Extended triplet set C_{343} of DNA sequences and its application to the $p53$ gene*

Yan Yan-Yan(闫艳艳) and Zhu Ping(朱 平)†

School of Science, Jiangnan University, Wuxi 214122, *China*

(Received 16 May 2010; revised manuscript received 9 August 2010)

Recently, much research has indicated that more and more cancers pose a threat to human life. Cancers are caused by oncogenes. Many human oncogenes have been found and most of them are located on chromosomes. The discovery of the oncogene plays a significant role in the treatment of cancer. The $p53$ tumor suppressor gene has received much attention because it frequently mutates or deletes in tumor cells of most people. Thus, the study of oncogenes is significant. In order to establish the Galois field ($GF(7)$), the indefinite gene is introduced as D and oncogene is introduced as O, and P. Taking the polynomial coefficients a_0, a_1, $a_2 \in GF(7)$ and the bijective function $f : GF(7) \to \{\mathrm{D}, \mathrm{A}, \mathrm{C}, \mathrm{O}, \mathrm{G}, \mathrm{T}, \mathrm{P}\}$, where $f(0) = \mathrm{D}$, $f(1) = \mathrm{A}$, $f(2) = \mathrm{C}$, $f(3) = \mathrm{O}$, $f(4) = \mathrm{G}$, $f(5) = \mathrm{T}$, and $f(6) = \mathrm{P}$, the bijective φ may be written as $\varphi(a_0 + a_1x + a_2x^2)$. Based on the algebraic structure, we can not only analyse the DNA sequence of oncogenes, but also predict possible new cancers.

Keywords: oncogene, gene encoding algebraic, polynomial, $p53$

PACS: 87.15.B–, 02.10.Dc **DOI:** 10.1088/1674-1056/20/1/018701

1. Introduction

In bioinformatics, codons have a great effect on genome evolution, protein function and the interaction of heredity and environment. Currently, bioinformatics lacks extensive theoretical study. Therefore, it is important to study general mathematical formulas to explain the structures and trends of proteins in bioinformatics. Many results on DNA sequences have been obtained using algebra and coarse-grained methods.[1−18] Firstly, Sánchez *et al.*[3] proposed a novel Lie algebra of the genetic code over the Galois field (GF) of four DNA bases and they[7] also suggested the genetic code Boolean lattice. The codons make up the Galois field. A vector space and the equivalent class of its quotient spaces can be obtained. According to the Lie algebra of the genetic code over the Galois field of four DNA bases, Sánchez *et al* obtained the physicochemical properties and biological information contained in the DNA sequence. Secondly, in order to establish the evolution model, a new base 'D' was introduced[2] which symbolises one (or more) alternative hypothetical base(s) or dummy variables with non-specific pairings present in primeval RNA and DNA molecules. Then, they pointed out that gene mutations were linear transformations or translations of wild type gene[1] and described the mutation pathways in the molecular evolution process.[1] Finally, they predicted and determined the origin and the evolution of the genetic code. These results show that the human genetic codon can build a deterministic algebraic structure, and the study of this algebraic structure can be used to explore the origins of life.

Recently, many researches have indicated that more and more cancers pose a threat to human life. Cancers seriously affect people at all ages, with increasing age, and they are caused by oncogenes. So far, human beings have been found to have more than 100 kinds of oncogenes. The predecessors of oncogenes are proto-oncogenes. According to their function, proto-oncogenes can be divided into the following three categories [19]. 1) The *src* family, *src*, *abl*, *yes* ect: these belong to the receptor and signaling protein classes. Their product is the growth of signaling proteins in a cell. 2) The *ras* family, *N-ras*, *K-ras*: these belong to G proteins. Their product is the G protein which is present in the cytomembrane and can transmit growth signals. 3) The *myc* family, *myc*, *ski*, etc.: these are transcription factors. Their product is the transcription factor which is present in the nucleus. Oncogenes are widely present in the biosphere. In the process of evolution, these genes are highly conserved. Actually, the proto-oncogene is a normal cell gene. Abnormalities of the amount and the structure result from the effects of carcinogens,

*Project supported in part by the Program for Innovative Research Team of Jiangnan University, China (Grant No. 2008 CX002).
†Corresponding author. E-mail: zptl2002@yahoo.com.cn
© 2011 Chinese Physical Society and IOP Publishing Ltd

such as tobacco smoke, radiation, chemicals, or infectious agents, so that normal cells transmute into tumor cells. Many human oncogenes have been found and most of them are located on chromosomes. The discovery of the oncogenes has played a significant role in the treatment of cancer.

In human beings the $p53$ is encoded by the *TP53* gene located on the short arm of the chromosome 17(17p13.1).[20,21] Since more than 50 percent of human tumors contain a mutation or deletion of the *TP53* gene, the $p53$ tumor suppressor gene has attracted much attention. In recent years, with the study of the $p53$ gene, other family members of the $p53$ have been found. The $p53$ is important in multicellular organisms, where it regulates the cell cycle and thus functions as a tumor suppressor.[22] However, when the $p53$ mutates, it becomes an oncogene. The $p53$ has many anticancer mechanisms, and plays a role in apoptosis, genetic stability, and inhibition of angiogenesis.[19]

The main aim of this paper is to establish an algebraic model and analyse the DNA sequences of oncogenes, which has a positive role in the treatment of cancer. In particular, the oncogene can be considered as the linear translation or transformation of a normal gene.

2. Theoretical model

2.1. Algebraic structure

Recently, the algebraic structure of DNA sequences over the Galois field of 64 codons was reported.[1−7] A new n-dimensional vector space of the DNA sequences on the Galois field of the codons ($GF(125)$) was proposed. In vector space gene mutation can be considered to be the liner transformation or translation of the wild type gene,[1] which allows us to define oncogene algebraic structures similar to those recently published.[1]

In order to analyse the oncogene, the alphabet $\{\mathrm{O,A,C,G,U}\}$[1] and $\{\mathrm{D,A,C,G,U}\}$[2] can be extended. They include the new bases $'\mathrm{O}' : \{\mathrm{G,C}\} \to \{\mathrm{A,T}\}; '\mathrm{P}' : \{\mathrm{A,T}\} \to \{\mathrm{G,C}\}$; $'\mathrm{D}'$: others. The extended DNA alphabet naturally leads us to build a new triplet set with elements $X_1X_2X_3$ where $X_i \in \{\mathrm{D,A,C,O,G,T,P}\}$. This set can be called the extended triplet set C_{343} with the elements $X_1X_2X_3$.

Now, considering the order in the set $\{\mathrm{D,A,C,O,G,T,P}\}$ and the biological importance of the base positions in the codons, it is possible to establish an order in the extend triplet set, i.e. from triplet DDD to PPP. The ordered triplet set can be seen in column III of Table 1.

2.2. Nexus between the Galois field element and the set of codons

In Table 1, a bijective function is suggested between the orders in the extended triplet set and the $GF(7^3)$ elements. In particular, there is a bijective function f: $GF(7) \to \{\mathrm{D,A,C,O,G,T,P}\}$ that is given by the equalities: $f(0) = \mathrm{D}$, $f(1) = \mathrm{A}$, $f(2) = \mathrm{C}$, $f(3) = \mathrm{O}$, $f(4) = \mathrm{G}$, $f(5) = \mathrm{T}$, $f(6) = \mathrm{P}$. Here, the bijective function φ is used to state the biological importance of base positions in the codons. The bijective function φ is between the extended triplet set and the polynomial representation of the $GF(7^3)$ elements. According to Ref. [1], the following formula can be obtained:

$$\begin{aligned}\varphi : \varphi(a_0 + a_1x + a_2x^2) &= (f(a_1)f(a_2)f(a_0)) \\ &= (X_1X_2X_3). \qquad (*)\end{aligned}$$

Note that the polynomial coefficient a_2 of the term with maximal degree a_2x^2 corresponds to the base in the second codon position. Next, the coefficient a_1 corresponds to the first base and finally a_0 corresponds to the third base. That is, the degree of polynomial term decreases from the most to the least biologically important base.

According to Table (1) and septenary digits of integer number, the following formula can be obtained:

$$\begin{aligned}&8 \leftrightarrow 011 \leftrightarrow 110 \leftrightarrow 1 + x \leftrightarrow \mathrm{ADA},\\ &71 \leftrightarrow 131 \leftrightarrow 131 \leftrightarrow 1 + 3x + x^2 \leftrightarrow \mathrm{OAA},\\ &223 \leftrightarrow 436 \leftrightarrow 634 \leftrightarrow 6 + 3x + 4x^2 \leftrightarrow \mathrm{OGP}.\end{aligned}$$

The bijective function $f : S \to GF(7^3)$, where $S \in \{0, 1, \ldots, 342\}$, is defined, then $f[8] = 1 + x$, $f[71] = 1 + 3x + x^2$, and $f[223] = 6 + 3x + 4x^2$ can be obtained.

Now, the product operation and the sum operation between two extended triplets and the product of a codon by the element $\alpha \in GF(7^3)$ can be defined.

Definition 1 For all $X_1Y_1Z_1 \in \mathrm{C}_{343}$, $X_2Y_2Z_2 \in \mathrm{C}_{343}$ and all $\alpha \in GF(7^3)$,

$$X_1Y_1Z_1 \cdot X_2Y_2Z_2 = \varphi[\varphi^{-1}(X_1Y_1Z_1)\varphi^{-1}(X_2Y_2Z_2) \bmod g(x)], \quad (1)$$

$$X_1Y_1Z_1 + X_2Y_2Z_2 = \varphi[(\varphi^{-1}(X_1Y_1Z_1) + \varphi^{-1}(X_2Y_2Z_2)) \bmod 7], \quad (2)$$

$$\alpha X_1Y_1Z_1 = \varphi[\alpha\varphi^{-1}(X_1Y_1Z_1) \bmod 7] \quad (3)$$

can be respectively called the product operation, the sum operation, and the multiplication over C_{343} and $(C_{343}, +, \bullet) \cong GF(7^3)$, where φ is defined by (*).

Table 1. Ordered set of extended triplets corresponding to the elements of $GF(7^3)$.

	D			A			C			O			G			T			P		
	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
D	0	000	DDD	49	001	DAD	98	002	DCD	147	003	DOD	196	004	DGD	245	005	DTD	294	006	DPD
	1	100	DDA	50	101	DAA	99	102	DCA	148	103	DOA	197	104	DGA	246	105	DTA	295	106	DPA
	2	200	DDC	51	201	DAC	100	202	DCC	149	203	DOC	198	204	DGC	247	205	DTC	296	206	DPC
	3	300	DDO	52	301	DAO	101	302	DCO	150	303	DOO	199	304	DGO	248	305	DTO	297	306	DPO
	4	400	DDG	53	401	DAG	102	402	DCG	151	403	DOG	200	404	DGG	249	405	DTG	298	406	DPG
	5	500	DDT	54	501	DAT	103	502	DCT	152	503	DOT	201	504	DGT	250	505	DTT	299	506	DPT
	6	600	DDP	55	601	DAP	104	602	DCP	153	603	DOP	202	604	DGP	251	605	DTP	300	606	DPP
A	7	010	ADD	56	011	AAD	105	012	ACD	154	013	AOD	203	014	AGD	252	015	ATD	301	016	APD
	8	110	ADA	57	111	AAA	106	112	ACA	155	113	AOA	204	114	AGA	253	115	ATA	302	116	APA
	9	210	ADC	58	211	AAC	107	212	ACC	156	213	AOC	205	214	AGC	254	215	ATC	303	216	APC
	10	310	ADO	59	311	AAO	108	312	ACO	157	313	AOO	206	314	AGO	255	315	ATO	304	316	APO
	11	410	ADG	60	411	AAG	109	412	ACG	158	413	AOG	207	414	AGG	256	415	ATG	305	416	APG
	12	510	ADT	61	511	AAT	110	512	ACT	159	513	AOT	208	514	AGT	257	515	ATT	306	516	APT
	13	610	ADP	62	611	AAP	111	612	ACP	160	613	AOP	209	614	AGP	258	615	ATP	307	616	APP
C	14	020	CDD	63	021	CAD	112	022	CCD	161	023	COD	210	024	CGD	259	025	CDD	308	026	CPD
	15	120	CDA	64	121	CAA	113	122	CCA	162	123	COA	211	124	CGA	260	125	CDA	309	126	CPA
	16	220	CDC	65	221	CAC	114	222	CCC	163	223	COC	212	224	CGC	261	225	CTC	310	226	CPC
	17	320	CDO	66	321	CAO	115	322	CCO	164	323	COO	213	324	CGO	262	325	CTO	311	326	CPO
	18	420	CDG	67	421	CAG	116	422	CCG	165	423	COG	214	424	CGG	263	425	CTG	312	426	CPG
	19	520	CDT	68	521	CAT	117	522	CCT	166	523	COT	215	524	CGT	264	525	CTT	313	526	CPT
	20	620	CDP	69	621	CAP	118	622	CCP	167	623	COP	216	624	CGP	265	625	CTP	314	626	CPP
O	21	030	ODD	70	031	OAD	119	032	OCD	168	033	OOD	217	034	OGD	266	035	OTD	315	036	OPD
	22	130	ODA	71	131	OAA	120	132	OCA	169	133	OOA	218	134	OGA	267	135	OTA	316	136	OPA
	23	230	ODC	72	231	OAC	121	232	OCC	170	233	OOC	219	234	OGC	268	235	OTC	317	236	OPC
	24	330	ODO	73	331	OAO	122	332	OCO	171	333	OOO	220	334	OGO	269	335	OTO	318	336	OPO
	25	430	ODG	74	431	OAG	123	432	OCG	172	433	OOG	221	434	OGG	270	435	OTG	319	436	OPG
	26	530	ODT	75	531	OAT	124	532	OCT	173	533	OOT	222	534	OGT	271	535	OTT	320	536	OPT
	27	630	ODP	76	631	OAP	125	632	OCP	174	633	OOP	223	634	OGP	272	635	OTP	321	636	OPP
G	28	040	GDD	77	041	GAD	126	042	GCD	175	043	GOD	224	044	GGD	273	045	GTD	322	046	GPD
	29	140	GDA	78	141	GAA	127	142	GCA	176	143	GOA	225	144	GGA	274	145	GTA	323	146	GPA
	30	240	GDC	79	241	GAC	128	242	GCC	177	243	GOC	226	244	GGC	275	245	GTC	324	246	GPC
	31	340	GDO	80	341	GAO	129	342	GCO	178	343	GOO	227	344	GGO	276	345	GTO	325	346	GPO
	32	440	GDG	81	441	GAG	130	442	GCG	179	443	GOG	228	444	GGG	277	445	GTG	326	446	GPG
	33	540	GDT	82	541	GAT	131	542	GCT	180	543	GOT	229	544	GGT	278	545	GTT	327	546	GPT
	34	640	GDP	83	641	GAP	132	642	GCP	181	643	GOP	230	644	GGP	279	645	GTP	328	646	GPP
T	35	050	TDD	84	051	TAD	133	052	TCD	182	053	TOD	231	054	TGD	280	055	TTD	329	056	TPD
	36	150	TDA	85	151	TAA	134	152	TCA	183	153	TOA	232	154	TGA	281	155	TTA	330	156	TPA
	37	250	TDC	86	251	TAC	135	252	TCC	184	253	TOC	233	254	TGC	282	255	TTC	331	256	TPC
	38	350	TDO	87	351	TAO	136	352	TCO	185	353	TOO	234	354	TGO	283	355	TTO	332	356	TPO
	39	450	TDG	88	451	TAG	137	452	TCG	186	453	TOG	235	454	TGG	284	455	TTG	333	456	TPG
	40	550	TDT	89	551	TAT	138	552	TCT	187	553	TOT	236	554	TGT	285	555	TTT	334	556	TPT
	41	650	TDP	90	651	TAP	139	652	TCP	188	653	TOP	237	654	TGP	286	655	TTP	335	656	TPP
P	42	060	PDD	91	061	PAD	140	062	PCD	189	063	POD	238	064	PGD	287	065	PTD	336	066	PPD
	43	160	PDA	92	161	PAA	141	162	PCA	190	163	POA	239	164	PGA	288	165	PTA	337	166	PPA
	44	260	PDC	93	261	PAC	142	262	PCC	191	263	POC	240	264	PGC	289	265	PTC	338	266	PPC
	45	360	PDO	94	361	PAO	143	362	PCO	192	363	POO	241	364	PGO	290	365	PTO	339	366	PPO
	46	460	PDG	95	461	PAG	144	462	PCG	193	463	POG	242	464	PGG	291	465	PTG	340	466	PPG
	47	560	PDT	96	561	PAT	145	562	PCT	194	563	POT	243	564	PGT	292	565	PTT	341	566	PPT
	48	660	PDP	97	661	PAP	146	662	PCP	195	663	POP	244	664	PGP	293	665	PTP	342	666	PPP

Column I is the triplet index number, column II the polynomical coefficient and column III the extended triplet.

Notes: (1) The product between two triplets is obtained from the product of their corresponding polynomial module $g(x)$ values, where $g(x)$ is an irreducible polynomial of second degree over $GF(7)$. Since there are many irreducible polynomials of second degree, we have many possible options of choosing the product between two extend triplets. (2) The summation is carried out by means of the polynomial sum by module 7. (3) This operation is analogous to the multiplication rule of a vector by a scalar.

Theorem 1 The set of codons $C_{343}\backslash\{DDD\} = C^*_{343}$ with the operation product '$\bullet$' is an Abelian group$(C^*_{343}, \bullet)$, which is not difficult to prove.

Theorem 2 The set of codons $(C_{343}, +)$ with '+' is an Abelian group and $(C_{343}, +) \cong (Z_7)^3$.

So, C_{343} is considered to be a one-dimensional vector space over $GF(7^3)$. The canonical base of this space is the triplet DDA. This structure can be referred to as the space of extended triplets over $GF(7^3)$. Such a structure can be extended to N-dimensional sequence space S, and it is composed of the set all 343^N DNA alignment sequences with N extended triplets. Obviously, it is isomorphic to the set of all N-tuples$(x_1, \ldots, x_N)$, where $x_i \in C_{343}$.

Next, set S can be represented by all N-tuples $(x_1, \ldots, x_N)$. As a result, the N-dimensional vector space of S over $GF(7^3)$ will be the direct sum

$$S = (C_{343})^N = C_{343} \oplus C_{343} \oplus \cdots \oplus C_{343}(N \text{ times}).$$

The sum and the product in S are carried out by components,[23] therefore, for all $s, s' \in S, \alpha \in GF(7^3)$, there are

$$\begin{aligned} s + s' &= (s_1, \ldots, s_N) + (s'_1, \ldots, s'_N) \\ &= (s_1 + s'_1, \ldots, s_N + s'_N) \end{aligned}$$

$$\alpha s = \alpha(s_1, \ldots, s_N) = (\alpha s_1, \ldots, \alpha s_N). \quad (4)$$

It can be proved that $(S, +)$ is an Abelian group with (DDD,DDD, ... DDD) as its neutral elements. The canonical base of this space is the set of vectors: e_1 = (DDA, DDD, ..., DDD), e_2 = (DDD, DDA, ..., DDD), ..., and e_N = (DDD, DDD, ..., DDA) As a result, for any $s \in S$ there is a unique representation

$$s = a_1e_1 + a_2e_2 + \ldots a_Ne_N(a_i \in GF(7^3)).$$

Theorem 3 There are 112 irreducible polynomials and 36 primitive polynomials over $GF(7)$.

Since there are many irreducible polynomials of second degree, the Galois field of codons is not unique. Actually, the 112 isomorphic Galois fields of codons have been obtained, each with the product operation defined from one of the 112 irreducible polynomials. The most attractive irreducible polynomials are the primitive polynomials. Through the following calculation, 112 irreducible polynomials and 36 primitive polynomials over $GF(7)$ can be obtained.

The number of irreducible polynomials over $GF(7)$ is

$$L_7(3) = \frac{1}{3}\sum_{d/3} \mu\left(\frac{3}{d}\right)7^d = 112,$$

where $\mu(n)$ is a Möbius function.[24]

The number of primitive polynomials over $GF(7)$ is

$$a_7(3) = \frac{\Phi(7^3 - 1)}{3} = 36,$$

where $\Phi(n)$ is a totient function.[24]

According to the algorithm[25] all second degree primitive polynomials over $GF(7)$ can be obtained (see Table 2).

Table 2. All second degree primitive polynomials over $GF(7)$.

1	x^3+3x+2	13	x^3+3x^2+x+4	25	x^3+5x^2+3x+2
2	x^3+5x+2	14	x^3+3x^2+2x+2	26	x^3+5x^2+3x+4
3	x^3+6x+2	15	x^3+3x^2+3x+4	27	x^3+5x^2+4x+2
4	x^3+x^2+x+2	16	x^3+3x^2+5x+2	28	x^3+5x^2+5x+2
5	x^3+x^2+2x+4	17	x^3+3x^2+5x+4	29	x^3+5x^2+6x+4
6	x^3+x^2+5x+2	18	x^3+3x^2+6x+2	30	x^3+6x^2+4
7	x^3+x^2+5x+4	19	x^3+4x^2+2x+2	31	x^3+6x^2+x+2
8	x^3+2x^2+x+4	20	x^3+4x^2+3x+2	32	x^3+6x^2+3x+2
9	x^3+2x^2+4x+2	21	x^3+4x^2+3x+4	33	x^3+6x^2+4x+4
10	x^3+2x^2+6x+2	22	x^3+4x^2+4x+4	34	x^3+6x^2+5x+4
11	x^3+2x^2+6x+4	23	x^3+5x^2+4	35	x^3+6x^2+6x+2
12	x^3+3x^2+4	24	x^3+5x^2+2x+4	36	x^3+6x^2+6x+4

2.3. Oncogene linear transformation

Theorem 4[1] For every single point mutation changing the codons α_i of the wild-type gene $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots, \alpha_N)$ (α is different from the null vector) by the codon β_i of the mutant gene $\beta = (\alpha_1, \alpha_2, \ldots, \beta_i, \ldots, \alpha_N)$, there is

(i) at least one local endomorphism f such that $f(\alpha) = \beta$;

(ii) at least one local automorphism f such that $f(\alpha) = \beta$;

(iii) a unique diagonal automorphism f such that $f(\alpha) = \beta$, if and only if the codons α_i and β_i of the wild-type gene and the mutant gene, respectively, have different GGG.

Then we have the following corollaries.

Corollary 1 For normal gene α and mutation gene β of the $p53$, there is at least one local endomorphism or automorphism f such that $f(\alpha) = \beta$.

Corollary 2 From the normal gene sequence α to the mutation gene sequence β of the $p53$, there is a unique diagonal automorphism g such that $g(\alpha) = \beta$, if and only if DDD $\notin \alpha$ and DDD $\notin \beta$.

3. Method and results

3.1. Method

Here, in order to obtain all codons from the codons AAO, those primitive polynomials with root $\alpha = x^2 + x + 3$ corresponding to codon AAO can be chosen. Only the primitive polynomial $g(x) = x^3 + 5x^2 + 3x + 4$ has this root, and in this way the product operation is unique.

By definition, $\alpha = x^2 + x + 3$ is root of $g(x) = x^3 + 5x^2 + 3x + 4$ if and only if, $g(\alpha) = 0$.

Note: $g(\alpha) = 0$, i.e., $[g(\alpha) \bmod g(x)] \bmod 7 = 0$.

The product operation $X_1Y_1Z_1 \cdot X_2Y_2Z_2 = \varphi[\varphi^{-1}(X_1Y_1Z_1)\varphi^{-1}(X_2Y_2Z_2) \bmod g(x)]$ is a method to obtain all codons from the codon AAO (see Table 3).

Table 3. $GF(7^3)$ generated by the primitive polynomial $g(x) = x^3 + 5x^2 + 3x + 4$.

$f[s]^*$	$f[1]$	$f[2]$	$f[3]$	$f[4]$	$f[5]$	$f[6]$	$f[7]$	$f[8]$	$f[9]$	$f[10]$	$f[11]$	$f[12]$	$f[13]$
n	0	114	57	228	285	171	49	260	208	60	93	10	222
$f[s]$	$f[14]$	$f[15]$	$f[16]$	$f[17]$	$f[18]$	$f[19]$	$f[20]$	$f[21]$	$f[22]$	$f[23]$	$f[24]$	$f[25]$	$f[26]$
n	163	207	32	124	322	336	174	106	67	117	317	279	150
$f[s]$	$f[27]$	$f[28]$	$f[29]$	$f[30]$	$f[31]$	$f[32]$	$f[33]$	$f[34]$	$f[35]$	$f[36]$	$f[37]$	$f[38]$	$f[39]$
n	265	277	94	321	108	146	288	238	334	3	165	151	295
$f[s]$	$f[40]$	$f[41]$	$f[42]$	$f[43]$	$f[44]$	$f[45]$	$f[46]$	$f[47]$	$f[48]$	$f[49]$	$f[50]$	$f[51]$	$f[52]$
n	203	36	220	51	181	264	231	37	89	98	335	79	218
$f[s]$	$f[53]$	$f[54]$	$f[55]$	$f[56]$	$f[57]$	$f[58]$	$f[59]$	$f[60]$	$f[61]$	$f[62]$	$f[63]$	$f[64]$	$f[65]$
n	198	153	140	309	70	186	1	110	88	78	257	178	105
$f[s]$	f [66]	f [67]	f [68]	$f[69]$	$f[70]$	$f[71]$	$f[72]$	$f[73]$	$f[74]$	$f[75]$	$f[76]$	$f[77]$	$f[78]$
n	319	282	33	103	109	80	126	315	20	14	341	142	62
$f[s]$	f [79]	f [80]	f [81]	$f[82]$	$f[83]$	$f[84]$	$f[85]$	$f[86]$	$f[87]$	$f[88]$	$f[89]$	$f[90]$	$f[91]$
n	232	320	74	133	196	59	102	215	8	11	75	268	271
$f[s]$	$f[92]$	$f[93]$	$f[94]$	$f[95]$	$f[96]$	$f[97]$	$f[98]$	$f[99]$	$f[100]$	$f[101]$	$f[102]$	$f[103]$	$f[104]$
n	301	120	169	294	270	266	212	312	107	267	193	254	332
$f[s]$	f [105]	f [106]	f [107]	f [108]	f [109]	f [110]	f [111]	f [112]	f [113]	f [114]	f [115]	f [116]	f [117]
n	256	188	76	247	4	310	92	81	224	184	202	300	192
$f[s]$	f [118]	f [119]	f [120]	f [121]	f [122]	f [123]	f [124]	f [125]	f [126]	f [127]	f [128]	f [129]	f [130]
n	115	173	125	216	189	329	40	122	29	54	292	147	219
$f[s]$	f [131]	f [132]	f [133]	f [134]	f [135]	f [136]	f [137]	f [138]	f [139]	f [140]	f [141]	f [142]	f [143]
n	217	91	43	66	73	42	234	38	283	223	134	194	128
$f[s]$	f [144]	f [145]	f [146]	f [147]	f [148]	f [149]	f [150]	f [151]	f [152]	f [153]	f [154]	f [155]	f [156]
n	240	113	87	155	210	275	50	197	255	136	116	132	65
$f[s]$	f [157]	f [158]	f [159]	f [160]	f [161]	f [162]	f [163]	f [164]	f [165]	f [166]	f [167]	f [168]	f [169]
n	159	325	68	272	166	71	30	137	56	77	183	24	145

Table 3. (*Continued*).

$f[s]$*	$f[1]$	$f[2]$	$f[3]$	$f[4]$	$f[5]$	$f[6]$	$f[7]$	$f[8]$	$f[9]$	$f[10]$	$f[11]$	$f[12]$	$f[13]$
$f[s]$	f [170]	f 171]	f [172]	f [173]	f [174]	f [175]	f [176]	f [177]	f [178]	f [179]	f [180]	f [181]	f [182]
n	58	127	135	167	243	328	327	226	16	323	9	177	199
$f[s]$	f [183]	f [184]	f [185]	f [186]	f [187]	f [188]	f [189]	f [190]	f [191]	f [192]	f [193]	f [194]	f [195]
n	190	35	119	253	131	289	314	90	34	235	160	339	162
$f[s]$	f [196]	f [197]	f [198]	f [199]	f [200]	f [201]	f [202]	f [203]	f [204]	f 205]	f [206]	f [207]	f [208]
n	326	307	84	26	221	104	39	143	333	168	331	64	205
$f[s]$	f [209]	f [210]	f [211]	f [212]	f [213]	f [214]	f [215]	f [216]	f [217]	f [218]	f [219]	f [220]	f [221]
n	261	28	118	302	82	290	206	197	157	6	180	152	187
$f[s]$	f [222]	f [223]	f [224]	f [225]	f [226]	f [227]	f [228]	f [229]	f [230]	f [231]	f [232]	f [233]	f [234]
n	55	156	195	72	338	306	298	229	316	337	12	248	227
$f[s]$	f [235]	f [236]	f [237]	f [238]	f [239]	f [240]	f [241]	f [242]	f [243]	f [244]	f [245]	f [246]	f [247]
n	308 f 22	201	242	287	101	239	154	330	236	303	41	161	83
$f[s]$	f 248]	f [249]	f [250]	f [251]	f [252]	f [253]	f [254]	f [255]	f [256]	f [257]	f [258]	f [259]	f [260]
n	22	96	278	141	52	258	284	69	299	23	305	214	112
$f[s]$	f [261]	f [262]	f [263]	f [264]	f [265]	f [266]	f [267]	f [268]	f [269]	f [270]	f [271]	f [272]	f [273]
n	209	63	213	244	237	200	262	46	48	318	121	225	2
$f[s]$	f [274]	f [275]	f [276]	f [277]	f [278]	f [279]	f [280]	f [281]	f [282]	f [283]	f [284]	f [285]	f [286]
n	293	211	158	18	45	296	252	286	21	129	31	13	53
$f[s]$	f [287]	f [288]	f [289]	f [290]	f [291]	f [292]	f [293]	f [294]	f [295]	f [296]	f [297]	f [298]	f [299]
n	85	263	139	175	76	5	17	269	311	324	27	47	250
$f[s]$	f [300]	f [301]	f [302]	f [303]	f [304]	f [305]	f [306]	f [307]	f [308]	f [309]	f [310]	f [311]	f [312]
n	164	100	95	99	123	340	291	130	230	97	246	182	179
$f[s]$	f [313]	f [314]	f [315]	f [316]	f [317]	f [318]	f [319]	f [320]	f [321]	f [322]	f [323]	f [324]	f [325]
n	44	273	313	25	304	245	149	61	233	280	170	185	191
$f[s]$	f [326]	f [327]	f [328]	f [329]	f [330]	f [331]	f [332]	f [333]	f [334]	f [335]	f [336]	f [337]	f [338]
n	144	297	251	86	274	204	111	148	276	7	138	249	259
$f[s]$	f [339]	f [340]	f [341]	f [342]									
n	281	172	15	241									

* $f[s] \to n_s$, $\begin{cases} f[x] \cdot f[y] \to (n_x + n_y) \mod 342; \\ \alpha \cdot f[x] \to (\alpha \cdot n_x) \mod 342; \\ f[x]/f[y] \to (n_x - n_y) \mod 342. \end{cases}$

The algorithm is given as follows.

```
Function model=F ()
m=341;
f=cell (m, 1);
q=cell (m, 1);
r=cell (m, 1);
Res=cell (m, 1);
RRes=cell (m, 1);
f{1}=[1,1,3];
RRes{1}=f{1};
g= [1, 5, 3, 4];
for i=2:341
f{i}=conv(RRes{i-1},f{1});
[q{i},r{ i } ]=deconv(f{i},g)
Res{i}=mod(r{i},7);
[x(i),y(i)]=size(Res{i});
RRes{i}=Res{i}(y(i)-2:y(i));
end
model=RRes;
```

Table 2 of Ref. [1] can be obtained using the above algorithm.

3.2. Results

When the $p53$ gene mutates, it becomes an oncogene. According to corollary 1, we can consider, for example, the missense mutation of the $p53$ in ovarian cancer, esophageal cancer, lung cancer and B-cell lymphoma.

According to Tables 4–7, the following results can be obtained.

(i) The majority of the $p53$ missense mutations are present in the second base of the codon.

(ii) The majority of the $p53$ missense mutations are of {G, C}→{A, T} transition.

Table 4. Missense mutations of the $p53$ in ovarian cancer.

codon[26]	nucleotide[26]	In$GF(7^3)$\$	value in Z_{343} $x \to y$	$f[s]*$
135	TGC→TAC	TGC→TOC	233→184	f[283]
163	TAC→TGC	TAC→TPC	86→331	f[206]
165	CAG→CGG	CAG→CPG	67→312	f[240]
175	CGC→CAC	CGC→COC	212→163	f[57]
176	TGC→CGC	TGC→PGC	233→240	f[204]
192	CAG→TAG	CAG→OAG	67→74	f[71]
195	ATC→ACC	ATC→APC	254→303	f[217]
214	CAT→CGT	CAT→CPT	68→313	f[88]
220	TAT→TGT	TAT→TPT	89→334	f[236]
234	TAC→TGC	TAC→TPC	86→331	f[206]
238	TGT→TAT	TGT→TOT	236→187	f[160]
245	GGC→AGC	GGC→OGC	226→219	f[114]
248	CGG→TGG	CGG→OGG	214→221	f[240]
251	ATC→AGC	ATC→APC	254→303	f[217]
273	CGT→CAT	CGT→COT	215→166	f[263]
277	TGT→TTT	TGT→TOT	236→187	f[160]
342	CGA→TGA	CGA→OGA	211→218	f[308]

\$ 'O':{G, C} → {A, T}; 'P':{A, T} → {G, C}; 'D': others * $f[y] \to n_y = (n_x + n_s) \mod 342 \to f[x] \cdot f[s]$

Table 5. Missense mutations of the $p53$ in esophageal cancer.

codon[27,28]	nucleotide[27,28]	In$GF(7^3)$\$	value in Z_{343} $x \to y$	$f[s]*$
152	CCG→CTG	CCG→COG	116→165	f[49]
154	GGC→GTC	GGC→GOC	226→177	f[308]
192	CAG→TAG	CAG→OAG	67→74	f[71]
213	CGA→TGA	CGA→OGA	211→218	f[308]
245	GGC→GTC	GGC→GOC	226→177	f[308]
248	CGG→CAG	CGG→COG	214→165	f[31]
258	GAA→AAA	GAA→OAA	78→71	f[277]
273	CGT→CAT	CGT→COT	215→166	f[263]
278	CCT→CTT	CCT→COT	117→166	f[234]
278	CCT→TCT	CCT→OCT	117→124	f[183]
282	CGG→TGG	CGG→OGG	214→221	f[240]

\$ 'O':{G, C} → {A, T}; 'P':{A, T} → {G, C}; 'D': others * $f[y] \to n_y = (n_x + n_s) \mod 342 \to f[x] \cdot f[s]$

Table 6. Missense mutations of the $p53$ in lung cancer.

codon[29]	nucleotide[29]	In$GF(7^3)$\$	value in Z_{343} $x \to y$	$f[s]*$
180	GAG→TAG	GAG→OAG	81→74	f[33]
157	GTC→TTC	GTC→OTC	275→268	f[181]
141	TGC→TAC	TGC→TOC	233→184	f[283]
278	CCT→CGT	CCT→CDT	117→19	f[326]
132	AAG→AGG	AAG→APG	60→305	f[308]
179	CAT→CTT	CAT→CDT	68→19	f[244]
249	AGG→GGG	AGG→PGG	207→242	f[97]
266	GGA→AGA	GGA→OGA	225→218	f[334]
152	CCG→CTG	CCG→COG	116→165	f[49]
130	CTC→GTC	CTC→DTC	261→247	f[121]
213	CGA→CTG	CGA→COG	211→165	f[322]
245	GGC→GTC	GGC→GOC	226→177	f[308]
273	CGT→CAT	CGT→COT	215→166	f[263]
122	GTT→GTG	GTT→GTP	278→279	f[328]

\$ 'O':{G, C} → {A, T}; 'P':{A, T} → {G, C}; 'D': others * $f[y] \to n_y = (n_x + n_s) \mod 342 \to f[x] \cdot f[s]$

Table 7. Missense mutations of the $p53$ in B-cell lymphomas.

codon[30]	nucleotide[30]	In$GF(7^3)$\$	value in Z_{343} $x \to y$	$f[s]*$
135	TGC→TTC	TGC→TOC	233→184	$f[283]$
141	TGC→CGC	TGC→PGC	233→240	$f[204]$
157	GTC→GGC	GTC→GPC	275→324	$f[230]$
173	GTG→TTG	GTG→OTG	277→270	$f[116]$
175	CGC→CAC	CGC→COC	212→163	$f[57]$
179	CAT→CAA	CAT→CAD	68→63	$f[113]$
245	GGC→AGC	GGC→OGC	226→219	$f[114]$
245	GGC→GAC	GGC→GOC	226→177	$f[308]$
246	ATG→GTG	ATG→PTG	256→291	$f[185]$
248	CGG→TGG	CGG→OGG	214→221	$f[240]$
273	CGT→TGT	CGT→OGT	215→222	$f[325]$
282	CGG→TGG	CGG→OGG	214→221	$f[240]$
282	CGG→CAG	CGG→COG	214→165	$f[31]$
294	CAG→TAG	CAG→OAG	67→74	$f[71]$

\$ 'O':{G, C} → {A, T}; 'P':{A, T} → {G, C}; 'D': others * $f[y] \to n_y = (n_x + n_s) \mod 342 \to f[x] \cdot f[s]$

According to corollary 2, we can consider, for example, the segment DNA sequences of human $p53$ codon 277–287:[31]

$$
\begin{aligned}
a &= \mathrm{UGUCCUGGGAGAGACCGGCGCACAGAGGAAGAG},\\
b &= \mathrm{GAGGUUA++GUC+++UCGAUCCAAAUCACCAUC},\\
c &= \mathrm{CCUGAUUAG++UAG++++GCCCUACCUCGUGGU},\\
d &= \mathrm{UUUAGUCUGAAAUCGGAGUACACAGG++++GCG},\\
e &= \mathrm{UGUUCUGGGACACU++UGCGCACAGAG++AGAG},
\end{aligned}
$$

where a is normal gene sequence, according to the algebraic structure,

$$
\begin{aligned}
b' &= \mathrm{POPPOUPDDOPPDDDOPGPPCPPAOPPOPPOPP},\\
c' &= \mathrm{OPUPPUPOGDDPOODDDDPPCPOAPPPPOPGOP},\\
d' &= \mathrm{UPUPPUPPGAPAPPPPOGOOCACAGODDDDGPG},\\
e' &= \mathrm{UGUOCOGGGAPAPPDDPGCGCACAGAGDDAGAG}
\end{aligned}
$$

can be obtained. Here, 'D' denotes a dummy variable with non-specific pairings present in RNA and DNA molecules; 'O' represents pyrimidine-to-pyrimidine and purine-to-purine, and 'P' refers to pyrimidine-to-purine and purine-to-pyrimidine.

According to observation, the diagonal automorphism of the gene sequence which contains DDD is not unique. For sequence e, there is a unique diagonal automorphism mapping. According to Table 1 these sequences correspond, in the vector space $(\mathrm{C}_{343})^{10}$, to vectors $a = \left[f[236]\ f[117]\ \ldots\ f[81] \right]$ $e' = \left[f[236]\ f[122]\ \ldots\ f[81] \right]$. Hence, according to Table 3 the following expression can be obtained:

$$
a = e' \begin{bmatrix} f[1] & & & \\ & f[122] & & \\ & & \ddots & \\ & & & f[1] \end{bmatrix}.
$$

We have the following conclusions.

(i) Linear transformation results show that the diagonal automorphism from the oncogene sequence which contains codons ranging from DDD to the normal sequence exists but is not unique.

(ii) Linear transformation results show that the oncogene sequence which does not contain codon DDD can be mapped into the normal gene. The diagonal automorphism is unique.

(iii) With different diagonal automorphisms, different gene sequences can be obtained. The new oncogene can also be predicted.

For human $p53$ codon 277–287, the missense mutation has been studied. A genetic change involves the substitution of one base in the DNA for another which results in the substitution of one amino acid in a polypeptide for another. A missense mutation is a 'readable' genetic message although its 'sense' (its meaning) is changed.[32] Beside missense mutation, there are other mutations such as nonsense mutations, synonymous mutations and frameshift mutations. This method can be used to study the cases mentioned above.

4. Conclusions

The algebraic structure of oncogene has been established using the nature extension of the DNA or RNA alphabet. The Galois field of the extended triplet set is defined using the bijection, and the bijection is between the extended triplet set and the polynomial representation of $GF(7^3)$ elements. Over this new field, the oncogene can be considered as the linear transformation or translation of a normal gene. It is proved that for all oncogenes that do not include DDD there is a unique diagonal automorphism that transforms the normal gene. Besides these, the theoretical results in this paper will positively be conducive to the treatment of cancer. This algebraic structure can be used to study the $p63$ and the $p73$.

It is now well recognized that genetic polymorphisms in drug-metabolizing enzymes, transporters, receptors and other drug targets are related to interindividual differences in the efficacy and toxicity of many medications. These conclusions are important for developing the relationship between the genetic polymorphisms and the diversity of clinical phenotypes. The algebraic structure for drug disposition and response should make it possible to select many medications and their dosages on the basis of each patient's inherited ability to metabolize, eliminate and respond to specific drugs. The relationship between phenotype and genotype is fundamental to the understanding of the mechanism of the disease. According to conclusions (i) and (ii), the linear transformations between genotype and phenotype can be built and studied.

The new oncogene can be discovered from conclusion (iii), so new diseases can be prevented. In addition, this algebraic structure can also be used to study the biological evolution.

References

[1] Sánchez R, Perfetti L A, Grau R and Morgado E 2005 *MATCH Commun. Math. Comput. Chem.* **54** 3
[2] Sánchez R and Grau R 2009 *Math. Biosci.* **221** 60
[3] Sánchez R, Grau R and Morgado E 2006 *Math. Biosci.* **202** 156
[4] Hornos J E and Hornos Y M 2006 *Phys. Rev. Lett.* **71** 4401
[5] Sánchez R and Grau R 2006 *Acta Biotheor.* **54** 27
[6] Sánchez R, Morgado E and Grau R 2005 *J. Math. Biol.* **51** 431
[7] Sánchez R, Morgado E and Grau R 2004 *MATCH Commun. Math. Comput. Chem.* **52** 29
[8] Zhu P, Tang X Q and Xu Z Y 2009 *Chin. Phys.* B **18** 363
[9] Tang X Q, Zhu P and Cheng J X 2010 *Pattern Recognition* **43** 3768
[10] Zhu P, Gao L and Xu Z Y 2009 *Acta Phys. Sin.* **58** 4295 (in Chinese)
[11] Liu Q, Tang C and Ou Y Q 2010 *Chin. Phys.* B **19** 040202
[12] Wang Q H, Zhang Y Y, Lai J C, Li Z H and He A Z 2007 *Acta Phys. Sin.* **56** 1203 (in Chinese)
[13] Liu X F and Wang Y 2009 *Chin. Phys.* B **18** 2690
[14] Bashford J D and Jarvis P D 2000 *Biosystems* **57** 147
[15] Jako E, Ari E, Ittzes P, Horvath A and Podani J 2009 *Molecular Phylogenetics and Evolution* **52** 887
[16] José M V, Morgado Ey R and Govezensky T 2005 *Bull. Math. Biol.* **67** 1
[17] Piccirilli J A., Krauch T, Moroney S E and Beener S A 1990 *Nature* **343** 33
[18] Switzer C Y, Moroney S E and Beener S A 1989 *J. Am. Chem. Soc.* **111** 8322
[19] Jiang P Z, Shen X M and Huang H 2001 *Volumes of International Oncology* **28** 89 (in Chinese)
[20] Isobe M, Emanuel B S, Givol D, Oren M and Croce C M 1986 *Nature* **6057** 84
[21] Matlashewski G, Lamb P, Pim D, Peacock J, Crawford L and Benchimol S 1984 *Embo. J.* **13** 3257
[22] http://en.wikipedia.org/wiki/P53#Additional_images 2010
[23] Redei L 1967 *Algebra Budapest: Akademiai Kiado* vol.1
[24] http://mathworld.wolfram.com/PrimitivePolynomial.html 2010
[25] http://wims.unice.fr/wims/wims.cgi?session=JM7F6CF054.2&+lang=cn&+module=tool%2Falgebra%2Fprimpoly.cn&+cmd=reply&+job=menu 2007
[26] Kupryjanczyk J, Bell D A, Dimeo D, Beauchamp R, Thor A D and Yandell D W 1993 *Medical Sciences* **99** 4961
[27] Hollstein M C, Metcalf R A and Welsh J A 1990 *Medical Sciences* **87** 9958
[28] Biramijamal F 2005 *Journal of Sciences* **16** 3
[29] Yoshimasa M, Masahiro Y, Chiho O ,Mayumi I, Kauzue K, Makiko K and Yutaka O 2001 *Chest* **20** 589
[30] Leroy K, Haioun C and Lepage E 2002 *Annals of Oncology* **13** 1108
[31] http://$p53$.bii.a-star.edu.sg/about$p53$/dnaseq 2005
[32] http://www.medterms.com/script/main/art.asp?articlekey=4396 2010