American Journal of Bioinformatics

Original Research Paper

Reducing Redundancy of Codons through Total Graph

Nisha Gohain, Tazid Ali and Adil Akhtar

Department of Mathematics, Dibrugarh University, Dibrugarh-786004, India

Article history
Received: 25-02-2015
Revised: 20-04-2015
Accepted: 17-06-2015

Corresponding Author:
Nisha Gohain
Department of Mathematics,
Dibrugarh University, Dibrugarh-786004, India
Email: gohainnisha@gmail.com

Abstract: In this study some algebraic connections in genetic code is being discussed. The genetic code is the rule by which DNA stores the genetic information about formation of protein molecule. Based on the physico-chemical properties of four RNA (or DNA) bases, two orders in the base sets are obtained. This ordering allows us to define a ring structure on the set of 64 codons. Then the total graph in the genetic code algebra is being discussed. It is shown that transition mutations (purine (A, G) to purine or pyrimidine (C, T) to pyrimidine) on the third base position of codons partitions the whole set of codons into disjoint graphs and thereby generates the total graph of the genetic code. The redundancy of the 64 codons coding for the 20 amino acids is reduced by the total graph.

Keywords: Amino Acid, Genetic Code, Mutation, Total Graph, Zero Divisor

Introduction

The genetic code is a series of codons that specify which amino acids are required to make up specific protein. Deoxyribonucleic Acid (DNA) stores genetic information about how to construct or synthesize proteins. Amino acids are the building blocks of proteins. There are 20 different amino acids being found till now that occurs in proteins. Each amino acid is a triplet code of four possible bases (nucleotides). Each nucleotide contains one sugar group (deoxyribose), one phosphate group and one nitrogenous base (Adenine (A), Cytosine (C), Guanine (G) or Thymine (T)). The bases are paired and joined together by hydrogen bonds. Two hydrogen bonds attach Adenine (A) to Thymine (T) and three hydrogen bonds attach Cytosine (C) to Guanine (G). Therefore, DNA consists of two complementary long chains of nucleotides. According to Watson-Crick, a purine of one chain is always paired with a pyrimidine of the other. The sequence of one side is enough to deduce the other. Mathematically, DNA can be considered as a sequence of four letters: A, G, C and T(or U). As there are four bases, this gives us 64 codons. Out of these 64, the three triplets UAA, UAG and UGA are known as stop codons or nonsense codons and their role is to stop the biosynthesis. The codon AUG codes for the initiation of the translation process and is therefore also known as start codon. In the evolutionary importance of genetic code, the second base is considered as biologically most significant base, whereas third base is least significant base in a codon. Different kinds of mutations are possible in codons namely, point mutation, deletion, insertion, inversion. In this paper we will consider only the case of point mutation. In case of a point mutation, there is a simple change in one base of the gene sequence. It replaces a single base nucleotide with another nucleotide of the genetic material, DNA or mRNA. This mutation may be at single point, mutation at two points etc. Point mutation from purine (A, G) to purine or a pyrimidine (C, T) to pyrimidine is known as transition mutation and the point mutation from a purine to pyrimidine or vice-versa is known as transversion mutation. Point mutations usually take place during DNA replication. In DNA (or RNA), 64 codons make up the genetic code, though there are only 20 amino acids. This means that there are some overlap i.e., more than one codon code for the same amino acid. The codons that code for the same amino acids are known as synonymous codons. We can consider this as a function of many to one carrying codons to amino acids. It is therefore of interest to find out if the genetic code has any mathematical property which gets optimized when the number of codons becomes nearly thrice the number of the amino acids (Balakrishnan, 2002).

Many authors (Antoneli *et al*., 2003; Balakrishnan, 2002; Bashford *et al*., 1998; Bashford and Jarvis, 2000; Beland and Allen, 1994; Siemion *et al*., 1995 and others) worked on this field and tried to give some algebraic formulation of the structure of the genetic code. Sanchez *et al*. (2005) brought a new idea for describing the quantitative relationship between DNA genomic sequences.

Different types of graph structures can be introduced corresponding to a given algebraic structure. Cameron and Ghosh (2011) introduced the power graph of a finite

group. The power graph of a group is the graph whose vertex set is the group and two elements being adjacent if one is a power of the other. Bertholf *et al*. (1978), introduced graphs of finite abelian groups whose vertices are in one-to-one correspondence with the non-identity subgroups of *G* and two vertices are joined by an edge if and only if the corresponding subgroups intersect. Anderson and Badawi (2008), introduced the total graph of a commutative ring and denoted by $T(\Gamma(R))$, where *R* is the commutative ring. It is the (undirected) graph with all elements of *R* as vertices and for distinct $x, y \in R$, the vertices *x* and *y* are adjacent iff $x + y \in Z(R)$, where *Z*(*R*) is the set of zero-divisors of *R*. They also discussed the three (induced) subgraphs namely, $\mathrm{Nil}(\Gamma(R))$, $Z(\Gamma(R))$, and $Reg(\Gamma(R))$ of $T(\Gamma(R))$, with vertices *Nil*(*R*), Z(*R*) and *Reg*(*R*), respectively, where *Nil*(*R*) is the ideal of nilpotent elements of *R* and *Reg*(*R*) is the set of regular elements of *R*. Beck (1988) introduced the concept of the graph of the zero divisors of *R*, where he was mainly interested in colorings. In his work all elements of the ring were vertices of the graph. The investigation of colorings of a commutative ring was then continued by Anderson and Naseer (1993). Anderson and Livingston (1999) associate a graph, Γ(*R*), to *R* with vertices $Z(R)\backslash\{0\}$, where *Z* (R) is the set of zero-divisors of *R* and for distinct $x, y \in Z(R)\backslash\{0\}$, the vertices *x* and *y* are adjacent if and only if *xy* = 0. Akbari *et al*. (2009) proved that the total graph is a Hamiltonian graph if it is connected.

The main motivation of this work is for exploring mathematical structures viz., graph structures that may naturally occur in genetic code. In this paper, we have discussed an algebraic structure of the genetic code. The hydrogen bond number and the chemical type (purine and pyrimidine) of bases play an important role in this. From this two orders of the base sets: {A, G, C, U} and {U, C, G, A} are obtained. Further an ordering of the codons are obtained. It is being found that transition mutation in codons partitioned the whole set of codons into two disjoint sets which gives two disjoint graphs, both of which are individually connected.

Genetic Code Algebraic Structure

The 64 codons code the 20 amino acids. It can be arranged or ordered in such a way that going from the codons that code to hydrophobic amino acids (e.g., AUA, ...) to the codons that code to hydrophilic amino acids (e.g., UAU, ...). This ordering of the 64 codons implicitly gives an ordering of the four RNA (or DNA) bases.

Based on the physico-chemical properties of four RNA (or DNA) bases two orders in the base set can be obtained. To determine the order, similar chemical type (purines/pyrimidines) of bases are taken together and the starting base needs a minimum hydrogen bond number. As a result, the two orders {*A*, *G*, *C*, *U*} and {*U*, C, *G*, *A*} in these base set are specified. Next, these two ordering of the bases will help us to obtain an ordering in the set of 64 codons. At first the ordering is applied in the third codon position, then the first codon position. Finally to the second base with a reverse ordering as shown in Table 1 and 2. That is from the less biologically relevant base to the most relevant base in the codon. In Table 1, the codons are arranged from codons that code hydrophobic amino acids to hydrophilic amino acids. The codons from 0 to 15 (codons XUZ) codes the most hydrophobic amino acids and the codons from 48 to 63(codons XAZ) codes the most hydrophilic amino acids.

A sum operation has been introduced in the set of codons in a manner to consecutively obtain all codons from the start codon AUG. Then the resulting inequalities of the sum of the codons with the codon AUG implicitly gives a sum operation in the set of bases {A, G, C, U}. Then next a sum operation is defined on the whole codon set. It was observed that the group obtained on the set of codons is isomorphic to the group of integer module 64, $(Z_{64}, +)$. Further a product operation can also be defined on the set of codons such that the set of all codons gives a ring structure isomorphic to the ring of $(Z_{64}, +, \cdot)$.

Base Group and Codon Group

Sanchez *et al*. (2005) defines an algebraic structure of the genetic code. Following Sanchez, an algebraic structure of the set of codons is obtained here such that it is isomorphic to the group of integers modulo 64 $(Z_{64}, +)$.

A sum operation is introduced in the set of codons in such a manner that all codons will be consecutively obtained from the start codon AUG. With this sum operation, a group structure will be defined on the set of codons. From Table 1, we observed that the codon AUA acts as the neutral element. The following equalities are obtained from the successive ordering of the codons in Table 1:

$$AUA + AUG = AUG$$
$$AUG + AUG = AUC$$
$$AUC + AUG = AUU$$
$$AUU + AUG = GUA$$

More generally:

$$XYA + AUG = XYG$$
$$XYG + AUG = XYC$$
$$XYC + AUG = XYU$$
$$AYU + AUG = GYA$$

where, $X, Y \in \{A, G, C, U\}$ denote the first and the second bases respectively.

Table 1. The primal genetic code table induced by the order {A, G, C, U}. The bijection between the group of codons with Z_{64} is shown in the table

	U			C			G			A			
	No	(1)	(2)	No	(1)	(2)	No	(1)	(2)	No	(1)	(2)	
A	0	AUA	I	16	ACA	T	32	AGA	R	48	AAA	K	A
	1	AUG	M	17	ACG	T	33	AGG	R	49	AAG	K	G
	2	AUC	I	18	ACC	T	34	AGC	S	50	AAC	N	C
	3	AUU	I	19	ACU	T	35	AGU	S	51	AAU	N	U
G	4	GUA	V	20	GCA	A	36	GGA	G	52	GAA	E	A
	5	GUG	V	21	GCG	A	37	GGG	G	53	GAG	E	G
	6	GUC	V	22	GCC	A	38	GGC	G	54	GAC	D	C
	7	GUU	V	23	GCU	A	39	GGU	G	55	GAU	D	U
C	8	CUA	L	24	CCA	P	40	CGA	R	56	CAA	Q	A
	9	CUG	L	25	CCG	P	41	CGG	R	57	CAG	Q	G
	10	CUC	L	26	CCC	P	42	CGC	R	58	CAC	H	C
	11	CUU	L	27	CCU	P	43	CGU	R	59	CAU	H	U
U	12	UUA	L	28	UCA	S	44	UGA	-	60	UAA	-	A
	13	UUG	L	29	UCG	S	45	UGG	W	61	UAG	-	G
	14	UUC	F	30	UCC	S	46	UGC	C	62	UAC	Y	C
	15	UUU	F	31	UCU	S	47	UGU	C	63	UAU	Y	U

Table 2. The dual genetic code table induced by the order {U, C, G, A}. The bijection between the group of codons with Z_{64} is shown in the table

	A			G			C			U			
	No	(1)	(2)	No	(1)	(2)	No	(1)	(2)	No	(1)	(2)	
U	0	*UAU*	*Y*	16	*UGU*	*C*	32	*UCU*	*S*	48	*UUU*	*F*	U
	1	*UAC*	*Y*	17	*UGC*	*C*	33	*UCC*	*S*	49	*UUC*	*F*	C
	2	*UAG*	-	18	*UGG*	*W*	34	*UCG*	*S*	50	*UUG*	*L*	G
	3	*UAA*	-	19	*UGA*	-	35	*UCA*	*S*	51	*UUA*	*L*	A
C	4	*CAU*	*H*	20	*CGU*	*R*	36	*CCU*	*P*	52	*CUU*	*L*	U
	5	*CAC*	*H*	21	*CGC*	*R*	37	*CCC*	*P*	53	*CUC*	*L*	C
	6	*CAG*	*Q*	22	*CGG*	*R*	38	*CCG*	*P*	54	*CUG*	*L*	G
	7	*CAA*	*Q*	23	*CGA*	*R*	39	*CCA*	*P*	55	*CUA*	*L*	A
G	8	*GAU*	*D*	24	*GGU*	*G*	40	*GCU*	*A*	56	*GUU*	*V*	U
	9	*GAC*	*D*	25	*GGC*	*G*	41	*GCC*	*A*	57	*GUC*	*V*	C
	10	*GAG*	*E*	26	*GGG*	*G*	42	*GCG*	*A*	58	*GUG*	*V*	G
	11	*GAA*	*E*	27	*GGA*	*G*	43	*GCA*	*A*	59	*GUA*	*V*	A
A	12	*AAU*	*N*	28	*AGU*	*S*	44	*ACU*	*T*	60	*AUU*	*I*	U
	13	*AAC*	*N*	29	*AGC*	*S*	45	*ACC*	*T*	61	*AUC*	*I*	C
	14	*AAG*	*K*	30	*AGG*	*R*	46	*ACG*	*T*	62	*AUG*	*M*	G
	15	*AAA*	*K*	31	*AGA*	*R*	47	*ACA*	*T*	63	*AUA*	*I*	A

Table 3. Sum operation tables defined on the four bases

3a

+	A	G	C	U
A	A	G	C	U
G	G	C	U	A
C	C	U	A	G
U	U	A	G	C

3b

+	U	G	C	A
U	U	G	C	A
G	G	C	A	U
C	C	A	U	G
A	A	U	G	C

These equalities allow us to define a sum operation in the set of bases {A, G, C, U}. Table 3a and 3b represents the sum operation of the bases obtained from the two possible orders. This sum operation together with the base set represents a cyclic group, say $(P, +)$ and is isomorphic to $(Z_{64}, +)$. Also by using the same algorithm given by Sanchez *et al*. (2005), a sum operation between the codons is obtained and the codon sum together with the codon set represents a cyclic group, say $(C_g, +)$ and it is isomorphic to $(Z_{64}, +)$.

In the group structure $(C_g, +)$, the neutral element and the identity elements are AUA and AUG

respectively. Both these codons are essentially different from all other codons. In case of protein synthesis, proteins accurate translation of the genetic code is essential. Determination of the mechanisms by which tRNAs decode all sense codons on mRNAs is essential to our understanding of this basic principle in all living organisms. Base modifications at the first (wobble) anticodon position of tRNAs play critical roles in deciphering cognate codons. In all sense codons, with the exception of AUA and AUG, two codon sets ending in a purine specify identical amino acids because modified uridines at the wobble positions pair not only with A but also with G by classical wobble pairing. But for these two codons AUA and AUG, ending in purines specify a different amino acid (Taniguchi *et al*., 2013).

A definition of "codon sum" has been given based on the distinction between the base positions in the codon, the order of the four bases set and the base sum operation. In this algorithm, the cyclic character of the sum of codons is hereditarily obtained from the base sum. As mentioned earlier, the second base position of a codon is biologically most important base followed by the first and then the third position. The order of importance of the bases is taken into account to define the sum of codons. The sum operation between two codons is obtained from the less biologically important base (third codon position) to the most important base (second codon position):

- The corresponding bases in the third position are added according to the sum table
- If the resultant base of the sum operation is previous in order to the added bases (the orders in the set of bases), then the new value is written and the base G (or C if the dual group of bases is used) is added to the next position
- The other bases are added according to the sum table, step (ii), going from the first base to the second base

For example, the sum of codons *UCC* and *GGU* is performed by using the group of bases (Table 1 A) in the following manner

> $C+U = G$, the third bases are added and the base is added to the next position because base G precedes bases C and U in the set of ordered bases $\{A, G, C, U\}$.
>
> $U+G+G = A+G = G$, the first bases and the base C obtained in the first step are added. Again, base C is added to the next position.
> $C+G = U$, the second bases are added.
> Finally, it gives $UCC + GGU = GUG$

This sum operation satisfies the sum operation of group axioms. As a result two cyclic Abelian groups with operation sum "+" can be defined in the set of codons C_g and denoted as $(C_g, +)$. Since all finite cyclic groups with the same number of elements are isomorphic, the codon sum together with the codon set represents a cyclic group, say $(C_g, +)$ and is isomorphic to $(Z_{64}, +)$.

Furthermore, this isomorphism allows us to refer to even or odd elements of C_g. In Table 1, the codons with bases A and C in the third position are even codons and the codons with bases U and G are odd codons.

The dual genetic code table induced by the dual order $\{U, C, G, A\}$ is given in Table 2. The bijection between the dual genetic code Abelian group with Z_{64} is also shown in the Table 2.

Graphical Representation of the Codons

In this section we attempt to give a graph structure to the set of codons based on the algebraic structure discussed in the previous section.

In the ring Z_{64}, $Z(Z_{64})$ = {AUC, GUA, CUA, ACA, AGA} is the set of zero divisors. Out of all the codons belonging to this set, three of them are hydrophobic codons (i.e., codes to hydrophobic amino acids) and the other two are hydrophilic codons (i.e., codes to hydrophilic amino acids). The three hydrophobic codons code all the aliphatic amino acids. Now we construct the total graph of the codon set. Here by definition two codons are adjacent if and only if their sum is a zero divisor.

For example, the codons AUU and GUG are adjacent as their sum is CUA (a zero divisor). By using Table-2, we can obtain the total graph of C_g as shown in Fig. 1.

In the total graph of C_g (Fig. 1), we observe that there are two disjoint components. One component G_1 (say) contain vertices representing all odd codons and the other G_2 (say) contains vertices representing all even codons. So:

G_1={AAU,AAG,CAU,CAG,GAU,GAG,UAU,UAG,ACU,ACG,CCU,CCG,GCU,GCG,UCU,UCG,AGU,AGG,CGU,CGG,GGU,GGG,UGU,UGG,AUU,AUG,CUU,CUG,GUU,GUG,UUU,UUU}

And

G_2={AAC,AAA,CAC,CAA,GAC,GAA,UAC,UAA,ACC,ACA,CCC,CCA,GCC,GCA,UCC,UCA,AGC,AGA,CGC,CGA,GGC,GGA,UGC,UGA,AUC,AUA,CUC,CUA,GUC,GUA,UUC,UUA}.

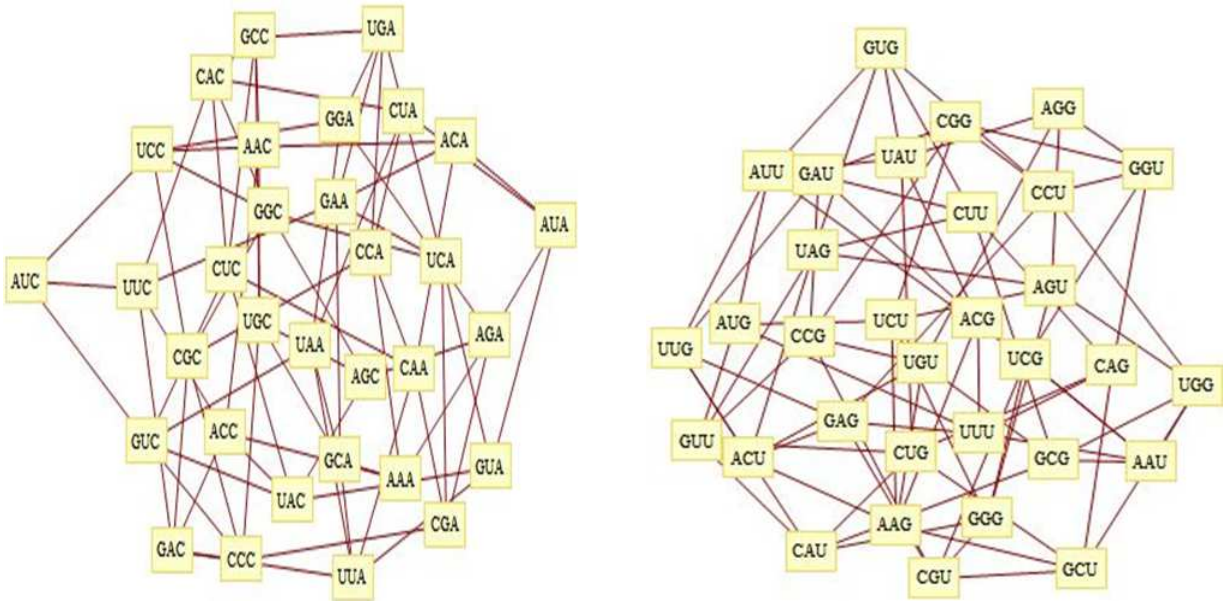

Fig. 1. Total graph of Z_{64}

We may note in passing that the vertices of G_1 form a group under multiplication and the vertices of G_2 form a group under addition. Also the codons of G_1 are sufficient to code for all the 20 amino acids. We can define a function f from the set of vertices of G_1 (odd codon) to the set of vertices to G_2 (even codon) given by $f: G_1 \rightarrow G_2$, such that:

$$f(XYZ) = \begin{cases} XY\text{A} & \text{if } Z = \text{G} \\ XY\text{C} & \text{if } Z = \text{U} \end{cases} \qquad \forall XYZ \in G_1$$

This function f can also be represented as:

$$f_1(XYZ) = XYZ + UAU$$

For example,

if we consider codon AAU which is an element of G_1 (here Z = U) then by our defined function f we get the image of AAU as AAC which is an element of G_2.

Similarly in the another function we have f_1(AAU) = AAU+ UAU = 51+63 = 50 = AAC.

Conversely we have a function g: $G_2 \rightarrow G_1$, such that

$$g(XYZ) = \begin{cases} XY\text{G} & \text{if } Z = \text{A} \\ XY\text{U} & \text{if } Z = \text{C} \end{cases} \qquad \forall XYZ \in G_2$$

This function g can also be represented as:

$$g_1(XYZ) = XYZ + AUG$$

It is clear that the functions f, f_1, g and g_1 are all bijective mappings. Also the mapping f and g are inverse of one another and f_1 and g_1 are inverse of one another. Also for all the functions the image and pre-image codes for the same amino acids with the exception of AUA and AUG. From a biological point of view it is observed that these functions represent the transition mutation of the third base of a codon.

Hence we can conclude that the transition of the third base of codons can be represented in terms of total graph of the genetic code.

It is observed that the codons in the set G_1 and G_2 keep the codon parity but the codon order may vary. Also from our defined function it is clear that the transition of the third base of all codons gives a bijective map.

Conclusion

The 64 codons are arranged in a genetic code table. By taking into account the hydrogen bond number and chemical types of the bases (purine/pyrimidine),we have obtained an ordering of the four bases {A, G, C, U}. This ordering of the bases induces an ordering of the 64 codons. This arrangement of the codons is such that the first codon is most hydrophobic and hydrophobicity decreases as we move towards the end of the table 1. Then a sum operation is introduced which results in a group structure (C_g, +) on the set of codons, isomorphic to (Z_{64}, +). Based on this group structure (C_g, +), a graph structure is introduced on the set of codons. Thus the total graph of the genetic code is being investigated. The total graph structure partitions the set of codons into two disjoint

graphs: One containing all vertices representing odd codons and the other containing vertices with even codons. We have also observed that this partition can be equivalently obtained by considering third base transition mutation of the codons. As the 64 codons code for the 20 amino acids there is a redundancy in the set of codons. The total graph discussed in this paper reduces this redundancy by providing a set of 32 codons, i.e., the vertices of G_1, which codes for all the 20 amino acids.

Acknowledgement

The authors gratefully acknowledge the suggestions and comments of the anonymous referee and the editor which helped immensely to make substantial improvements to the content and presentation of the paper.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Akbari, S., D. Kiani, F. Mohammadi and S. Moradi, 2009. The total graph and regular graph of a commutative ring. J. Pure Applied Algebra, 213: 2224-2228. DOI: 10.1016/j.jpaa.2009.03.013

Anderson, D.D. and M. Naseer, 1993. Beck′s coloring of a commutative ring. J. Algebra, 159: 500-514. DOI: 10.1006/jabr.1993.1171

Anderson, D.F. and A. Badawi, 2008. The total graph of a commutative ring. J. Algebra, 320: 2706-2719. DOI: 10.1016/j.jalgebra.2008.06.028

Anderson, D.F. and P.S. Livingston, 1999. The zero-divisor graph of a commutative ring. J. Algebra, 217: 434-447. DOI: 10.1006/jabr.1998.7840

Antoneli, F., L. Braggion, M. Forger and J.E.M. Hornos, 2003. Extending the search for symmetries in the genetic code. Int. J. Mod. Phys. B, 17: 3135-3204. DOI: 10.1142/S0217979203020764

Balakrishnan, J., 2002. Symmetry scheme for amino acid codons. Phys. Rev., 65, 021912-5. DOI: 10.1103/PhysRevE.65.021912

Bashford, J.D. and P.D. Jarvis, 2000. The genetic code as a periodic table: Algebraic aspects. Biosystems, 57: 147-161. DOI: 10.1016/S0303-2647(00)00097-6

Bashford, J.D., I. Tsohantjis and P.D. Jarvis, 1998. A supersymmetric model for the evolution of the genetic code. Proc. Natl. Acad. Sci. USA, 95: 987-992.

Beck, I., 1988. Coloring of commutative rings. J. Algebra, 116: 208-226. DOI: 10.1016/0021-8693(88)90202-5

Bertholf, D., Stillwater, Hattiesburg and G. Walls, 1978. Graphs of finite abelian groups. Czechoslovak Math. J., 28: 365-368.

Cameron, J.P. and S. Ghosh, 2011. The power graph of a finite group. Discrete Math., 311: 1220-1222. DOI: 10.1016/j.disc.2010.02.011

Sanchez, R., E. Morgado and R. Grau, 2005. Gene algebra from a genetic code algebraic structure. J. Math. Biol., 51: 431-457. DOI: 10.1007/s00285-005-0332-8

Siemion, I.Z., P.J. Siemion and K. Krajewski, 1995. Chou-Fasman conformational amino acid parameters and the genetic code. Biosystems, 36: 231-238. DOI: 10.1016/0303-2647(95)01559-4

Taniguchi, T., K. Miyauchi, D. Nakane, M. Miyata and A. Muto *et al*., 2013. Decoding system for the AUA codon by tRNAIle with the UAU anticodon in Mycoplasma mobile. Nucl. Acids Res., 41: 2621-2631. DOI: 10.1093/nar/gks1344