MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **85** (2021) 545-553

ISSN 0340 - 6253

A Lie Algebra Approach on Maximal Self Complementary C^3-Codes

F. Fayazi[1,*], A. Gholami[1] , A. R. Ashrafi[2]

[1]*Department of Mathematics, Faculty of Science, University of Qom, Qom, I. R. Iran*
fayazifariba@yahoo.com

[2]*Department of Pure Mathematics, Faculty of Mathematical Sciences, University of Kashan, Kashan 87317-53153, I. R. Iran*

(Received May 22, 2020)

Abstract

In this paper, we will obtain two equal-size equivalence classes for the set of all codons under the action of the group L which is corresponding to the maximal self-complementary C^3-codes. An $n-$dimensional vector space on the field $GF(2^3)$ will be introduced based on its Boolean lattice structure and then obtain its associated Lie algebra. The commutator between codons will recognize collinear codons and codon graph is defined for all codons of the desired set that are collinear and it will illustrate the relation between codons using Hamming distances. This establishes a relationship between the codons of the defined genetic code and the algebraic structure.

1 Basic definitions

Throughout this paper the DNA bases will be denoted by A, G, C and T and in RNA T will be changed by U. The set of all DNA bases is denoted by Ω and the notations Ω^*, Ω^+ and Ω^n, $n \geq 1$, stand for the set of all words, non-empty words and words of length n on Ω, respectively. A sequence of length three in Ω is called a *codon*. It is well-known that each codon can be converted to one of twenty amino acids or one terminated signal.

It is well-known that for each prime power p^n there exists a unique finite field of this order that will be denoted by $GF(p^n)$, the Galois field of order p^n. A non-empty set B

*Corresponding author

equipped with two binary operations $\wedge$ and $\vee$, a unary operation $\neg$ and two special elements 0 and 1 is called a Boolean algebra if $(B, \wedge, \vee, \neg, 0, 1)$ is a distributive complemented lattice with minimum element 0 and maximum element 1.

The pairs (G, C) and (A, T) of bases in Ω are called the *complementary bases* of DNA. This notations help us to define a Boolean algebra on Ω and extends this structure to Ω^n, where $n \geq 1$ is a positive integer. Sanchez et al. [14] introduced two Boolean algebra structures on the set of all codons. These Boolean algebra structures are dual and since they have the same number of elements, they are isomorphic. Furthermore, The Boolean algebras constructed from the Galois field $GF(64)$ and the authors calculated the Hamming distance between codons which reflects the different hydrophobicities between their respective coded amino acids. We refer the interested readers to consults other interesting papers of Sanchez and his co-authors on the genetic code Boolean lattice [12], DNA sequences vector space on a genetic code Galois field [13] and the symmetry group of the genetic-code cubes [15].

We recall that Ω^3 is the set of all codons in which $\Omega = \{C, T, A, G\}$. Define $\phi : \Omega^3 \longrightarrow \Omega^3$ by $\phi(rst) = str$. Following [3,4], four permutations I, c, p and r on Ω can be defined by $I = ()$, $c = (A, T)(C, G)$, $p = (A, G)(C, T)$ and $r = (A, C)(G, T)$. The mappings I, c, p and r are called, the identity, the nucleotide complementary, the pyrimidine/purine and the keto/amino mappings, respectively.

Suppose $\Omega^\star$ and Ω^+ denote the set of all words and non-empty words on Ω, respectively. A subset S of Ω^+ is called a code if for $x_1, \ldots, x_n, y_1, \ldots, y_m \in S$, $n, m \geq 1$, the condition $x_1 \ldots x_n = y_1 \ldots y_m$ implies that $n = m$ and $x_i = y_i$, $1 \leq i \leq n$. A *trinucleotide code* is a non-empty subset X of Ω^3. It is called *self-complementary* if for each trinucleotide t from X, $c(t) \in X$. A trinucleotide code $X \subseteq \Omega^3$ is said to be *circular*, if for $x_1, \ldots, x_n, y_1, \ldots, y_m \in X$, $n, m \geq 1$, $r \in \Omega^*$ and $s \in \Omega^+$, the conditions $sx_2 \ldots x_n r = y_1 \ldots y_m$ and $x_1 = rs$ imply that $n = m$, $x_i = y_i$, $1 \leq i \leq n$, and $r = \varepsilon$, where ε denotes the empty word, see [10, Definitions 3, 4 and 5] for more details. A trinucleotide code X is called C^3 *self-complementary*, if $X, \phi(X)$ and $\phi^2(X)$ are circular, $c(X) = X$, $c(\phi(X)) = \phi^2(X)$ and $c(\phi^2(X)) = \phi(X)$. A circular trinucleotide code X is *maximal*, if for each $x \in \Omega^3 \setminus X$, $X \cup \{x\}$ is not a circular trinucleotide code. The codons *abc*, *bca* and *cab* are called *cyclic* and a *periodic* code is a code of type *xxx* in which $x \in \Omega$.

There are 60 non-periodic codons that can be partitioned into 20 parts each of which

contains cyclic codons. Arquès and Michel [1] proved that there are 216 C^3 maximal self-complementary trinucleotide code and Fimmel et al. [4] proved that these 216 codes can be partitioned into 27 equivalence classes of the same length 8 and defined the group $L = \{I, c, r, p, \pi_{AT}, \pi_{CG}, \pi_{ACTG}, \pi_{AGTC}\}$ isomorphic to the dihedral group of order 8. Here, $\pi_{AT} = (A, T)$, $\pi_{CG} = (C, G)$, $\pi_{ACTG} = (A, C, T, G)$ and $\pi_{AGTC} = (A, G, T, C)$. They also noted that a trinucleotide circular code can contain at most one element from every equivalence class and cannot contain the codons AAA, CCC, GGG and TTT. As a consequence, a trinucleotide circular code can contain at most 20 codons and there are at most 3^{20} maximal trinucleotide circular codes.

Following Arquès and Michel [1], the set Ω^3 of all codons can be partitioned into three subsets X_0, X_1 and X_2 as follows:

$$\begin{aligned}
X_0 = \ \{ \ & AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, \\
& GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}; \\
X_1 = \ \{ \ & AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG, \\
& GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG\}; \\
X_2 = \ \{ \ & AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT, \\
& CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT\}.
\end{aligned}$$

in which X_0 is a maximal and circular C^3 self-complementary trinucleotide code. Furthermore, $c(X_0) = X_0$, $c(X_1) = X_2$, $c(X_2) = X_1$, $X_1 = P(X_0)$ and $X_2 = P(X_1)$. Here, $P : \Omega^3 \longrightarrow \Omega^3$ is a one to one correspondence given by $P(XYZ) = YZX$.

The aim of this paper is to study the genetic-code architecture by an Lie algebra approach. We follow the procedure used by Sanchez et al. in [14] . First we obtain two equal-size equivalence classes for the set of all codons under the action of the group L which is corresponding to the maximal self-complementary C^3-codes. An $n-$dimensional vector space on the field $GF(2^3)$ will be introduced based on its Boolean lattice structure and then obtain its associated Lie algebra. Note that by [14], the algebraic properties of the codons in the constructed Boolean algebra are related to the hydrophobic properties of the amino acids. The commutator between codons will recognize collinear codons and codon graph is defined for all codons of the desired set that are collinear and it will illustrate the relation between codons using Hamming distances. The Hamming distances between codons are calculated. Knowing the distance between two codons is important

in identifying the physico-chemical properties of the code, the position of the base and the genetic mutation.

Throughout this paper, our notation is standard. The finite field of order $q = p^n$, p is prime, is denoted by $GF(q)$. We refer to [8, 9] for our notations in code biology and to [7] for applications of this topic in physics. Our algebraic notions and notations are taken from [2, 6, 11]. Our calculations are done with the aid of computer algebra system GAP [16].

2 Lie algebra on GF(8)

Following Sanchez and Grau [12], it is possible to construct a Boolean algebra on Ω in which A is minimum, G is maximum and C, T are complement of each other and so there are incomparable. If $Z_2 = \{0, 1\}$ then by correspondence $C \leftrightarrow 00$, $A \leftrightarrow 01$, $T \leftrightarrow 10$ and $G \leftrightarrow 11$, they introduced a geometric description of this Boolean algebra with vertices of a symmetric square. This shows that each codon $XYZ \in \Omega^3$ can be written in a unique way as $a_0a_1a_2a_3a_4a_5$ such that $a_i \in \{0, 1\}$, $0 \leq i \leq 5$. This representation of XYZ is denoted by $\xi(XYZ)$.

Suppose F is a field of prime order p and $f(x) \in F[x]$ is an irreducible polynomial of degree n in x. It is well-known that $\frac{F[x]}{\langle f(x)\rangle}$ is a field of order p^n and each member of this field can be represented uniquely by a polynomial of degree $< n$. Sanchez et al. [13] used this known result in algebra to present a one-to-one correspondence $\alpha : \Omega^3 \longrightarrow GF(64)$ given by $\alpha(XYZ) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$, where $\xi(XYZ) = a_0a_1a_2a_3a_4a_5$. By [2, Example 3.2], there are nine irreducible polynomials of degree 6 over the field $GF(2)$. One of these irreducible polynomials is $g(x) = x^6 + x + 1$ and so $\frac{Z_2[x]}{\langle g(x)\rangle}$ is a field of order 64.

It is easy to see that $x^3 + x^2 + 1$ is an irreducible polynomial on $GF(2)$. Set $M = \{b_0b_1b_2 \mid b_i \in \{0, 1\}\}$ and $\beta : M \longrightarrow GF(8)$ is given by $\beta(b_0b_1b_2) = b_0 + b_1x + b_2x^2$. Set $M = \{m_0 = 111, m_1 = 110, m_2 = 101, m_3 = 011, m_4 = 100, m_5 = 010, m_6 = 001, m_7 = 000\}$ and $G_j = \{m_im_j \mid m_i \in M, 0 \leq i \leq 7\}$, where $0 \leq j \leq 7$. It can easily be seen that $|G_j| = 8$, $0 \leq j \leq 7$, $\cap_{j=0}^{7} G_j = \emptyset$ and $\Omega^3 = \cup_{j=0}^{7} G_j$. The subsets G_j of Ω^3, $0 \leq j \leq 7$, are as follows:

$$
\begin{aligned}
G_0 &= \{111111, 110111, 101111, 011111, 100111, 010111, 001111, 000111\},\\
G_1 &= \{111110, 110110, 101110, 011110, 100110, 010110, 001110, 000110\},\\
G_2 &= \{111101, 110101, 101101, 011101, 100101, 010101, 001101, 000101\},\\
G_3 &= \{111011, 110011, 101011, 011011, 100011, 010011, 001011, 000011\},\\
G_4 &= \{111100, 110100, 101100, 011100, 100100, 010100, 001100, 000100\},\\
G_5 &= \{111010, 110010, 101010, 011010, 100010, 010010, 001010, 000010\},\\
G_6 &= \{111001, 110001, 101001, 011001, 100001, 010001, 001001, 000001\},\\
G_7 &= \{111000, 110000, 101000, 011000, 100000, 010000, 001000, 000000\}.
\end{aligned}
$$

We now define an action of the group L on Ω^3 by $XYZ^{\alpha} = (X\alpha)(Y\alpha)(Z\alpha)$, $XYZ \in \Omega^3$ and $\alpha \in L$, and extend this action to $\{G_0, G_1, G_2, G_3, G_4, G_5, G_6, G_7\}$. By some tedious calculations, one can see that $c(G_1) = G_6$, $r(G_1) = G_4$, $p(G_1) = G_3$, $c(G_7) = G_0$, $r(G_7) = G_5$ and $p(G_7) = G_2$. This proves that the action of L on Ω^3 has exactly two orbits as $[G_0] = \{G_0, G_2, G_5, G_7\}$ and $[G_1] = \{G_1, G_3, G_4, G_6\}$. Note that in the notation of Sanchez et al. [13], $AAA \in G_2$, $CCC \in G_0$, $GGG \in G_7$ and $TTT \in G_5$.

Suppose $\Psi : G_0 \longrightarrow GF(8)$ is a mapping given by $\Psi(XYZ) = \Psi(a_0a_1a_2a_3a_4a_5) = (a_0 + a_5) + (a_1 + a_4)x + (a_2 + a_3)x^2$. Then $\Psi(CCC) = \Psi(111111) = 0$, $\Psi(CAC) = \Psi(110111) = x^2$, $\Psi(TCC) = \Psi(101111) = x$, $\Psi(ACC) = \Psi(011111) = 1$, $\Psi(TAC) = \Psi(100111) = x + x^2$, $\Psi(AAC) = \Psi(010111) = 1 + x^2$, $\Psi(GCC) = \Psi(001111) = 1 + x$ and $\Psi(GAC) = \Psi(000111) = 1 + x + x^2$.

Following Sanchez et al. [13], if XYZ and $X'Y'Z'$ are two codons and $\lambda \in GF(8)$, then we define:

$$
\begin{aligned}
XYZ + X'Y'Z' &= \Psi^{-1}([\Psi(XYZ) + \Psi(X'Y'Z') \ mod\ 2]),\\
XYZ \bullet X'Y'Z' &= \Psi^{-1}([f(XYZ) \bullet \Psi(X'Y'Z') \ mod\ q(x)]),\\
\lambda \star XYZ &= \Psi^{-1}([\lambda.\Psi(XYZ) \ mod\ q(x)]).
\end{aligned}
$$

It is easy to check that by above definitions, G_0 is a vector space of dimension 3 over the field $GF(2)$. To makes G_0 into a Lie algebra, we have to define the codon commutator $[XYZ, X'Y'Z']$. To do this we use a similar method as Sanchez et al. [14]. Define $[XYZ, X'Y'Z'] = (Z \times X' + X \times Z')(Z \times Y' + Y \times Z')(Z \times Z')$. Here, the operations+ and $\times$ between DNA bases are defined by Table 2.1.

+	C	T	A	G
C	C	T	A	G
T	T	C	G	A
A	A	G	C	T
G	G	A	T	C

×	C	T	A	G
C	C	C	C	C
T	C	T	A	G
A	C	A	G	T
G	C	G	T	A

Table 2.1. The operations + and × between DNA bases.

XYZ	$X'Y'Z'$	$XYZ + X'Y'Z'$
CAC	TAC	TCC
CAC	ACC	AAC
CAC	GCC	GAC
TCC	AAC	GAC
TCC	ACC	GCC
TAC	AAC	GCC
TAC	ACC	GAC

Table 2.2. Collinear codons of G_0.

Note that $[XYZ, X'Y'Z'] = [X'Y'Z', XYZ]$ and by this definition, G_0 is a Lie algebra on $GF(8)$. It is well-known that the symmetry of commutators is related to the homologous recombination.

Two codons $(XYZ), (X'Y'Z') \in \Omega^3 \setminus (CCC)$ is called collinear if $[XYZ, X'Y'Z'] = CCC$. It is easy to check that by above definition

$$[XYZ, X'Y'Z'] = [XYZ, XYZ + X'Y'Z'] = [X'Y'Z', XYZ + X'Y'Z'] \quad (1)$$

and $[XYZ, XYZ] = CCC$.

Based on $(Eq.1)$ it is obvious, if $[XYZ, X'Y'Z'] = CCC$, then $[XYZ, XYZ + X'Y'Z'] = CCC$.

We classified collinear codons of G_0 in Table 2.2.

By adding CCC to every rows of Table 2.2 , we have seven subgroups in form of $\{CCC, XYZ, X'Y'Z', XYZ + X'Y'Z'\}$. Based on Table 2.2, we have seven subgroups H_i, for $0 \leq i \leq 6$ as follows;

$$H_0 = \{CCC, CAC, TAC, TCC\}, H_1 = \{CCC, CAC, ACC, AAC\},$$

$$H_2 = \{CCC, CAC, GCC, GAC\}, H_3 = \{CCC, TCC, AAC, GAC\},$$

$$H_4 = \{CCC, TCC, ACC, GCC\}, H_5 = \{CCC, TAC, AAC, GCC\},$$

$$H_6 = \{CCC, TAC, ACC, GAC\}.$$

Collinear codons can be considered as one-dimensional subspace and can generate a vectorial line.Therefore, each of H_i for $0 \leq i \leq 6$ is a one-dimensional vector space.

Let $J \subseteq \Omega^3$. We define a directed codon graph $\mathcal{G}(J) = (V(J), E(J))$ as follows:

$$V(J) = \{N_1, N_2, ..., N_n | N_i \in \Omega^3 \setminus (CCC), \forall 1 \leq i \leq n, N_1 < N_2 < ... < N_n\},$$

$$E(J) = \{(N_j, N_i) | [N_j, N_i] = CCC, N_i, N_j \in V(J), \forall j, i \in \{1, 2, ..., n\}, j < i\},$$

where (N_j, N_i) is an edge that connect vertices N_j to N_i.

It's obvious that $|V(J)| = n$ and $|E(J)| = \frac{n(n-1)}{2} - |Noncol(J)|$, in which $|Noncol(J)|$ is number of codons in J that are not collinear.

Amino acid type	His	Ser	Trp	Thr	Asn	Ala	Asp	
$V(G_0)$	CAC	TCC	TAC	ACC	AAC	GCC	GAC	
$E(G_0)$	(CAC, GCC), (TAC ,ACC), (TCC, AAC)							Hamming distance=3
	(CAC, TCC), (CAC, ACC), (CAC, GAC), (TCC, GAC), (TAC, AAC), (TAC, GCC), (ACC, GAC), (TCC, ACC), (AAC, GCC)							Hamming distance=2
	(CAC, AAC), (TCC, TAC), (ACC, AAC), (TAC, GAC), (AAC, GAC), (GCC, GAC),(TCC, GCC) , (ACC, GCC), (CAC, TAC)							Hamming distance=1

Table 2.3. Vertices, edges and Hamming distance of codon Graph $\mathcal{G}(G_0)$.

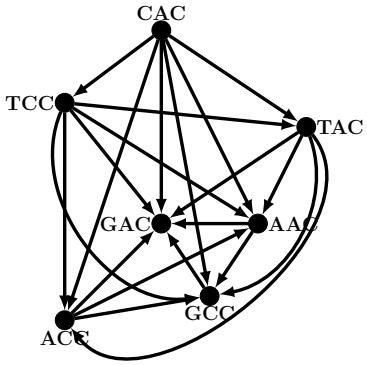

Figure 1. The codon graph $\mathcal{G}(G_0)$.

Codon Graph $\mathcal{G}(G_0)$ is consist of seven vertices and 21 edges (Figure 1). Now we define Hamming distance between edges of codon graph $\mathcal{G}(G_0)$ to illustrate the differences between codons.

Let $(XYZ), (X'Y'Z') \in \Omega^3$. The Hamming distance between two codons (XYZ) and $(X'Y'Z')$ which is indicated by $d(XYZ, X'Y'Z')$ is the number of places in which the two codon differ, i.e., have different characters and measure the differences between codons. There are three codons $\{TAA, TAG, TGA\}$ in Ω^3 that called stop codons and other 61 codons construct 20 types of standard amino acids in protein structures. seven vertices of codon graph $\mathcal{G}(G_0)$ is consist of seven different type of amino acids. In Table 2. 3, more information about vertices and edges of codon graph $\mathcal{G}(G_0)$ are given.

References

[1] D. G. Arquès, C. J. Michel, A complementary circular code in the protein coding genes, *J. Theor. Biol.* **182** (1996) 45–58.

[2] A. Das, *Computational Number Theory*, CRC Press, Boca Raton, 2013.

[3] E. Fimmel, A. Danielli, L. Strüngmann, On dichotomic classes and bijections of the genetic code, *J. Theor. Biol.* **336** (2013) 221–230.

[4] E. Fimmel, S. Giannerini, D. L. Gonzalez, L. Strüngmann, Circular codes, symmetries and transformations, *J. Math. Biol.* **70** (2015) 1623–1644.

[5] E. Fimmel, C. J. Michel, M. Starman and Strüngmann, Self–complementary circular codes in coding theory, *Theory Biosci.* **137** (2018) 51–65.

[6] S. Fomin, N. Reading, *Root Systems and Generalized Associahedra*, Am. Math. Soc., Providence, 2007.

[7] D. L. Gonzalez, S. Giannerini, R. Rosa, Strong short–range correlations and dichotomic codon classes in coding DNA sequences, *Phys. Rev. E* **78** (2008) #051918.

[8] D. L. Gonzalez, The mathematical structure of the genetic code, in: M. Barbieri (Ed.), *Codes of Life: The Rules of Microevolution*, Springer, 2008, pp. 111–152.

[9] W. Kauzmann, Some factors in the interpretation of protein denaturation, *Adv. Protein Chem.* **14** (1959) 1–63.

[10] C. J. Michel, A genetic scale of reading frame coding, *J. Theor. Bio.* **355** (2014) 83–94.

[11] J. J. Rotman, *An Introduction to the Theory of Groups*, Springer, Berlin, 1995.

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

[13] R. Sanchez, L. A. Perfetti, R. Grau, E. R. Morgado Morales, A new DNA sequences vector space on a genetic code Galois field, *MATCH Commun. Math. Compu. Chem.* **54** (2005) 3–28.

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

[15] R. Sanchez, Symmetric group of the genetic–code cubes. Effect of the genetic–code architecture on the evolutionary process, *MATCH Commun. Math. Comput. Chem.* **79** (2018) 527–560.

[16] The GAP Team, *GAP – Groups, Algorithms, and Programming*, Version 4.5.5, 2012, (`http://www.gap-system.org`).