Algebraic and toroidal representation of the genetic code

Rodrigo Rodríguez-Gutiérrez[1], Francisco Hernandez-Cabrera[1]*,
Francisco Javier Almaguer-Martínez[1] and
José De Jesús Bernal-Alvarado[2]

[1]Universidad Autónoma de Nuevo León, Facultad de Ciencias Físico Matemáticas, Av. Universidad s/n.
Ciudad Universitaria, San Nicolás de los Garza, Nuevo León, Mexico, [2]Departamento de Ingeniería
Física, División de Ciencias e Ingenierías, Universidad de Guanajuato, León, Mexico

The genetic code is a set of regulatory principles that control the translation of information encoded in messenger RNA (mRNA) into a sequence of amino acids. This study proposes a model starting from the relationships between its physicochemical properties of nucleobase in the codons. We employed a binary metric and the Kronecker product to represent the physicochemical properties of nucleobases in a vector space. This state space has a hierarchical order, with self-similarity and symmetry properties in the hydrogen bonds of triplets. The state space can be mapped under linear transformations to a toroidal geometry with allometric properties. Furthermore, this geometric representation highlights a charge symmetry that exists in amino acids and in the distribution of essential and non-essential amino acids. These results describe a state space between codon-anticodon interactions that can be interpreted as Bell states. This suggests that there is a quantum phenomenon involved in the mechanisms of information storage in DNA. In addition, toroidal geometry can be used to represent the sequences of codons of the mRNA that encode the sequence of amino acids of the proteins to find similarities or homologies between evolutionary related species.

1 Introduction

The deoxyribonucleic acid (DNA) molecule encodes the information that is involved in all organisms' development, traits, and functions. DNA is structured in a helical coil of two anti-parallel sequences of nucleotides, which, in turn, are made up of three fundamental blocks; a nitrogenous base (nucleobase), a sugar molecule (deoxyribose), and a phosphate group. Four types of nitrogenous bases are found in these nucleotide sequences: adenine (A), guanine (G), cytosine (C), and thymine (T). The information contained in each nucleobase can be classified into three physicochemical categories, according to its structure; Purines → P / Pyrimidines → Y, the functional group; Keto → K / Amino → M, and the type of Hydrogen bond; Strong → S / Weak → W (see Table 1).

Another property of the double strand of DNA is chirality, which obeys the shape and orientation of the anti-parallel and complementary sequences. This orientation is established by considering the position of the carbon atoms in the deoxyribose molecule. Thus, the sense strand of DNA will go from the 5' to 3' carbon and the antisense from the 3'

TABLE 1 Classification of nucleobases according to their three physicochemical properties.

Nucleobase	Structure (E)	Functional group (F)	Type of H bond (H)
Adenine (A)	Purine (R)	Amino (M)	Weak (W)
Guanine (G)	Purine (R)	Keto (K)	Strong (S)
Cytosine (C)	Pyrimidine (Y)	Amino (M)	Strong (S)
Thymine (T)	Pyrimidine (Y)	Keto (K)	Weak (W)

TABLE 2 Outcome of the AND (∧) operator applied to the Y/R and K/M binary states to represent the nucleobases.

Nucleobase	Y/R	K/M	Y/R ∧ K/M
A	0	0	0
G	0	1	0
C	1	0	0
T	1	1	1

to 5'. This orientation is important in DNA replication and transcription mechanisms. Inthe transcription process, the coding DNA strand is transcribed into a primary transcript called pre-messenger RNA (pre-mRNA) that, after processing result in a messenger RNA (mRNA), can be exported to the cytosol and traslated into a protein, where the Thymine nucleotide is replaced by Uracil (U) without loss or change in information. Uracil has the same physicochemical properties previously described for Thymine; it is a Pyrimidine with a Keto group that forms a double hydrogen bond with Adenine. The mRNA carries the information that will be read in triplets of nucleotides (codons) by the ribosome. Various tRNA molecules (46 different molecules in the standard code) with different anticodons are used to recognize 20 specific sets of mRNA codons (singlets, doublets, quadruplets etc.). This specific sets of synonymous codons are used to specify each one of the 20 amino acids.

The study of the genetic code has been the subject of research from various scientific perspectives using the physicochemical properties of the nucleotides in the coding DNA sequences [1–3]. In particular, the concept of genetic code degeneracy and codon formation were two of the most studied areas [4, 5]. Rumer et al. [4] studied the types of hydrogen bonds formed between codon-anticodon, as well as the stability associated with the third nucleobase of this interaction. In recent research, binary metrics have been used to represent the three main physicochemical properties of nucleobases, to propose matrix arrangements that justify the degeneracy of the genetic code [6–8]. A previous study suggests that a mathematical model could justify the stability of codons and degeneracy of the genetic code. Furthermore, symmetries have been found in the genetic code with number theory [9]. Previous studies suggest that the genetic code could be represented based on an algebraic structure, to model the physicochemical properties of codon–anticodon interactions. Other findings show that correlations exist between the positions of the nucleobases in each codon and the physicochemical properties of the corresponding amino acids. However, it could be helpful to establish a theoretical formalism that explains certain features associated with the transfer of information related to the genetic code.

In this study, a binary metric is applied to the physicochemical properties to generate a Hilbert space, which represents all the combinations of nucleobases that forms the 64 codons. This model suggests a hierarchical order in the physicochemical properties of the nucleobases associated with the codons and symmetric representations in the type of bond observed.

2 Methodology

The physicochemical properties of nucleobases (Table 1) can be used to analyze sequences in time series such as DNA walks [10–13]. In these studies, binary metrics have been used to represent physicochemical properties. For this research, we use a methodology similar to the study by Hernandez-Cabrera et al. [14], where the physicochemical properties of the nucleobases that form hydrogen bonds are studied. This structural (E), functional group (F), and hydrogen bond (H) properties can be represented by the Y/R, K/M, and W/S states (Equations 1–3), respectively. Each of the states that represent the nucleobases can be transformed into logical inputs, *False*(0) and *True*(1). Thus, the Y/R states have logical values; $Y \rightarrow 1$ and $R \rightarrow 0$, for the K/M states; $K \rightarrow 1$ and $M \rightarrow 0$, and finally the S/W states; $W \rightarrow 1$ and $S \rightarrow 0$.
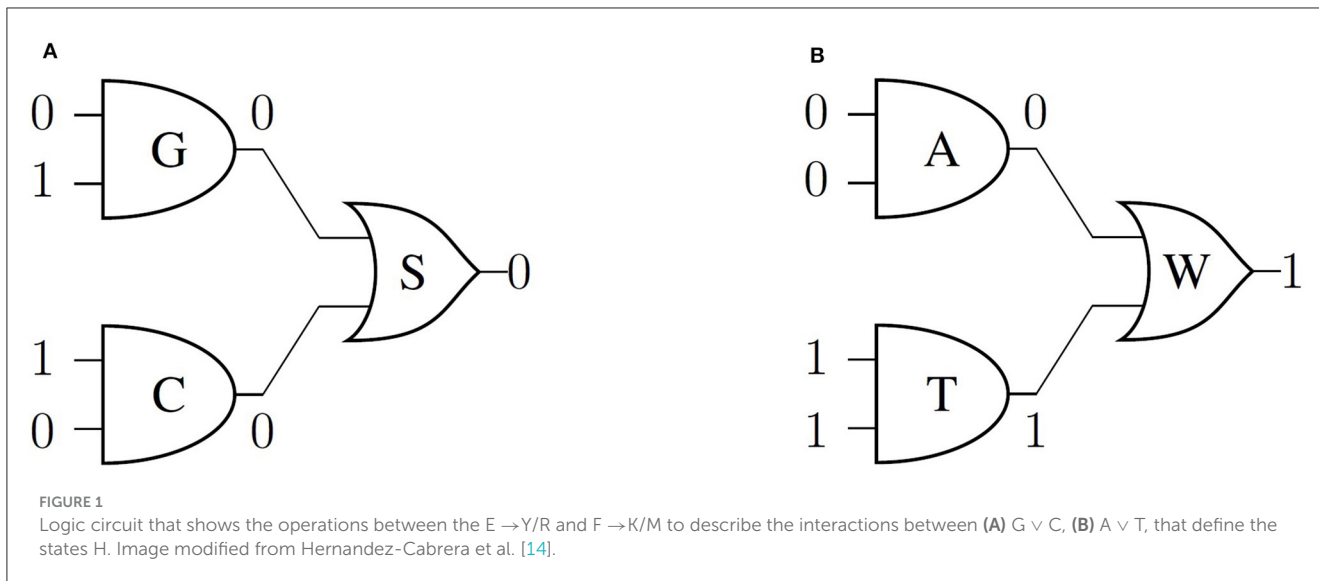
$$\text{Structure (E)} \begin{cases} 1 & \text{if Pyrimidine } (Y) = \{C, T\} \\ 0 & \text{if Purine } (R) = \{A, G\} \end{cases} \quad (1)$$

$$\text{Functional group (F)} \begin{cases} 1 & \text{if Keto } (K) = \{G, T\} \\ 0 & \text{if Amino } (M) = \{A, C\} \end{cases} \quad (2)$$

$$\text{Hydrogen bond (H)} \begin{cases} 1 & \text{if Weak } (W) = \{A, T\} \\ 0 & \text{if Strong } (S) = \{G, C\} \end{cases} \quad (3)$$

Considering the previous metric, we propose the logical operator AND (∧) to connect two independent propositions that must be satisfied in the nucleobases, considering the binary inputs of the E and F properties. In this way, the input data of each ∧ operator used to define G, C, A, and T in terms of the E and F properties are $(0 \wedge 1), (1 \wedge 0), (0 \wedge 0)$, and $(1 \wedge 1)$, respectively. Therefore, the logical operator will show as a result an output equal to 1 for T and 0 for G, C, and A. In this output state, the two fundamental characteristics (structure and functional group) associated with each nucleobase are combined, as shown in Table 2.

To model the double-stranded DNA stability, we propose an OR operator (∨) to represent the interaction between two nucleobases that can generate hydrogen bonds. The operator ∨ inputs the results of the AND operation of each nucleobase (Table 2), and this logical circuit can be represented by an XOR gate. Thus, the G and C states operated by ∨ will have H → 0 state, corresponding to the strong bond (Figure 1A). Then, we can prove that the interaction between A and T, defined by ∨, results in H →

FIGURE 1
Logic circuit that shows the operations between the E →Y/R and F →K/M to describe the interactions between **(A)** G ∨ C, **(B)** A ∨ T, that define the states H. Image modified from Hernandez-Cabrera et al. [14].

1 (Figure 1B). Moreover, we can establish the H states starting from the structure and functional group properties of complementary nucleobases.

Therefore, it is possible to represent the H states as a consequence of the interaction between two nucleobases using a logic circuit (Figure 1). All possible interactions permitted for the expressions (Equations 1, 2) are presented in Supplementary material. This model meets the rules of Boolean algebra and effectively depicts the interactions between physicochemical properties of complementary nucleobases. However, the logic circuit used could imply that there are no restrictions on the interactions between non-complementary nucleobases (A and G or C and T) to represent the permitted hydrogen bonds. To overcome this inconsistency, in the following section, we propose an expansion of the model's state space depicting the correct interactions.

2.1 Physicochemical orthogonal transformation properties of the nucleobases

A key feature of this system is that states of the E and F properties are independent and orthogonal in the Watson-Crick base pairing. Therefore, it is appropriate to represent the E and F properties as orthonormal vectors in $\mathbb{C}$. Hence, we proposed a finite dimensional complex vector space system with an inner product and associated norm. While the current analysis restricts the system to a finite-dimensional setting, it possesses the potential for extension to a complete infinite-dimensional system. This inherent expandability motivates our use of the Hilbert space framework. Moreover, in a previous study [14], the hydrogen bond representation using this metric suggest an analogy with the Pauli matrices. Then, we use this mathematical framework to represent the physicochemical states (Y/R, K/M, and S/W) using the following transformations: $|0> = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $|1> = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Then,

$R \to |0>, Y \to |1>$ for the structure and $M \to |0>, K \to |1>$ for the functional group.

Now, each nucleobase can be described as a linear combination of two independent variables (structure and functional group) in a complex vector space $\mathcal{H}$ (Hilbert space). Hence, the structure (ψ) and functional group (ϕ) can be expressed as follows:

$$|\psi> = a|0> + b|1> \tag{4}$$

$$|\phi> = c|0> + d|1> \tag{5}$$

where $|a|^2$ and $|b|^2$ are positive constants that represent the probability of having a Purine or Pyrimidine in the structural function (ψ) of each nucleobase, while, $|c|^2$ and $|d|^2$ are the probabilities of having an Amino or Keto functional group (ϕ) of each nucleobase. These constants must satisfy the normalization condition, $|a|^2 + |b|^2 = 1$ and $|c|^2 + |d|^2 = 1$ for Equations 4, 5, respectively.

2.2 Nucleobase representation using the Kronecker product

We use the Kronecker product ($\otimes$) in analogy to the AND ($\wedge$) operator to engage the complementary physicochemical properties of each nucleobase. The Kronecker product helps us to model nucleobase complementarity, since it is not a commutative operation. Moreover, there is a hierarchical order in operations between ψ and ϕ, where the structure is prioritized over the functional group. Then, we applied the Kronecker product to ψ and ϕ, as shown in Equation 6. Definition of the Kronecker product is provided in Supplementary material.

$$|\varepsilon> \to |\psi> \otimes |\phi> = ac|00> + ad|01> + bc|10> + bd|11> \tag{6}$$

Given that $a = b = c = d = 1/\sqrt{2}$. Then, $|ac|^2 = |ad|^2 = |bc|^2 = |bd|^2 = 1/4$ are the probabilities to select a combination of two allowed physicochemical properties to assemble each nucleobase. The Kronecker product between structure state space ($\mathcal{H}_E$) and functional group state space ($\mathcal{H}_F$) results in an assemble state space ($\mathcal{H}_{EF} = \mathcal{H}_E \otimes \mathcal{H}_F$). Therefore, $|\varepsilon >$ represents the state space of the nucleobases A, G, C, and T. In this state space, the four possible values represent the nucleobases (Equations 7–10):

$$A \rightarrow \frac{1}{\sqrt{2}}|0> \otimes \frac{1}{\sqrt{2}}|0> = \frac{1}{2}\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{2}|00> \quad (7)$$

$$G \rightarrow \frac{1}{\sqrt{2}}|0> \otimes \frac{1}{\sqrt{2}}|1> = \frac{1}{2}\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{2}|01> \quad (8)$$

$$C \rightarrow \frac{1}{\sqrt{2}}|1> \otimes \frac{1}{\sqrt{2}}|0> = \frac{1}{2}\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \frac{1}{2}|10> \quad (9)$$

$$T \rightarrow \frac{1}{\sqrt{2}}|1> \otimes \frac{1}{\sqrt{2}}|1> = \frac{1}{2}\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \frac{1}{2}|11> \quad (10)$$

Analogously, we can expand the results for two and three adjacent nucleobases, as shown in the Equations 11, 12.

$$|\varepsilon> \otimes |\varepsilon> \rightarrow \frac{1}{4}(|0000> + |0001> + \cdots + |1111>) \rightarrow$$
$$\frac{1}{4}(AA + AG + \cdots + TT) \quad (11)$$

$$|\varepsilon> \otimes |\varepsilon> \otimes |\varepsilon> \rightarrow \frac{1}{8}(|000000> + \cdots + |111111>) \rightarrow$$
$$\frac{1}{8}(AAA + \cdots + TTT) \quad (12)$$

$$\mathcal{H}_{codon} = (\mathcal{H}_E \otimes \mathcal{H}_F)_{N1} \otimes (\mathcal{H}_E \otimes \mathcal{H}_F)_{N2} \otimes (\mathcal{H}_E \otimes \mathcal{H}_F)_{N3} \quad (13)$$

The combination of physicochemical properties for the 3 adjacent nucleobases (Equation 13) results in a codon space containing 64 possible states. Where N1, N2, and N3 represent the first, second, and third positions of the nucleobases in a codon.

Remarking the codons sorted by their bit values (Table 3) reveals that a transversion mutation (change in purine to pyrimidine or vice versa) in the third position of a codon can result in a change in the amino acid encoded.

To model the interaction between codon–anticodon, we apply the Kronecker product to complementary antiparallel

TABLE 3 Table with the codons sorted in ascending order by their value in bits, Purine nucleobases in the third position are highlighted in red, while Pyrimidine nucleobases in the third position are highlighted in blue.

Binary representation	Codon	aa
00 00 00	AAA	K
00 00 01	AAG	K
00 00 10	AAC	N
00 00 11	AAT	N
⋮	⋮	⋮
11 11 00	TTA	L
11 11 01	TTG	L
11 11 10	TTC	F
11 11 11	TTT	F

nucleobases. Therefore, we add a $180°$ rotation using the phase shift operator $Ph(\delta) = e^{i\delta}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$; where i is the imaginary unit, $\delta = \pi$ is the rotation angle, and the I is the identity matrix (Equations 14–16). This operation depicts the antiparallel complementary nucleobases, rotated 180 degrees relative to the nucleobase in the 5' to 3' strand. Biologically, this is a justification for representing the steric interactions allowed in Watson-Crick base pairing.

$$Ph(\pi) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \quad (14)$$

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (15)$$

$$\mathcal{I}_k = I \otimes I \otimes \cdots \otimes I \quad , \quad k - times \quad (16)$$

The $\mathcal{I}_k$ matrix is given by I_{2x2} for $k = 1$; I_{4x4} for $k = 3$; and I_{8x8} for $k = 5$. Then, the $180°$ rotation of the complementary antiparallel nuclobaseses are given by $\mathcal{I}_k \otimes Ph(\pi)$, as shown in Equations 17–19 for one, two, and three nucleobases, respectively.

$$(\mathcal{I}_1 \otimes Ph(\pi))|\varepsilon> \rightarrow -\frac{1}{2}(|00> + |01> + |10> + |11>) \quad (17)$$

$$(\mathcal{I}_3 \otimes Ph(\pi))(|\varepsilon> \otimes |\varepsilon>) \rightarrow -\frac{1}{4}(|0000> + |0001> + \ldots$$
$$+ |1111>) \quad (18)$$

$$(\mathcal{I}_5 \otimes Ph(\pi)(|\varepsilon> \otimes |\varepsilon> \otimes |\varepsilon>)) \rightarrow -\frac{1}{8}(|000000> + \ldots$$
$$+ |111111>) \quad (19)$$

Finally, the nucleobases of the sense strands are added with their complementary nucleobases to the antisense strand, analogously to the OR operation in the logical circuit. We noted that the antisense nucleobases ($A, G, C,$ and T) are the negation (NOT operator) of the nucleobases in the sense strand. Then, we can represent the hydrogen bond of complementary nucleobases using the symbol (-) as shown in the examples for one and three complementary antiparallel nucleobases (Equations 20–23).

$$A - T \rightarrow \frac{1}{2}(|00> -|11>) \tag{20}$$

$$T - A \rightarrow \frac{1}{2}(|11> -|00>) \tag{21}$$

$$5' \quad AAA \quad 3' - 3' \quad TTT \quad 5' \rightarrow \frac{1}{8}(|000000> -|111111>) \tag{22}$$

$$5' \quad TTT \quad 3' - 3' \quad AAA \quad 5' \rightarrow \frac{1}{8}(|111111> -|000000>) \tag{23}$$

The Equations 20–23 describe a symmetric property between binary states of codon–anticodon interactions. Moreover, the stability of the codon–anticodon interaction depends on the complementary physicochemical properties of the nucleobases. Specifically, hydrogen bonds are formed between nucleobases possessing opposite properties; $(R/M) \leftrightarrow (Y/K)$ or $(R/K) \leftrightarrow (Y/M)$. Therefore, changes in the physicochemical properties of a codon must be compensated by state changes in the corresponding properties of the anticodon. This suggests an interdependent relationship to maintain hydrogen bond stability. In addition, these states cannot be factorized with respect to the Kronecker product and have the same probability of occurring. Hence, the interactions between the physicochemical properties that describe the hydrogen bond formation between complementary antiparallel nucleobases, including the codon–anticodon state space, could be interpreted as an entangled system. A more detailed analysis on the physicochemical interaction between nucleobases is shown in the study by Hernandez-Cabrera et al. [14].

2.3 Metric representation of the physicochemical state space

Another representation of the state space can be obtained using the Kronecker product between $|\psi>$ and $<\phi|$ to generate a matrix representation of the physicochemical properties of each nucleobase, as shown in the Equation 24, where each summand represents the combination of states of a nucleobase. Therefore, the nucleobases can be represented as Equations 25–28.

$$\eta \rightarrow |\psi> \otimes <\phi| = ac|0> \otimes <0| + ad|0> \otimes <1|$$
$$+ bc|1> \otimes <0| + bd|1> \otimes <1| \tag{24}$$

$$A \rightarrow \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{25}$$

$$G \rightarrow \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{26}$$

$$C \rightarrow \frac{1}{2}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \tag{27}$$

$$T \rightarrow \frac{1}{2}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \tag{28}$$

Now, we apply the Kronecker product between two adjacent nucleobases (Equation 29), and the obtained individual matrices can be expressed as Equation 30.

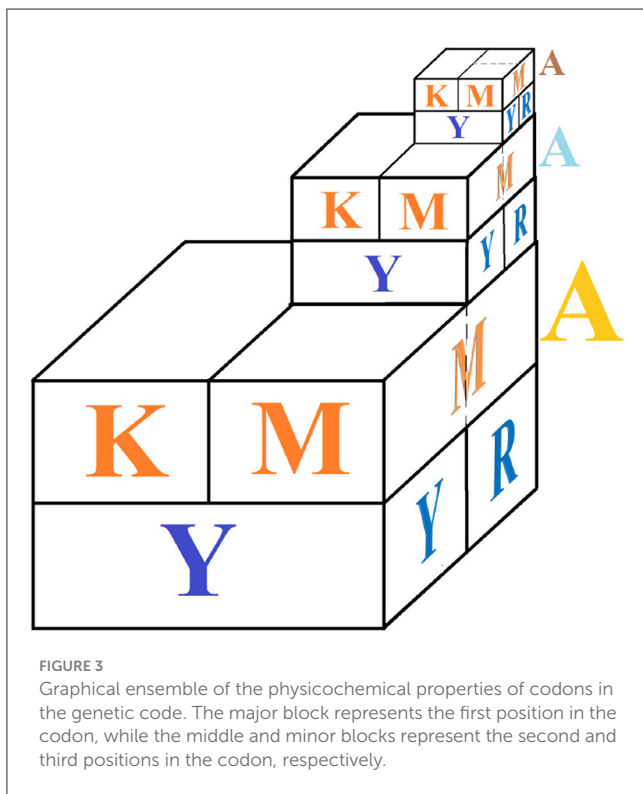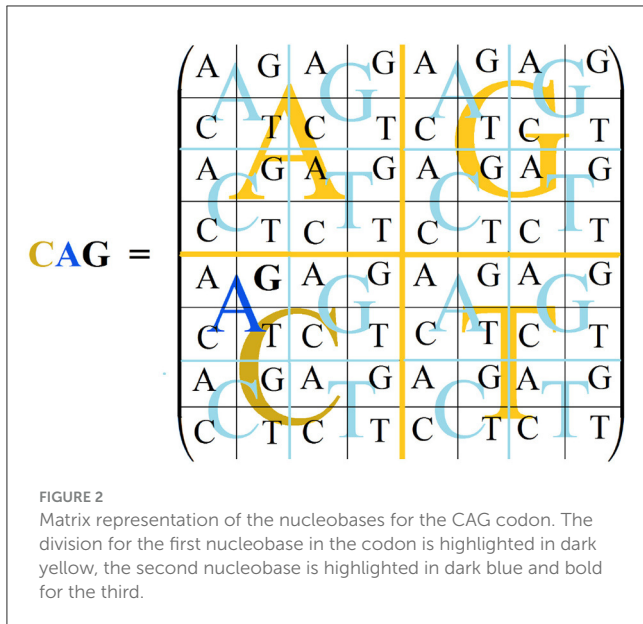$$\eta \otimes \eta \rightarrow \frac{1}{4}(A \otimes A + A \otimes G + \cdots + T \otimes T) \tag{29}$$

$$A \otimes G \rightarrow \frac{1}{4}\begin{pmatrix} 0 & 1 & | & 0 & 0 \\ 0 & 0 & | & 0 & 0 \\ 0 & 0 & | & 0 & 0 \\ 0 & 0 & | & 0 & 0 \end{pmatrix} \tag{30}$$

The 16 elements of Equation 29 can be observed as a matrix of four major blocks with four elements in each one. In Equation 30, the first major block represents A, and their four elements are given by the G matrix (Equation 26). Hence, there are two hierarchical levels to describe the interaction of two adjacent nucleobases in a sequence of mRNA.

To obtain a matrix representation of the state space for the genetic code, we apply the Kronecker product to three adjacent nucleobases, as shown in Equation 31. Where the $N_i, N_j,$ and N_k are the nucleobases and $|c_{ijk}|^2$ is the probability of selecting 1 of the 64 codons. This probability can be related to preferential use of the codons in an organism. This matrix representation has 64 elements divided into 3 hierarchical levels, according to the nucleobase position in the codon (Figure 2). The matrix representation for the CAG codon is shown in Equation 32 and Figure 2.

$$\omega = \eta \otimes \eta \otimes \eta \rightarrow \sum_{i=1}^{4}\sum_{j=1}^{4}\sum_{k=1}^{4} c_{ijk}(N_i \otimes N_j \otimes N_k) \tag{31}$$

$$C \otimes A \otimes G \rightarrow c_{ijk}\begin{pmatrix} 0 & 0 & 0 & 0 & | & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & | & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & | & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & | & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & | & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & | & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & | & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & | & 0 & 0 & 0 & 0 \end{pmatrix} \tag{32}$$

FIGURE 2
Matrix representation of the nucleobases for the CAG codon. The division for the first nucleobase in the codon is highlighted in dark yellow, the second nucleobase is highlighted in dark blue and bold for the third.

FIGURE 3
Graphical ensemble of the physicochemical properties of codons in the genetic code. The major block represents the first position in the codon, while the middle and minor blocks represent the second and third positions in the codon, respectively.

The representation of the Kronecker product between the matrices that model the codon state space has self-similarity properties, as shown in Figure 3. The hierarchical order of the representations in Figures 2, 3 is associated with the position of the nucleobases of each codon, implying that the first nucleobase in the codon has the greatest effect on the anticodon recognition.

3 Results

In agreement to Equations 31, 32, the complete representation of the genetic code is given by expression (Equation 33), and the corresponding amino acid matrix is given by Equation 34.

In this algebraic framework, linear transformations ($\mathcal{R}_A$) may be applied to the matrix without affecting its underlying properties (Equation 35B). Preserving the arrangement of the first block, reflections can be performed on the first and second nucleobases to generate the elements of the second, third, and fourth major blocks (Equation 35A). Biologically, this represents transitions or transversions between nucleobases.

We proposed a new symmetric arrange that reflects the last two nucleobases of each codon with respect to the horizontal, vertical, and central axes. Equation 36 represents this matrix arrangement, where the red dotted lines serve as indicators of the vertical and horizontal axes. In addition, this arrangement also shows symmetries between the weak bond triplets in the principal diagonal of the matrix, e.g., AAA, TAA, and TTT. Moreover, a correlation can be established between the preferential codon usage ($|c_{ijk}|^2$) and the stabilization-free energy for the trimeric duplexes formed by the complementary antiparallel codons, as demonstrated in the study by Klump et al. [15]. Figures 4A, B illustrate the distribution of the normalized free energy for both the original (Equation 33) and reflected codon matrix (Equation 36), respectively. These figures reveal that the free energy for the first two nucleobases in a codon, i.e., GC, CG, GG, and CC pairs (strong H bond) exhibit greater stability compared with the AT, TA, AA, and TT pairs (Weak H bond). Furthermore, an entropy map can be constructed to represent the Shannon entropy of the binary physicochemical states [like the Ising model, Taroni [16]] for each codon in Equations 33, 36, as shown in Figures 4C, D, respectively. To construct an entropy map, we use the appearance probability of the codons depending on their binary representation (Table 3), given by their physicochemical properties, as shown in Figure 5. Furthermore, the Shannon entropy state space for codons can be observed in a configuration consistent with a Karnaugh map in the framework of Boolean algebra. Therefore, the matrix representation of the genetic code can be mapped to a toroidal structure.

$$
\omega \to c_{ijk}
\begin{pmatrix}
AAA & AAG & | & AGA & AGG & | & GAA & GAG & | & GGA & GGG \\
AAC & AAT & | & AGC & AGT & | & GAC & GAT & | & GGC & GGT \\
ACA & ACG & | & ATA & ATG & | & GCA & GCG & | & GTA & GTG \\
ACC & ACT & | & ATC & ATT & | & GCC & GCT & | & GTC & GTT \\
CAA & CAG & | & CGA & CGG & | & TAA & TAG & | & TGA & TGG \\
CAC & CAT & | & CGC & CGT & | & TAC & TAT & | & TGC & TGT \\
CCA & CCG & | & CTA & CTG & | & TCA & TCG & | & TTA & TTG \\
CCC & CCT & | & CTC & CTT & | & TCC & TCT & | & TTC & TTT
\end{pmatrix}
\quad (33)
$$

$$
AA \to c_{ijk}
\begin{pmatrix}
LYS & LYS & | & STP & STP & | & GLU & GLU & | & GLY & GLY \\
ASN & ASN & | & SER & SER & | & ASP & ASP & | & GLY & GLY \\
THR & THR & | & MET & MET & | & ALA & ALA & | & VAL & VAL \\
THR & THR & | & ILE & ILE & | & ALA & ALA & | & VAL & VAL \\
GLN & GLN & | & ARG & ARG & | & STP & STP & | & TRP & TRP \\
HIS & HIS & | & ARG & ARG & | & TYR & TYR & | & CYS & CYS \\
PRO & PRO & | & LEU & LEU & | & SER & SER & | & LEU & LEU \\
PRO & PRO & | & LEU & LEU & | & SER & SER & | & PHE & PHE
\end{pmatrix}
\quad (34)
$$
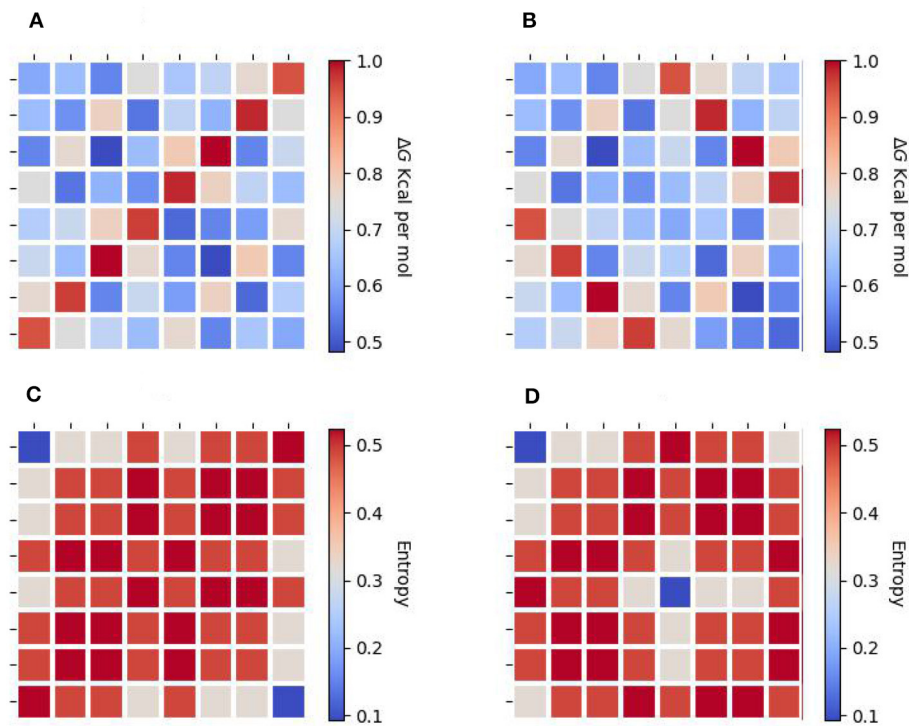
FIGURE 4
Energy maps depicting the normalized free energy for the trimeric duplexes formed by the complementary antiparallel codons in human mitochondrial DNA (up). **(A)** shows the original matrix, and **(B)** shows the reflected matrix. The heatmap reveals that the free energy mixture in the first two codons for guanine (G) and cytosine (C) is more stable than the mixture of adenine (A) and thymine (T). Entropy map depicting the Shannon entropy of the physicochemical states for each codon (down). **(C)** shows the original matrix, and **(D)** shows the reflected matrix. We can observe that the entropy map of **(C, D)** can be folded to form a torus.

$$a) \quad \begin{pmatrix} I & | & II \\ -- & -- & -- \\ III & | & IV \end{pmatrix} \quad b) \quad \begin{pmatrix} A & G & | & G & A \\ & & | & & \\ C & T & | & T & C \\ - & - & - & - & - \\ C & T & | & T & C \\ & & | & & \\ A & G & | & G & A \end{pmatrix} \quad (35)$$

$$\mathcal{R}_A\omega \to c_{ijk} \quad \begin{pmatrix} AAA & AAG & | & AGA & AGG & | & GGG & GGA & | & GAG & GAA \\ AAC & AAT & | & AGC & AGT & | & GGT & GGC & | & GAT & GAC \\ ACA & ACG & | & ATA & ATG & | & GTG & GTA & | & GCG & GCA \\ ACC & ACT & | & ATC & ATT & | & GTT & GTC & | & GCT & GCC \\ -- & -- & - & -- & -- & - & -- & -- & - & -- & -- \\ CCC & CCT & | & CTC & CTT & | & TTT & TTC & | & TCT & TCC \\ CCA & CCG & | & CTA & CTG & | & TTG & TTA & | & TCG & TCA \\ CAC & CAT & | & CGC & CGT & | & TGT & TGC & | & TAT & TAC \\ CAA & CAG & | & CGA & CGG & | & TGG & TGA & | & TAG & TAA \end{pmatrix} \quad (36)$$

$$\mathcal{R}_A(AA) \to c_{ijk} \quad \begin{pmatrix} LYS & LYS & | & STP & STP & | & GLY & GLY & | & GLU & GLU \\ ASN & ASN & | & SER & SER & | & GLY & GLY & | & ASP & ASP \\ THR & THR & | & MET & MET & | & VAL & VAL & | & ALA & ALA \\ THR & THR & | & ILE & ILE & | & VAL & VAL & | & ALA & ALA \\ PRO & PRO & | & LEU & LEU & | & PHE & PHE & | & SER & SER \\ PRO & PRO & | & LEU & LEU & | & LEU & LEU & | & SER & SER \\ HIS & HIS & | & ARG & ARG & | & CYS & CYS & | & TYR & TYR \\ GLN & GLN & | & ARG & ARG & | & TRP & TRP & | & STP & STP \end{pmatrix} \quad (37)$$

3.1 Toroidal geometry of the genetic code

After applying this transformation, we model the continuous codon states of Equations 36 and 37 using a toroidal geometry under an isomorphic transformation (Figure 7). In this mitochondrial toroidal model, we can assign angular coordinates ($\alpha, \theta \in [0, 2\pi)$) to each codon, where α is the toroidal angle that goes from the center of the torus to the center of the toroidal cavity, while θ is the poloidal angle corresponding to the transversal area of the torus cavity (Figure 6). We suggest that the glycine codon (GGG) could be the starting point, represented by (0,0). In a 2D mapping, we can observe a vortex over the amino acids that have three strong hydrogen bond (SSS) interactions between codon and anticodon. In addition, another vortex is on direction of amino acids that have three weak hydrogen bonds (WWW). Additionally, the difference between angles of vortex lines ($\theta_{SSS} - \theta_{WWW}$) is always π radians (Figure 8).

The upper region of Figure 7 is predominantly occupied by polar amino acids, while in the lower region, non-polar amino acids are found. In addition, there is a symmetry in the distribution of the charged amino acids ($K^+, H^+, R^+, D^-,$ and E^-), which are found at the top. Moreover, D^- and E^- are opposed to R^+ ($\alpha = \pi$). In addition, the four stop codons are facing each other at the top of the torus. The essential amino acids (highlighted in orange) are in the bottom half of the torus, except for K, H,

FIGURE 5
Distribution function for the appearance probability of codons depending on their binary representation (Table 3). **(a)** codons with 6 zeros (000000), **(b)** codons with 5 zeros (000001, 000010, ...), **(c)** codons with 4 zeros (000011, 000101, ...), **(d)** codons with 3 zeros (000111, 001011, ...), **(e)** codons with 2 zeros (001111, 010111, ...), **(f)** codons with 1 zero (011111, 101111, ...), and **(g)** codons without zero (111111).

and W. Additionally, the polar amino acids (K, H, and T) are all located in the same quadrant (Figure 8). Otherwise, most of the non-essential amino acids (highlighted in bold) are in the fourth quadrant.

4 Discussion

The genetic code is highly conserved throughout evolution across all known life forms, and its redundancy provides some protection against mutations. Previous investigations of the genetic code have been focused on understanding its evolution and how this has led to its degeneration [6–8]. Moreover, binary metrics and group theory have been used to generate graphical representations in 2D [17, 18] and 3D [19, 20] for phylogenetic analysis.

Our proposal is based on several tools, such as binary metrics, logical circuits, linear algebra, and topological representations, to model the genetic code. In addition, we have used previous research studies to understand the evolution of the genetic code [6, 7] and matrix representations of the codons using the Kronecker product [8]. Something in common with previous studies is the binary metric applied to the physicochemical properties of nucleobases. The results show that the physicochemical interactions that describe the codon–anticodon state space could be represented by Bell states. This does not imply that the nucleobases themselves exhibit the property of entanglement. However, hydrogen bonds formed by complementary antiparallel nucleobases could be entangled. This proposal is in accordance with the results obtained by Hernandez-Cabrera et al. [14]. There are several examples of biological systems that may exhibit quantum behaviors at nanoscales, such as quantum coherence in photoreceptors, magnetoreception in birds, and quantum tunneling in biomolecules among others [21]. This could suggest that quantum phenomena could be involved in the mechanisms of information storage in DNA. Over and above that, this study
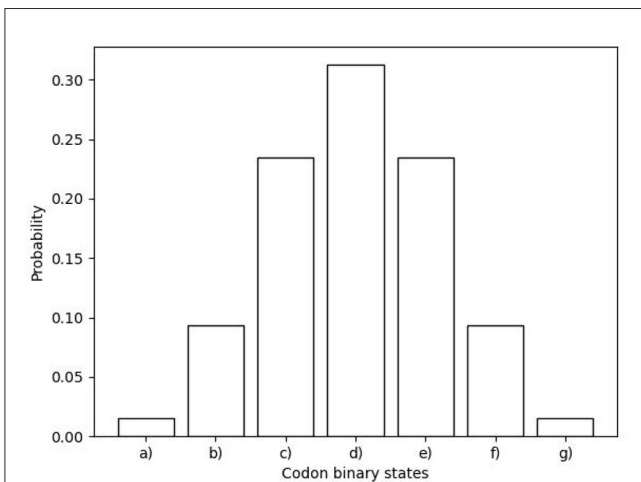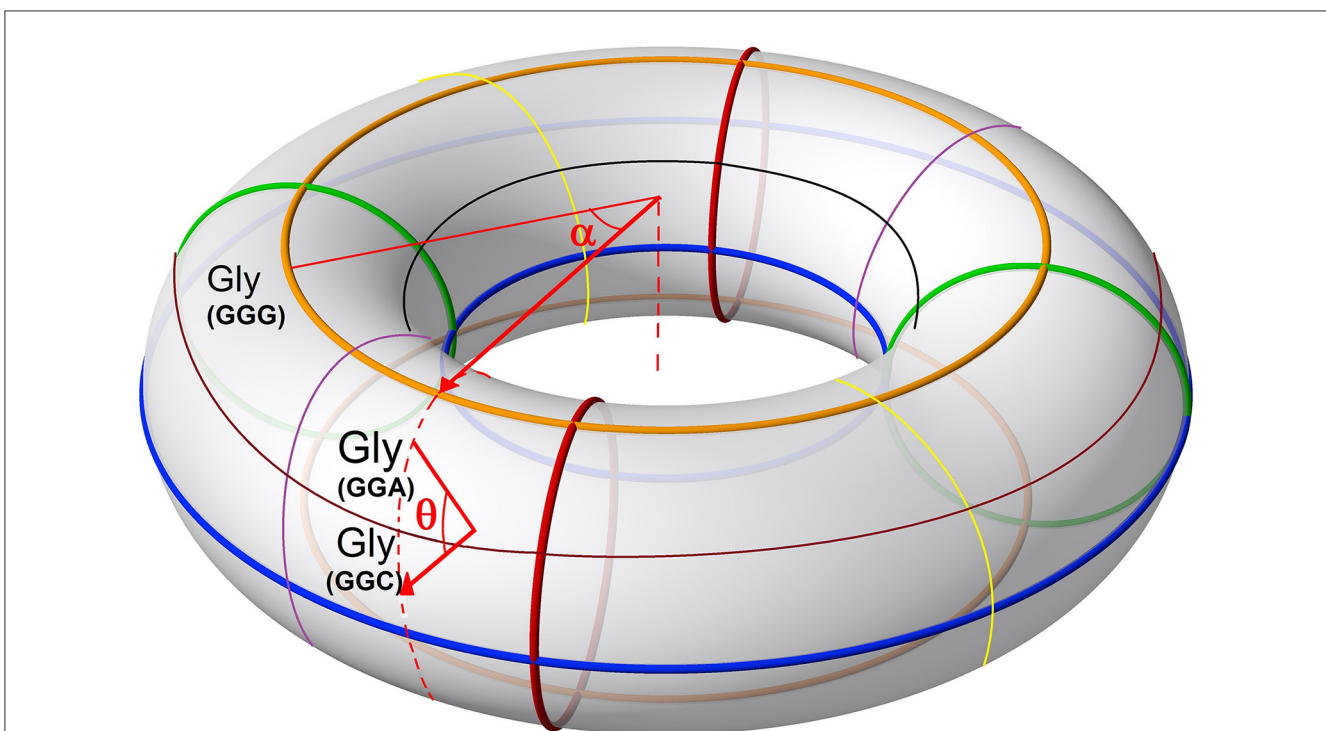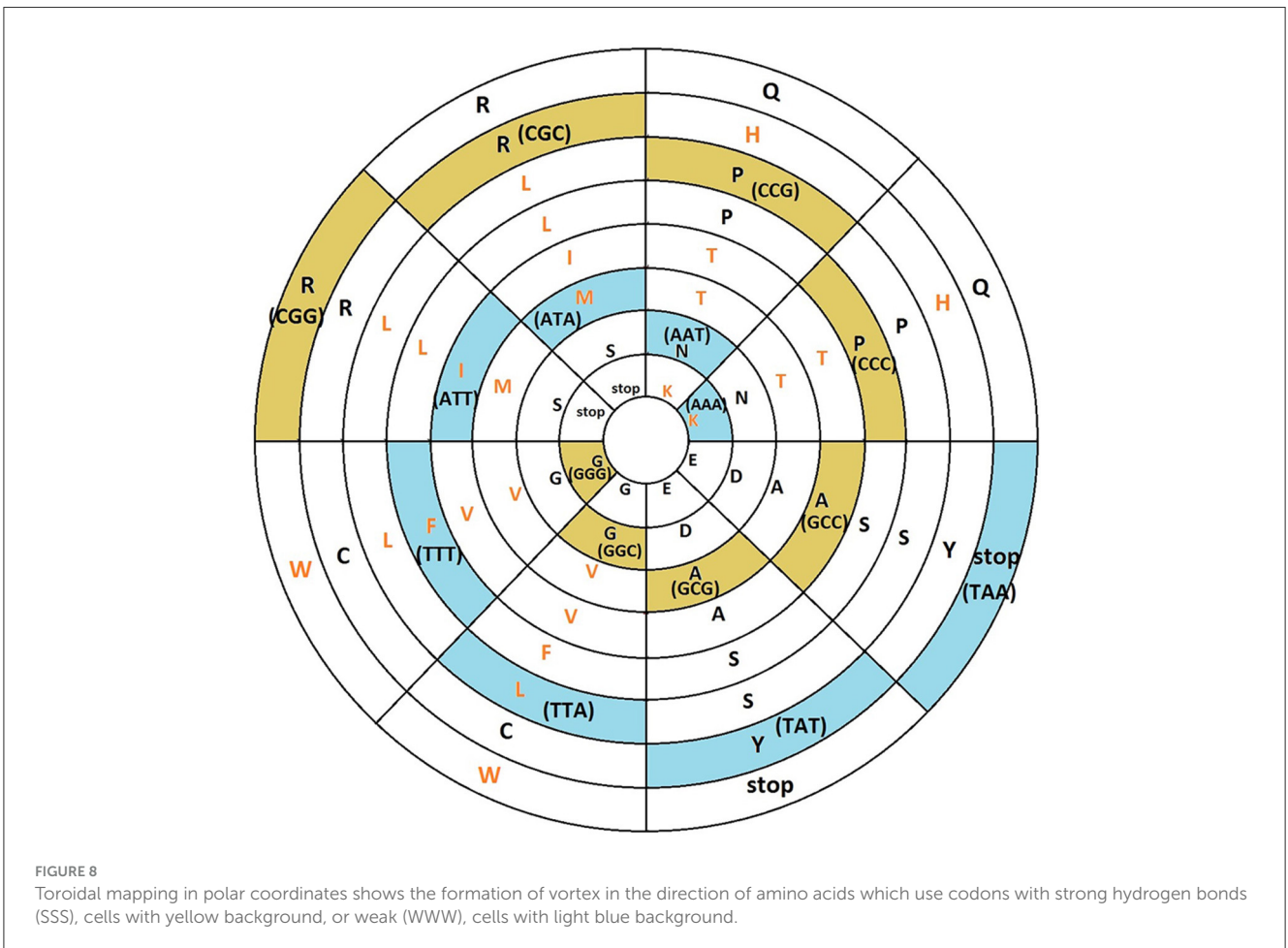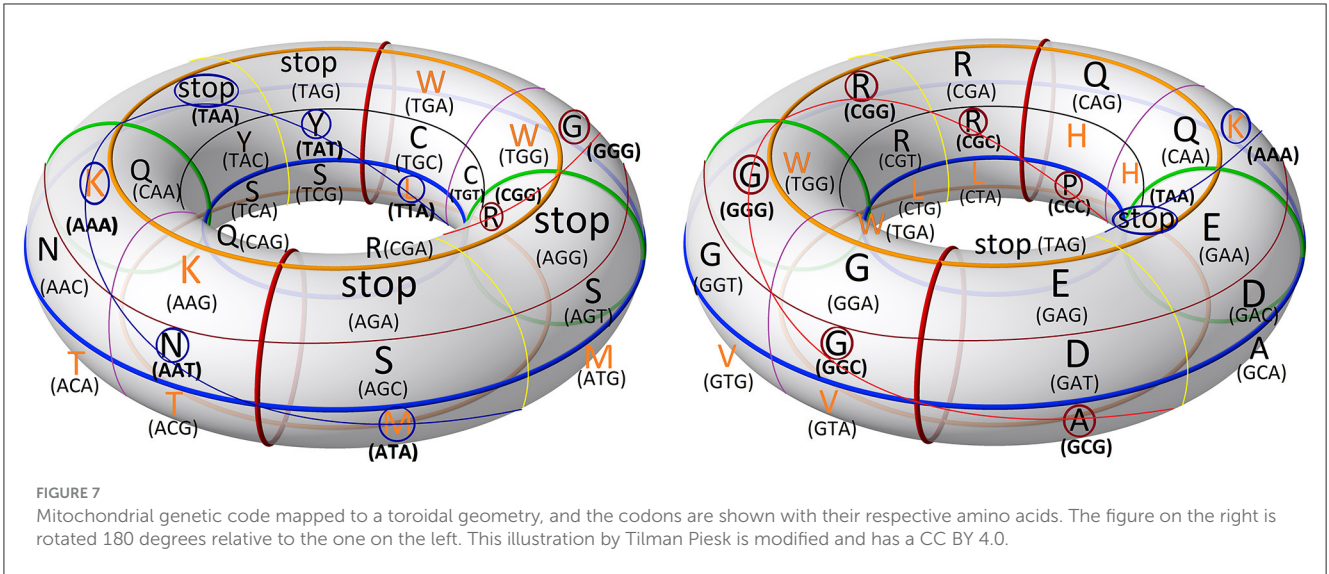

FIGURE 6
Ring torus with arrows indicating the poloidal direction (θ) and the toroidal direction (α). As an example, the positions of the GGG ($\alpha = 0, \theta = 0$), GGA ($\alpha = \frac{\pi}{4}, \theta = 0$), and GGC ($\alpha = \frac{\pi}{4}, \theta = \frac{\pi}{4}$) codons that code for the amino acid glycine (GLY) are shown.

FIGURE 7
Mitochondrial genetic code mapped to a toroidal geometry, and the codons are shown with their respective amino acids. The figure on the right is rotated 180 degrees relative to the one on the left. This illustration by Tilman Piesk is modified and has a CC BY 4.0.

FIGURE 8
Toroidal mapping in polar coordinates shows the formation of vortex in the direction of amino acids which use codons with strong hydrogen bonds (SSS), cells with yellow background, or weak (WWW), cells with light blue background.

strongly suggest that an intrinsic hierarchical order exists in the genetic code. We can represent the codon code as a self-similarity object (block matrices). This algebraic model could allow linear transformations of the genetic code into compatible geometric manifolds. This is an advantage that could be used to analyze sequences as time series.

In this study, we propose a geometric model of the genetic code that has an allometry property, which depends only on the rotation group. This can be useful to make comparative studies of DNA or amino acid sequences for phylogenetic analysis, evolutionary relationships, or codon usage. In addition, this model could help to find conserved motifs or relevant changes in proteins.

5 Conclusion

The genetic code is the cornerstone of molecular biology and indispensable to comprehend the cellular processes and their evolution. In this study, we suggest that the hydrogen bonds between paired nucleobases can be represented using Bell states due to their structural and functional group properties. In addition, we proposed a new geometric structure to represent the genetic code using a toroidal space. This model can be used as a tool to research the evolution of the genetic code, mutations in proteins, and the evolutionary relationships between organisms.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

RR-G: Conceptualization, Formal analysis, Investigation, Methodology, Writing—original draft. FH-C: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing—original draft. FA: Conceptualization, Methodology, Visualization, Writing—original draft, Writing—review & editing. JB-A: Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2024.1341158/full#supplementary-material

References

1. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*. (1953) 171:737–8. doi: 10.1038/171737a0

2. Nirenberg MW, Matthaei JH. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A*. (1961) 47:1588–602. doi: 10.1073/pnas.47.10.1588

3. Hoagland MB, Stephenson ML, Scott JF, Hecht LI, Zamecnik PC. A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem*. (1958) 231:241–57. doi: 10.1016/S0021-9258(19)77302-5

4. Rumer Y. Translation of 'Systematization of codons in the genetic code [II]' by Yu. B. Rumer (1968). *Philos Trans A Mathem Phys Eng Sci*. (2016) 374:20150447. doi: 10.1098/rsta.2015.0447

5. Lagerkvist U. Unorthodox codon reading and the evolution of the genetic code. *Cell*. (1981) 23:305–6. doi: 10.1016/0092-8674(81)90124-0

6. Wilhelm T, Friedel S. A new classification scheme of the genetic code. *J Mol Evol*. (2004) 59:598–605. doi: 10.1007/s00239-004-2650-7

7. Nemzer L. A binary representation of the genetic code. *Biosystems*. (2016) 07:155. doi: 10.1016/j.biosystems.2017.03.001

8. Petoukhov S, Petukhova E. On genetic unitary matrices and quantum-algorithmic genetics. In *Advances in Artificial Systems for Medicine and Education II 2*, Springer International Publishing (2020). p. 103–115. doi: 10.1007/978-3-030-12082-5_10

9. Ngadi T. The genetic codes: mathematical formulae and an inverse symmetry-information relationship. *Information*. (2017) 8:6. doi: 10.3390/info8010006

10. Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, et al. Long-range correlations in nucleotide sequences. *Nature*. (1992) 356:168–70. doi: 10.1038/356168a0

11. Stanley HE, Buldyrev SV, Goldberger AL, Havlin S, Peng C -K, Sciortino F, et al. Scaling concepts and complex fluids : long-range power-law correlations in DNA. *J Phys IV France*. (1993) 3:C1-15–C1-25. doi: 10.1051/jp4:1993102

12. Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL. Mosaic organization of DNA nucleotides. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*. (1994) 49:1685–9. doi: 10.1103/PhysRevE.49.1685

13. Corona-Ruiz M, Hernandez-Cabrera F, Cantú-González JR, González-Amezcua O, Javier Almaguer F. A stochastic phylogenetic algorithm for mitochondrial DNA analysis. *Front Genet*. (2019) 10:66. doi: 10.3389/fgene.2019.00066

14. Hernandez-Cabrera F, Rodríguez-Gutiérrez R, Javier Almaguer F, Grimaldo-Reyna ME. Exclusion principle between the physicochemical properties of complementary nucleobases and symmetry breaking in double-stranded DNA conformations. *Phys A*. (2024) 634:129477. doi: 10.1016/j.physa.2023.129477

15. Klump HH, Völker J, Breslauer KJ. Energy mapping of the genetic code and genomic domains: implications for code evolution and molecular Darwinism. *Q Rev Biophys*. (2020) 53:e11. doi: 10.1017/S0033583520000098

16. Taroni A. 90 years of the Ising model. *Nat Phys*. (2015) 11:997–997. doi: 10.1038/nphys3595

17. Liao B, Wang TM. New 2D graphical representation of DNA sequences. *J Comput Chem*. (2004) 25:1364–8. doi: 10.1002/jcc.20060

18. Liu XQ Dai Q, Xiu Z, Wang T. PNN-curve: a new 2D graphical representation of DNA sequences and its application. *J Theor Biol*. (2006) 243:555–61. doi: 10.1016/j.jtbi.2006.07.018

19. Xie G, Mo Z. Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. *J Theor Biol*. (2011) 269:123–30. doi: 10.1016/j.jtbi.2010.10.018

20. Zhang Y, Liao B, Ding K. On 3DD-curves of DNA sequences. *Mol Simul*. (2006) 32:29–34. doi: 10.1080/08927020500517223

21. Lambert N, Chen YN, Cheng YC, Li CM, Chen GY, Nori F. Quantum biology. *Nat Phys*. (2013) 9:10–8. doi: 10.1038/nphys2474