified Kyte-Doolittle hydrophobicity scale (Juretic et al., 1998)	0.849743
Average side chain orientation angle (Meirovitch et al., 1980)	0.848508
Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)	0.844863
Average reduced distance for side chain (Meirovitch et al., 1980)	-0.832254
Polarity (Grantham, 1974)	-0.830742
Side chain hydropathy, corrected for solvation (Roseman, 1988)	-0.830353
Polar requirement (Woese, 1973)	-0.809848
Hydropathy scale based on self-information values in the two-state n	0.807113



Cube ACGU

	1st base	2nd base	3rd base
Weights 1	0.9	1	0.29
Weights 2	0.5	1	0.81

Physico-Chemical AAindex	Correlation
Average side chain orientation angle (Meirovitch et al., 1980)	-0.910431
Effective partition energy (Miyazawa-Jernigan, 1985)	-0.896586
Inf. value for accessibility; average fraction 23% (Biou et al., 1988)	-0.885604
Average reduced distance for side chain (Meirovitch et al., 1980)	0.883107
Optimized relative partition energies - D (Miyazawa-Jernigan, 1999)	0.881177
Inf. value for accessibility; average fraction 35% (Biou et al., 1988)	-0.87335
Mean fractional area loss (Rose et al., 1985)	-0.872642
Av. gain ratio in surrounding hydrophobicity (Ponnuswamy, 1980)	-0.871645
Long range non-bonded energy per atom (Oobatake-Ooi, 1977)	0.870597
Optimized relative partition energies - A (Miyazawa-Jernigan, 1999)	0.869324

Figure 5. Principal component analysis of amino acid scales derived from GC cubes from the Klein four subgroup $\left(SW_K^{\mathrm{ACGU}},\circ\right)$ and from cube ACGU. In the case of cubes from the set SW_K^{ACGU}, each amino acid is represented as vector of 4 (cubes)x4(statistics)x2(weights vectors) = 32 coordinates; while eight coordinates were used in the case of cube ACGU. In both analysis the first three principal components carry most of the 80% of the sample variance.

This and additional analyses can be repeated/accomplished by using the CDF-2 available in Appendix C. An example of this analysis is presented in Fig. 5.

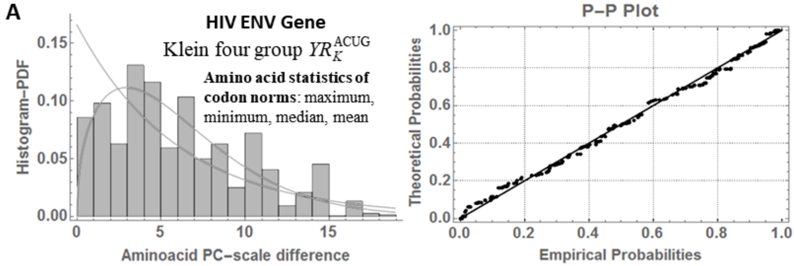
4 Evolutionary encoded mutational cost (EMC)

Since we have reasons to believe that the genetic code architecture is optimized to minimize the transcription and translation errors [14–16], we would expect that at least one of the numerous possible amino-acid PC-scales would model the molecular evolutionary cost of new mutational variants fixed in the organismal population. Let x_0 and x_t be the amino-acid PC-scale values for a given position in a gene at the evolutionary times 0 and *t*, respectively. Then, the encoded cost of the mutational event that involves the change from x_0 to x_t can be expressed by the difference $\Delta x = \left|x_t - x_0\right|$ (6). A Weibull model for the cost Δx was deduced on thermodynamic/biophysical basis with cumulative probability distribution $F\left(\Delta x \middle| \alpha,\beta\right) = 1 - e^{-\left(\frac{\Delta x}{\beta(l)}\right)^{\alpha}}$ $\Delta x > 0$ (7) (Appendix D).

For a set of aligned protein sequences, a set of weights w_i to estimate the cost Δx based on Eq. 5 can be approached by the application of an optimization algorithm. The application of the above ideas to concrete datasets of mutational variants reported in three proteins is presented in Fig. 6 (data analyses available in CDF-2, Appendix C). For each protein, two codon norms were derived from Eq. 5 by using two sets of weights. For the sake of simplification (to reduce computational time), the cost Δx was estimated for each amino acid position in respect to a reference protein sequence (the first protein found in the sequence alignment). A genetic algorithm from the R package GA [17] was used to approach the weights that maximize the goodness of fit (gof) of Eq. 7 (minimization of Kolmogorov–Smirnov statistic, Fig. 6). We shall call the cost Δx estimated according to this approach as *evolutionary encoded mutational cost* or simply *evolutionary mutational cost* (EMC).

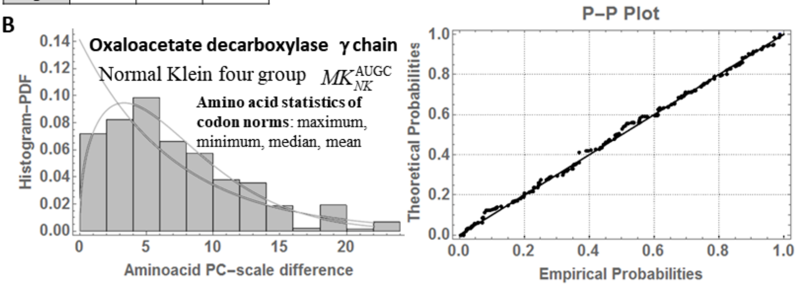
A wider analysis was performed on 105 alignment of different protein sequences from distinct species. The kernel density plots for the estimated α and β parameters from Eq. 7 are given in Fig. 7. It is worthy to observe the small variation between the estimated values of parameter α. Since these estimations were made in datasets of unrelated protein sequences (except for HIV *Env* and *Gag*) with completely different evolutionary history, different amino acid PC-scales estimated in different subsets of genetic code-cubes, we should expect larger variations between these estimations. To verify whether this behavior is a general regularity of the molecular evolutionary process goes beyond the limits of current study.

An evolutionary implication on the conservation of parameter α value derives from Eq. A6 (Appendix D): $Nq = (\Delta x/\beta(l))^{\alpha-1}$. After applying the logarithm in both side of this equation we have: $\dfrac{log(Nq_i)}{log(\Delta x_i/\beta(l))} = \alpha - 1$ (8), where Nq_i is the expected number of times that an evolutionary cost Δx_i can be observed in N mutational events, while $\Delta x_i/\beta(l)$ is the normalized cost (non-dimensional cost) estimated for a given set of aligned protein sequences. In other words, the ratio of the logarithm of the expected number of times that an evolutionary cost Δx_i can be observed in N mutational events to the logarithm of the normalized cost $\Delta x_i/\beta(l)$ is constant and independent of the protein sequence. Notice that both parameters, $\beta(l)$ and the cost Δx_i, depend on the set of homologous protein sequences under scrutiny. Moreover, given the parameter α, $\beta(l)$ depends on the complete set of Δx_i values.



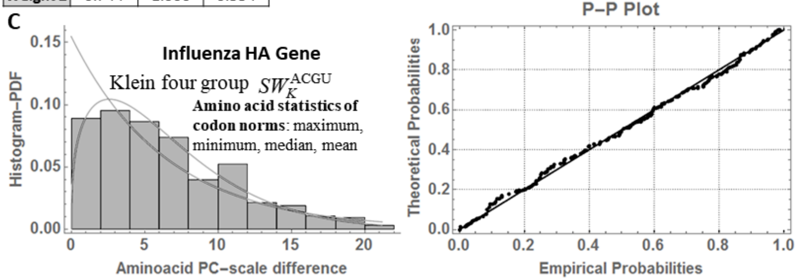
	1rst base	2nd base	3th base
Weight 1	0.630	1.000	0.842
Weight 2	0.581	1.000	0.906

Distribution	Test statistic	Bootstrap samplings	MC sample size	Test-pvalue
Weibull	0.0399	1000	200	0.898



	1rst base	2nd base	3th base
Weight 1	0.875	1.000	0.244
Weight 2	0.744	1.000	0.334

Distribution	Test statistic	Bootstrap samplings	MC sample size	Test-pvalue
Weibull	0.0336	1000	200	0.97



	1rst base	2nd base	3th base
Weight 1	0.871	1.000	0.278
Weight 2	0.476	1.000	0.464

Distribution	Test statistic	Bootstrap samplings	MC sample size	Test-pvalue
Weibull	0.027591	1000	200	0.995

Figure 6. Fitting of the Weibull distribution model for the fixation probability of amino acid mutational variants based on the evolutionary mutational cost (EMC) Δx. Panels **A** to **C**, provide following: 1) weights used to compute codon norms according to Eq. 5, 2) subset of genetic-code cubes where the estimation was performed, 3) histogram, exponential decay and Weibull distribution for the corresponding EMC Δx, 4) probability plots, and 5) results of Monte Carlo (MC) Kolmogorov–Smirnov (KS) test. There is not enough reason to reject the null hypothesis: Weibull distribution (*p*-value >> 0.05). These analyses (and others, e.g., different options for MC-KS or MC-Kuiper goodness-of-fit) can be verified in CDF-2 given in Appendix C.

Assuming $q_i = \frac{\Delta x_i^{\alpha-1}}{\sum_i^N \Delta x_i^{\alpha-1}}$ and after replacing q_i in Eq. A6, we have $\beta(l)^{\alpha-1} = \frac{1}{N}\sum_i^N \Delta x_i^{\alpha-1}$ (9), which is the maximum likelihood estimator of the parameter $\beta(l)$ from the Weibull probability distribution given in Eq. A12. Notice that $log_2(Nq_i) = -log_2(1/N) - (-log_2 q_i)$, i.e., $log_2(Nq_i)$ is the difference between two entropies corresponding to a mutational event with probability distribution $1/N$: $S_U = -log_2(1/N)$ and that one with probability $q_i : S_Q = -log_2 q_i$. In other words, $log_2(Nq_i) = S_U - S_Q$ expresses the uncertainty reduction or the information gain for a mutational event with cost Δx_i and success probability q_i in respect to the event with uniform chance over N trials. Hence, according with Eq. 8 the observation of $\alpha > 1$ in a gene population indicates a gain of information. Results presented in Fig. 7 suggests that the fixation of new mutational variants in natural gene populations is governed by a stochastic process that leads to gain of information.

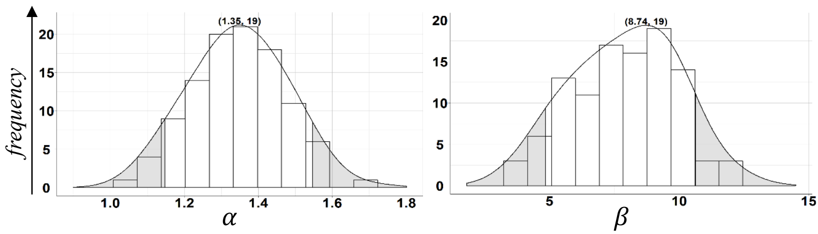


Figure 7. Kernel density plots for the α and β parameters from Eq. 7 estimated on 105 alignments of different sets of homologous proteins from distinct species. For each protein sequence alignment two codon norms were derived from Eq. 5 by using two sets of weights. A genetic algorithm from the R package GA [17] was used to approach the weights that maximize the gof of Eq. 7 (minimization of Kolmogorov–Smirnov statistic). The areas under the curve in blue cover the regions between the 5% and 95% percentiles, i.e., 90% of the estimated α parameters have values between 1.14 and 1.54, while 90% of the β parameters have values between 4.83 and 10.71.

4.1 Application to immunoescape variants prediction

The analytical procedure described in the last section has a straightforward application to predict immunoescape mutational variants originated in populations of pathogenic microorganisms and viruses and to improve de novo vaccine design. As suggested in Fig. 1, the immune epitopes of interest are found in the subset of mutational variants with high probability of fixation, provided that the Weibull model is built on a set of protein multiple sequence alignment of mutational variants fixed in a given organismal population.

A more complex analysis involves the examination of functional or structural dependences between proteins. We should check if the immunoescape epitopes from the protein under scrutiny are independents or not with respect to some mutational variants of another essential protein required for the adaptation and propagation of the pathogen in the host (recall that lack of correlation does not necessarily implies independence). Herein, an example with HIV proteins *Env* and *Gag* is given. Currently is not possible to track the pairwise association of simultaneous mutations *in situ* of *Env* and *Gag* proteins in patients. However, it is possible to track the sum of EMC values from protein sequences isolated from the same patient (i.e., to match the sum of EMC pairwise values from *Env* and *Gag* proteins isolated from the same patient). Results indicate that the sum of EMC values in both proteins, *Env* and *Gag*, has bimodal probability density (Fig. 8A to D). In addition, the sum of EMCs from *Env* is statistically significant correlated with the sum of EMCs from *Gag* with a Kendal’s tau value of 0.52. This correlation is emphasized by the joint probability density of these variables, which implies that the total evolutionary mutational cost estimated for these proteins are not independent (Fig. 8E to F). This result is consistent with a published report that HIV-1 evolution in *Gag* and *Env* are highly correlated [18].

In addition, the joint probability density of these variables indicates the grouping of the 1051 HIV mutational variants under analysis into two classes: i) those with simultaneous high values of total EMC in *Env* and *Gag*, and ii) those with simultaneous low EMC cost. The unsupervised classification into these two classes is easily detected by applying K-means algorithm implemented in R [19], which was used to derive the mixture of probability densities of McKay's bivariate gamma distribution model presented in Fig. 8F. It turned out that 661 (63%) of the 1051 HIV mutational variants under analysis are classified in the group with simultaneous highest values of total EMC.

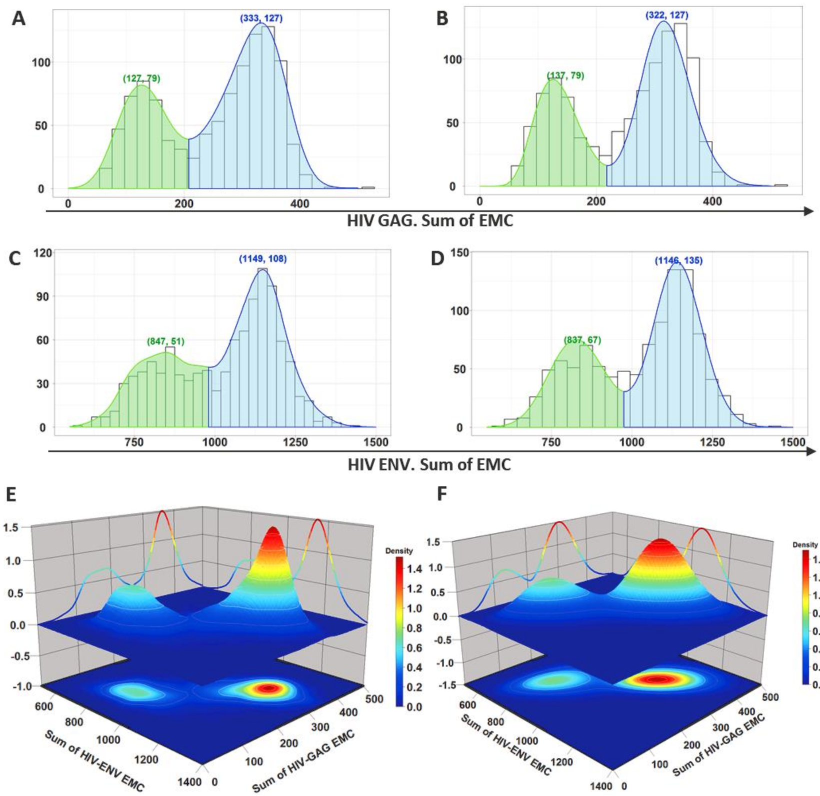


Figure 8. Density plots for sum of evolutionary mutational cost (EMC) estimated for HIV *Env* and *Gag* proteins. **A** and **C**, based on kernel density estimations. **B** and **D**, based on mixture of gamma distributions estimated in R [19]. **E**, based on 2D kernel density estimation performed with the R package *KernSmooth* [20]. **F**, based on mixtures of McKay's bivariate gamma distributions estimated by maximum likelihood estimation using the R package *VGAM* [21,22]. The 3D graphics were built with the R package *plot3D* [23]. The analyses are based on multiple aligned sequences of *Env* and *Gag* proteins taken from 1052 HIV mutational variants. That is, in panels E and F, each experimental pair of coordinate in the plane *xy* corresponds to the estimations of the sums of EMCs for the *Env* and *Gag* proteins found in a HIV mutational variant isolated from one patient.

5 Discussion

The symmetric group of the genetic-code cubes $(GC,\circ)$ integrates the studies of genetic code architectures based on a single genetic code cube. Each subgroup of $(GC,\circ)$ and left cosets of its dihedral and Klein groups are associated to fundamental physicochemical properties of the DNA bases. That is, the *ad hoc* and intuitive classification based on IUPAC codes for nucleotides [7] that was introduced in reference [5] is now mathematically derived. Results indicate that the multiple facets linking the genetic code architectures to the molecular evolutionary process [2,8] are not merely a set of simple observations derived from human curiosity, but an objective set of quantitative relationships physicochemically determined, which can be mathematically described and integrated in the symmetric group of the genetic-code cubes $(GC,\circ)$.

Results also indicate that for group $(GC,\circ)$, and most of its subgroups and cosets, the PCA performed as described in section 3 leads to a strong classification of the amino acids into four groups. Furthermore, there is a one-to-one correspondence between amino acid classification based on PCA and the four vertical planes of the genetic-code cubes (Fig. 3 and CDF-1, Appendix C). This result implies that the current genetic-code architecture is not the result of a random assignment of codons to amino acids, which is consistent with current beliefs that the genetic code architecture is optimized to minimize the transcriptional and translational errors [14–16].

Amino-acid PC-scales derived from the PCA of codon norms on cubes from different subgroup or cosets from $(GC,\circ)$ are correlated with physicochemical indexes reported by studies on protein folding and protein interactions. Amino-acid PC-scales derived for the dihedral group *SW* and its Klein four group $\left(SW_K^{\mathrm{ACGU}},\circ\right)$ are strongly correlated with the amino acid hydrophobic scales (Fig. 3). Amino acid hydrophobicity has been considered one of the most important physicochemical characteristics of amino acids and the major driving force of protein folding [13,24]. This result suggests the existence of a link between the genetic code architecture and a major driving force in protein folding, the hydrophobic effects. The results presented in Fig. 3 (and in the CDF-2, Appendix C) indicate that the information carried by these physicochemical properties are already encoded in the genetic-code architecture and quantitatively unveiled in the symmetric group of the genetic-code cubes. These correlations, however, are only unveiled in cubes inserted in $\mathbb{R}^3$ throughout bijections $\phi_{\mathrm{ACGU}}(X)=(x_1,x_2,x_3)$.

Results support the hypothesis that amino-acid PC-scales are linked to the evolutionary cost of the new mutational variants fixed in the organismal populations. For each gene set or subset, it is possible to fit a Weibull distribution model that predicts the amino acid mutation probability based on the variation the EMC Δx. The molecular evolutionary process is an optimization process that ultimately leads to species adaptation and survival. In mathematical terms, this optimization process can be quantitatively expressed by the set of weights w_i from Eq. 5 that maximize the gof of the empirical cumulative distribution values of the cost Δx. In consequence, under the hypothesis expressed by Eq. 7, amino-acid PC-scales derived through an optimization algorithm carry information on the molecular evolutionary process, which is specific for each gene population. Therefore, theses scales simultaneously carry the physicochemical information inherited from the encoded genetic code-cube architectures and the evolutionary one derived from the phylogenetic development of the gene population under study. So, the physicochemical information encoded in the genetic code-cube architectures just provides a general deterministic framework modulated/enriched by the action of the evolutionary pressure. On this scenario, the stochastic nature of the mutational process leads to consider the evolutionary process as a stochastic deterministic process [25,26].

As suggested in Fig 6 (and the CDF-2, Appendix C), the mutational process at different sets of homologous genes will be described by different amino-acid PC-scales and, consequently, will fit different Weibull distribution models given by Eqs. 7 (A12). This observation would suggest that the evolutionary pressure on each gene induces specific adaptive evolutionary paths, conserving molecular biophysical and biochemical features specific for the biological function of each encoded protein. The adaptive evolutionary paths lead to a gain of information in the population of 102 sets of homologous proteins analyzed (Fig. 7), which is quantitatively expressed by the parameter α of Weibull distribution with values $\alpha > 1$ and small variation around 1.35. In consequence, the ratio of the information gained to the normalized evolutionary cost $\Delta x_i / \beta(l)$ tend to be nearly constants (with small variation) and independent of the protein sequence (within the limits of the experimental and numerical errors, see Eq. 8, section 4). It is worthy to notice that the results presented in Fig. 7 can be improved, since these estimations depend on optimization of the sets of weights and on the selection of the best subset of genetic-code cubes that better describe the evolutionary process in each set of homologous protein under study. Such a study in hundreds or thousands alignments sets of homologous proteins is not impossible but computationally expensive.

The estimated Weibull model will expose the mutational events with high and low probabilities of fixation in the given gene population. This is a valuable information to predict immunoescape epitope variants originated in populations of pathogenic microorganisms and viruses and to improve *de novo* vaccine design. For example, attenuated vaccines against pathogenic microorganisms can be designed based on the immunogenicity of exposed immunoescape epitopes to the host. To facilitate this design, we could simply estimate the probability of fixation of such exposed immune-epitopes.

An immunoescape epitope carried by an external protein could have a relativity high probability of fixation, but if the mutational events are not independent from mutational variants found, for example, in an inner essential pathogen protein, then the overall fixation success will be determined by the joint probabilities for the mutational variants found in the immune epitope and in the essential pathogen proteins. This is the case for HIV *Gag* and *Env* proteins presented in Fig 8, which are consistent with published report [18]. Hence, if the mutational variants in the proteins under scrutiny are not independent, then a low joint probability of fixation will indicate that the vaccine candidate might not be needed, since a natural strain carrying the immunoescape epitope has low probability of adaptation in the host. Such a joint probability can be estimated by applying the state of the art in density estimation (as illustrated in Fig. 8) and copula distribution [27]. The *in-silico* prediction of immunoescape mutational variants as suggested here is feasible, can save time, and it would considerably reduce the cost of vaccine clinical trials.

6 Concluding remarks

The derivation of the algebraic structure of the symmetric group of the genetic-code cubes $(GC,\circ)$ is given in the manuscript. A deep complexity of the quantitative relationships between codons and their encoded amino acids is unveiled by group $(GC,\circ)$. These quantitative relationships expressed by group $(GC,\circ)$, its subgroups and cosets were quantitatively manifested in the amino-acid PC-scales derived from codon norms. These scales are strongly correlated with the physicochemical indexes reported by studies on protein folding and protein interactions.

The effect of the genetic code architecture on the evolutionary process was exposed by a Weibull distribution model inferred for the mutational process. For a set of homologous protein different amino PC-scales can be estimated in different subsets of genetic code-cubes through the application of an optimization algorithm. The size of the set of all possible amino-acid PC-scales is large enough to reflect the huge diversity of evolutionary strategies found in natural encoded proteins. A small variation of the estimated values of α parameter from Weibull distribution would suggest that, in the gene populations under scrutiny, the ratio of the information gained to the normalized evolutionary cost $\Delta x_i/\beta(l)$ tend to be nearly constants (with small variation) and independent of the protein sequence.

The result presented here would be particularly relevant to predict immunoescape epitope variants originated in populations of pathogenic microorganisms and viruses. This knowledge would improve the lifespan of *de novo* vaccines as well as the neutralization of potential *superbugs*. Current results indicate that, on thermodynamic basis, a stochastic deterministic mutational process [25,26] is constrained by the genetic code architecture.

Appendix A. The extended genetic–code cube

Analysis of the primordial chemistry led to the development of an algebraic structure for a plausible ancestral genetic code [2]. This code is founded on the plausible existence of one or more nucleotide bases in the primeval DNA protein-coding regions with nonspecific (non-Watson-Crick) base-pairings. The existence of these ancestral nucleotide bases is likely the simplest explanation to overcome the difficulties for the origin of life discussed in references [2,28]. Prebiotic chemistry studies suggest that the current DNA bases could have populated the early terrestrial environment together with other nucleotide bases [29–35]. Thus, it is feasible that the standard genetic code could have been derived from an ancestral code architecture with five or more bases [2]. A larger DNA alphabet with geologically stable bases would ensure thermal stability of the DNA molecule in the inhospitable prebiotic landscape [2,28]. Consistently with last hypothesis, the current DNA methylation could be considered a relic footprint left by ancient DNA molecules. It was recently shown that most methylation changes occurring within cells are likely induced by thermal fluctuations to ensure thermal stability of the DNA molecules, seemingly explainable by statistical mechanics laws [36]. Perhaps the more significant role of the fifth base in the current DNA molecules is played by the epigenetics role of cytosine DNA methylation (CDM). CDM patterning represents one feature of the epigenome that is highly responsive to environmental stress and associates with transgenerational adaptation in plants and in animals [36].

The natural extension of the DNA alphabet permitted the definition of a genetic code algebraic structure over an extended triplet set (see Table 2). In particular, a Galois field ($GF(5)$) was defined over the set of an extended RNA alphabet $B = \{D, A, C, G, U\}$, where the letter D symbolizes one (or more) alternative hypothetical base(s) or a dummy variable with non-specific pairings in primeval RNA and DNA molecules (Table A1) [2]. Based on the Watson-Crick DNA base-pairing and the codon order according to the evolutionary importance of their bases, it was shown that the extended genetic code is mathematically equivalent to a cube inserted in $\mathbb{R}^3$ (see Figure 1) [2].

Table A1. Extended base-triplet set for the genetic code cube ACGU

	No	D	No	A	*aa*[1]	No	C	aa	No	G	aa	No	U	aa	
	0	DDD	25	DAD		50	DCD		75	DGD		100	DUD		D
	1	DDA	26	DAA		51	DCA		76	DGA		101	DUA		A
D	2	DDC	27	DAC		52	DCC		77	DGC		102	DUC		C
	3	DDG	28	DAG		53	DCG		78	DGG		103	DUG		G
	4	DDU	29	DAU		54	DCU		79	DGU		104	DUU		U
	5	ADD	30	AAD		55	ACD		80	AGD		105	AUD		D
	6	ADA	31	AAA	K	56	ACA	T	81	AGA	R	106	AUA	I	A
A	7	ADC	32	AAC	N	57	ACC	T	82	AGC	S	107	AUC	I	C
	8	ADG	33	AAG	K	58	ACG	T	83	AGG	R	108	AUG	M	G
	9	ADU	34	AAU	N	59	ACU	T	84	AGU	S	109	AUU	I	U
	10	CDD	35	CAD		60	CCD		85	CGD		110	CUD		D
	11	CDA	36	CAA	Q	61	CCA	P	86	CGA	R	111	CUA	L	A
C	12	CDC	37	CAC	H	62	CCC	P	87	CGC	R	112	CUC	L	C
	13	CDG	38	CAG	Q	63	CCG	P	88	CGG	R	113	CUG	L	G
	14	CDU	39	CAU	H	64	CCU	P	89	CGU	R	114	CUU	L	U
	15	GDD	40	GAD		65	GCD		90	GGD		115	GUD		D
	16	GDA	41	GAA	E	66	GCA	A	91	GGA	G	116	GUA	V	A
G	17	GDC	42	GAC	D	67	GCC	A	92	GGC	G	117	GUC	V	C
	18	GDG	43	*Gag*	E	68	GCG	A	93	GGG	G	118	GUG	V	G
	19	GDU	44	GAU	D	69	GCU	A	94	GGU	G	119	GUU	V	U
	20	UDD	45	UAD		70	UCD		95	UGD		120	UUD		D
	21	UDA	46	UAA	Stop	71	UCA	S	96	UGA	Stop	121	UUA	L	A
U	22	UDC	47	UAC	Y	72	UCC	S	97	UGC	C	122	UUC	F	C
	23	UDG	48	UAG	Stop	73	UCG	S	98	UGG	W	123	UUG	L	G
	24	UDU	49	UAU	Y	74	UCU	S	99	UGU	C	124	UUU	F	U

Consistently with, but independently from, the organic chemistry experiments that support the necessity of five or more DNA bases in the primordial genetic system [28], the formal development of the algebraic theory necessarily leads to an extension of the DNA base alphabet. The introduction of an alternative hypothetical base D, as a variable in the mathematical model, leds to consistent phylogenetic results based on a weighted Manhattan distance [8]. It was demonstrated that the distance between codons is mathematically equivalent to the codon order according to the evolutionary importance of their DNA nucleotide bases [8]. The relationship between the genetic code architecture (expressed in the genetic-code cube) and the evolutionary mutational event have been reported [2]. Consistent phylogenetic analysis of DNA protein-coding regions can be obtained based on the genetic-code cube inserted in the 3D space R^3.

Appendix B. Classification of the 24 algebraic representations of the genetic code

Each DNA/RNA base can be classified into three main classes according to three criteria: number of hydrogen bonds (strong-weak), chemical type (purine-pyrimidine), and chemical groups (amino versus keto) [7]. Each criterion produces a partition of the set of bases [37]:

1) According to the number of hydrogen bonds (on DNA/RNA double helix): strong S={C,G} (three hydrogen bonds) and weak W={A,U} (two hydrogen bonds).
2) According to the chemical type: purines R={A, G} and pyrimidines Y={C,U}.
3) According to the presence of amino or keto groups on the base rings: amino M={C,A} and keto K={G,U}.

The ordered sets for each partition criterion are:

1) strong-weak (*SW*): {A,C,G,U}, {A,G,C,U}, {U,C,GA}, {U,G,C,A}, {C,A,U,G}, {C,U,A,G}, {G,A,U,C}, and {G,U,A,C}.
2) purine-pyrimidine (*YR*): {A,C,U,G}, {A,U,C,G}, {G,U,C,A}, {G,C,U,A}, {C,A,G,U}, {U,A,G,C}, {U,G,A,C}, and {C,G,A,U}.
3) amino-keto (*MK*): {A,G,U,C}, {A,U,G,C}, {C,U,G,A}, {C,G,U,A}, {G,C,A,U}, {G,A,C,U}, {U,A,C,G}, and {U,C,A,G}.

The 24 ordered base sets can be used to derive 24 ordered codon sets (*GC*), and 24 possible cubes for the standard genetic code [5]. For brevity, the set of 24 genetic-code cubes is denoted as $GC = \{SW, YR, MK\}$. Codon ordering in these sets is not arbitrary, but sorted out according to the evolutionary importance of base positions. Herein, we aim to show that a group structure $(GC, \circ)$ isomorphic to the well-known symmetric group S_4 can be defined on biophysical basis on the set *GC*.

Appendix C. Supporting material. Computational document format files

Computational document format (CDF) with graphic user interfaces to facilitate the comprehension of the theory exposed in the main text, as well as, its applications are provided as supplemental materials. A CDF is a standalone computable document created by using the software Wolfram Mathematica. The interaction with the CDF requires the installation of the software CDF player, which is freely available at http://www.wolfram.com/cdf/. A compressed zip file containing CDFs is provided at: https://drive.google.com/open?id=0B-4gzFH012dqc3NGOWl3R2s3a3c.

Inside the zip file, readers will find the files:

1) CDF-1: *IntroductionToZ5GeneticCodeVectorSpace.cdf*

   This CDF contains an interactive didactic introduction to the $\mathbb{Z}_5 -$vector space B^3 over the field $(\mathbb{Z}_5, +, \bullet)$ and to the general mathematical biology background used in this manuscript, as well as, tools to verify all the algebraic claims presented in the main text. Since the genetic-code algebras are found in the intersection of molecular biology and abstract algebras, I encourage the readers not familiar with this subject to see this CDF to get a fast and didactic introduction to the subject.

2) CDF-2: *Genetic-Code-Scales_of_Amino-Acids.cdf*

   This is a CDF containing an interactive graphical user interface tool to generate genetic code based PC-scales. The subjacent sets from the subgroups of the symmetric group of genetic-code cubes are given to explore different options to generate PC scales of amino acids correlated with physicochemical properties found in AAindex database [12]. The analysis for six protein sequence alignments is provided as well: 1) Repeat domain of breast cancer type 2 susceptibility protein, 2) Oxaloacetate decarboxylase, gamma chain, 3) p53 DNA binding domain, 4) Photosynthesis system II assembly factor YCF48 (PSII BNR repeat protein), 5) Influenza HA protein, 6) ENV and 7) GAG proteins from HIV1.

3) *GeneticCodeScales.wl*. File required to run "*Genetic-Code-Scales_of_Amino-Acids.cdf*".

4) *GeneticCodePC-scales&Weibull-fit_snapshots.pdf*

Appendix D. Deduction of the Weibull distribution for EMC

In a parsimony model framework, we would expect that mutational events with high Δx values should be less frequent than those with low values. In particular, if Δx is linked to the thermodynamics of organismal populations, then a natural statistical mechanical assumption considers the probability density function (PDF) $f(\Delta x)$ of Δx proportional to the Boltzmann factor $e^{-\left(\frac{\Delta x}{\beta(l)}\right)}$, i.e., $f(\Delta x) \propto e^{-\left(\frac{\Delta x}{\beta(l)}\right)}$ (A2), where $\beta(l) = \lambda(l) k_B T$ is a scaling parameter that depends on the Boltzmann constant k_B, the absolute temperature T and a proportionally constant $\lambda(l)$ that depends on the population size. The Boltzmann factor, $e^{-\left(\frac{\Delta x}{\beta(l)}\right)}$ reveals the relative probability of an arrangement for a given evolutionary cost. That is, on average, after a considerable number *N* of mutational events, the proportion of mutations with at least certain mutational cost Δx is constant and equal to the Boltzmann factor given by the formula: $\frac{n}{N} = e^{-\left(\frac{\Delta x}{\beta(l)}\right)}$, where *n* is the number of particles with mutational cost above Δx.

Since each mutational event is independent of the previous event and in a very small interval of time the chance of two or more mutations is negligible, mutational events usually are modelled by a Poisson process [38]. That is, given a Poisson process, the probability that an evolutionary cost Δx can be observed exactly *n* times in *N* mutational events is given by the binomial distribution: $B(n|N,q) = \frac{N!}{n!(N-n)!} q^n (1-q)^{N-n}$ (A3), where q is the probability of mutation success. Since it is expected that, under normal conditions, high values of Δx have low success probability q ($0 \le q \le 1$), it can be estimated subject to the constraint $ln(q) = (\alpha - 1) ln\left(\frac{\Delta x}{\beta(l)}\right) + c(l)$ ($\Delta x > 0$) (A4), where $c(l)$ is a constant parameter that depends on the population size l. This equation leads to equalities $ln(q) = c(l)$ for $\Delta x = \beta(l)$ and $c(l) = -(\alpha - 1) ln\left(\frac{\Delta x^0}{\beta(l)}\right)$ for $q = 1$, where Δx^0 is the cost with probability 1 (see below). Next, the scaling factor $\beta(l)$ can be estimated subject to the constraint $(\alpha - 1) ln\left(\frac{\beta(l)}{\Delta x^0}\right) = -ln(N)$ or $ln(N) = (\alpha - 1) ln\left(\frac{\Delta x^0}{\beta(l)}\right)$ (A5); then $c(l) = -ln(N)$. Thus, it can be assumed that $Nq = (\Delta x/\beta(l))^{\alpha - 1}$ (A6).

After large enough number of mutational events, the probability that an evolutionary cost Δx can be observed exactly *n* times in *N* mutational events approaches Poisson distribution $P(n|\nu) = \frac{\nu^n}{n!} e^{-\nu}$ (A7), where ν is the expected number of times that an evolutionary cost Δx can be observed in *N* mutational events, i.e., $\nu = (\Delta x/\beta(l))^{\alpha - 1}$ (A8). Next, the probability that a cost Δx would be observed at least one time in *N* mutational events will be $P(1|\nu) = \nu e^{-\nu}$ (A9). It should then be expected that mutational events with high probabilities $P(1|\nu)$ will be observed more frequently, i.e., $f(\Delta x) \propto \nu e^{-\nu}$ (A10). As a result, we can write $f(\Delta x) = k \nu e^{-\nu} e^{-\left(\frac{\Delta x}{\beta(l)}\right)}$ (A11), where k is a proportionality constant. By assuming $k = \frac{\alpha}{\beta(l)}$ this leads to the Weibull PDF:

$$f(\Delta x | \beta(l), \alpha) = \begin{cases} \frac{\alpha}{\beta(l)} \left(\frac{\Delta x}{\beta(l)}\right)^{\alpha-1} e^{-\left(\frac{\Delta x}{\beta(l)}\right)^{\alpha}} & \Delta x > 0 \\ 0 & \Delta x \le 0 \end{cases} \tag{A12}$$

References

[1] R. Sanchez, R. Grau, E. Morgado, A novel Lie algebra of the genetic code over the Galois field of four DNA bases, *Math. Biosci.* **202** (2006) 156–174.

[2] R. Sanchez, R. Grau, An algebraic hypothesis about the primeval genetic code architecture, *Math. Biosci.* **221** (2009) 60–76.

[3] R. Sanchez, E. Morgado, R. Grau, The genetic code Boolean lattice, *MATCH Commun. Math. Comput. Chem.* **52** (2004) 29–46.

[4] R. Sanchez, E. Morgado, R. Grau, Gene algebra from a genetic code algebraic structure, *J. Math. Biol.* **51** (2005) 431–457.

[5] M. V José, E.R. Morgado, R. Sánchez, T. Govezensky, The 24 possible algebraic representations of the standard genetic code in six or in three dimensions, *Adv. Stud. Biol.* **4** (2012) 119–152.

[6] F. H. C. Crick, The origin of the genetic code, *J. Mol. Biol*. **38** (1968) 367–379.

[7] A. Cornish-Bowden, Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984, *Nucleic Acids Res.* **13** (1985) 3021–3030.

[8] R. Sanchez, Evolutionary Analysis of DNA-protein-coding regions based on a genetic code cube metric, *Curr. Top. Med. Chem.* **14** (2014) 407–417.

[9] R. Sanchez, E. Morgado, R. Grau, R. Sánchez, A genetic code Boolean structure. I. The meaning of Boolean deductions, *Bull. Math. Biol.* **67** (2005) 1–14.

[10] J. P. Stevens, *Applied Multivariate Statistics for the Social Sciences*, *Routledge Academic*, 2009.

[11] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, *Nucleic Acids Res.* **36** (2008) D202–D205.

[12] S. Kawashima, H. Ogata, M. Kanehisa, AAindex: Amino acid index database, *Nucleic Acids Res.* **27** (1999) 368–369.

[13] H. J. Dyson, P. E. Wright, H. A. Scheraga, The role of hydrophobic interactions in initiation and propagation of protein folding., *Proc. Natl. Acad. Sci. U. S. A.* **103** (2006) 13057–13061.

[14] S. J. Freeland, R. D. Knight, L. F. Landweber, L. D. Hurst, Early fixation of an optimal genetic code, *Mol. Biol. Evol.* **17** (2000) 511–518.

[15] S. Itzkovitz, U. Alon, The genetic code is nearly optimal for allowing additional information within protein-coding sequences, *Genome Res.* **17** (2007) 405–412.

[16] A. Guilloux, J. L. Jestin, The genetic code and its optimization for kinetic energy conservation in polypeptide chains, *Biosystems.* **109** (2012) 141–144.

[17] L. Scrucca, GA: a package for genetic algorithms in R, *J. Stat. Softw.* **53** (2013) 1–37.

[18] A. Piantadosi, B. Chohan, D. Panteleeff, J. M. Baeten, K. Mandaliya, J. O. Ndinya-Achola, J. Overbaugh, HIV-1 evolution in gag and env is highly correlated but exhibits different relationships with viral load and the immune response, *AIDS* **23** (2009) 579–587.

[19] R Core Team, A language and environment for statistical computing, (2017).

[20] M. Wand, KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995). R package version 2.23-15, (2015).

[21] T. W. Yee, *Vector Generalized Linear and Additive Models: With an Implementation in R,* Springer, New York, 2015.

[22] T. W. Yee, VGAM: Vector Generalized Linear and Additive Models. R package version 1.0-3, (2017).

[23] K. Soetaert, plot3D: Plotting Multi-Dimensional Data. R package version 1.1, (2016).

[24] C. Camilloni, D. Bonetti, A. Morrone, R. Giri, C. M. Dobson, M. Brunori, S. Gianni, M. Vendruscolo, Towards a structural biology of the hydrophobic effect in protein folding, *Sci. Rep.* **6** (2016) 28285.

[25] K. H. Cheong, Z. X. Tan, N. Xie, M. C. Jones, A paradoxical evolutionary mechanism in stochastically switching environments, *Sci. Rep.* **6** (2016) #34889.

[26] K. Mineta, T. Matsumoto, N. Osada, H. Araki, Population genetics of non-genetic traits: Evolutionary roles of stochasticity in gene expression, *Gene* **562** (2015) 16–21.

[27] G. Christian, A. Favre, Everything you always wanted to know about Copula modeling but were afraid to ask, *J. Hydrol. Engin*. **12** (2007) 347–368.

[28] M. Levy, S. L. Miller, The stability of the RNA bases: implications for the origin of life, *Proc. Natl. Acad. Sci. U. S. A.* **95** (1998) 7933–7938.

[29] M. Bernstein, Prebiotic materials from on and off the early Earth., *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **361** (2006) #1689-700–2.

[30] M. W. Powner, J. D. Sutherland, Prebiotic chemistry: a new modus operandi., *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366** (2011) #2870–7.

[31] L. E. Orgel, Prebiotic chemistry and the origin of the RNA world., *Crit. Rev. Biochem. Mol. Biol.* **39** (2004) 99–123.

[32] S. Benner, H. J. Kim, M. J. Kim, A. Ricardo, Planetary organic chemistry and the origins of biomolecules., *Cold Spring Harb. Perspect. Biol.* **2** (2010) #a003467.

[33] A. C. Rios, Y. Tor, Refining the genetic alphabet: a late-period selection pressure? *Astrobiology* **12** (2012) 884–91.

[34] C. N. Birts, A. P. Sanzone, A. H. El-Sagheer, J. P. Blaydes, T. Brown, A. Tavassoli, Transcription of click-linked DNA in human cells, *Angew. Chem. Int. Ed.* **53** (2014) 2362–2365.

[35] J. Barciszewski, M. Barciszewska, G. Siboska, S. S. Rattan, B.C. Clark, Some unusual nucleic acid bases are products of hydroxyl radical oxidation of DNA and RNA, *Mol. Biol. Rep.* **26** (1999) 231–238.

[36] R. Sanchez, S.A. Mackenzie, Information thermodynamics of cytosine DNA methylation, *PLoS One*. **11** (2016) #e0150427.

[37] M. A. Jimenez-Montano, C. R. de la Mora-Basanez, T. Poschel, M. Jiménez-Montaño, C. R. de la Mora-Basánez, T. Pöschel, The hypercube structure of the genetic code explains conservative and non-conservative aminoacid substitutions in vivo and in vitro, *Biosystems* **39** (1996) 117–125.

[38] S. J. Balin, M. Cascalho, The rate of mutation of a single gene, *Nucleic Acids Res.* **38** (2009) 1575–1582.