International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:5, 2007

Vector Space of the Extended Base-triplets over the Galois Field of five DNA Bases Alphabet

Robersy Sánchez and Ricardo Grau

***Abstract*—**A plausible architecture of an ancient genetic code is derived from an extended base triplet vector space over the Galois field of the extended base alphabet {D, G, A, U, C}, where the letter D represent one or more hypothetical bases with unspecific pairing. We hypothesized that the high degeneration of a primeval genetic code with five bases and the gradual origin and improvements of a primitive DNA repair system could make possible the transition from the ancient to the modern genetic code. Our results suggest that the Watson-Crick base pairing and the non-specific base pairing of the hypothetical ancestral base D used to define the sum and product operations are enough features to determine the coding constraints of the primeval and the modern genetic code, as well as the transition from the former to the later. Geometrical and algebraic properties of this vector space reveal that the present codon assignment of the standard genetic code could be induced from a primeval codon assignment. Besides, the Fourier spectrum of the extended DNA genome sequences derived from the multiple sequence alignment suggests that the called period-3 property of the present coding DNA sequences could also exist in the ancient coding DNA sequences.

***Keywords*—**genetic code vector space, primeval genetic code, power spectrum.

I. Introduction

MUTATIONS continuously occur in the genomes of all living organisms at a very low frequency, which tends to be constant for a specific species. Genome instability caused by the great variety of DNA-damaging agents would be an overwhelming problem for cells and organisms if it were not for DNA repair. So, we can ask how was conserved the prebiotic "genetic information" without a repair system. A plausible explanation could be the existence of extra DNA bases with unspecific pairing with the present bases and the possible high degeneration of the primeval genetic code. If we suppose that there was a primeval genetic code then there are not biological reasons to restrict the base alphabet of the primeval RNA and DNA to the present DNA bases. Likewise, there are not biological reasons to restrict the number of codons which code to amino acids. Actually, the environment conditions of the primeval environment could make possible the relative abundance of different bases and their isomers [1]. This richness of bases should increase the success probability of life. The Watson-Crick base pairs $G \equiv C$ and $A = U$ (where each '–' symbolizes a hydrogen bond) characterize the present DNA molecule. The accessibility to five or more bases in the "primeval soup" make plausible the non-Watson-Crick base pairing in a primitive RNA world and later in a primeval DNA molecule.

Piccirilli et al. [2] demonstrated that the alphabet can in principle be larger. C. Switzer et al. [3] have shown an enzymatic incorporation of new functionalized bases into RNA and DNA. This expanded the genetic alphabet from 4 to 6 letters, permits new base pairs, and provides RNA molecules with the potential for greatly increased catalytic power. Actually, a number of alternative base pairs have been proposed. These include isoguanine and isocytosine [2]-[4], diaminopurine and U [2], pseudodiaminopyrimidine (ribose bound to C5 rather than N1) and xanthine [2], A and urazole (1,2,4-triazole-3,5-dione) [5]. According to Levy and Miller, however, to get greater stability it may be necessary to use bases other than these pyrimidines [6].

The existence of uncommon bases in the present RNA and DNA molecules could be the relict genetic fingerprint of a molecular evolution process from primordial cells with an extended DNA base alphabet. At present, minor bases of DNA, 5-Methylcytidine occurs in the DNA of animals and higher plants, N-6-methyladenosine in bacterial DNA, and 5-hydroxymethylcytidine in the DNA of bacteria infected with certain bacteriophages. Ribosomal RNAs characteristically contain a number of specially modified nucleotides, including pseudouridine residues, ribothymidylic acid, and methylated bases. The abundance of bases with non-specific pairing in a primeval DNA molecule could increase the degeneration of coding apparatus and diminished the error frequency during transcription and translation process [2].

On other hand, different algebraic structures of the genetic code have been recently published [7]-[9]. The natural extension of these algebraic structures to gene sequences allows derive substitution mutations as lineal transformations and translations in the *N*-dimensional DNA sequence vector space (or *N*-dimensional modulo). In these structures, however, deletion and insertion mutations (indels) can not be analyzed. Now, in order to include these kinds of mutations, we have defined a new Galois field ($GF(5)$) over the set of extended DNA alphabet $X_i \in$ {D, A, C, G, U}, where the letter D can symbolize a deletion or can be considered as a hypothetical base with non-specific pairing present in a

R Sánchez is with the Research Institute of Tropical Roots, Tuber Crops and Bananas (INIVIT). Bioinformatic Group. Santo Domingo. Villa Clara. Cuba. e-mail: robersy@uclv.edu.cu, robersy@inivit.co.cu).

R. Grau is with Center of Studies on Informatics. Central University of Las Villas. Villa Clara. Cuba (e-mail: rgrau@uclv.edu.cu).

primeval RNA and DNA molecules.

The extended DNA alphabet naturally leads us to an extended set of base-triplets $X_1X_2X_3$, i.e. an "extended genetic code", where $X_i \in$ {D, A, C, G, U}. At this point a new base-triplet vector space over $GF(5)$ can be derived in analogous way to the genetic code vector space over $GF(4)$ as appears in [8] and [9]. The extension of this vector space to the DNA sequences should lead, however, to new biological-algebraic insights. The extended DNA sequences appear, normally, during the multiple sequence analysis of genes or genomic sequences. The aligned DNA sequences with gaps resulting of the alignment procedure could be analyzed, integrally, as element of the new vector space.

Here, we shall show that the standard genetic code architecture could be derived from an ancient coding apparatus with an extended alphabet of five bases.

II. Theoretical Model

A Galois field $GF(5)$ is defined on the ordered set B = {D, G, A, U, C}. Next, the order in the extended base-triplet is induced by the order in the base set and the vector space of the extended base-triplet set over the Galois field $GF(5)$ will be defined.

A. The extended base-triplet vector space over the Galois field GF(5)

If a Galois field algebraic structure is defined on the extended base alphabet subject to the constraint A + U = U + A = D and A • U = U • A = G then the sum and product operations can be defined on the sets {D, G, A, U, C} and {D, G, A, U, C}. That is, it is required that bases A and U will be inverses in the sum and product operations with the base G as neutral element for product. So, these definitions reflect the Watson-Crick base pairs G ≡ C and A = U distinctive of the present DNA molecule and the non-specific pairing of the ancient hypothetical base(s) D.

The definitions of sum and product operations are presented in Table I. By construction, the field defined in the extended base alphabet B = {D, G, A, U, C} is isomorphic to the field of integer's remainder modulo 5, a simple representation of GF(5). Explicitly, there is the bijection: D↔0, G↔1, A↔2, U↔3, C↔4.

An abelian group on the set of extended triplets set B^3 = $\{X_1X_2X_3\}$ (see Table II) can be defined as the direct third power of $(B^3, +) = (B, +) \times (B, +) \times (B, +)$ of the group $(B, +)$, where the operation "+" is given by coordinated as appear in Table I. Next, for all elements $\alpha \in \{0, 1, 2, 3, 4\}$ and for all codons $XYZ \in (B^3, +)$, the element $\alpha \bullet XYZ = \overbrace{XYZ + XYZ + ... + XYZ}^{\alpha \text{ times}} \in \left(B^3, +\right)$ is well defined. As a result, group $(B^3, +)$ is a three-dimensional vector space over $GF(5)$. Likewise, the N-dimensional vector space $(B^3)^N$ is obtained.

TABLE I
Operation Tables of the Galois Field on the Ordered Set of the Extended Bases Alphabet B = {D, G, A, U, C}

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

•	D	G	A	U	C
D	D	D	D	D	D
G	D	G	A	U	C
A	D	A	C	G	U
U	D	U	G	C	A
C	D	C	U	A	G

III. Geometrical and Algebraic Features of the Primordial Genetic Code Model

There is a unique abelian group up to isomorphism with 5 elements. As a result, the abelian group $(B, +)$ can be represented in many ways by means of isomorphisms. For instance, as was said before this group is isomorphic to the group of integer modulo 5 and the complex representation $C_5 = \left\{ Exp\left(-\frac{2\pi i x}{5}\right) \middle| x \in Z \right\}$ of Z_5 is also isomorphic group to $(B, +)$. i.e. there is the bijection: D ↔ 1, $G \leftrightarrow Exp\left(-\frac{2\pi i}{5}\right)$, $A \leftrightarrow Exp\left(-\frac{4\pi i}{5}\right)$, $U \leftrightarrow Exp\left(-\frac{6\pi i}{5}\right)$ and $C \leftrightarrow Exp\left(-\frac{8\pi i}{5}\right)$. By means of this representation we can insert the set B^3 into the ordinary three-dimensional vector space P^3 and it can be represented as an ordinary cube or regular hexahedron with three of its faces contained in the coordinated planes XY, XZ and YZ (see Fig. 1).

Notice that the cube presented in Fig. 1 encloses the cubic representation of the genetic code, discussed in [8]. Now, the planes XY, XZ and YZ are the subsets of extended triplets $\{XY\text{D}\}$, $\{X\text{D}Z\}$ and $\{\text{D}YZ\}$, where X, $Y \in B$, i.e. in the cube only the codons are located out of the coordinated planes.

A. Geometrical model of the primeval genetic code

The geometric features of the cubic representation of the extended triplet set (see Fig. 1) suggest the highly degenerated nature of a primeval coding apparatus. Let D ∉ {G, A, U, C} represent the base (or bases) with non-specific pairing with at least two bases of the set {G, A, U, C} in the DNA molecule. The non-specific pairing of base D is reflected in the model making it the neutral element of $(B, +)$.

In the vector space of the extended triplets, the subset with non-specific pairing S_{XDZ} = $\{X\text{D}Z\}$ conform a two-dimensional vector subspace, which is contained in the XZ plane. Likewise, in the coordinated planes XY and YZ are contained the bidimensional vector subspaces corresponding to subset with non-specific pairing S_{XYD} = $\{XY\text{D}\}$ and S_{DYZ} = $\{\text{D}YZ\}$, respectively (see Fig. 1). In particular, the quotient space B^3/S_{XDZ} has the elements:

$$\{ S_{XDZ},\ S_{XDZ} + X_G\ ,\ S_{XDZ} + X_A\ ,\ S_{XDZ} + X_U,\ S_{XDZ} + X_C\ \}$$

Where X_G, X_A, X_U and X_C are arbitrary elements of the extended triplet subsets: $\{X\text{G}Y\}$, $\{X\text{A}Y\}$, $\{X\text{U}Y\}$ and $\{X\text{C}Y\}$ respectively. For instance, codons that belong to the subset {XUY} may be represented by some of the sums,

TABLE II
EXTENDED BASE-TRIPLET SET

	No	D	No	G	a[1]	No	A	a	No	U	a	No	C	a	
	0	DDD	25	DGD		50	DAD		75	DUD		100	DCD		D
	1	DDG	26	DGG		51	DAG		76	DUG		101	DCG		G
D	2	DDA	27	DGA		52	DAA		77	DUA		102	DCA		A
	3	DDU	28	DGU		53	DAU		78	DUU		103	DCU		U
	4	DDC	29	DGC		54	DAC		79	DUC		104	DCC		C
	5	GDD	30	GGD		55	GAD		80	GUD		105	GCD		D
	6	GDG	31	GGG	G	56	GAG	E	81	GUG	V	106	GCG	A	G
G	7	GDA	32	GGA	G	57	GAA	E	82	GUA	V	107	GCA	A	A
	8	GDU	33	GGU	G	58	GAU	D	83	GUU	V	108	GCU	A	U
	9	GDC	34	GGC	G	59	GAC	D	84	GUC	V	109	GCC	A	C
	10	ADD	35	AGD		60	AAD		85	AUD		110	ACD		D
	11	ADG	36	AGG	R	61	AAG	K	86	AUG	M	111	ACG	T	G
A	12	ADA	37	AGA	R	62	AAA	K	87	AUA	I	112	ACA	T	A
	13	ADU	38	AGU	S	63	AAU	N	88	AUU	I	113	ACU	T	U
	14	ADC	39	AGC	S	64	AAC	N	89	AUC	I	114	ACC	T	C
	15	UDD	40	UGD		65	UAD		90	UUD		115	UCD		D
	16	UDG	41	UGG	W	66	UAG	-	91	UUG	L	116	UCG	S	G
U	17	UDA	42	UGA	-	67	UAA	-	92	UUA	L	117	UCA	S	A
	18	UDU	43	UGU	C	68	UAU	Y	93	UUU	F	118	UCU	S	U
	19	UDC	44	UGC	C	69	UAC	Y	94	UUC	F	119	UCC	S	C
	20	CDD	45	CGD		70	CAD		95	CUD		120	CCD		D
	21	CDG	46	CGG	R	71	CAG	Q	96	CUG	L	121	CCG	P	G
C	22	CDA	47	CGA	R	72	CAA	Q	97	CUA	L	122	CCA	P	A
	23	CDU	48	CGU	R	73	CAU	H	98	CUU	L	123	CCU	P	U
	24	CDC	49	CGC	R	74	CAC	H	99	CUC	L	124	CCC	P	C

[1]The one-letter symbol of amino acids

$\{XUY\}$ = AUC + S_{XDZ} = AUC + $\{XDY\}$ or $\{XUY\}$ = GUA+ + S_{XDZ} = GUA + $\{XDY\}$

which in particular, for codon AUG, means the translations:

AUG = AUC + DDA or AUG = GUA + GDC

where AUC$\in\{XUY\}$, GUA$\in\{XUY\}$, DDA $\in\{XDY\}$ and GDC $\in\{XDY\}$

The quotient space B^3/S_{XDZ} is a partition of the set of extended triplets into 5 equivalence classes or cosets. Every class has the same number of elements S_{XDZ}, i.e. 25 extended triplets, which correspond to the 5 main columns of Table II. Subsets $S_{XDZ} + X_G = \{XGZ\}$, $S_{XDZ} + X_A = \{XAZ\}$, $S_{XDZ} + X_U = \{XUZ\}$ and $S_{XDZ} + X_C = \{XCZ\}$ are cosets of the vector subspace S_{XDZ} and consequently, they are affine subspaces of the vector space B^3, i.e. they are vertical planes with respect to the horizontal plane XY in Fig. 1. That is, codons that code to amino acids with similar properties belong to the same affine subspace and can be obtained from the extended triplets with non-specific pairing XDZ by means of simple translations. So, if the mutation process is described by means of translations then the geometrical-algebraic features of extended triplet set suggest the transition during the molecular evolution process from a high degenerated primeval code to the present code. The extended triplet of every vertical plane should code, in general, to amino acid with similar physicochemical properties. In particular, the extended triplets with non-specific pairing DX_2D ($X_2 \neq$ D) should code to any amino acids coded by triplets XYZ ($Y = X_2$) from the same vertical plane; while the extended triplets DX_2X_3 (X_2, $X_3 \neq$ D) should code to any amino acid coded by triplets XYZ ($X \neq$ D, $Y = X_2$, $Z = X_3$). The subset extended triplet $\{XDZ\}$ ($X, Z \in B^3$) should not to code to any amino acid.

The subset S_{DDZ} = $\{DDZ\}$ conforms a one-dimensional vector subspace, one of the cube edges, which is inserted in the coordinated Z-axis. So, the S_{DDZ} is a vectorial line, which is generated by the any extended triplets DDZ with $Z \neq$ D. While the quotient vector space B^3/S_{DDZ} of the vector space B^3 is conformed by extended triplet subsets that have fixed the first and the second nucleotides. As can be noticed in Fig. 1, cosets of S_{DDZ} are vertical lines of the cube, which are orthogonal to the plane of XYD extended triplets, i.e. the face inserted in the coordinated plane XY. As a result, there are 25 equivalent classes, having every class 5 extended triplets with the first two bases constant. Such arrangement can be noticed in Table II. It can be noticed that, in most of the cases, codons that code for the same or similar amino acid belong to the same vertical line. For instance,

AUG + S_{DDZ} = {AUD, AUG, AUA, AUU, AUC}
CAG + S_{DDZ} = {CAD, CAG, CAA, CAU, CAC}

for two of these classes (see Table II and Fig. 1). Thus, the extended triplet XYD should code to amino acids coded by codons that belong to its vertical line. This conjecture is based on the non-specific pairing of base D in the third codon position. The first two letters of each codon are the primary determinants of specificity, a feature that has some interesting consequences. Analogue situation can be found in nature. The anticodons in some tRNAs include the nucleotide inosinate, which contains the uncommon base hypoxanthine. Inosinate can form hydrogen bonds with three different nucleotides U,

C, and A.

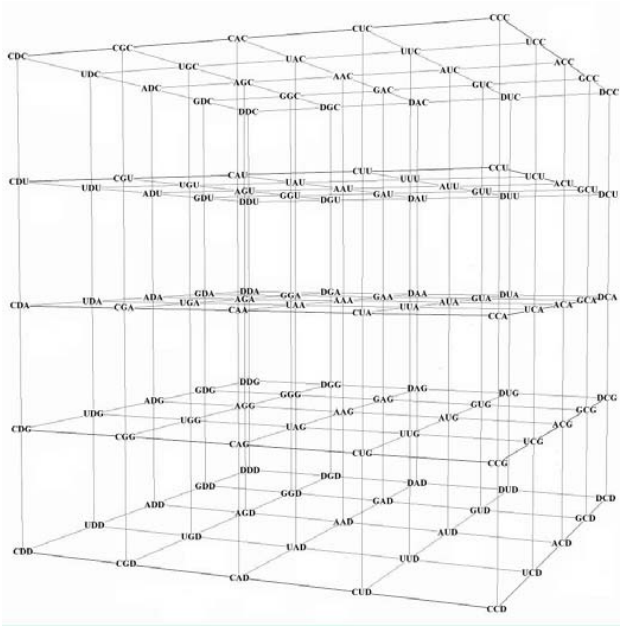

Fig. 1 Cubic representation of the extended genetic code

There are 61 different encoding codons. Organisms do not have, however, the 61 tRNA species with all possible anticodons. In 1966, Crick proposed the famous "wobble hypothesis": through non-Watson-Crick base-pairing rules, less tRNA species are needed [10]. Based on observations of characterized yeast tRNAs and modified base-pairing rules, Guthrie and Abelson updated and revised the wobble hypothesis and predicted that 46 different tRNA species would be found in yeast, and perhaps in all eukaryotes [11]. Now, it is known most Eukaryotic cells with sequenced genome follow the revised wobble hypothesis almost perfectly [12]. In vertebrate mitochondria, however, an unusual set of wobble rules allows the 22 tRNAs to decode all 64 possible codon triplets.

Analogous situations can be derived from the two-dimensional vector subspaces on S_{XYD} and on S_{DYZ}. The geometrical arrangement of their extended triplet are also connected with the coding features of standard genetic code and the codon set can be also derived from these subspaces by means of translations. For instance the horizontal lines {DAG, GAG, AAG, UAG, CAG} and {ADG, AGG, AAG, AUG, ACG} can be obtained from the vectorial lines S_{XDD} and S_{DYD}, respectively, by means of the translations:

$$\mathrm{UAG} + S_{XDD} = \{\mathrm{DAG, GAG, AAG, UAG, CAG}\}$$
$$\mathrm{AUG} + S_{DYD} = \{\mathrm{ADG, AGG, AAG, AUG, ACG}\}$$

In the first case codons code to similar amino acid, which belong to the same column in Table II, and, in the second case, each codon belong to a different column of this table.

The primeval life environment could favor the relative abundance of bases different from G, A, U and C. A gradual origin of a primordial coding apparatus without a loss risk of the ancient "genetic information" could be possible by way of the non-specific pairing of these bases.

The present architecture of the standard genetic code is not in disagreement with the plausible architecture of an ancient genetic code like to those proposed in Table II and in Fig. 1. The gradual origin and improvements of a primitive DNA repair system could make possible the transition from the ancient to the modern genetic code.

B. Algebraic partitions of the extended triplet set and architecture of the modern genetic codes

Likewise to the three-dimensional vector space over P^3, the elements X, $Y \in B^3$ are called collinear if $Y_1Y_2Y_3 = \lambda\, X_1X_2X_3$, where $\lambda \neq \mathrm{D}$. The set of all extended triplets can be sorted into 31 subsets of collinear elements (see Table III). The subsets of collinear extended triplet have the general $S_{Co} = \{X_1X_2X_3, X_{1A}X_{2A}X_{3A}, X_{1U}X_{2U}X_{3U}, X_{1C}X_{2C}X_{3C}\}$, where $X_{i\lambda} = \lambda\, X_i$, $\lambda \in B\backslash\mathrm{D}$ and $X_i \in B$. In particular, the subsets of collinear codons have the general form $S_{CC} = \{X_1GX_3, X_{1A}AX_{3A}, X_{1U}UX_{3U}, X_{1C}CX_{3C}\}$, where $X_{i\lambda} = \lambda\, X_i$ and $\lambda, X_i \in B\backslash\mathrm{D}$. That is, the principal partitions of the standard genetic code $\{\{X_1GX_3\}, \{X_1AX_3\}, \{X_1UX_3\}, \{X_1CX_3\}\}$ are represented in every collinear codon subset. In other words: $X_1GX_3 \in \{X_1GX_3\}$, $X_{1A}AX_{3A} \in \{X_1AX_3\}$, $X_{1U}UX_{3U} \in \{X_1UX_3\}$ and $X_{1C}CX_{3C} \in \{X_1CX_3\}$. Moreover, the sum operation defined in Table I induces symmetry in every collinear subset (see Table III),

$$X_1GX_3 + X_{1C}CX_{3C} = \mathrm{DDD} \text{ and } X_{1A}AX_{3A} + X_{1U}UX_{3U} = \mathrm{DDD}$$

If $X_1X_2X_3 + Y_1Y_2Y_3 = \mathrm{DDD}$, we shall say that the extended triplet $X_1X_2X_3$ and $Y_1Y_2Y_3$ are symmetrical. Likewise, in the N-dimensional vector space $(B^3)^N$ of DNA aligned sequences with length N, we shall say that two DNA sequences $\alpha, \beta \in (\mathrm{B}^3)^N$ are symmetrical if $\alpha + \beta = \{\mathrm{DDD, DDD,..,DDD}\} \in (B^3)^N$.

The set of collinear extended triplet subsets can be sorted into two subsets with specific and non-specific pairing, respectively (see Table III). The subset with specific pairing encloses all codons and every collinear subset contains a codon from a different affine subspace. Next, if we define the product operation between extended triplet as the product coordinated by coordinated then the subset of all codons $S_C = \left\{XYZ \,\middle|\, X,Y,Z \in \{\mathrm{G,A,U,C}\}\right\}$, is closed to product operation, i.e. $(S_C, \bullet)$ is a multiplicative group; while the collinear subset S_{XXX} = {GGG, AAA, UUU, CCC} determine the subgroup $(S_{XXX}, \bullet) \subset (S_C, \bullet)$. The quotient subgroup S_C/S_{XXX} split the codon subset S_C into the subset of collinear codons presented in Table III. That is, every collinear subset $\{XYZ\} \in S_{Co}$ can be represented as

TABLE III
SUBSETS OF COLLINEAR EXTENDED BASE TRIPLETS[1].

Specific pairing (codons)				Non-specific pairing			
GGG	AAA	UUU	CCC	DDG	DDA	DDU	DDC
GGA	AAC	UUG	CCU	GDD	ADD	UDD	CDD
GGU	AAG	UUC	CCA	GDG	ADA	UDU	CDC
GGC	AAU	UUA	CCG	GDA	ADC	UDG	CDU
AGG	CAA	GUU	UCC	GDU	ADG	UDC	CDA
AGA	CAC	GUG	UCU	GDC	ADU	UDA	CDG
AGU	CAG	GUC	UCA	DGD	DAD	DUD	DCD
AGC	CAU	GUA	UCG	DGG	DAA	DUU	DCC
UGG	GAA	CUU	ACC	DGA	DAC	DUG	DCU
UGA	GAC	CUG	ACU	DGU	DAG	DUC	DCA
UGU	GAG	CUC	ACA	DGC	DAU	DUA	DCG
UGC	GAU	CUA	ACG	GGD	AAD	UUD	CCD
CGG	UAA	AUU	GCC	GAD	ACD	UGD	CUD
CGA	UAC	AUG	GCU	GUD	AGD	UCD	CAD
CGU	UAG	AUC	GCA	GCD	AUD	UAD	CGD
CGC	UAU	AUA	GCG				

[1] The subsets have been sorted into subsets of extended triplets with specific and non-specific pairing. In each subset four extended triplet are found.

$$\{XYZ\} = X_1X_2X_3 \bullet S_{XXX}$$

where $X_1X_2X_3$ is an arbitrary element of $\{XYZ\}$. For instance, the subset of collinear codons {AGC, CAU, GUA, UCG} and {CGA, UAC, AUG, GCU} can be written, respectively, as

$$\text{AGC} \bullet S_{XXX} \text{ and } \text{CGA} \bullet S_{XXX}$$

These results suggest that the architecture of the modern genetic codes could be determined from the architecture of an ancient genetic code as those proposed here. Actually, the partition of the modern genetic codes is implicit in the structure (SC, •). The subset of codons $S_{XGZ} = \{XGZ\}$ determine a subgroup $(S_{XGZ}, \bullet) \subset (S_C, \bullet)$.

The quotient subgroup S_C/S_{XGZ} is a partition of the set of codons into 4 equivalence classes:

$$\{S_{XGZ}, XAZ \bullet \text{S}_{XGZ}, XUZ \bullet S_{XGZ}, XCZ \bullet S_{XGZ}\}$$

where, XAZ, XUZ and XCZ are arbitrary elements of the codon subsets: $\{XAZ\}$, $\{XUZ\}$ and $\{XCZ\}$, respectively, with $X, Z \neq$ D. Every class has the same number of elements S_{XGZ}, i.e. 16 codons, which correspond to the 4 main columns of codons Table III. These are the 4 main columns of the modern genetic codes tables (see [9]), where each column enclose codons that code, in general, to similar amino acids. For instance, codons that belong to the set $\{XUZ\}$ may be represented by some of the products,

$$\{XUZ\} = \text{AUC} \bullet S_{XGZ} \subset \text{AUC} + S_{XDZ}$$

or

$$\{XUZ\} = \text{GUA} \bullet S_{XGZ} \subset \text{GUA} + S_{XDZ}$$

which in particular, for codon AUG, means:

$$\text{AUG} = \text{AUC} \bullet \text{GGC or AUG} = \text{GUA} \bullet \text{AGU}$$

where $\text{AUC} \in \{XUZ\}$, $\text{GUA} \in \{XUZ\}$, $\text{GGC} \in \{XGZ\}$ and $\text{AGU} \in \{XGZ\}$

As was pointed out before, codons from the same vertical line code for the same or similar amino acid. Recall that every vertical line encloses five extended triplets one of them having non-specific pairing. The rest of four codons from every vertical line can be getting as classes from the quotient group S_C/S_{GGZ}, where

$$(S_{GGZ}, \bullet) = (\{\text{GGG, GGA, GGU, GGC}\}, \bullet) \subset (S_{XGZ}, \bullet) \subset (S_C, \bullet)$$

As a result, the partition of the modern genetic code tables into 16 subsets of codons that code to the same or similar amino acids is also derived from the ancient architecture proposed here (see [9]). That is, it can be written, for instance,

$$\text{AUG} \bullet S_{GGZ} = \{\text{AUG, AUA, AUU, AUC}\} \subset \text{AUG} + S_{DDZ} = \{\text{AUD, AUG, AUA, AUU, AUC}\}$$

$$\text{CAG} \bullet S_{GGZ} = \{\text{CAG, CAA, CAU, CAC}\} \subset \text{CAG} + S_{DDZ} = \{\text{CAD, CAG, CAA, CAU, CAC}\}$$

Since, the Watson-Crick base pairing and the non-specific base pairing of the hypothetical ancestral base D are the fundamental features to define the operations of sum and product, our results suggest that these features are enough to determine the coding constraints of the primeval and the modern genetic code, as well as the transition from the former to the later.

C. Some biological remarks

The simpler translation machinery of the mammalian mitochondrial genome suggests that the universal code as we understand it might not have existed at the beginning of the life. Ohno and Epplen [13] propose that life started with the simpler mitochondria-like code involving fewer species of tRNAs and, therefore, fewer anticodons of greater infidelity with respect to their codon recognition. The last suggestions still presume that the Watson–Crick pairing of A with U and of G with C is retained as the basis of genetic template recognition and that these bases were readily available on early Earth. Shapiro [14] argued that presumption is not supported by the existing knowledge of the basic chemistry of these substances. Levy and Miller [6] pointed out that the rates of decomposition of the nucleobases A, U, G, C, and T clearly shows that these compounds are not stable on a geologic time scale at temperatures much above 0°C. Even at 25°C, the rates of hydrolysis of the compounds are fast on the geologic time scale. They conclude that unless the origin of life took place extremely rapidly (<100 yr), a high-temperature origin of life may be possible, but it cannot involve adenine, uracil, guanine, or cytosine.

In our model the letter D means any alternative base with non-specific pairing. The newly emerged prebiotic translation

machinery had to cope with base sequences that were not preselected to be coding sequences [13]. Such regions could increase the length of the primeval coding DNA regions. Following Susumu and Epplen, in a primitive mitochondria-like code the number of chain terminator codons in randomly generated 300-base-long sequences can only specify a number of oligopeptides. Obviously, the number of chain terminators in randomly generated 300-base-long sequences must be decreased in the primeval genes with extended triplets. In addition, the free code regions should decrease the frequency of transcription and translation errors. So, it should not be strange that the primeval genes could have small regions with free code. Translation of every primeval mRNA should produce a set of homolog proteins and the most frequent synthesized amino acid sequence should depend on the amino acids cell concentration.

The primordial cell was the "natural lab test" to design the primitive coding sequences starting from large regions of free code with non-specific pairing. Those primordial cells supporting genome architectures with relatively low frequency of translations and transcription errors could have an advantage over those cells with less fidelity in these molecular processes. The regions of free code with non-specific pairing should confer to the primordial cells high mutability capacities. This feature may plausible that the first speciation in the history of life could take place in a relative very short geological time.

The primordial cells enclosing regions of free code with non-specific pairing could come into being the progenotes –the common ancestral forms of the eukaryotes and the two prokaryotic groups: eubacteria and archaea [15]-[16]. The origin and development of the enzymatic mechanisms for recombination and repairs could make plausible the transition from this progenote to the primitive eukaryotes and prokaryotes. The free coding region could have different destinations. In one way, the repair and recombination mechanisms could gradually eliminate the regions of free coding to come into been the primitive prokaryote cells. While, in another way, the repair and recombination mechanisms could gradually replace the free coding regions with bases G, A, T and C to come into been new coding regions or the primitive introns found in eukaryotes cells.

IV. The Power Spectra of Extended DNA Genomic Sequences

The present highly variable DNA regions with insertions and deletion could be the relict finger printing of the ancient DNA regions of free code. The gaps observed in the alignment of the DNA genome sequences resemble the hypothesized ancient bases. If this hypothesis is right then we must expect that the coding pattern observed in the present DNA genome sequences ought to be conserved in the DNA sequences produced in the multiple alignments, where the gaps are taken as the mark of the ancient bases. In order to verify this conjecture we have obtained the Fourier spectrum from the multiple alignments of 12 HIV-1 genomes and 14 mammal mitochondrial genomes. In every aligned sequence the hypothetical base D of our model replaced the gaps.

It has been pointed out that the relative height of the peak at frequency f = 1/3 in the Fourier spectrum is a good discriminator of coding potential [17]-[19]. This feature is the called period-3 property of a DNA sequence and it has been used to detect probable coding regions in DNA sequences. Here, we show evidences that the called period-3 property of DNA sequences it is also present in the aligned coding regions of DNA genome sequences including the gaps, which are replaced by the hypothetical base D.

Thus, we can represent the bases of the DNA aligned sequences by the elements of C_5 (see the beginning of section 3). The discrete Fourier transform of a complex representation of a DNA genomic sequence of length N with parameters a and b is defined to be

$$F(s/N) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} g_n e^{\frac{2\pi i}{N} sn} \quad (1)$$

where s=0,…,N-1 and the frequency $f = s/N$. The power spectrum is given by $S(f) = |F(s/N)|^2$. In Fig 2 the power spectra of the aligned mitochondrial genomes are presented. The hypothetical base D replaced the gaps produced by the multiple alignments. It is notable that the spectrums of those species phylogenetically connected are similar. The power spectra have significant spectral peak at frequency 1/3 and 2/3 not founds in random sequences. In 10100 random sequences, the peak magnitudes at these frequencies are at the white noise level, significantly lesser than the peaks observed in mitochondrial genes. Thus, the extended DNA genomes sequences derived from multiple alignment conserve the coding pattern of original sequences, the period-3 property.

The period-3 property was also found in HIV-1 whole genomics sequences. The power spectra of the three aligned HIV-1 genomes are presented in Fig 3. The three aligned sequences keep the coding pattern with a peak at frequency 2/3. Thus, despite to the non-protein-coding regions (at the beginning and at the end of these sequences) and the superposition of genes, the period-3 property can be detected.

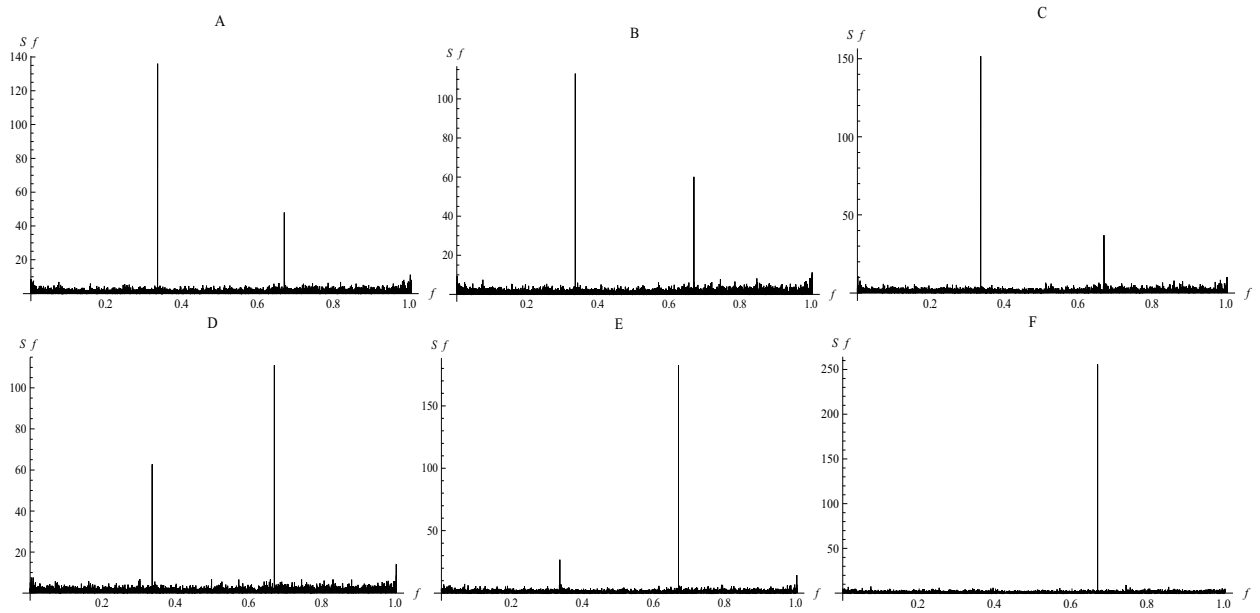

Fig. 2 Power spectra of 13 concatenated and aligned mitochondrial genes of 14 mammals. With similar spectrum A: Homo sapiens (gi|115315570); B: Chimpanzee (gi|5835121), Pygmy chimpanzee (gi|5835135) and Gorilla (gi|5835149; C: Sumatran orangutan (gi|5835834), Orangutan (gi|5835163) and Common gibbon (gi|5835820); D: Bos taurus (gi|60101824), Finback whale (gi|5819095, Blue whale (gi|5834995), Harbor seal (gi|5834857) and Gray seal (gi|5835009); C Mouse (gi|34538597) and D: Opossum (gi|5835037).

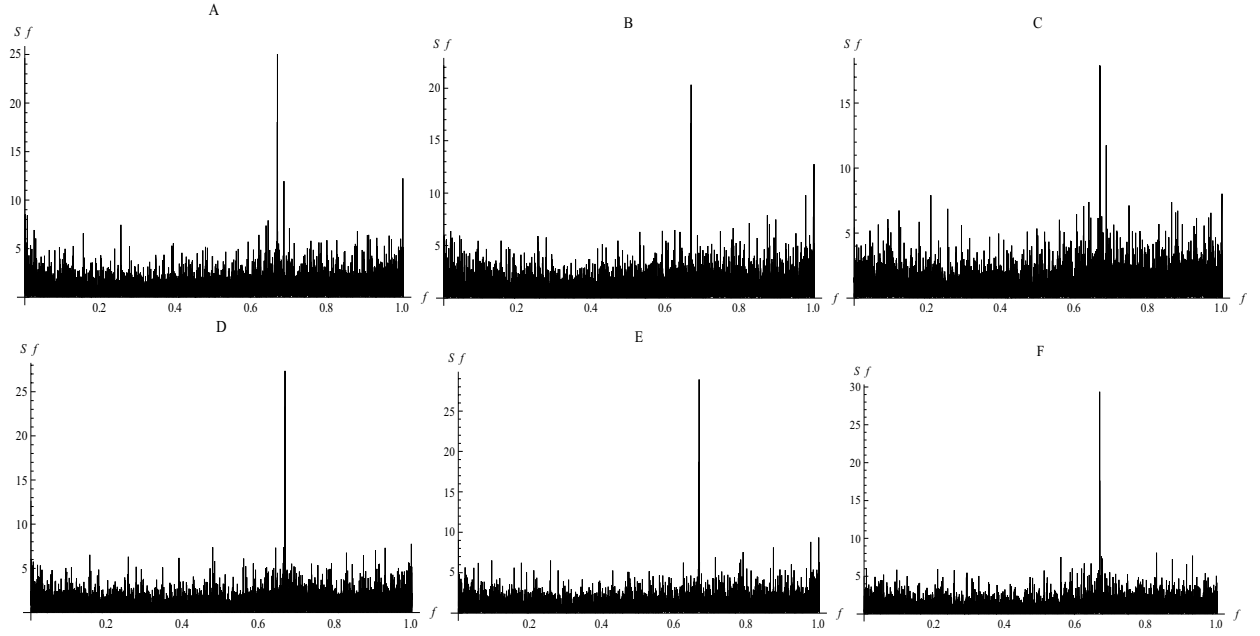

Fig. 3 Power spectra of three aligned HIV-1 whole genomes. A, B and C: the power spectra corresponding to the HIV-1 whole genomes sequences with GenBank accession numbers K03455.1 (HXB2), AB023804.1 and U51188.1, respectively. D, E and F: the power spectra corresponding to the same HIV-1 genomes taken from the base position 700 to 9580.

V. Conclusions

The present genetic code architecture could be derived from an ancient coding apparatus with an extended alphabet of five bases. The Watson-Crick base pairing and the non-specific base pairing of the hypothetical ancestral base D used to define the sum and product operations are enough features to determine the coding constraints of the primeval and the modern genetic code, as well as the transition from the former to the later.

The transition from an ancient DNA coding sequence to the present could be biologically plausible and mathematically determined by an ancient coding apparatus highly degenerated minimizing the transcription and translation errors.

Acknowledgment

. This research was supported within the framework of a VLIR-IUS Collaboration Program.

References

[1] L. E. Orgel, Prebiotic Chemistry and the Origin of the RNA World. Critical Reviews in Biochemistry and Molecular Biology 39 (2004) 99–123,.
[2] J. A. Piccirilli, Krauch T., Moroney, S. E. & Benner, S. A. Nature 343 (1990) 33–37.
[3] C. Switzer, S. E. Moroney. and S. A. Benner, Enzymatic incorporation of a new base pair into DNA and RNA. J. Am. Chem. Sot. I I I (1989) 8322 -8323.
[4] A. Rich, in Horizons in Biochemistry, eds Kasha, M. & Pullman, B. (Academic, New York) (1962) 103–126.
[5] V. M. Kolb, Dworkin, J. P. & Miller, S. L. J. Mol. Evol. 38 (1994) 549–5573.
[6] M. Levy and M. L. Stanley, The stability of the RNA bases: Implications for the origin of life. Proc. Natl. Acad. Sci. USA 95 (1998) 7933–7938
[7] R.. Sánchez, E. Morgado and R. Grau, Gene algebra from a genetic code algebraic structure. J. Math. Biol. 51 (2005) 431 - 457
[8] Sánchez R., Grau R., Morgado E. A Novel Lie Algebra of the Genetic Code over the Galois Field of Four DNA Bases. Mathematical Biosciences 202 (2006) 156–174
[9] R. Sánchez, R. Grau, A Novel Algebraic Structure of the Genetic Code over the Galois Field of Four DNA Bases. Acta Biotheoretica 54 (2006) 27-42
[10] F. Crick, Codon-anticodon pairings: the wobble hypothesis, J. Mol. Biol. 19 (1966) 548.
[11] C. Guthrie and J. Abelson, Organization and expression of tRNA genes in Saccharomyces cerevisiae, in The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression (eds. Strathern J., et al.), New York: Cold Spring Harbor Laboratory Press, (1982) 487.
[12] W. Xiyin, S. Xiaoli and H. Bailin, The transfer RNA genes in Oryza sativa L. ssp. Indica. Science in China C. 45 (2006) 504-511.
[13] S. Ohno and J. T. Epplen, The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proc. Natd Acad. Sci. USA 80 (1983) 3391-3395
[14] R. Shapiro, Prebiotic cytosine synthesis: A critical analysis and implications for the origin of life. Proc. Natl. Acad. Sci. USA 96 (1999) 4396–4401,
[15] C.R. Woese and G. E. Fox. The concept of cellular evolution. J. Mol. Evol. 10 (1962) 1–6.
[16] H.. Hartman and A. Fedorov, The origin of the eukaryotic cell: A genomic investigation. Proc. Natl. Acad. Sci. USA, 99 (2002) 1420–1425
[17] J.W. Fickett. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 10 (1982) 5303-5318
[18] A. A. Tsonis, J. B. Elsner and P.A. Tsonis,.Periodicity in DNA coding sequences: implications in gene evolution. J. Theor. Biol. 151 (1991) 323
[19] T. Shrish, S. Ramachandran, A. Bhattacharya, S. Bhattacharya and R. Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences. CABIOS, 13 (1997) 263-270,

Robersy Sánchez graduated in Biochemistry at the Havana University in 1991. He hold a Master degree in Applied Mathematics at the Universidad Central de Las Villas (UCLV) and a PhD in Biological Sciences at the Havana University. He is assistant professor at the UCLV and lead Bioinformatic group in INIVIT. His research areas include Mathematical and Computational Biology (genome algebra) and Bioinformatics.

Ricardo Grau received his PhD degrees in Mathematical Physic at the UCLV in 1986. Currently, he is a full professor at the UCLV. His research areas include Mathematical and Computational Biology (genome algebra), Bioinformatics, artificial intelligence, pattern.