Polynomial Representation of Codons by Galois Field Extensions and Labeling in the Study of the Genetic Code

A. J. Oliveira , R. Palazzo Jr

State University of Campinas

Center for Science of Information
NSF Science and Technology Center

Introduction

The study of the genetic code, its characteristics, properties and functions is an issue of utmost importance. The application of mathematical structures has been a subject of several works. The polynomial representation of codons is an important application, since it may be relevant in the study of mutations. There are **24** mappings or permutations from the set representing the nucleotides (adenine, cytosine, guanine, thymine / uracil), denoted by $\boldsymbol{N = \{A, C, G, T/U\}}$, into the set of the code alphabet denoted by $\mathbb{Z}_2 \times \mathbb{Z}_2 = \{00, 01, 10, 11\}$. These **24** permutations may be grouped into three subsets called *labelings* and denoted by A, B and C, associated with the geometrical shapes that they generate. The aim of this work is to present a polynomial representation of the codons in each one of the labelings, in order to determine a vector space, and relate this representation with a constellation of **64** signals (codons) of the genetic code. The polynomial representation of these codons uses the elements of $\boldsymbol{GF(64)}$ obtained by Galois extension, as well as the importance of the nitrogenous bases in the codons. As a result, we obtain polynomial representation for each one of the three labelings and their association with a constellation of **64** signals in a digital communication system.

Representation Model of Polynomial Codons

In [3], Sanchez et al. used the representation $\boldsymbol{G-00}$, $\boldsymbol{U-10}$, $\boldsymbol{A-01}$ and $\boldsymbol{C-11}$ to make the construction of the genetic code table, which reflects one of the **24** permutations of the labeling proposed in [1] and [2]. This representation is the result of an isomorphism of two Boolean lattices, $\varphi : \boldsymbol{B(X)} \rightarrow ((\mathbb{Z}_2)^2, \wedge, \vee)$ where $\boldsymbol{B(X) = \{A, C, G, T/U\}}$, $\mathbb{Z}_2 = \{0, 1\}$, $\wedge$ - logical connective and (conjunction) and $\vee$ - connective logical or (disjunction).
The coefficients in the polynomial representation of the codons obey an important order of the position of bases in the codons and the isomorphism $\varphi : \boldsymbol{B(X)} \rightarrow (\mathbb{Z}_2)^2$ allows the presentation of a function $\psi : \boldsymbol{GF(64)} \rightarrow \boldsymbol{C_g}$, where $\boldsymbol{GF(64)}$ represents the elements of the Galois extension and $\boldsymbol{C_g}$ the elements of the genetic code, so that:

$$\boldsymbol{a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 \rightarrow (f_1(a_2 a_3), f_2(a_4 a_5), f_3(a_0 a_1))}$$

For all $\boldsymbol{X_1 X_2 X_3} \in \boldsymbol{C_g}$ there is a polynomial $\boldsymbol{p(x)} \in \boldsymbol{GF(64)}$ such that:

$$\psi(\boldsymbol{p(x)}) = \boldsymbol{X_1 X_2 X_3}$$

Figure 1 shows the construction of the genetic code with the polynomial representation for each one of the **64** codons.

	G				U				A				C				
	n°	GF(64)	I	II	n°	GF(64)	I	II	n°	GF(64)	I	II	n°	GF(64)	I	II	
G	0	000000	GGG	G	16	000010	GUG	V	32	000001	GAG	E	48	000011	GCG	A	G
	1	100000	GGU	G	17	100010	GUU	V	33	100001	GAU	D	49	100011	GCU	A	U
	2	010000	GGA	G	18	010010	GUA	V	34	010001	GAA	E	50	010011	GCA	A	A
	3	110000	GGC	G	19	110010	GUC	V	35	110001	GAC	D	51	110011	GCC	A	C
U	4	001000	UGG	W	20	001010	UUG	L	36	001001	UAG	STOP	52	001011	UCG	S	G
	5	101000	UGU	C	21	101010	UUU	F	37	101001	UAU	Y	53	101011	UCU	S	U
	6	011000	UGA	STOP	22	011010	UUA	L	38	011001	UAA	STOP	54	011011	UCA	S	A
	7	111000	UGC	C	23	111010	UUC	F	39	111001	UAC	Y	55	111011	UCC	S	C
A	8	000100	AGG	R	24	000110	AUG	M	40	000101	AAG	K	56	000111	ACG	T	G
	9	100100	AGU	S	25	100110	AUU	I	41	100101	AAU	N	57	100111	ACU	T	U
	10	010100	AGA	R	26	010110	AUA	I	42	010101	AAA	K	58	010111	ACA	T	A
	11	110100	AGC	S	27	110110	AUC	I	43	110101	AAC	N	59	110111	ACC	T	C
C	12	001100	CGG	R	28	001110	CUG	L	44	001101	CAG	Q	60	001111	CCG	P	G
	13	101100	CGU	R	29	101110	CUU	L	45	101101	CAU	H	61	101111	CCU	P	U
	14	011100	CGA	R	30	011110	CUA	L	46	011101	CAA	Q	62	011111	CCA	P	A
	15	111100	CGC	R	31	111110	CUC	L	47	111101	CAC	H	63	111111	CCC	P	C

Figure: Genetic code under labeling A

Proposed Model under Labeling B and C

Next, we show the polynomial representation for labelings B and C. Among the 8 possible permutation within the subset with labeling B one may choose anyone of them. Hence, if one selects **1230**, the corresponding binary representation in $\mathbb{Z}_2 \times \mathbb{Z}_2$ is $\boldsymbol{U-00}$, $\boldsymbol{A-10}$, $\boldsymbol{G-01}$ and $\boldsymbol{C-11}$. Figure 2 shows the genetic code under labeling B with the polynomial representation for each one of the codons.

	U				A				G				C				
	n°	GF(64)	I	II	n°	GF(64)	I	II	n°	GF(64)	I	II	n°	GF(64)	I	II	
U	0	000000	UUU	F	16	000010	UAU	Y	32	000001	UGU	C	48	000011	UCU	S	U
	1	100000	UUA	L	17	100010	UAA	STOP	33	100001	UGA	STOP	49	100011	UCA	S	A
	2	010000	UUG	L	18	010010	UAG	STOP	34	010001	UGG	W	50	010011	UCG	S	G
	3	110000	UUC	F	19	110010	UAC	Y	35	110001	UGC	C	51	110011	UCC	S	C
A	4	001000	AUU	I	20	001010	AAU	N	36	001001	AGU	S	52	001011	ACU	T	U
	5	101000	AUA	I	21	101010	AAA	K	37	101001	AGA	R	53	101011	ACA	T	A
	6	011000	AUG	M	22	011010	AAG	K	38	011001	AGG	R	54	011011	ACG	T	G
	7	111000	AUC	I	23	111010	AAC	N	39	111001	AGC	S	55	111011	ACC	T	C
G	8	000100	GUU	V	24	000110	GAU	D	40	000101	GGU	G	56	000111	GCU	A	U
	9	100100	GUA	V	25	100110	GAA	E	41	100101	GGA	G	57	100111	GCA	A	A
	10	010100	GUG	V	26	010110	GAG	E	42	010101	GGG	G	58	010111	GCG	A	G
	11	110100	GUC	V	27	110110	GAC	D	43	110101	GGC	G	59	110111	GCC	A	C
C	12	001100	CUU	L	28	001110	CAU	H	44	001101	CGU	R	60	001111	CCU	P	U
	13	101100	CUA	L	29	101110	CAA	Q	45	101101	CGA	R	61	101111	CCA	P	A
	14	011100	CUG	L	30	011110	CAG	Q	46	011101	CGG	R	62	011111	CCG	P	G
	15	111100	CUC	L	31	111110	CAC	H	47	111101	CGC	R	63	111111	CCC	P	C

Figure: Genetic code under labeling B

In the following example, the table features are detailed.

Example

Let us consider the number **54** in the previous table. This number is represented in binary form (from left to right) by **011011**. However, this binary representation is associated with the codon $\boldsymbol{ACG}$ and its polynomial representation is $\boldsymbol{x + x^2 + x^4 + x^5}$.

$$\boldsymbol{54 \rightarrow 110110 \rightarrow 011011 \rightarrow ACG \rightarrow x + x^2 + x^4 + x^5 \rightarrow T}$$

The number **54** is the representation in $\mathbb{Z}_{64}$ of the codon $\boldsymbol{ACG}$ as well as **110110** is its binary representation, which read from right to left identifies the polynomial coefficients related to this codon, ie $\boldsymbol{x + x^2 + x^4 + x^5}$. The order of importance of the bases identifies the codon $\boldsymbol{ACG}$. The amino acid is represented by $\boldsymbol{T}$ is threonine.

For the labeling C, we have used the label **1320**, which in $\mathbb{Z}_2 \times \mathbb{Z}_2$ is $\boldsymbol{U-00}$, $\boldsymbol{A-10}$, $\boldsymbol{C-01}$, $\boldsymbol{G-11}$.

Figure 3 shows the polynomial representation under labeling C.

	U				A				C				G				
	n°	GF(64)	I	II	n°	GF(64)	I	II	n°	GF(64)	I	II	n°	GF(64)	I	II	
U	0	000000	UUU	F	16	000010	UAU	Y	32	000001	UCU	C	48	000011	UGU	S	U
	1	100000	UUA	L	17	100010	UAA	STOP	33	100001	UCA	STOP	49	100011	UGA	S	A
	2	010000	UUC	F	18	010010	UAC	Y	34	010001	UCC	C	50	010011	UGC	S	C
	3	110000	UUG	L	19	110010	UAG	STOP	35	110001	UCG	W	51	110011	UGG	S	G
A	4	001000	AUU	I	20	001010	AAU	N	36	001001	ACU	S	52	001011	AGU	T	U
	5	101000	AUA	I	21	101010	AAA	K	37	101001	ACA	R	53	101011	AGA	T	A
	6	011000	AUC	I	22	011010	AAC	N	38	011001	ACC	S	54	011011	AGC	T	C
	7	111000	AUG	M	23	111010	AAG	K	39	111001	ACG	R	55	111011	AGG	T	G
C	8	000100	CUU	L	24	000110	CAU	H	40	000101	CCU	R	56	000111	CGU	P	U
	9	100100	CUA	L	25	100110	CAA	Q	41	100101	CCA	R	57	100111	CGA	P	A
	10	010100	CUC	L	26	010110	CAC	H	42	010101	CCC	R	58	010111	CGC	P	C
	11	110100	CUG	L	27	110110	CAG	Q	43	110101	CCG	R	59	110111	CGG	P	G
G	12	001100	GUU	V	28	001110	GAU	D	44	001101	GCU	G	60	001111	GGU	A	U
	13	101100	GUA	V	29	101110	GAA	E	45	101101	GCA	G	61	101111	GGA	A	A
	14	011100	GUC	V	30	011110	GAC	D	46	011101	GCC	A	62	011111	GGC	G	C
	15	111100	GUG	V	31	111110	GAG	E	47	111101	GCG	A	63	111111	GGG	A	G

Figure: Genetic Code under labeling C

The same analysis used in the example of the labeling B can be applied to the case of the labeling C.

Conclusions

We have constructed tables of the genetic code for the labelings B and C using a structure of the vector space via an extension of the field $\boldsymbol{GF(2)}$ into $\boldsymbol{GF(64)}$, englobing the **64** codons of the genetic code. We can relate this structure to the context of a digital communication system, where the **64** codons represent the signals of a signal constellation.
Thus, we have constructed geometric representations for the genetic code by use of the vector space approach, that is, an algebraic model for the labeling A, B and C. This was possible since the degree of the Galois field extension $\boldsymbol{[GF(64) : GF(2)]}$ provides such a vector space whose dimension is **6**. This vector space approach is one of the many ways of representing geometrically the genetic code.

References

- Faria, L.C.B., Palazzo, R.: Existências de códigos corretores de erros e protocolos de comunicação em sequências de dna. In: Tese de Doutorado (2011)
- Rocha, A.S.L., Palazzo R., J.: Modelo de sistema de comunicacões digital para o mecanismo de importacâo de proteínas mitocondriais através de códigos corretores de erros. In: Tese de Doutorado (2010)
- Sánchez R. Grau, R.M.E.: A new dna sequences vector space on a genetic code galois field. In: MATCH Commun. Math. Comput. Chem (2005)

Acknowledgements

CAPES CNPq FAPESP