Article

Algebraic structures and distance based analysis of genetic code

Chandra Borah, **Tazid Ali**

Department of Mathematics, Dibrugarh University, Assam 786004, India

E-mail: chandra92borah@gmail.com, tazid@dibru.ac.in

Received 26 November 2022; Accepted 18 December 2022; Published online 28 December 2022; Published 1 March 2023

Abstract

This paper explores the genetic code’s algebraic structures associated with the four mRNA (or DNA) bases A, G, C, and U. We have obtained quotient group structures of codons by considering the transition and substitution mutation. In these quotient group structures, cosets (codon members) explain intriguing interactions between the algebraic properties of codons and the physico-chemical properties (polarity, hydrophilicity, and hydrophobicity) of amino acids. Considering the evolutionary impacts of base locations in a codon, the base’s hydrogen bond number, and the base’s chemical form distinctions, we have generated a distance-based amino acids matrix. This matrix exhibits a fascinating association between distance measurements and amino acids’ physico-chemical aspects. Also, we have obtained multiple amino acid graphs relating to this distance-giving matrix, which explores the evolutionary organization of amino acids.

Keywords genetic code; amino acid; quotient group; coset; distance matrix; graph.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

The ability to preserve and transfer genetic material in terms of nucleic acids that pass one generation to the next is a primary requirement in a living body. DNA and RNA are the nucleic acids present in cells. A cell contains several thousand genes, where a gene is a fragment of DNA molecule containing the information needed for protein synthesis. The genetic code is a biochemical mechanism that sets the rules for transcription of the gene sequence into the mRNA sequence and then translated into the ordered amino acid sequence. A codon is a sequence of three DNA bases from the bases: A → Adenine, C → Cytosine, G → Guanine, and T → Thymine or U → Uracil (in RNA), which defines one amino acid out of twenty amino acids in proteins. The string of bases is not replicated accurately from the DNA chain due to mutation that influences protein formation. We are only considering the case of the transition and substitution mutations of codon in this paper.

The 64 genetic codes that result in 20 amino acids and the end signal can be viewed as a many-to-one mapping. Alanine, for example, is given by the codons GCA, GCC, GCG, and GCU. Balakrishnan (2002) noted that some mathematical structures might be present in the genetic code since the codon number is around three times the amino acids number.

The Standard Genetic Code is scientifically considered to be optimized to lessen the consequence of translational errors generated either by the insertion of incorrect amino acids or by the out of-frame stop codon encounter. It appears that the genetic code evolved to reduce the consequences of transcription and translation mistakes (Crick, 1968; Epstein, 1966; Gillis et al., 2001). Some authors assert that the genetic code is optimized and fixed (Freeland and Hurst, 1998; Freeland et al., 2000), which suggests the existence of an optimal codon order.

It has been noticed that genetic code is closely related to the physical and chemical properties of the bases, such as chemical types (pyrimidine {A, G} and purine {C, U}) and the hydrogen bonds number (A = U and G ≡ C). A genetic code has three bases, and the influence of each varies depending on its location in a codon. The second base is the most biologically significant, as the rate of codon errors drops from third base to first and then to the second in a codon (Friedman and Weinstein, 1964; Lehmann, 2000; Woese, 1965). The base U-containing amino acids in the second codon place are hydrophobic (nonpolar), whereas the base A-containing are hydrophilic (polar) (Watson and Crick, 1953). The polarities of the amino acids given by the codons with C as the second base lie in the middle, between the last two classes, whereas those with G as the second base position do not follow any regularity.

Hornos and Hornos (1993) first developed group theoretical methods to study genetic code, demonstrating the genetic code degeneracy by breaking up symmetry. Many researchers such as Bashford et al. (1998), Lehmann et al. (2000), Jimenez Montano et al. (1999), Schuster et al. (1994), Sanchez et al. (2004, 2005a, 2005b, 2005c), Jose et al. (2012), and Sanchez (2014, 2018) aimed to present systematic genetic code characterization algebraically. Their research focuses on the quantitative affinity between codons expressed via hydrogen bonds and chemical classes of bases and suggests the hydrogen bond number and chemical type should be enough to obtain a “natural order” in the 64 genetic code set. Sanchez et al. (2004, 2005a) recently presented a Boolean structure of the genetic code, where the partial order of the codon set and Boolean deductions between codons are associated with the amino acid physicochemical aspects.

Considering two primary factors associated with the codon-anticodon interplays, the chemical type of bases, and hydrogen bonds number, Sanchez et al. (2005c) obtained an array of the 64 genetic codes. They introduced a sum operation in this codon array to get one-by-one all the codons starting from AAC. The consequent codon set group $(C_g, +)$ is isomorphic to the integer modulo 64 group $(Z_{64}, +)$. They notice that the genetic code Abelian groups render algebraic symmetry in the genetic code table by associating the hydrophobic properties of coded amino acids to the algebraic properties of the corresponding codons. Ali et al. (2016) studied the transition/transversion mutation of codons with algebraic structures and found fascinating relationships between the distance matrix and physico-chemical properties of amino acids.

Over the years, numerous researchers have strived to explore different genetic code enigmas: why there is codon degeneracy, finding the most significant base location in a codon, the codon-anticodon interaction, the H-bonding count versus amino acid physicochemical aspects, and so on (Beland and Allen, 1994; Freeland and Hurst, 1998; Bashford and Jarvis, 2000).

In recent years, network analysis has emerged as one of the most significant fields of study in many disciplines, including biological systems, to comprehend complex networks of interrelated entities. Numerous studies have been conducted over the years in the biological networks field to obtain a detailed description of the genetic code (Bertman and Jungck, 1979; Jiao et al., 2007; Ali et al., 2016; Bora et al., 2020; Yan et al., 2020; Ali and Borah., 2021).

In this paper, we use algebraic structures including groups, subgroups, quotient groups, cosets, and so on to show the quantitative connections among codons. The primary aim of this paper is to obtain different quotient group structures by considering transition and substitution mutations of codons and then observe the intriguing relationship between genetic codes algebraic structures and amino acid’s physicochemical characters. The next goal is to generate an amino acid distance matrix that explores the evolutionary trend of amino acids employing network structures.

2 Algebraic Structures of Genetic Code

Sanchez et al. (2005c) investigated the RNA (or DNA) base order as a consequence of the base's chemical type (purine and pyrimidine) and hydrogen bond numbers. The base set is ordered as B = {A, C, G, U}, and on this set, an addition operation is defined as in Table 1. The obtained base set (B, +) is isomorphic to integer modulo 4 group $(Z_4, +)$. We are employing the sum operation table for set B = {A, C, G, U}, as indicated in Table 1, described earlier by Sanchez et al. (2005c).

Table 1 Sum operation on the set *B*.

	+	**A**	**C**	**G**	**U**
SUM	A	A	C	G	U
	C	C	G	U	A
	G	G	U	A	C
	U	U	A	C	G

We consider the cartesian product of group set B and organize all the 64 genetic codes in the following way. i.e., $B \times B \times B$ and name it as C_G,

$$B \times B \times B = \{(X_1, X_2, X_3): X_1, X_2, X_3 \in \{\text{A, C, G, U}\}\}$$

$$\text{i.e., } C_G = \{(X_1X_2X_3): X_1, X_2, X_3 \in \{\text{A, C, G, U}\}\}$$

with the sum operation between the codons as

$$X_1X_2X_3 + Y_1Y_2Y_3 = (X_1 + Y_1)(X_2 + Y_2)(X_3 + Y_3),$$

C_G possesses group structure and is isomorphic to $Z_4 \times Z_4 \times Z_4$.

Table 2 Genetic code table, $C_G \approx Z_4 \times Z_4 \times Z_4$.

	A			**C**			**G**			**U**			
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	
A	000	AAA	K	010	ACA	T	020	AGA	R	030	AUA	I	A
	001	AAC	N	011	ACC	T	021	AGC	S	031	AUC	I	C
	002	AAG	K	012	ACG	T	022	AGG	R	032	AUG	M	G
	003	AAU	N	013	ACU	T	023	AGU	S	033	AUU	I	U
C	100	CAA	Q	110	CCA	P	120	CGA	R	130	CUA	L	A
	101	CAC	H	111	CCC	P	121	CGC	R	131	CUC	L	C
	102	CAG	Q	112	CCG	P	122	CGG	R	132	CUG	L	G
	103	CAU	H	113	CCU	P	123	CGU	R	133	CUU	L	U
G	200	GAA	E	210	GCA	A	220	GGA	G	230	GUA	V	A
	201	GAC	D	211	GCC	A	221	GGC	G	231	GUC	V	C
	202	GAG	E	212	GCG	A	222	GGG	G	232	GUG	V	G
	203	GAU	D	213	GCU	A	223	GGU	G	233	GUU	V	U
U	300	UAA	-	310	UCA	S	320	UGA	-	330	UUA	L	A
	301	UAC	Y	311	UCC	S	321	UGC	C	331	UUC	F	C
	302	UAG	-	312	UCG	S	322	UGG	W	332	UUG	L	G
	303	UAU	Y	313	UCU	S	323	UGU	C	333	UUU	F	U

Here, we are taking the genetic code group structure $C_G \approx Z_4 \times Z_4 \times Z_4$. This group structureis also one of the 24 algebraic representations of the genetic-code cube reported in Jose et al. (2012) with the base order ACGU (it also denotes cube ACGU in (Sanchez, 2018)). José et al. (2012) explored the 24 four algebraic representations of the genetic code, as well as the rules for transforming one genetic code cube into any other.

In Table 2, we have displayed genetic code group structure $C_G \approx Z_4 \times Z_4 \times Z_4$.

An effort has been made to obtain genetic code algebraic structures exhibiting fascinating biological properties by considering the transition and substitution mutation of the codon *AAA* at various base positions.

2.1 Transition mutation

Transition mutations are those mutations in which purines are interchanged by purine $(A \leftrightarrow G)$ or pyrimidines are interchanged by pyrimidine $(C \leftrightarrow U)$. We consider the transition mutation of the codon *AAA* at different base positions.

First, we consider the transition mutation of codon *AAA* at single base positions. We have obtained the following sets:

$$T_1 = \{AAA, GAA\},\ T_2 = \{AAA, AGA\}, T_3 = \{AAA, AAG\}.$$

If we consider the transition mutation of the codon AAA at double base positions and triple base positions at a time, we have obtained the following sets:

$$T_4 = \{AAA, GGA\}, T_5 = \{AAA, GAG\}, T_6 = \{AAA, AGG\}, T_7 = \{AAA, GGG\}.$$

Again, considering the transition mutation for the codon AAA at first or second, first or third, second or third and first or second or third base positions, we get the following sets:

$$T_8 = \{AAA, AGA, GAA, GGA\}, T_9 = \{AAA, GAA, GAG, AAG\}, T_{10} = \{AAA, AAG, AGA, AGG\},$$
$$T_{11} = \{AAA, GAA, AGA, AAG, GGA, AGG, GAG, GGG\}.$$

Here, $T_1, T_2, T_3, \cdots\cdots\cdots, T_{11}$ are all subgroups of C_G. Since C_G is an Abelian group, so all the subgroups of C_G will be normal. So, we can consider the quotient groups of C_G corresponding to these normal subgroups.

2.1.1 For the subgroup, $T_1 = \{AAA, GAA\}$

We have, ${}^{C_G}/_{T_1} = \{\{AAA, GAA\}, \{AAC, GAC\}, \{AAG, GAG\}, \{AAU, GAU\}, \{CAA, UAA\}, \cdots\cdots\cdots$

$$\cdots, \{CUG, UUG\}, \{CUU, UUU\}\}$$

The quotient group ${}^{C_G}/_{T_1}$ has 32 cosets (members of codons) and each coset contains 2 codons.

For every amino acid, its corresponding synonymous codons occur in different cosets except, {CUA, UUA} and {CUG, UUG}. We have 4 cosets: {AAA, GAA}, {AAC, GAC}, {AAG, GAG}, {AAU, GAU} which give hydrophilic and polar amino acids. The cosets {CAC, UAC} and {CAU, UAU} give neutral and polar amino acids. The cosets {AUA, GUA}, {AUC, GUC}, {AUG, GUG}, {AUU, GUU}, {CUA, UUA}, {CUC, UUC}, {CUG, UUG}, {CUU, UUU} give hydrophobic and nonpolar amino acids.

So, we note that for about half of the cases, cosets contain codons encoding amino acids that do not change its polarity as well as its hydrophobicity/hydrophilicity.

2.1.2 For the subgroup, $T_2 = \{AAA, AGA\}$

We have ${}^{C_G}/_{T_2} = \{\{AAA, AGA\}, \{AAC, AGC\}, \{AAG, AGG\}, \{AAU, AGU\}, \{CAA, CGA\}, \cdots\cdots\cdots$

$$\cdots, \{UCG, UUG\}, \{UCU, UUU\}\}$$

The quotient group ${}^{C_G}/_{T_2}$ has 32 cosets and each coset contains 2 codons. Each coset consists of either a *XAY* type codon encoding a hydrophilic amino acid or *XUY* type codon encoding a hydrophobic amino acid, except {ACA, AUA}. For every amino acid, its corresponding coded synonymous codons occur in different cosets. We have 8 cosets: {AAA, AGA}, {AAC, AGC}, ⋯⋯⋯,{CAU, CGU} which give only polar amino acids and 8 cosets: {CCA, CUA}, {CCC, CUC},⋯⋯⋯,{GCU, GUU} which give only non-polar amino acids.

2.1.3 For the subgroup, $\mathrm{T}_3 = \{\mathrm{AAA}, \mathrm{AAG}\}$

We have $^{C_G}/_{T_3} = \{\{AAA, AAG\}, \{AAC, AAU\}, \{CAA, CAG\}, \{CAC, CAU\}, \{GAA, GAG\}, \cdots\cdots\cdots$

$$\cdots, \{UUA, UUG\}, \{UUC, UUU\}\}$$

The quotient group $^{C_G}/_{T_3}$ has 32 cosets and each coset contains 2 codons. Here, every coset except {UAA, UAG}, {UGA, UGG} and {AUA, AUG} contains synonymous codons (codons encoded for the same amino acid).

Similarly, we get the quotient groups structures $^{C_G}/_{T_4}, {}^{C_G}/_{T_5}, {}^{C_G}/_{T_6}, {}^{C_G}/_{T_7}$ and then observe the cosets describing the connections between algebraic properties of codons and amino acid physico-chemical characters.

$$^{C_G}/_{T_4} = \{\{AAA, GGA\}, \{AAC, GGC\}, \{AAG, GGG\}, \{AAU, GGU\}, \{CAA, UGA\}, \cdots\cdots$$

$$\cdots, \{UCG, CUG\}, \{UCU, CUU\}\}$$

$$^{C_G}/_{T_5} = \{\{AAA, GAG\}, \{AAC, GAU\}, \{AAG, GAA\}, \{AAU, GAC\}, \{CAA, UAG\}, \cdots\cdots$$

$$\cdots, \{CUG, UUA\}, \{CUU, UUC\}\}$$

$$^{C_G}/_{T_6} = \{\{AAA, AGG\}, \{AAC, AGU\}, \{AAG, AGA\}, \{AAU, AGC\}, \{CAA, CGG\}, \cdots\cdots\cdots$$

$$\cdots, \{UCG, UUA\}, \{UCU, UUG\}\}$$

$$^{C_G}/_{T_7} = \{\{AAA, GGG\}, \{AAC, GGU\}, \{AAG, GGA\}, \{AAU, GGC\}, \{CAA, UGG\}, \cdots\cdots$$

$$\cdots, \{UCG, CUA\}, \{UCU, CUC\}\}$$

Next, we take the subgroup, $T_8 = \{AAA, AGA, GAA, GGA\}$ comprises of four codons.

2.1.4 For the subgroup, $\mathrm{T}_8 = \{\mathrm{AAA}, \mathrm{AGA}, \mathrm{GAA}, \mathrm{GGA}\}$

We have $^{C_G}/_{T_8} = \{\{AAA, GAA, AGA, GGA\}, \{AAC, GAC, AGC, GGC\}, \{AAG, GAG, AGG, GGG\},$

$\{AAU, GAU, AGU, GGU\}, \cdots\cdots\cdots\cdots\cdots\cdots, \{CCG, UCG, CUG, UUG\}, \{CCU, UCU, CUU, UUU\}\}$

The quotient group $^{C_G}/_{T_8}$ has 16 cosets and each coset contains 4 codons. Every coset contains at least one codon encoded for polar amino acid and one codon encoded for non-polar amino acid, except {CAA, UAA, CGA, UGA}.

So, we note that for about half of the cases, cosets contain codons encoding amino acids that do not change its polarity as well as its hydrophobicity/hydrophilicity.

Similarly, we have the quotient groups structures $^{C_G}/_{T_9}$ and $^{C_G}/_{T_{10}}$.

$$^{C_G}/_{T_9} = \{\{AAA, GAA, AAG, GAG\}, \{AAC, GAC, AAU, GAU\}, \{CAA, UAA, CAG, UAG\},$$

$$\{CAC, UAC, CAU, UAU\}\cdots\cdots\cdots\cdots, \{CUA, UUA, CUG, UUG\}, \{CUC, UUC, CUU, UUU\}\}$$

$$^{C_G}/_{T_{10}} = \{\{AAA, AGA, AAG, AGG\}, \{AAC, AGC, AAU, AGU\}, \{CAA, CGA, CAG, CGG\},$$

$$\{CAC, CGC, CAU, CGU\}\cdots\cdots\cdots\cdots, \{UCA, UUA, UCG, UUG\}, \{UCC, UUC, UCU, UUU\}\}$$

2.1.5 For the subgroup, $\mathrm{T}_{11} = \{\mathrm{AAA}, \mathrm{GAA}, \mathrm{AGA}, \mathrm{AAG}, \mathrm{GGA}, \mathrm{AGG}, \mathrm{GAG}, \mathrm{GGG}\}$

We have $^{C_G}/_{T_{11}} =$

$$\{\{AAA, AAG, GAA, AGA, GAG, AGG, GGA, GGG\}, \{AAC, AAU, GAC, AGC, GAU, AGU, GGC, GGU\},$$

$$\{CAA, CAG, UAA, CGA, UAG, CGG, UGA, UGG\}, \{CAC, CAU, UAC, CGC, UAU, CGU, UGC, UGU\},$$
$$\{ACA, ACG, GCA, AUA, GCG, AUG, GUA, GUG\}, \{ACC, ACU, GCC, AUC, GCU, AUU, GUC, GUU\},$$
$$\{CCA, CCG, UCA, CUA, CUG, UUA, CUG, UUG\}, \{CCC, CCU, UCC, CUC, UCU, CUU, UUC, UUU\}\}.$$

The quotient group ${}^{C_G}/_{T_{11}}$ has 8 cosets and each coset contains 8 codons. The coset $\{AAA, AAG, GAA, AGA, GAG, AGG, GGA, GGG\}$ contains all the stop codons and if we consider replacing the Watson-Crick base pairs $(A \leftrightarrow U, G \leftrightarrow C)$ of codons, we shall obtain the coset $\{CCC, CCU, UCC, CUC, UCU, CUU, UUC, UUU\}$ (which are the anti-codons).

Discussion and observations:

- In all cases, we have obtained algebraic structures giving different cosets. For each coset, the extreme physicochemical properties of the amino acids given by the corresponding codons are not observed i.e., the codons giving most hydrophilic and most hydrophobic amino acids do not belong to the same coset.
- In some cases, we have obtained cosets of synonymous codons (i.e., giving the same amino acid). Considering all the 11 cases of quotient group structures, we have obtained a total of 280 cosets.

  i.e., 32+32+32+32+32+32+32+16+16+16+8=280.

  Out of these, in case of 74 cosets we have observed synonymous codons. For example, the coset {CCG, UCG, CUG, UUG} in ${}^{C_G}/_{T_8}$ contains codons CUG and UUG which encode for the amino acid Leucine (L).
- In all the cases, we have obtained cosets encoded for amino acids without altering polarity and hydrophilicity/hydrophobicity/neutrality. Out of 280, in the case of 131 cosets we have observed codons encode for the amino acids without changing polarity. For example, the coset {AAA, AGA, AAG, AGG} (in ${}^{C_G}/_{T_{10}}$) give polar amino acids Lysine (coded by AAA, AAG) and Arginine (coded by AGA, AGG).
- We have obtained quotient group structures for all instances, dividing the set of 64 codons into disjoint codon cosets. For each quotient group structures, the codon coset has its corresponding anticodon coset. For example, in case of ${}^{C_G}/_{T_1}$, the coset {AAA, GAA} has its corresponding anticodon coset {UUU, CUU} (considering Watson-Crick base pairs $(A \leftrightarrow U, G \leftrightarrow C)$).

2.2 Substitution mutation

A substitution mutation swaps one base for another. Under substitution mutation, the first base A in the codon ACG can be changed to either of the bases C, G and U. We consider the substitution mutation of the codon AAA at different base positions.

First, we consider the substitution mutation of codon AAA at single base positions. We have obtained the following sets:

$$S_1 = \{AAA, CAA, GAA, UAA\},\ S_2 = \{AAA, ACA, AGA, AUA\}, S_3 = \{AAA, AAC, AAG, AAU\},$$

If we consider the substitution mutation of the codon AAA at double base positions and triple base positions at a time, we have obtained the following sets:

$$S_4 = \{AAA, CCA, GGA, UUA\}, \qquad S_5 = \{AAA, CAC, GAG, UAU\}, S_6 = \{AAA, ACC, AGG, AUU\},$$
$$S_7 = \{AAA, CCC, GGG, UUU\},$$

Again, considering the substitution mutation for the codon AAA at first or second, first or third, second or third and first or second or third base positions, we get the following sets:

$$S_8 = \begin{Bmatrix} AAA, CAA, GAA, UAA, ACA, CCA, GCA, UCA, \\ AGA, CGA, GGA, UGA, AUA, CUA, GUA, UUA \end{Bmatrix},$$

$$S_9 = \begin{Bmatrix} AAA, AAC, AAG, AAU, ACA, ACC, ACG, ACU, \\ AGA, AGC, AGG, AGU, AUA, AUC, AUG, AUU \end{Bmatrix},$$

$$S_{10} = \begin{Bmatrix} AAA, AAC, AAG, AAU, CAA, CAC, CAG, CAU, \\ GAA, GAC, GAG, GAU, UAA, UAC, UAG, UAU \end{Bmatrix}$$

$$S_{11} = \{AAA, AAC, AAG, AAU, CAA, CAC, CAG, CAU, GAA, \cdots\cdots\cdots, UUG, UUU\} = C_G$$

As in the case of transition mutation, the sets $S_1, S_2, S_3, \cdots\cdots\cdots, S_{11}$ are all subgroups of C_G. Since C_G is an Abelian group, so all the subgroups of C_G will be normal. So, we can consider the quotient groups of C_G for these normal subgroups.

2.2.1 For the subgroup, $\mathrm{S_1 = \{AAA, CAA, GAA, UAA\}}$

We have ${}^{C_G}/_{S_1} = \{\{AAA, CAA, GAA, UAA\}, \{AAC, CAC, GAC, UAC\}, \{AAG, CAG, GAG, UAG\},$

$\cdots\cdots\cdots, \{AUU, CUU, GUU, UUU\}\}$

The quotient group ${}^{C_G}/_{S_1}$ has 16 cosets and each coset consists of 4 codons. We observe that in the case of 6 cosets, the amino acids given by the codons do not change the polarity. For each amino acid, except for Leucine (L) and Arginine (R), the related coded synonymous codons appear in different cosets.

2.2.2 For the subgroup, $\mathrm{S_2 = \{AAA, ACA, AGA, AUA\}}$

We have ${}^{C_G}/_{S_2} = \{\{AAA, ACA, AGA, AUA\}, \{AAC, ACC, AGC, AUC\}, \{AAG, ACG, AGG, AUG\},$

$\cdots\cdots\cdots, \{UAU, UCU, UGU, UGU\}\}$

The quotient group ${}^{C_G}/_{S_2}$ has 16 cosets and each coset consists of 4 codons. For every amino acid, its corresponding coded synonymous codons appear in different cosets i.e., no two synonymous codons belong to the same cosets. Every coset contains a codon of the type XAY that give a polar amino acid (except UAA, UAG) and a codon of the type XUY that give a nonpolar amino acid.

2.2.3 For the subgroup, $\mathrm{S_3 = \{AAA, AAC, AAG, AAU\}}$

We have ${}^{C_G}/_{S_3} = \{\{AAA, AAC, AAG, AAU\}, \{ACA, ACC, ACG, ACU\}, \{AGA, AGC, AGG, AGU\},$

$\cdots\cdots\cdots, \{UUA, UUC, UUG, UUU\}\}$

The quotient group ${}^{C_G}/_{S_3}$ has 16 cosets and each coset consists of 4 codons. We have observed synonymous codons (i.e., encoding the same amino acid) for each coset. For each coset (except {UAA, UAC, UAG, UAU} and {UGA, UGC, UGG, UGU}), we have observed codons that give amino acids without altering polarity, hydrophilicity, hydrophobicity, and neutrality.

Similarly, we observe the quotient group structures: ${}^{C_G}/_{S_4}$, ${}^{C_G}/_{S_5}$, ${}^{C_G}/_{S_6}$ and ${}^{C_G}/_{S_7}$, each one consisting of 16 cosets.

2.2.4 For the subgroup, $\mathrm{S_8} = \begin{Bmatrix} \mathrm{AAA, CAA, GAA, UAA, ACA, CCA, GCA, UCA,} \\ \mathrm{AGA, CGA, GGA, UGA, AUA, CUA, GUA, UUA} \end{Bmatrix}$

We have ${}^{C_G}/_{S_8} =$

$$\begin{gathered} \{\{AAA, CAA, GAA, UAA, ACA, CCA, GCA, UCA, AGA, CGA, GGA, UGA, AUA, CUA, GUA, UUA\}, \\ \{AAC, CAC, GAC, UAC, ACC, CCC, GCC, UCC, AGC, CGC, GGC, UGC, AUC, CUC, GUC, UUC\}, \\ \{AAG, CAG, GAG, UAG, ACG, CCG, GCG, UCG, AGG, CGG, GGG, UGG, AUG, CUG, GUG, UUG\}, \\ \{AAU, CAU, GAU, UAU, ACU, CCU, GCU, UCU, AGU, CGU, GGU, UGU, AUU, CUU, GUU, UUU\}\} \end{gathered}$$

The quotient group ${}^{C_G}/_{S_8}$ is consist of 4 cosets and each coset consists of 16 codons. The whole codon set is divided into four subsets with respect to the third base. That is, in every coset the codons have the same third base position. Every amino acid coded by less than six synonymous codons are distributed in different cosets of the quotient group structure. Each coset contains codons that encode amino acids with different physico-chemical properties.

2.2.5 For the subgroup, $S_9 = \begin{Bmatrix} \text{AAA, AAC, AAG, AAU, ACA, ACC, ACG, ACU,} \\ \text{AGA, AGC, AGG, AGU, AUA, AUC, AUG, AUU} \end{Bmatrix}$

We have $^{C_G}/_{S_9} =$

$$\{\{AAA, AAC, AAG, AAU, ACA, ACC, ACG, ACU, AGA, AGC, AGG, AGU, AUA, AUC, AUG, AUU\},$$
$$\{CAA, CAC, CAG, CAU, CCA, CCC, CCG, CCU, CGA, CGC, CGG, CGU, CUA, CUC, CUG, CUU\},$$
$$\{GAA, GAC, GAG, GAU, GCA, GCC, GCG, GCU, GGA, GGC, GGG, GGU, GUA, GUC, GUG, GUU\},$$
$$\{UAA, UAC, UAG, UAU, UCA, UCC, UCG, UCU, UGA, UGC, UGG, UGU, UUA, UUC, UUG, UUU\}\}$$

The quotient group $^{C_G}/_{S_9}$ is consist of 4 cosets and each coset consists of 16 codons. The whole codon set is divided into four subsets with respect to the first base. That is, in every coset the codons have the same first base position. For every amino acid coded by less than six synonymous codons are belong to the same cosets of the quotient group structure. Each coset contains codons that encode for most distant amino acids in terms of polarity and hydrophilicity/hydrophobicity. All the stop codons belong to the same cosets.

2.2.6 For the subgroup, $S_{10} = \begin{Bmatrix} \text{AAA, AAC, AAG, AAU, CAA, CAC, CAG, CAU,} \\ \text{GAA, GAC, GAG, GAU, UAA, UAC, UAG, UAU} \end{Bmatrix}$

We have $^{C_G}/_{S_{10}} =$

$$\{\{AAA, AAC, AAG, AAU, CAA, CAC, CAG, CAU, GAA, GAC, GAG, GAU, UAA, UAC, UAG, UAU\},$$
$$\{ACA, ACC, ACG, ACU, CCA, CCC, CCG, CCU, GCA, GCC, GCG, GCU, UCA, UCC, UCG, UCU\},$$
$$\{AGA, AGC, AGG, AGU, CGA, CGC, CGG, CGU, GGA, GGC, GGG, GGU, UGA, UGC, UGG, UGU\},$$
$$\{AUA, AUC, AUG, AUU, CUA, CUC, CUG, CUU, GUA, GUC, GUG, GUU, UUA, UUC, UUG, UUU\}\}$$

The quotient group $^{C_G}/_{S_{10}}$ is consist of 4 cosets and each coset consists of 16 codons. In every coset the codons have the same second base position and they are located column wise in the genetic code table. For each amino acid (except Serine (S)), coded synonymous codons belong to the same cosets of the quotient group structure. Each coset contains codons that encode amino acids with almost identical polarity and hydrophilicity/hydrophobicity properties.

Discussions and observations:

- In all 10 instances, we have obtained quotient group structures, partitioning the set of 64 genetic codes into disjoint cosets.
- We have obtained quotient group structures giving different cosets. In certain cases, we have seen cosets encoding amino acids with extreme physicochemical properties i.e., codons giving most hydrophilic and most hydrophobic amino acids belong to the same coset. For example, the coset {AAA, ACC, AGG, AUU} in $^{C_G}/_{S_6}$ provides the most hydrophilic amino acid Arginine (coded by AGG) and the most hydrophobic amino acid Isoleucine (coded by AUU).
- In some cases, we have obtained cosets of synonymous codons (i.e. giving the same amino acid). Considering all the 10 cases of quotient group structures, we have obtained a total of 124 cosets.

  i.e., 16+16+16+16+16+16+16+4+4+4=124.

  Out of these, in case of 40 cosets we have observed synonymous codons.

  For example, the coset

  $\begin{Bmatrix} AUA, AUC, AUG, AUU, CUA, CUC, CUG, CUU, \\ GUA, GUC, GUG, GUU, UUA, UUC, UUG, UUU \end{Bmatrix}$ in $^{C_G}/_{S_{10}}$ contains synonymous codon

$UUA, UUG, CUA, CUC, CUG, CUU$ encode for the amino acid Leucine (L).

- In all the cases (except for the quotient group ${}^{C_G}/_{S_3}$), we have obtained cosets encoded for amino acids with altering polarity and hydrophilicity/hydrophobicity. Out of 124, in the case of 98 cosets we have observed codons encoding polar and nonpolar amino acids. For example, the coset {AAA, ACA, AGA, AUA} give polar amino acids Lysine (coded by AAA), Threonine (coded by ACA), Arginine (coded by AGA) and nonpolar amino acid Isoleucine (coded by AUA).

2.3 Biological significance

Here, we provide a biologically relevant explanation for our findings. For example, transitions like AAA ↔ GAA (K↔ E, T_1) and AAA ↔ AGA (K ↔ R, T_2) usually are the less probable transitions fixed in population from superior organisms, since in general they lead to conformational changes in the 3D structure of proteins (see example of diseases associated mutation in Steinmann et al. (2016) and Boer et al. (1994)). For this reason, however, they would be present in virus populations, since they would help the viruses to escape from the action from the host immune system and drug treatment (Telwatte et al., 2015).

3 Distance Based Analysis of Genetic Code

Sanchez et al. (2005c) note that four DNA bases can be organized or arranged by analyzing the codon-anticodon relationships. The physicochemical properties of bases such as chemical classes (purine, pyrimidine) and hydrogen bonds number are the principal factors, which are taken into consideration in codon-anticodon interactions to generate a sequence in the 4 DNA bases. These factors must be implemented in compliance with the following requirements.

1. Chemical types are responsible for key distinctions between bases.
2. The highest distinction from one element to the next is used as a basis for the selection of arrangements.
3. The starting base must have the least hydrogen bond number.

Accordingly, two sequences of the base set are obtained: {A, C, G, U} and {U, G, C, A}. An addition operation (Table 1) is introduced on the first base set in such a way that it is isomorphic to the cyclic group $(Z_4, +)$ (Sanchez et al., 2005c). Identifying each base with the corresponding integer in Z_4 as given by Table 1, we define the distance between any two bases X and Y as |X - Y|. For example, the distance between the bases A and G will be |A - G| = |0 - 2| = 2.

Table 3 Computing the distance between bases.

D = |X - Y|	**A**	**C**	**G**	**U**
A	0	1	2	3
C	1	0	1	2
G	2	1	0	1
U	3	2	1	0

As per evolutionary influence, the codon’s second base is the most biologically important, and the third one to be the least. A codon has three base positions, and each one has a distinctive contribution to the corresponding amino acid.

Considering (1) the evolutionary value of the base positions in the codon, (2) hydrogen bonding number in complementary bases and (3) chemical nature (purine, pyrimidine) of the base, the distance between the two codons $X_1X_2X_3$and $Y_1Y_2Y_3$ is defined in the following ways:

1. If the first bases of the two codons differ, assign a value of $2|X_1 - Y_1|$, otherwise a value of 0,
2. If the second bases of the two codons differ, assign a value of $3|X_2 - Y_2|$, otherwise a value of 0,
3. If the third bases of the two codons differ, assign a value of $1|X_3 - Y_3|$, otherwise a value of 0.

So, the distance between $X_1X_2X_3$and $Y_1Y_2Y_3$ is given by $2|X_1 - Y_1| + 3|X_2 - Y_2| + 1|X_3 - Y_3|$ and we denote it by $D_C(X_1X_2X_3, Y_1Y_2Y_3)$.

i.e., $$D_C(X_1X_2X_3, Y_1Y_2Y_3) = 2|X_1 - Y_1| + 3|X_2 - Y_2| + 1|X_3 - Y_3| \tag{1}$$

We find the distance between the codons ACC and UAC is

$$D_C(ACC, UAC) = 2|A - U| + 3|C - A| + 1|C - C| = 2|0 - 3| + 3|1 - 0| + 1|0 - 0| = 9$$

To measure the distance between any two amino acids, we compute the average distance among the respective codons. We compute the distance between the amino acids Lysine (provided by AAA, AAG) and Tyrosine (provided by UAU, UAC) in Table 4.

Table 4 The distance between the codons.

D_C	**UAU**	**UAG**
AAA	9	8
AAG	7	6

So, the distance between Lysine (K) and Tyrosine (Y) is the mean distance for the above codons, i.e., 7.50 (Table 4).

The following distance-giving matrix describes the distances among amino acids (Table 5).

We have observed that, the distinctions in physicochemical properties of amino acids rise as the distance values increase (table 5). There are great distance values between hydrophobic and hydrophilic amino acids. The distance between phenylalanine (strong hydrophobic) and Lysine (strong hydrophilic) is the highest: 16.50. A minimal distance value between the respective amino acids suggests small-scale mutations or discrepancies between amino acids. The weighted Manhattan distances we describe here are analogous to those presented in Sanchez et al. (2006) and Sanchez (2014).

Table 5 The distance matrix for amino acids pairs.

	R	K	E	Q	D	N	H	P	Y	S	T	G	W	A	M	C	F	L	V	I
R	0	8.50	9.83	7.83	10.00	8.50	8.00	4.91	11.83	6.89	5.41	3.92	5.67	6.91	5.33	6.00	9.60	6.33	6.83	5.67
K	8.50	0	5.00	3.00	5.50	1.50	3.50	6.25	7.50	9.33	4.25	11.25	13.00	8.33	10.00	13.50	16.50	13.50	14.25	10.33
E	9.83	5.00	0	3.00	1.50	5.50	3.50	6.25	3.50	8.00	8.25	7.25	9.00	4.25	14.00	9.50	12.50	12.00	10.25	14.33
Q	7.83	3.00	3.00	0	3.50	3.50	1.50	4.25	4.50	8.67	6.25	9.25	11.00	6.25	11.00	11.50	14.50	11.50	12.25	12.50
D	10.00	5.50	1.50	3.50	0	5.00	3.00	6.25	3.00	7.83	8.25	7.25	9.00	4.25	14.00	9.00	12.00	12.33	10.25	14.33
N	8.50	1.50	5.50	3.50	5.00	0	3.00	6.25	3.00	8.33	4.25	11.25	13.00	8.25	10.00	13.00	16.00	13.67	14.25	10.33
H	8.00	3.50	3.50	1.50	3.00	3.00	0	4.25	5.00	8.50	6.25	9.25	11.00	6.25	11.00	11.00	14.00	11.67	12.25	11.33
P	4.91	6.25	6.25	4.25	6.25	6.25	4.25	0	8.25	5.50	3.75	6.25	8.00	3.50	9.00	8.25	11.25	8.60	9.25	9.33
Y	11.83	7.50	3.50	4.50	3.00	3.00	5.00	8.25	0	7.17	10.25	9.25	7.00	6.25	16.00	7.00	10.00	11.67	12.25	16.33
S	6.89	9.33	8.00	8.67	7.83	8.33	8.50	5.50	7.17	0	6.17	5.83	5.33	4.83	10.33	5.17	8.17	9.11	8.83	9.20
T	5.41	4.25	8.25	6.25	8.25	4.25	6.25	3.75	10.25	6.17	0	8.25	13.00	5.25	7.00	10.25	13.25	10.08	11.25	7.33
G	3.92	11.25	7.25	9.25	7.25	11.25	9.25	6.25	9.25	5.83	8.25	0	3.00	4.25	8.00	3.25	6.25	6.25	4.25	8.25
W	5.67	13.00	9.00	11.00	9.00	13.00	11.00	8.00	7.00	5.33	13.00	3.00	0	6.50	9.00	1.00	4.00	6.67	6.00	10.00
A	6.91	8.33	4.25	6.25	4.25	8.25	6.25	3.50	6.25	4.83	5.25	4.25	6.50	0	11.00	6.25	9.25	9.25	7.25	11.25
M	5.33	10.00	14.00	11.00	14.00	14.00	11.00	9.00	16.00	10.33	7.00	8.00	9.00	11.00	0	10.00	7.00	4.33	5.50	1.33
C	6.00	13.50	9.50	11.50	9.00	13.00	11.00	8.25	7.00	5.17	10.25	3.25	1.00	6.25	10.00	0	4.00	5.67	6.25	10.33
F	9.00	16.50	12.50	14.50	12.00	16.00	14.00	11.25	10.00	8.17	13.25	6.25	4.00	9.25	7.00	4.00	0	4.00	3.25	7.33
L	6.33	13.50	12.00	11.50	12.33	13.67	11.67	8.60	11.67	9.11	10.08	6.25	6.67	9.25	4.33	5.67	4.00	0	3.25	4.00
V	6.83	14.25	10.25	12.25	10.25	14.25	12.25	9.25	12.25	8.83	11.25	4.25	6.00	7.25	5.50	6.25	3.25	3.25	0	3.33
I	5.67	10.33	12.50	12.50	14.33	10.33	11.33	9.33	16.33	9.20	7.33	8.25	10.00	11.25	1.33	10.33	7.33	4.00	3.33	0

3.1 Graphs of amino acids

We analyze the developmental tendencies of amino acids in this section by presenting a set of graphs developed from the distance matrix (Table 5). To obtain a graph structure, we consider each amino acid as a vertex, and any two amino acids a and b are linked if their distance is less or equal to some assigned value d and $d > 0$. We have obtained graphs of amino acids for different assigned values and observed interesting relationships among amino acids.

Case 1: $d = 2.00$

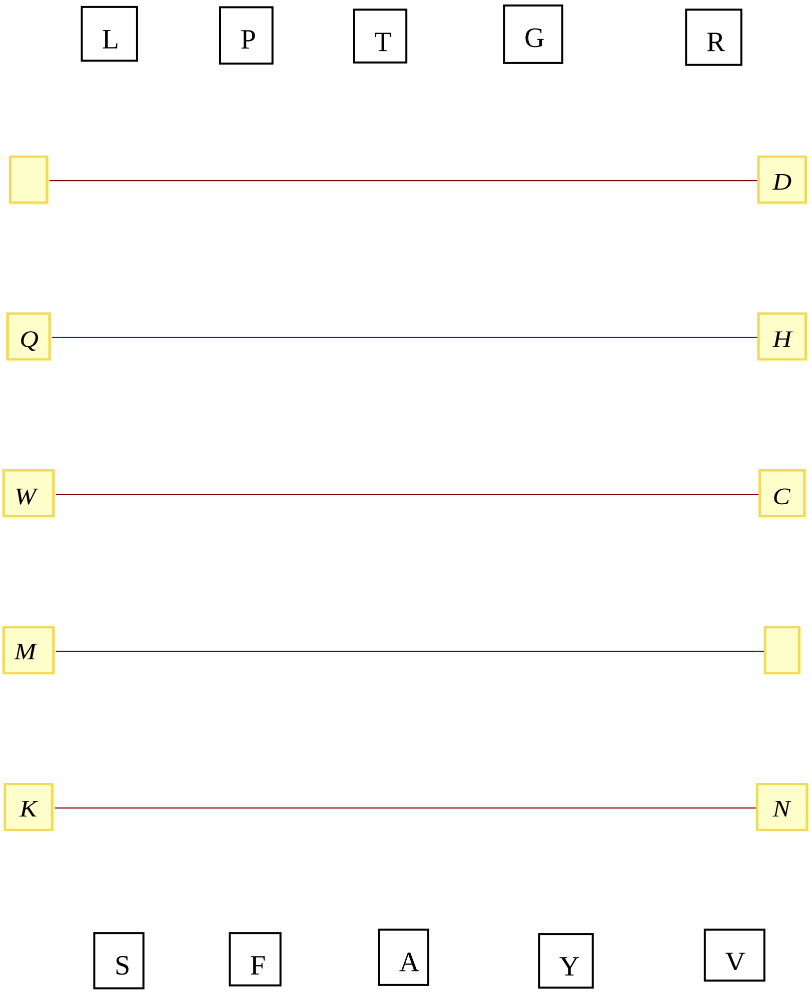

Fig. 3.1 (graph 1).

Here, we have noticed that amino acids E, Q, W, M, and K connect D, H, C, I, and N, respectively, and L, P, T, G, R, S, F, A, Y, V are isolated, as in Fig. 3.1. By a third base mutation in the corresponding codon, we can obtain one amino acid from the other for each pair of connected amino acids. The related amino acids have the same base in the first and second base locations for the respective codons, and each one has a degeneracy less or equal to 3.

Case 2: $d = 3.00$

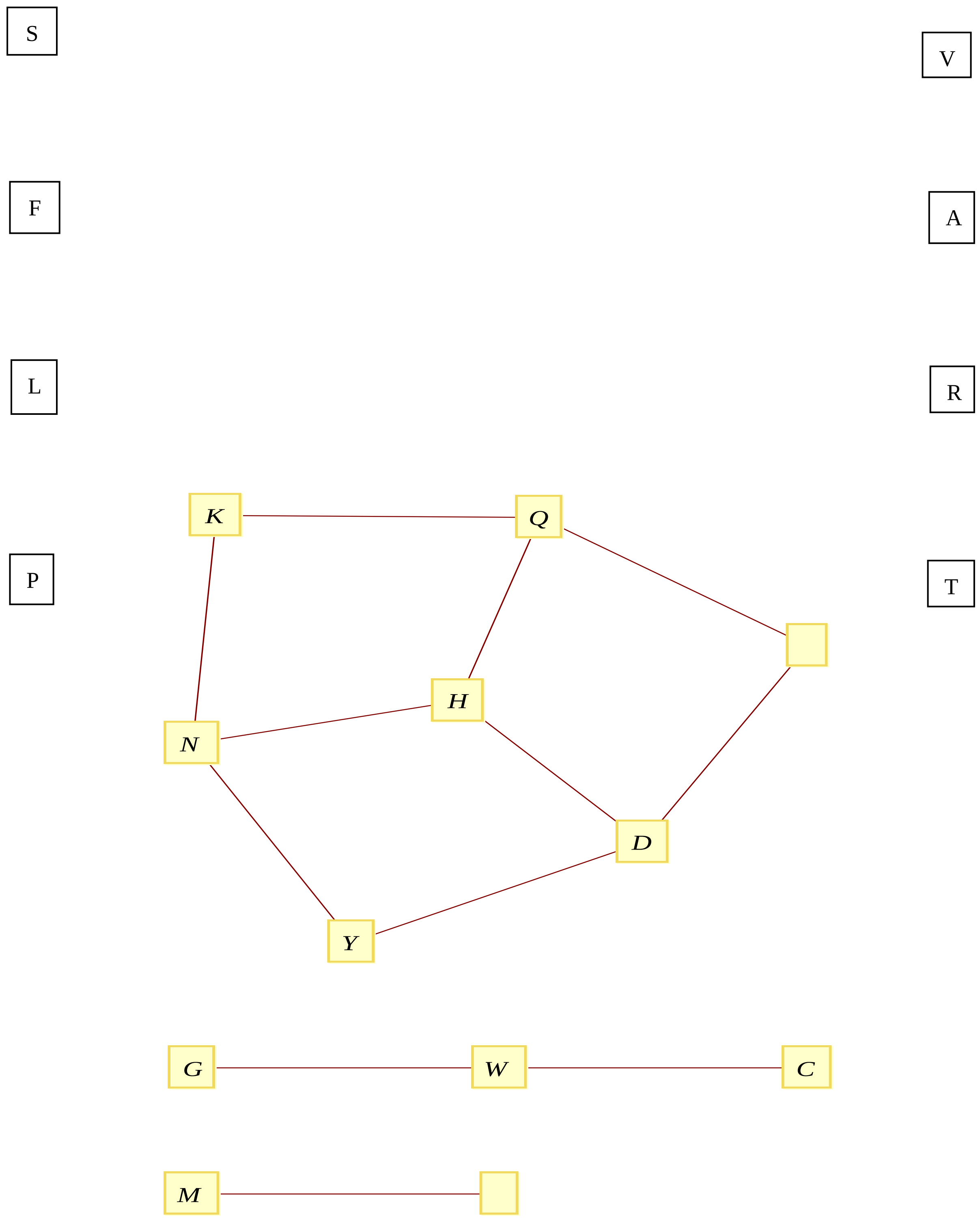

Fig. 3.2 (graph 2).

In Fig. 3.2, we have obtained a disconnected graph consisting of 11 components. The amino acids S, F, L, P, V, A, R and T are isolated and each of them contains at least four synonymous codons, except F. The polar amino acids K, N, Y, D, E, Q and H form a pentagon and each of them has exactly two synonymous codons, with A as second base. G, W and C are all nonpolar amino acids lying in a straight line. The amino acids M and I are related such that one of them giving the other through a third base mutation in the respective codons.

In Fig. 3.3, we have obtained a disconnected graph consisting of 4 components. The amino acid S (Serine) is isolated from the other 19 amino acids, and it is the only one encoded by six codons with a different second base. The rest 19 amino acids allocate to the other three components according to their encoded codons that share the same second base. The most hydrophobic amino acids I, V, L, F, C, M and the most hydrophilic H, D, Q, E, K, N, Y are in separate components. T, P, and A are in a straight line, and each of them is encoded by four synonymous codons with C as the second base.

Case 3: $d = 4.00$

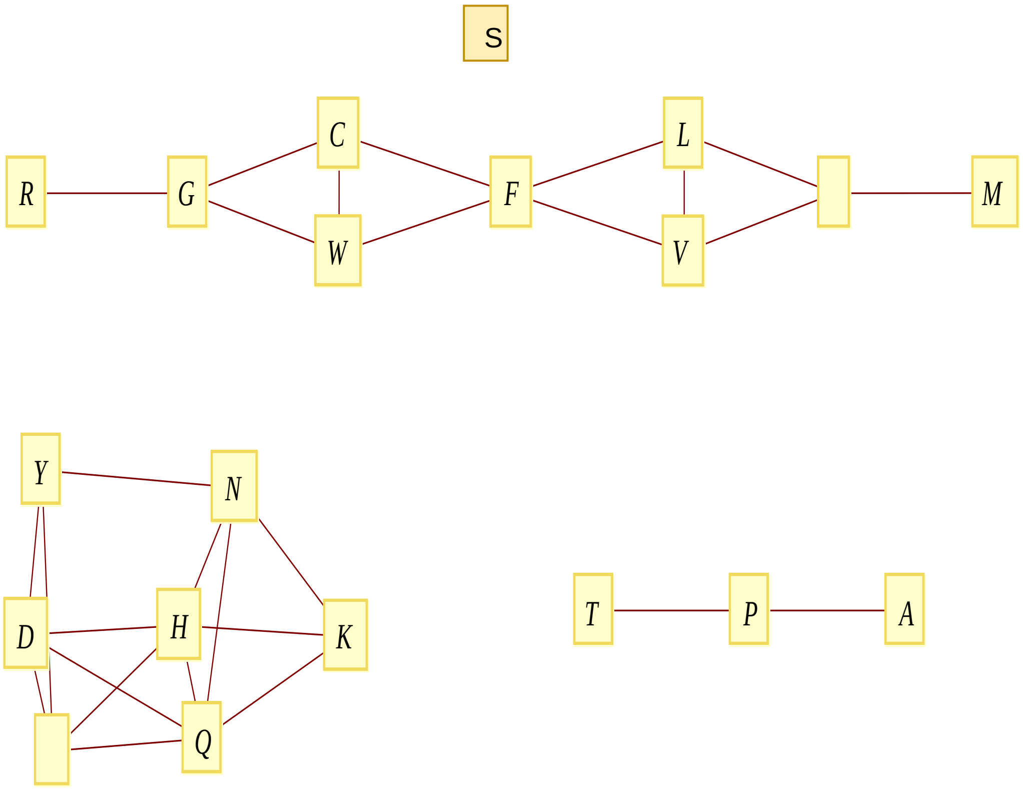

Fig. 3.3 (graph 3).

Case 4: $d = 5.00$

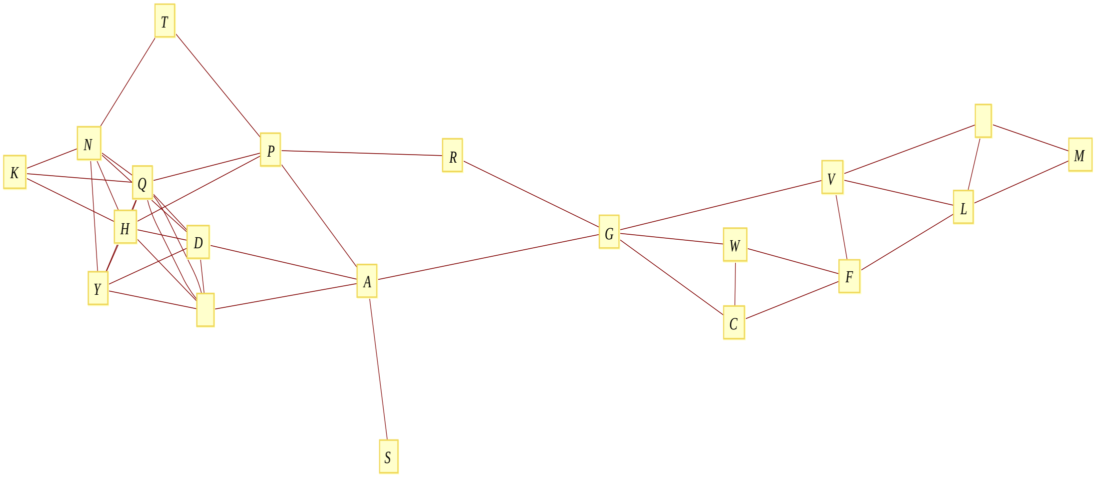

Fig. 3.4 (graph 4).

In Fig. 3.4, we have obtained a graph of 20 amino acids. Here, the most hydrophilic amino acids R, K, E, Q, D, N, H connect to the most hydrophobic I, V, L, F, C, M through the neutral amino acid G. The amino acid S differs from the rest of the graph, as the second base has a change concerning synonymous codons.

Case 5: $d = 7.50$

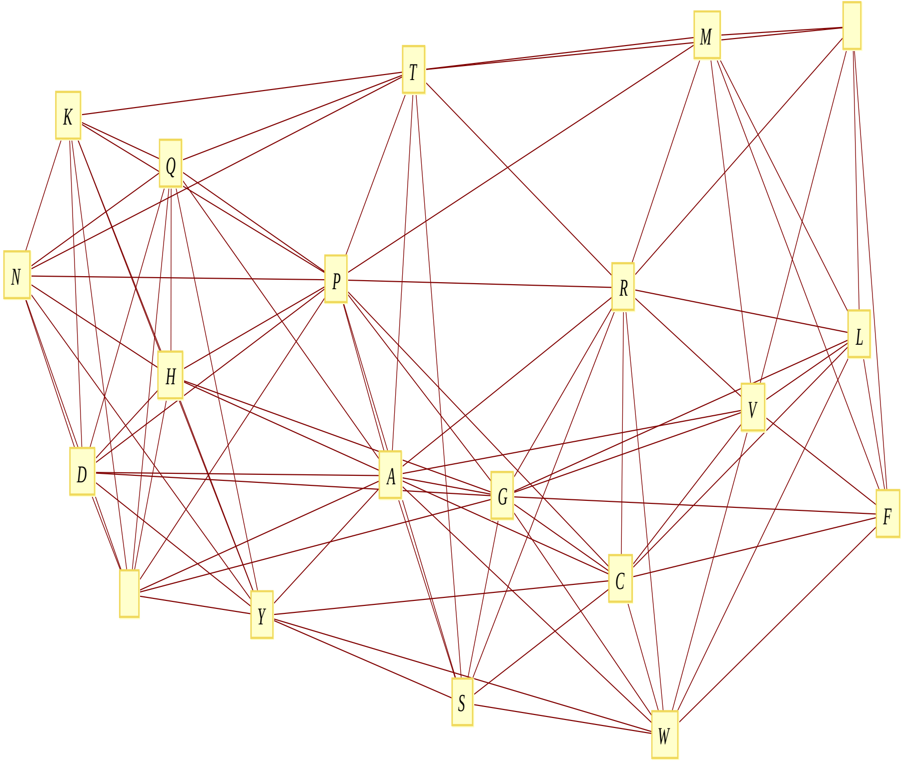

Fig. 3.5 (graph 5).

Here, we obtain a graph where all the 20 amino acids are connected. We observe that as the value of d increases, the likelihood that a mutational event transforming a codon into another encoding for a different amino acid increase. That is, the likelihood that a non-neutral or quasi-neutral mutational event would happen. It is noticed Sanchez et al. (2004), Gohain et al. (2015) attained similar results, using a different approach to obtain distance matrix tables.

3.2 Biological significance

Here, we consider a real-world scenario and observe that the distance between commonly occurring codon mutations is minimal. We operate the Hamming distance measure D_C (equation (1)) to calculate the distances between codon mutations in the HIV-1 protease gene (see Sanchez et al., 2005c, Table 8, Table 9).

In Table 6 and 7, we compute the distances between codon mutations in the HIV-1 protease gene. Concerning HXB2, the wild-type HIV, it confers drug resistance. We notice that in most cases, the distance between the wild-type codon and the mutant one has a value less or equal to 6. A small distance value usually suggests a slight variation in their biological activity. Table 7 displays the distances between codon mutations for the human beta-globin gene.

We observe that mutations V20M, V34F, and V111F retain the hydrophobic character of amino acids, whereas mutations H97Q, D99E, K82E, D99N, and H146P preserve the hydrophilicity. But all these mutations affect the oxygen affinity to hemoglobin. For the above-mentioned mutations, the distance between the wild-type codon and the mutant one has a value less or equal to 4.

It indicates that the mutational paths followed by genes throughout the molecular evolution process are likely to have a minimum distance value at each stage. We also see mutations where small changes in the physicochemical characteristics of the amino acids are enough to affect the biological activity of hemoglobin.

Table 6 The distance values of the mutations found in the HIV protease gene. It confers drug resistance with regard to the wild type of HXB2.

AMINO ACID MUTATIONS	CODON-MUTATION	D_C	ANTIVIRAL DRUG	AMINO ACID MUTATIONS	CODON-MUTATION	D_C	ANTIVIRAL DRUG
A71I	GCU→AUU	10	ABT-378	L10Y	CUC→UAC	13	BMS 2322632
A71L	GCU→CUC	10	ABT-378	L23I	CUA→AUA	2	BILA 2185 BS
A71T	GCU→ACU	4	Indinavir, Crixivan	L24I	UUA→AUA	6	Indinavir, Crixivan
A71V	GCU→GUU	6	Nelfinavir, Viracept	L24V	UUA→GUA	2	Telinavir
D30N	GAU→AAU	4	Nelfinavir, Viracept	L33F	UUA→UUC	1	ABT-378
D60E	GAU→GAA	3	DMP 450	L63P	CUC→CCC	6	ABT-378, AG1343
G16E	GGG→GAG	6	ABT-378	L90M	UUG→AUG	6	Nelfinavir, Viracept
G48V	GGG→GUG	3	Telinavir, MK-639	L97V	UUA→GUA	2	DMP-323
G52S	GGU→AGU	4	AG1343	M36I	AUG→AUA	2	Nelfinavir, Viracept
G73S	GGU→AGU	4	AG1343 MK-639	M46F	AUG→UUC	7	A-77009
H69Y	CAU→UAU	4	Aluviran, Lopinavir	M46I	AUG→AUA	2	Indinavir, Crixivan
I47V	AUA→GUA	4	ABT-378	M46L	AUG→UUC	7	Indinavir, Crixivan
I50L	AUU→CUU	2	BMS 232632	M46V	AUG→GUG	4	A-77006
I54L	AUC→CUC	2	ABT-378	N88D	AAU→GAU	4	Nelfinavir, Viracept
I54M	AUU→AUG	1	BILA 2185 BS	N88S	AAU→AGU	6	BMS 232632
I54T	AUC→ACC	6	ABT-378	P817	CCU→ACU	2	Telinavir
I54V	AUC→GUC	4	ABT-378, MK-639	R8K	CGA→AAA	8	A-77003

I82T	AUC→ACC	6	A-77003	R8Q	CGA→CAA	6	A-77004
I84A	AUA→GCA	10	BILA 1906 BS	R57K	AGA→AAA	6	AG1343
I84V	AUA→GUA	4	Nelfinavir, Viracept	T91S	ACU→UCU	6	ABT-378
K20M	AAG→AUG	9	Indinavir, Crixivan	V32I	GUA→AUA	4	A-77005, Telinavir
K20R	AAG→AGG	6	Indinavir, Crixivan	V75I	GUA→AUA	4	Telinavir
K45I	AAA→AUA	9	DMP-323	V77I	GUA→AUA	4	AG1343
K55R	AAA→AGA	6	AG1343	V82A	GUC→GCC	6	Ritonovir, Norvir
L10I	CUC→AUC	2	Indinavir, Crixivan	V82F	GUC→UUC	2	Ritonovir, Norvir
L10R	CUC→CGC	3	Indinavir, Crixivan	V82I	GUC→AUC	4	A-77011
L10V	CUC→GUC	2	Indinavir, Crixivan	V82S	GUC→UCC	8	Ritonovir, Norvir
L10F	CUC→UUC	4	Lopinavir	V82T	GUC→ACC	10	Ritonovir, Norvir

Table 7 Compute the distance measures of the mutations found in the human beta-globin gene.

AMINO ACID MUTATIONS	CODON MUTATION	D_C	BIOLOGICAL EFFECTS	REFERENCES [PMID]
P36H	CCT→CAT	3	High oxygen affinity	[11939509] Hemoglobin. 2002, 26, 21-31
T123I	ACC→ATC	6	Asymptomatic	[11300351] Hemoglobin. 2001, 25, 67-78
V20E	GTG→GAG	9	High oxygen affinity	[7914875] Eur J Haematol. 1994, 53, 21-25
V20M	GTG→ATG	4	High oxygen affinity	[7914875] Eur J Haematol. 1994, 53, 21-25
V126L	GTG→CTG	2	Neutral	[11939515] Hemoglobin. 2002, 26, 7-12
V111F	GTC→TTC	2	Low oxygen affinity	[10975442] Hemoglobin. 2000, 24, 227-237
H97Q	CAC→CAA	1	High oxygen affinity	[8571935] Am J Hematol. 1996, 51, 32-36
V34F	GTC→TTC	2	High oxygen affinity	[10846826] Int J Hematol. 2000, 71, 221-226
E121Q	GAA→CAA	2		[8095930] Hemoglobin. 1993, 17, 9-17
L114P	CTG→CCG	6	Non-functional	[11300352] Hemoglobin. 2001, 25, 79-89
A128V	GCT→GTT	6	Mild instability	[11300349] Hemoglobin. 2001, 25, 45-56
H97Q	CAC→CAG	1	High oxygen affinity	[8890707] Ann Hematol. 1996, 73, 183-188
D99E	GAT→GAA	3	High oxygen affinity	[1814856] Hemoglobin. 1991, 15, 487-496
D21N	GAT→AAT	4		[8507722] Hematol. 1993, 66, 269-272
N139Y	AAT→TAT	6	High oxygen affinity	[8718692] Hemoglobin. 1995, 19, 335-341
V34D	GTC→GAC	9	Unstable	[1260309] Hemoglobin. 2003, 27, 31-35
E121K	GAA→AAA	4		[790828] Hemoglobin. 1993, 17, 523-535
A140V	GCC→GTC	6	Mild polycythemia	[9028820] Hemoglobin. 1997, 21, 17-26
K82E	AAG→GAG	4	Altered oxygen affinity	[9255613] Hemoglobin. 1997, 21, 345-361
G83D	GGC→GAC	6	Hb Pyrgos (Normal)	[11843288] Int J Hematol. 2002, 75, 35-39
D99N	GAT→GAC	2	High oxygen affinity	[1427427] Haematologica. 1992, 77, 215-220
G15R	GGT→CGT	2	Neutral	[11939517] Hemoglobin. 2002, 26, 77-81
V111L	GTC→CTC	2	Fannin-lubbock variant	[7852084] Hemoglobin. 1994, 18, 297-306
G119D	GGC→GAC	6	Fannin-lubbock variant	[7852084] Hemoglobin. 1994, 18, 297-306
E26K	GAG→AAG	4		[9140717] Hemoglobin. 1997, 21, 205-218
N108I	AAC→ATC	9	Low Infinity	[12010673] Haematologica. 2002, 87, 553-554

H146P	CAC→CCC	3	High oxygen affinity	[11475152] Ann Hematol. 2001, 80, 365-367
H92Y	CAC→TAC	4	Cyanosis	[9494043] Hemoglobin. 1998, 22, 1-10
C112W	TGT→TGG	1	Silent and Unstable	[8936462] Hemoglobin. 1996, 20, 361-369
A111V	GCC→GTC	6	Silent	[7615398] Hemoglobin. 1995, 19, 1-6
A123S	GCC→TCC	2	Silent	[7615398] Hemoglobin. 1995, 19, 1-6
D52G	GAT→GGT	6	Silent	[9730366] Hemoglobin. 1998, 22, 355-371
V126G	GTG→GGG	3	Mild beta-thalassaemia	[1954392] Blood. 1991, 78, 3070-3075
W15STOP	TGG→TAG	6	Beta-thalassaemia	[10722110] Hemoglobin. 2000 Feb; 24(1):1-13
F42L	TTT→TTG	1	Hemolytic anemia	[11920235] Hematol J. 2001; 2(1):61-66

4 Conclusion

In this article, we observed that the cosets obtained using quotient group structures show a close relationship between the algebraic structures of the genetic code and the physicochemical aspects of amino acids. By considering the transition mutation of the codon AAA at different base positions, we have obtained 11 quotient group structures for the set of 64 codons. The property that transition mutation of codons does not cause extreme physicochemical properties changes in the amino acids, given by the respective cosets, is reflected in these quotient group structures. Again, considering the substitution mutation of the codon AAA at different base positions, we have obtained 10 quotient group structures for the set of 64 codons. The property that depending on the codon base positions, the substitution mutation of codons causes extreme physicochemical properties change to the amino acids, is reflected in the obtained quotient group structures.

Furthermore, considering the evolutionary rank of base locations plus the physicochemical properties, we have obtained a distance-based matrix incorporating the 20 amino acids. This distance-giving matrix reveals that the differentiation in the physicochemical characteristics of amino acids is associated with the distance between amino acids. The distances between the corresponding codons determine the possibilities of a mutational event changing one codon into another encoding for a different amino acid. Subsequently, we have introduced a set of graphs that shows distinctive associations between amino acids by taking some predetermined distance values. These graph structures roughly highlight the biochemical pathways of the amino acids: hydrophilic and hydrophobic affinity, as well as their polar and non-polar characteristics and the degeneracy distribution of codons.

Also, we consider a real-life example where we found that a small distance value between wild-type codon and mutant one indicates the slight difference between the biological activities in the human beta-globin gene and HIV protease gene.

Aknowledgement

I, Chandra Borah, the first author of this research work, a recipient of JRF, would like to express my sincere gratitude towards the University Grants Commission (India) for the financial assistance in doing my research. This work was supported by University Grants Commission (India). Grant Number: 1125/ (CSIR-UGC DEC. 2017).

References

Ali T, Borah C. 2021. Analysis of amino acids network based on transition and transversion mutation of codons. Network Biology, 11(3): 125-136

Ali T, Gohain N. 2016. Some Algebraic Aspects and Evolution of Genetic Code. In: Applied Analysis in Biological and Physical Sciences (Cushing JM, Saleem M, Srivastava HM, et al., eds). Springer

Ali T, Akhtar A, Gohain N. 2016. Analysis of amino acid network based on distance matrix. Physica A, 452: 69-78

Balakrishnan J. 2002. Symmetry scheme for amino acids codons. Physical Review E, 65: 021912-021915

Bashford JD, Jarvis PD. 2000. The genetic code as a periodic table. BioSystems, 57: 147-161

Bashford JD, Tsohantjis I, Jarvis PD. 1998. A supersymmetric model for the evolution of genetic code. Proceedings of the National Academy of Sciences of USA, 95: 987-992

Beland P, Allen TF. 1994. The origin and evolution of the genetic code. Journal of Theoretical Biology, 170: 359-365

Bertman MO, Jungck, JR. 1979. Group graph of the genetic code. Journal of Heredity, 70: 379-384

Bora PK, Hazarika P, Baruah AK. 2020. Distance based amino acids network analysis. Gene Reports, 21: 100933

Crick FHC. 1968. The origin of the genetic code. Journal of Molecular Biology, 38: 367-379

de Boer M, Hilarius-Stokman P, Hossle J, Verhoeven A, Graf N, Kenney R, et al. 1994. Autosomal recessive chronic granulomatous disease with absence of the 67-kD cytosolic NADPH oxidase component: identification of mutation and detection of carriers. Blood, American Society of Hematology, 83: 531-536

Epstein CJ. 1966. Role of the amino-acid “code” and of selection for conformation in the evolution of proteins. Nature, 210: 25-28

Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. Journal of Molecular Evolution, 47: 238-248

Freeland SJ, Knight RD, Landweber LF, Hurst LD. 2000. Early fixation of an optimal genetic code. Molecular Biology and Evolution, 17: 511-518

Friedman SJ, Weinstein IB. 1964. Lack of fidelity in the translation of ribopolynucleotides. Proceedings of the National Academy of Sciences of USA, 52: 988-996

Gohain N, Ali T, Akhtar A. 2015. Lattice structure and distance matrix of genetic code. Journal of Biological Systems, 23(3): 1-20

Gillis D, Massar S, Cerf NJ, Rooman M. 2001. Optimality of the genetic code with respect to protein stability and amino acid frequencies. Genome Biology, 2: research0049.1-research0049.12

Hornos JEM,Hornos YMM. 1993. Algebraic model for the evolution of the genetic code. Physical Review Letters, 71(26): 4401-4404

Jiao X, Chang S, Li C, Chen W, Wang C. 2007. Construction and application of the weighted amino acid network based on energy. Physical Review E, 75(5 Pt 1): 051903

Jimenez-Montano MA. 1999. Protein evolution drives the evolution of the genetic code and vice versa. BioSystems, 54: 47-64

José MV, Morgado ER, Sánchez R, Govezensky T. 2012. The 24 possible algebraic representations of the standard genetic code in six or in three dimensions. Advanced Studies in Biology, 4: 119-152

Lehmann J. 2000. Physico-chemical Constraints Connected with the Coding Properties of the Genetic Systems. Journal of Theoretical Biology, 202: 129-144

Sanchez R. 2014. Evolutionary analysis of DNA-Protein-Coding regions based on a genetic code cube metric. Current Topics In Medicinal Chemistry, 14: 407-417

Sanchez R. 2018. Symmetric group of the genetic-code cubes. effect of the genetic-code architecture on the evolutionary process. MATCH Communications in Mathematical and in Computer Chemistry, 79: 527-560

Sanchez R, Morgado E, Grau R. 2004. The genetic code boolean lattice. MATCH Communications in Mathematical and in Computer Chemistry, 52: 29-46

Sanchez R, Moragdo E, Grau R. 2005a. A genetic code Boolean Structure. I. The meaning of Boolean deductions. Bulletin of Mathematical Biology, 67: 1-14

Sanchez R, Perfetti LA, Morgado E, Grau R. 2005b. A new DNA sequences vector space on a genetic code Galois field. MATCH Communications in Mathematical and in Computer Chemistry, 54(1): 3-28

Sanchez R, Morgado E, Grau R. 2005c. Gene algebra from a genetic code algebraic structure. Journal of Mathematical Biology, 51: 431-457

Schuster P, Fontana W, Hofacker IL. 1994. From sequences to shapes and back: a case study in RNA secondary structures. Proceedings: Biological Sciences, 255: 279-284

Steinmann D, Bremer M, Rades D, Skawran B, Siebrands C, Karstens JH, et al. 2001. Mutations of the BRCA1 and BRCA2 genes in patients with bilateral breast cancer. Britain Journal of Cancer, 85: 850-858

Telwatte S, Hearps AC, Johnson A, Latham CF, Moore K, Agius P, et al. 2015. Silent mutations at codons 65 and 66 in reverse transcriptase alleviate indel formation and restore fitness in subtype B HIV-1 containing D67N and K70R drug resistance mutations. Nucleic Acids Research, 43: 3256-3271

Watson JD, Crick FHC. 1953. A structure for deoxyribose nucleic acid. Nature, 171(3): 737-738

Woese CR. 1965. On the evolution of the genetic code. Proceedings of the National Academy of Sciences of USA, 54: 1546-1552

Yan W, Yu C, Chen J, Zhou J, Shen B. 2020. ANCA: A web server for amino acid networks construction and analysis. Frontiers in Molecular Biosciences, 19(7): 582702