Symmetry: Culture and Science
Vol. 23, Nos. 3-4, 303-322, 2012

FRACTAL GENETIC NETS AND SYMMETRY PRINCIPLES IN LONG NUCLEOTIDE SEQUENCES

S.V. Petoukhov*, V.I. Svirin**

* Biophysicist, bioinformatician (b. Moscow, Russia, 1946).

Address: Laboratory of Biomechanical Systems, Mechanical Engineering Research Institute of Russian Academy of Sciences; Malyi Kharitonievskiy pereulok, 4, Moscow, 101990, Russia. E-mail: spetoukhov@gmail.com.

Fields of interest: genetics, bioinformatics, biosymmetries, multidimensional numbers, musical harmony, mathematical crystallography (also history of sciences, oriental medicine).

Awards: Gold medal of the Exhibition of Economic Achievements of the USSR, 1974; State Prize of the USSR, 1986; Honorary diplomas of a few international conferences and organizations, 2005-2012.

Publications: *Biomechanics, Bionics and Symmetry*, Moscow, Nauka, (1981), 239 pp. (in Russian); *Biosolitons. Fundamentals of Soliton Biology*, Moscow, GPKT, (1999), 288 pp. (in Russian); *Matrix Genetics, Algebras of the Genetic Code, Noise-immunity*, Moscow, RCD, (2008), 316 pp. (in Russian); with M. He: *Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications*, Hershey, USA: IGI Global, (2010), 271 pp.; with He M.: *Mathematics of Bioinformatics: Theory, Practice, and Applications*, USA: John Wiley & Sons, Inc., (2011), 295 pp.

** Biophysicist, bioinformatician (b. Nizhnekamsk, Russia, 1987).

Address: Laboratory of Biomechanical Systems, Mechanical Engineering Research Institute of Russian Academy of Sciences; Malyi Kharitonievskiy pereulok, 4, Moscow, 101990, Russia. E-mail: vitaly.i.svirin@gmail.com.

Fields of interest: genetics, bioinformatics, biosymmetries, multidimensional numbers, musical harmony, descrete mathematics, cybernetics, neural networks.

Publications : Stepanyan I.V., Cygankov V.D., Svirin V.I. and Golovanyova G.V. (2012) Neurophysiological approaches to medical cybernetics based on the creative heritage by Academician P.K. In: Anokhin: *Biozashchita i Biobezopastnost'*, IV, №1, 28-42 (in Russian).

Abstract: *This article is devoted to hidden regularities of long nucleotide sequences. It contains a description and a thematic application of a new research tool that is termed as «fractal genetic nets». Described results testify in favor of existence of new Symmetry Principles of long nucleotide sequence as an addition to the known Symmetry Principle on the base of the generalized Chargaff's second parity rule. Our results provide new materials to the Chargaff's problem about a grammar of biology and to the idea about an algebraic essence of the genetic coding system.*

Keywords: genetic code, Chargaff's rule, long nucleotide sequence, grammar, fractal, genetic nets.

1. ON A GRAMMAR OF BIOLOGY AND THE NOTION OF FRACTAL GENETIC NETS

Fantastic successes of molecular genetics were defined in particular by a disclosure of phenomenological facts of symmetry in molecular constructions of genetic code and by a skillful implementation of these facts in theoretical modeling. A bright example is a disclosure of a symmetrological fact, reflected in the famous Chargaff's first parity rule, which says that in any double-stranded DNA segment, the quantities (or frequencies) of adenine and thymine are equal, and so are the frequencies of cytosine and guanine (Chargaff, 1950). This rule was used by Watson and Crick to support their famous DNA double-helix structure model (Watson & Crick, 1953).

In his works, Chargaff pursued goal of searching for a grammar of biology that defines hidden regularities of genetic texts to construct living cells with their "confounded multi-dimensionality", etc. One of his works "Preface to a Grammar of Biology" (Chargaff, 1971) which was devoted to a hundred years of nucleic acid research, reflects his thoughts that all achievements of molecular genetics are only the first steps in to discovering such grammar.

Besides his first parity rule for double-stranded DNA, Chargaff also perceived that the parity rule approximately holds in the sufficient long single-stranded DNA segment. This last rule is known as Chargaff's second parity rule (CSPR), and it has been confirmed in several organisms (Mitchell & Bride, 2006). Originally, CSPR is meant to be valid only to mononucleotide frequencies (that is quantities of monoplets) in the single-stranded DNA. "*But, it occurs that oligonucleotide frequencies follow a generalized Chargaff's second parity rule (GCSPR) where the frequency of an oligonucleotide is approximately equal to its complement reverse oligonucleotide frequency (Prahbu, 1993). This is known in the literature as the Symmetry Principle*" (Yamagishi, Herai, 2011, p. 2). The work of Prahbu (1993) shows the implementation of the Symmetry Principle in long DNA-sequences for cases of complementary reverse n-plets with n = 2, 3, 4, 5 at least. In scientific publications, long genetic sequences are those sequences that contain no less that 50.000 nucleotides (see for example (Yamagishi, Herai, 2011)). "*As correctly pointed out by Forsdyke, higher order equifrequency does imply lower order, and he therefore conjectured that the original*

CSPR was actually a particular case of a higher order parity rule" (Yamagishi, Herai, 2011). This Symmetry Principle was studied or described in many other publications (Albrecht-Buehler, 2006; Chargaff, 1971, 1975; Dong, Cuticchia, 2001; Forsdyke, 2002; Forsdyke, Bell, 2004; Kong, et al. 2009; Mitchell, Bridge, 2006; Sueoka, 1999). Due to this Symmetry Principle, the work (Yamagishi, Herai, 2011) has uncovered new rules of long nucleotide sequences and emphasized a fractal-like property of such sequences across a large set of genomes "*since no matter of scale, the same pattern is observed (self-similarity)*".

Our previous research in the field of "matrix genetics" (Petoukhov, 2008a-d, 2011, 2012a,b,c; Petoukhov, He, 2009; Petoukhov, Svirin, 2012) has led to the hypothesis that structures of long nucleotide sequences of different organisms are connected with so called "fractal genetic nets" (FGN). In this work we are proposing a novel approach to discover new Symmetry Principles in such sequences. In general case, each variant of FGN is constructed by means of the author's "method of a positional convolution (or positional splitting) of long genetic sequences" to get a cluster of long sequences, each of which, respectively, shorter than the original sequence. In the particular case considered in our article, the method lies in the positional convolution (or splitting) of long sequences of triplets through the removal or retention of individual positions (items) in each triplet.

1.1. Methodology

Let us explain a construction of FGN of various types on an example of FGN for sequences of triplets (Figure 1). In each triplet, its three positions are numbered by 0, 1 and 2 correspondingly. At the first level of a convolution, an initial long sequence S_0 of triplets is transformed by means of a positional convolution into three new sequences of nucleotides $S_{1/0}$, $S_{1/1}$, $S_{1/2}$, each of which is 3 times shorter in comparison with the initial sequence (numerator of the index in this notation of sequences shows the level of the convolution, and the denominator - the position of the triplets, which is used for the convolution): the sequence $S_{1/0}$ includes one by one all the nucleotides that are in the initial position "0" of triplets of the original sequence S_0; the sequence $S_{1/1}$ includes one by one all the nucleotides that are in the middle position "1" of triplets of the original sequence S_0; the sequence $S_{1/2}$ includes one by one all the nucleotides that are in the last position "2" of triplets of the original sequence S_0. At the final stage of the first level of the positional convolution, each of the sequences of nucleotides $S_{1/0}$, $S_{1/1}$, $S_{1/2}$ is represented as a sequence of triplets, where three positions inside each of triplets are

numbered again by 0, 1 and 2. To construct the second level of the convolution, each of the sequences $S_{1/0}$, $S_{1/1}$, $S_{1/2}$ is transformed by means of the same positional convolution in three new sequences: $S_{1/0}$ is convolved in $S_{2/00}$, $S_{2/01}$, $S_{2/02}$; $S_{1/1}$ – in $S_{2/10}$,$S_{2/11}$, $S_{2/12}$; $S_{1/2}$ – in $S_{2/20}$, $S_{2/21}$, $S_{2/22}$. Similarly, the third level and subsequent levels of the convolution are constructed to form a multi-level net of sequences of nucleotides called "the fractal genetic net for the triplet convolution" or briefly "FGN-3" (Figure 1).

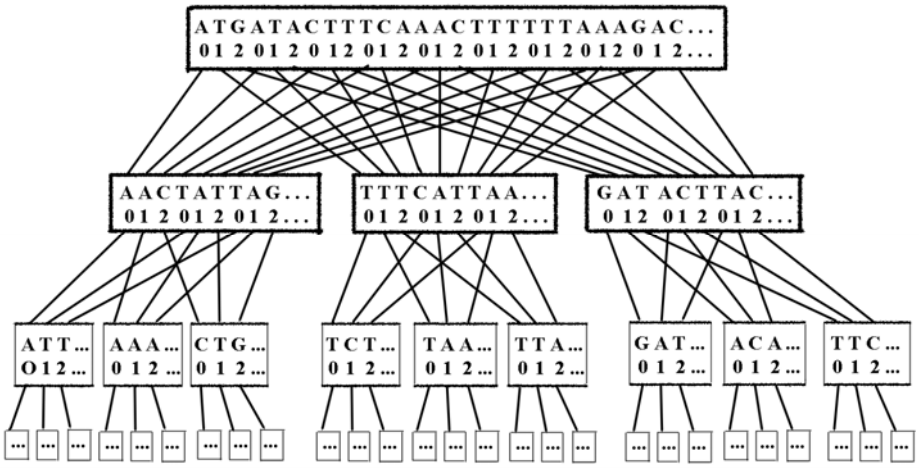


Figure 1: The scheme of the fractal genetic net (FGN-3) for a sequence of triplets

This FGN possesses a fractal-like character if the enumeration of positions is only taken into account: each of long sequences of this FGN can be taken as an initial sequence to form a similar genetic net on its basis (Figure 1). In general case, the FGN can be built not only for triplets, but also for other n-plets (n = 2, 4, 5, ...) or oligonucleotides by means of a repeated positional convolution of each of sequences from the previous level into "n" sequences of the next level of the convolution. This way one can built FGN-2, FGN-4, FGN-5, etc. for n=2, 3, 4, 5,… correspondingly. (Each of these FGN-2, FGN-3, FGN-4, FGN-5, etc. is a tree, but all of them form a net of separate trees; in a wide sense, FGN is the complete set of such separate trees). This article, on the other hand, concentrates only on the results related to the FGN-3.

2. ON SYMMETRY PRINCIPLES IN LONG NUCLEOTIDE SEQUENCES AND THE FGN-3

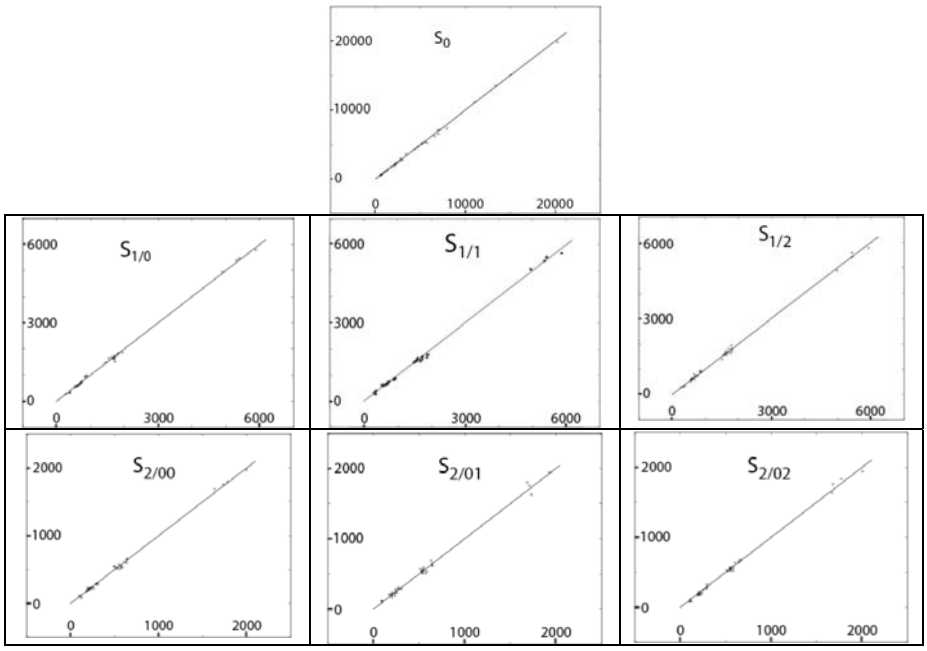
To test the author's hypothesis that structures of long nucleotide sequences of different organisms are connected with fractal genetic nets (first of all with FGN-3), we analyze an implementation of the known Symmetry Principle for long nucleotide sequences at different levels of a positional convolution in the fractal genetic net for the triplet convolution (FGN-3). In our article we use two different notions of complementary oligonucleotides (or n-plets): 1) complementary oligonucleotides in a traditional sense (for example ACGTG and TGCAC are the pair of complementary oligonucleotides in a traditional sense); 2) complementary reverse oligonucleotides (Prahbu, 1993) briefly called CR-oligonucleotides or reverse complements (for example ACGTG and CACGT are the pair of CR-oligonucleotides). The mentioned Symmetry Principle has been revealed for pairs of CR-oligonucleotides. Taking this into account we began testing the author's hypothesis by means of analyzing frequencies (or quantities) of all variants of pairs of CR-oligonucleotides in long DNA-sequences of different organisms at different levels of their FGN-3. We test frequencies of n-plets in the FGN-3 with n = 1, 2, 3, 4, 5 only because of our computer limitations, but we assumed that our described results for FGN-3 hold true also for n > 5. Initial nucleotide sequences for testing are taken from (NCBI, 2012a). To test the proposed hypothesis, we use special software written by V.I.Svirin using programming language Python.

In our preliminary studies we have revealed the following: 1) the Symmetry Principle for pairs of CR-oligonucleotides is realized in each of long nucleotide sequences at different levels of the convolution in FGN-3 (the length of oligonucleotides or n-plets under consideration is equal to n = 1, 2, 3, 4, 5 at least); 2) a series of new Symmetry Principles exists in those initial long nucleotide sequences where the famous Symmetry Principle for pairs of CR-oligonucleotides is performed; 3) each of these new Symmetry Principles is performed for n-plets in each of long nucleotide sequences at different levels of the convolution in FGN-3 (n = 1, 2, 3, 4, 5 at least).

Let us take, for example, the long nucleotide sequence of Mycoplasma crocodyli MP145 chromosome, complete genome (NCBI Reference Sequence: NC_014014.1 (NCBI, 2012b). This sequence contains 934379 nucleotides. Figure 2 shows realisations of the mentioned Symmetry Principle (we'll name it as the Symmetry Principle №1) in the 13 sequences at the first three levels of convolution in the FGN-3 of this sequence. It displays the number of occurrences of 32 triplets (AAA, AAC, AAG, AAT, ACA, ACC, ACG, ACT, AGA, AGC, AGG, ATA, ATC, ATG, CAA,

CAC, CAG, CCA, CCC, CCG, CGA, CGC, CTA, CTC, GAA, GAC, GCA, GCC, GGA, GTA, TAA, TCA) and their 32 CR-triplets (TTT, GTT, CTT, ATT, TGT, GGT, CGT, AGT, TCT, GCT, CCT, TAT, GAT, CAT, TTG, GTG, CTG, TGG, GGG, CGG, TCG, GCG, TAG, GAG, TTC, GTC, TGC, GGC, TCC, TAC, TTA, TGA) in the long sequences S_0, $S_{1/0}$, $S_{1/1}$, $S_{1/2}$, $S_{2/00}$, $S_{2/01}$, $S_{2/02}$, $S_{2/10}$, $S_{2/11}$, $S_{2/12}$, $S_{2/20}$, $S_{2/21}$, $S_{2/22}$ at the first three levels of the FGN-3 (a limited volume of the article doesn't allow demonstration of other levels of this FGN).

The straight line in each frame is a slope 1 (it is a bisector of the coordinate angle). Each dot in a frame represents one pair "triplet and CR-triplet"; its coordinate X shows number of occurrences (or the frequency) of the triplet, and its coordinate Y shows number the frequency of its CR-triplet on the same strand of the sequence. Each frame contains all 32 pairs «triplet and its CR-triplet». The dots agglutinate at the line of slope 1, demonstrating that amounts of occurrences (or frequencies) of two members of each of 32 pairs «triplet and its CR-triplet» are approximately equal in each of the sequences at each of the levels of convolution in the FGN-3. It means that the Symmetry Principle №1 is performed for each of these sequences.



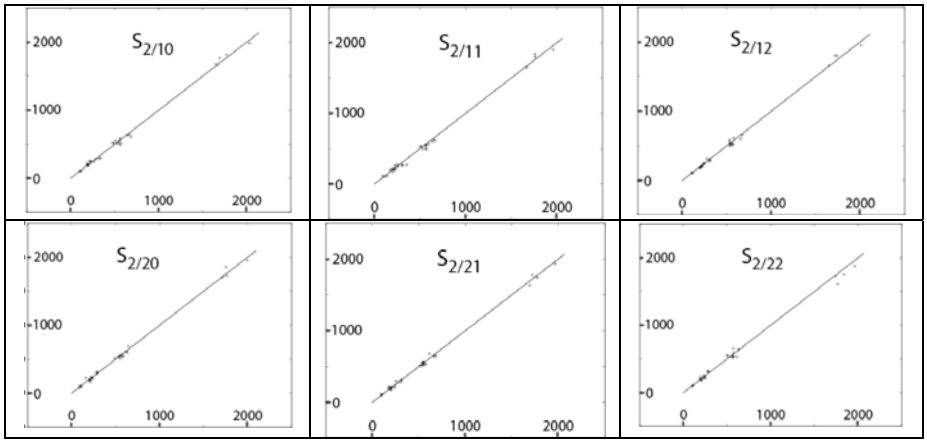


Figure 2: Realizations of the Symmetry Principle №1 in the long sequences S_0, $S_{1/0}$, $S_{1/1}$, $S_{1/2}$, $S_{2/00}$, $S_{2/01}$, $S_{2/02}$, $S_{2/10}$, $S_{2/11}$, $S_{2/12}$, $S_{2/20}$, $S_{2/21}$, $S_{2/22}$ at the first three levels of the FGN-3 for Mycoplasma crocodyli MP145 chromosome, complete genome (NCBI Reference Sequence: NC_014014.1 (NCBI, 2012b)). The initial sequence S_0 contains 934379 nucleotides.

These results show the effectiveness of the proposed fractal genetic nets as a research tool and they also testify in favour of existence of the generalized **Symmetry Principle № 1**: in long nucleotide sequences at different levels of convolution in FGN-3, oligonucleotide frequencies follow a generalized Chargaff's second parity rule where the frequency of each oligonucleotide is approximately equal to its complement reverse oligonucleotide frequency.

Now let us present our research results that testify in favor of the existence of new Symmetry Principles of long nucleotide sequences. Below, we formulate these new Symmetry Principles directly and then provide some data confirming their existence.

The Symmetry Principle № 2 (concerning FGN): the frequency of each oligonucleotide is approximately the same in all the long nucleotide sequences of each of levels of FGN-3.

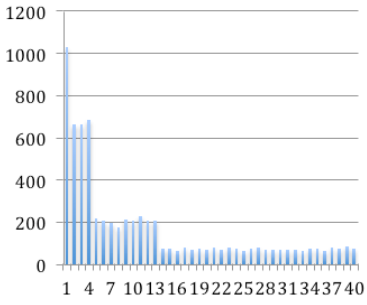


Figure 3: Frequencies of the triplet ACG in 40 long nucleotide sequences S_0, $S_{1/0}$, $S_{1/1}$, $S_{1/2}$, $S_{2/00}$, $S_{2/01}$, $S_{2/02}$, $S_{2/10}$, $S_{2/11}$, $S_{2/12}$, $S_{2/20}$, $S_{2/21}$, $S_{2/22}$, ….., $S_{3/221}$, $S_{3/222}$ at the first four levels of the FGN-3 of Mycoplasma crocodyli MP145 chromosome, complete genome (NCBI Reference Sequence: NC_014014.1 (NCBI, 2012b)). Coordinate X shows the 40 sequences and coordinate Y shows appropriate frequencies of the triplet ACG in them.

Figure 3 demonstrates an example of frequencies of the triplet ACG in 40 long nucleotide sequences S_0, $S_{1/0}$, $S_{1/1}$, $S_{1/2}$, $S_{2/00}$, $S_{2/01}$, $S_{2/02}$, $S_{2/10}$, $S_{2/11}$, $S_{2/12}$, $S_{2/20}$, $S_{2/21}$, $S_{2/22}$, ….., $S_{3/221}$, $S_{3/222}$ at the first four levels of the FGN-3 of the same initial sequence shown on Figure 2.

Figure 4 shows examples of frequencies of all 64 triplets in 12 long nucleotide sequences at the first three levels of FGN-3 of the same initial sequence as on Figure 2.

	S_0	$S_{1/0}$	$S_{1/1}$	$S_{1/2}$	$S_{2/00}$	$S_{2/01}$	$S_{2/02}$	$S_{2/10}$	$S_{2/11}$	$S_{2/12}$	$S_{2/20}$	$S_{2/21}$
AAA	19832	5786	5679	5768	1975	1944	1944	1986	1899	1952	1954	1935
AAC	6246	1709	1707	1643	550	560	567	587	543	557	504	531
AAG	7087	1859	1783	1940	607	619	651	607	630	611	615	679
AAT	15037	5320	5352	5428	1784	1685	1769	1770	1758	1743	1757	1775
ACA	5049	1527	1564	1635	542	513	546	492	566	521	537	517
ACC	2363	747	755	747	233	253	266	241	273	253	231	203
ACG	1029	660	663	684	214	205	197	175	210	208	228	203
ACT	4714	1713	1745	1702	526	548	553	536	568	544	552	537
AGA	5272	1784	1688	1737	657	635	640	631	600	598	691	644
AGC	2754	590	623	586	226	216	200	220	199	181	198	223
AGG	2150	973	880	912	293	294	322	292	259	287	314	288
AGT	4713	1700	1704	1820	530	595	543	513	492	523	562	549
ATA	11126	4952	5051	4886	1688	1629	1637	1671	1655	1659	1706	1635
ATC	5250	1619	1570	1568	507	537	511	518	527	583	531	522
ATG	5499	1450	1488	1505	494	547	582	510	529	538	551	514
ATT	15079	5397	5390	5419	1797	1802	1835	1816	1832	1799	1857	1745
CAA	7427	1620	1615	1661	526	538	568	565	502	508	540	561
CAC	1872	700	746	733	225	236	201	255	262	243	222	238
CAG	2105	553	653	605	212	206	211	205	203	221	197	203
CAT	5375	1473	1497	1388	547	546	519	548	497	539	543	507
CCA	2750	727	712	664	235	248	252	267	252	246	218	255
CCC	569	522	622	513	181	166	170	188	198	183	164	196
CCG	681	336	390	326	81	113	84	105	109	118	94	109
CCT	2181	887	945	885	292	277	293	323	302	307	281	292
CGA	1171	635	607	600	189	208	178	196	207	194	204	208
CGC	508	321	341	319	106	113	89	104	101	107	109	116
CGG	693	402	365	366	124	93	102	121	137	105	85	97
CGT	989	671	663	684	214	175	194	199	189	215	164	204
CTA	4326	1664	1659	1562	545	515	520	526	532	542	543	524
CTC	1786	832	893	841	310	306	295	289	306	316	283	312
CTG	2115	577	647	636	215	236	206	207	172	211	196	173
CTT	6917	1950	1913	1785	635	646	641	684	653	577	617	614
GAA	7190	1823	1801	1812	651	689	675	640	610	655	602	654
GAC	1404	585	598	611	208	204	178	195	223	208	189	201
GAG	1820	932	833	930	289	289	274	284	275	295	291	291
GAT	5225	1664	1563	1555	555	530	523	488	507	534	577	557
GCA	2974	572	580	631	228	195	197	241	210	196	196	198
GCC	710	276	353	288	97	91	110	102	92	118	106	99
GCG	497	377	321	361	113	97	121	109	108	109	103	104
GCT	2973	622	582	585	227	233	218	181	227	192	207	178

GGA	2330	958	875	888	283	302	296	297	272	316	308	318
GGC	676	286	275	283	115	110	100	112	112	108	93	104
GGG	616	551	546	555	185	200	193	187	170	200	215	174
GGT	2446	733	755	728	246	240	281	232	249	246	229	243
GTA	3636	1680	1606	1709	516	537	532	518	544	501	513	546
GTC	1374	601	577	578	199	198	218	194	217	226	202	203
GTG	2083	711	720	777	265	257	238	219	226	237	244	239
GTT	6587	1694	1736	1770	575	541	553	560	559	530	551	577
TAA	13334	5401	5418	5430	1732	1708	1678	1693	1760	1717	1762	1723
TAC	3369	1625	1685	1624	520	528	509	571	571	535	492	554
TAG	4368	1685	1596	1641	498	538	556	475	505	561	587	544
TAT	11019	4895	4950	4973	1635	1730	1667	1648	1672	1649	1711	1695
TCA	6993	1542	1679	1586	549	514	546	542	559	575	573	542
TCC	2302	872	904	854	291	276	293	340	353	277	295	316
TCG	1240	681	662	669	183	201	204	193	204	206	220	212
TCT	5710	1794	1866	1798	639	645	600	646	634	658	647	669
TGA	6550	1642	1533	1602	577	570	560	530	557	525	568	556
TGC	2790	616	610	611	190	233	201	219	219	213	205	191
TGG	2658	720	689	796	222	285	245	246	241	240	239	293
TGT	5123	1727	1583	1619	585	568	547	565	592	546	531	536
TTA	13519	5470	5525	5585	1755	1753	1760	1765	1793	1802	1733	1777
TTC	7131	1839	1864	1809	644	631	659	669	660	672	632	680
TTG	7918	1700	1745	1689	576	579	583	563	562	559	541	553
TTT	20219	5886	5876	5921	1997	1929	2004	2034	1960	2010	1995	1969

Figure 4: the table of frequencies of 64 triplets in long nucleotide sequences S_0, $S_{1/0}$, $S_{1/1}$, $S_{1/2}$, $S_{2/00}$, $S_{2/01}$, $S_{2/02}$, $S_{2/10}$, $S_{2/11}$, $S_{2/12}$, $S_{2/20}$, $S_{2/21}$ at the first three levels of FGN-3 of Mycoplasma crocodyli MP145 chromosome, complete genome (NCBI Reference Sequence: NC_014014.1 (NCBI, 2012b)).

The Symmetry Principle № 3: for each of long nucleotide sequences at each level of FGN-3 the following rules hold true: sum of the frequencies of all the oligonucleotides, that begin with the letter A, approximately equal to the sum of the frequencies of all the oligonucleotides that begin with the letter T; sum of the frequencies of all the oligonucleotides, that begin with the letter C, approximately equal to the sum of the frequencies of all the oligonucleotides that begin with the letter T.

In particularly, these rules hold not only for long sequences at lower levels of FGN-3 but also for an initial long sequence S_0. Figure 5 illustrates the Symmetry Principle № 3 using examples of *n*-plets (*n*=2, 3, 4, 5) in sequences S_0, $S_{1/0}$,…, $S_{2/22}$ at the first levels in FGN-3 of the same sequence as on Figure 2.

The total frequencies of the sets of duplets in sequences of FGN-3:

	S_0	$S_{1/0}$	$S_{1/1}$	$S_{1/2}$	$S_{2/00}$	$S_{2/01}$	$S_{2/02}$	$S_{2/10}$	$S_{2/11}$	$S_{2/12}$
F(A)	169757	56674	56197	56747	19065	18857	18836	18764	18756	18713
F(T)	171420	57022	57247	57179	18870	18937	19067	19102	19155	19040
F(C)	62531	20763	21486	20600	6929	6903	6882	7113	7148	7178
F(G)	63471	21265	20794	21198	7044	7211	7123	6929	6849	6977

The total frequencies of the sets of triplets in sequences of FGN-3:

	S_0	$S_{1/0}$	$S_{1/1}$	$S_{1/2}$	$S_{2/00}$	$S_{2/01}$	$S_{2/02}$	$S_{2/10}$	$S_{2/11}$	$S_{2/12}$
F(A)	113200	37786	37642	37980	12623	12582	12763	12565	12540	12557
F(T)	114243	38095	38185	38207	12593	12688	12612	12699	12842	12745
F(C)	41465	13870	14268	13568	4637	4622	4523	4782	4622	4632
F(G)	42541	14065	13721	14061	4752	4713	4707	4559	4601	4671

The total frequencies of the sets of 4-plets in sequences of FGN-3:

	S_0	$S_{1/0}$	$S_{1/1}$	$S_{1/2}$	$S_{2/00}$	$S_{2/01}$	$S_{2/02}$	$S_{2/10}$	$S_{2/11}$	$S_{2/12}$
F(A)	84955	28475	27999	28493	9522	9391	9274	9342	9417	9434
F(T)	85573	28439	28601	28639	9531	9532	9624	9552	9570	9538
F(C)	31189	10286	10758	10270	3409	3438	3499	3573	3578	3594
F(G)	31867	10662	10504	10460	3492	3593	3557	3487	3389	3388

The total frequencies of the sets of 5-plets in sequences of FGN-3:

	S_0	$S_{1/0}$	$S_{1/1}$	$S_{1/2}$	$S_{2/00}$	$S_{2/01}$	$S_{2/02}$	$S_{2/10}$	$S_{2/11}$	$S_{2/12}$
F(A)	67729	22626	22512	22788	7551	7506	7542	7491	7417	7431
F(T)	68688	22951	22918	22764	7620	7613	7684	7626	7668	7677
F(C)	25144	8242	8503	8217	2780	2762	2701	2851	2907	2849
F(G)	25304	8470	8356	8520	2812	2882	2836	2795	2771	2806

Figure 5: The illustration of the Symmetry Principle № 3 in the case of Mycoplasma crocodyli MP145 chromosome, complete genome (NCBI Reference Sequence: NC_014014.1 (NCBI, 2012b)). Here F(A), F(T), F(C) and F(G) mean sum of the frequencies of oligonucleotides (or n-plets) that begin with the letters A, T, C or G correspondingly. The tables show the F(A) ≈ F(T) and F(C) ≈ F(G) for sets of *n*-plets (*n*=2, 3, 4, 5) in each of long nucleotide sequences S_0, $S_{1/0}$, $S_{1/1}$, $S_{1/2}$, $S_{2/00}$, $S_{2/01}$, $S_{2/02}$, $S_{2/10}$, $S_{2/11}$, $S_{2/12}$ at the first three levels of FGN-3 of this sequence

This result was obtained in connection with studies related to genetic matrices $[C\ T;\ A\ G]^{(n)}$ (see (Petoukhov, 2012c) in this issue). Each of 4 quadrants of such genomatrices contains all oligonucleotides that begin with one of 4 letters C, T, A or G. In these genomatrices, each oligonucleotide and its complementary oligonucleotide are disposed inverse-symmetrical relative to the center of the appropriate matrix. In accordance with the Symmetry Principle № 3, the total frequencies of oligonucleotides in both quadrants along the main diagonal of these genomatrices are approximately equal each other (F(C) ≈ F(G)); the total frequencies of oligonucleotides in both quadrants along the second diagonal of these genomatrices are also approximately equal each other (F(A) ≈ F(T)).

An additional illustration of the Symmetry Principle № 3 is obtained based on the initial data about frequencies of separate triplets in the whole human genome from the work (Perez, 2010). This genome contains 2843411612 triplets. Figure 6 shows the total frequencies F_C, F_G, F_A and F_T of sets of triplets that begin with one of four letters C, G,

A or T. The percentage difference between the total frequencies F_C and F_G is equal to 0.05% and between F_A and F_T is equal to 0.16%. But we don't have data about the quantities of separate triplets in convoluted sequences $S_{1/0}$, $S_{1/1}$,… at the lower levels of FGN-3 for this genome because we have no information about the order of triplets in its huge sequence S_0.

The total frequencies of the sets of triplets, which begin with C:	The total frequencies of the sets of triplets, which begin with G:
F_C = F(CCC+CCT+CCA+CCG+ CTC+CTT+CTA+CTG+ CAC+CAT+CAA+CAG+ CGC+CGT+CGA+CGG) = **581026275**	F_G = F(GGG+GGA+GGT+GGC+ GAG+GAA+GAT+GAC+ GTG+GTA+GTT+GTC+ GCG+GCA+GCT+GCC)= **581343106**
The total frequencies of the sets of triplets, which begin with A:	The total frequencies of the sets of triplets, which begin with T:
F_A = F(ACC+ACT+ACA+ACG+ ATC+ATT+ATA+ATG+ AAC+AAT+AAA+AAG+ AGC+AGT+AGA+AGG)= **839827642**	F_T = F(TGG+TGA+TGT+TGC+ TAG+TAA+TAT+TAC+ TTG+TTA+TTT+TTC+ TCG+TCA+TCT+TCC)= **841214589**

Figure 6: the approximate equality of the total frequencies of sets of triplets that begin with letters C and G (upper table) and with letters A and T (bottom table) in the case of the sequence S_0 of the whole human genome. Initial data about frequencies of separate triplets are taken from the work (Perez, 2010).

Now let us introduce the Symmetry Principle № 4, which deals with reading frame shifts, deletion mutations, and also positional permutations in oligonucleotides. Concerning those DNA sequences (including the mentioned sequence on Figures 2-5), that have been tested till today in our laboratory, we have discovered the following phenomenological facts (this study is continued now for a wide list of DNA-sequences of different organisms and organelles):

- a transformation of long nucleotide sequences by means of a reading frame shift in them preserves implementations of all described Symmetry Principles inside new long nucleotide sequences (in our tests, a reading frame shift means that the reading of sequence does not begin with its first position, but with one of subsequent positions; the missing fragment of the sequence can be moved into the end of the sequence, and in this case a reading frame shift leads to a simple change of order of all sequences at each of lower levels of FGN);

- a transformation of long nucleotide sequences by means of a deletion mutation (when their short parts are missing) preserves implementations of all described Symmetry Principles in new long nucleotide sequences.

One should separately consider the question about positional permutations in oligonucleotides. The theory of noise-immunity coding pays a special attention to permutations of elements of transmitted signals. It is obvious that for different n-plets different quantities of variants of permutation of their positions exist:

for duplets two variants of positional permutations exist (1-2 and 2-1);

for triplets six variants of positional permutations exist (1-2-3, 2-3-1, 3-1-2, 3-2-1, 2-1-3, 1-3-2);

for 4-plets 24 variants of positional permutations exist (1-2-3-4, 2-3-4-1, …..);

for 5-plets 120 variants of positional permutations exist (1-2-3-4-5, 2-3-4-5-1, …..).

It is also evident that if a long nucleotide sequence is interpreted as a sequence of a certain type of oligonucleotides (duplets, or triplets, or 4-plets, or 5-plets, …), and one of possible positional permutations is done simultaneously inside all of its oligonucleotides, then a quite new long nucleotide sequence appears (we named simultaneous positional permutations inside all oligonucleotides of a certain type as "collective positional permutations" inside these oligonucleotides). For example, if we have initially a sequence of triplets CGA-TAA-AGC-GTC-TAG-CGC-ATC -…, then after changing of the positional order from the initial order 1-2-3 to new order 2-3-1 inside each of triplets, we obtain the quite different sequence GAC-AAT-GCA-TCG-AGT-GCC-TCA -… . But our studies of a wide set of long nucleotide sequences (including the sequence on Figure 2-5) demonstrated that the FGN-3 for such new long nucleotide sequence has obeyed the same Symmetry Principles №№ 1-3 described above. These results attest to possible existence of the Symmetry Principle № 4, which can be briefly formulated as:

The Symmetry Principle № 4:

- reading frame shifts and deletion mutations in long nucleotide sequences, and also collective positional permutations inside their oligonucleotides don't essentially violate implementations of all Symmetry Principles for long nucleotide sequences and their fractal genetic net (FGN-3).

The final part of this article illustrates an additional application of the FGN-3 approach to study hidden regularities of long nucleotide sequences from the point of view of the

black-and-white mosaics of the genetic matrix [C T; A G]$^{(3)}$ from the article (Petoukhov, 2012c, Figure 20). Figure 7 shows this matrix, which reflects phenomenological properties of the genetic coding system and which contains the complete set of 64 triplets in a strong order. The mosaic of this matrix is identical to the mosaic of one of Hadamard (8*8)-matrices that are widely used in noise-immunity coding (for example, codes based on Hadamard matrices have been used on spacecraft «Mariner» and «Voyadger», which allowed obtaining high-quality photos of Mars, Jupiter, Saturn, Uranus and Neptune in spite of the distortion and weakening of the incoming signals; Hadamard matrices are used to create quantum computers, which are based on Hadamard gates, etc.). In addition, this Hadamard representation of the genetic matrix [C T; A G]$^{(3)}$ is the biquaternion by Hamilton with unit coordinates (see details in the article (Petoukhov, 2012c)). A possible connection between the black-and-white mosaic of this genetic matrix and hidden regularities of long nucleotide sequences has a special interest. Below we present our initial results of studying this connection.

CCC	CCT	CTC	CTT	TCC	TCT	TTC	TTT
CCA	CCG	CTA	CTG	TCA	TCG	TTA	TTG
CAC	CAT	CGC	CGT	TAC	TAT	TGC	TGT
CAA	CAG	CGA	CGG	TAA	TAG	TGA	TGG
ACC	ACT	ATC	ATT	GCC	GCT	GTC	GTT
ACA	ACG	ATA	ATG	GCA	GCG	GTA	GTG
AAC	AAT	AGC	AGT	GAC	GAT	GGC	GGT
AAA	AAG	AGA	AGG	GAA	GAG	GGA	GGG

Figure 7: the genetic matrix [C T; A G]$^{(3)}$ of 64 triplets with its black-and-white mosaic, which reflects phenomenological properties of the genetic coding system (from the work (Petoukhov, 2012c)).

This matrix [C T; A G]$^{(3)}$ in Figure 7 contains two subsets with 28 kinds of white triplets and 36 kinds of black triplets. The authors calculate total quantities (frequencies F_{WHITE} and F_{BLACK}) of members of these two subsets in long nucleotide sequences. For example, we calculated the total frequencies for the whole human genome, which contains the huge number 2843411612 (about three billion) of triplets. The initial data about this genome are shown on Figure 8 from the article (Perez, 2010). Very different frequencies of different triplets are represented in this genome. For example, the frequency of the triplet CGA is equal to 6251611 and the frequency of the triplet TTT is equal to 109951342; they differ in 18 times approximately. But our result of the calculation shows that in this genome the percentage difference between F_{WHITE} and F_{BLACK} is approximately equal to 0.1% because the total quantity F_{WHITE} of white triplets is equal to 1422456641 and the total quantity F_{BLACK} of black triplets is equal to 1420954971.

TRIPLET	TRIPLET FREQUENCY	TRIPLET	TRIPLET FREQUENCY	TRIPLET	TRIPLET FREQUENCY	TRIPLET	TRIPLET FREQUENCY
AAA	109143641	CAA	53776608	GAA	56018645	TAA	59167883
AAC	41380831	CAC	42634617	GAC	26820898	TAC	32272009
AAG	56701727	CAG	57544367	GAG	47821818	TAG	36718434
AAT	70880610	CAT	52236743	GAT	37990593	TAT	58718182
ACA	57234565	CCA	52352507	GCA	40907730	TCA	55697529
ACC	33024323	CCC	37290873	GCC	33788267	TCC	43850042
ACG	7117535	CCG	7815619	GCG	6744112	TCG	6265386
ACT	45731927	CCT	50494519	GCT	39746348	TCT	62964984
AGA	62837294	CGA	6251611	GGA	43853584	TGA	55709222
AGC	39724813	CGC	6737724	GGC	33774033	TGC	40949883
AGG	50430220	CGG	7815677	GGG	37333942	TGG	52453369
AGT	45794017	CGT	7137644	GGT	33071650	TGT	57468177
ATA	58649060	CTA	36671812	GTA	32292235	TTA	59263408
ATC	37952376	CTC	47838959	GTC	26866216	TTC	56120623
ATG	52222957	CTG	57598215	GTG	42755364	TTG	54004116
ATT	71001746	CTT	56828780	GTT	41557671	TTT	109591342

Figure 8: quantities of repetitions of each triplet in the whole human genome (from [Perez, 2010])

Similar results about approximate equality of F_{WHITE} and F_{BLACK} were obtained for all 811 long fragments of the human genome studied in his student's thesis by one of the authors – V.Svirin, who became the pioneer of this comparative analyses of F_{WHITE} and F_{BLACK} in long nucleotide sequences from the point of view of the phenomenological genomatrix shown in Figure 7.

What conclusion can be made about an application of the method of the FGN-3 to study the total quantities F_{WHITE} and F_{BLACK} in long nucleotide sequences? Figure 9 shows typical results of the comparison analysis of F_{WHITE} and F_{BLACK} in sequences on different levels of the FGN-3 for the same initial sequence S_0 of Mycoplasma crocodyli MP145 chromosome.

	S_0	$S_{1/0}$	$S_{1/1}$	$S_{1/2}$	$S_{2/00}$	$S_{2/01}$	$S_{2/02}$	$S_{2/10}$	$S_{2/11}$	$S_{2/12}$
F_{WHITE} %	52	49	49	49	49	49	50	50	49	49
F_{BLACK} %	48	51	51	51	51	51	50	50	51	51

	$S_{2/20}$	$S_{2/21}$	$S_{2/22}$	$S_{3/000}$	$S_{3/001}$	$S_{3/002}$	$S_{3/010}$	$S_{3/011}$	$S_{3/012}$	$S_{3/020}$
F_{WHITE} %	49	49	49	50	49	50	49	50	50	49
F_{BLACK} %	51	51	51	50	51	50	51	50	50	51

	$S_{3/021}$	$S_{3/022}$	$S_{3/100}$	$S_{3/101}$	$S_{3/102}$	$S_{3/110}$	$S_{3/111}$	$S_{3/112}$	$S_{3/120}$	$S_{3/121}$
F_{WHITE} %	49	50	50	50	50	49	50	49	50	50
F_{BLACK} %	51	50	50	50	50	51	50	51	50	50

	$S_{3/122}$	$S_{3/200}$	$S_{3/201}$	$S_{3/202}$	$S_{3/210}$	$S_{3/211}$	$S_{3/212}$	$S_{3/220}$	$S_{3/221}$	$S_{3/222}$
F_{WHITE} %	49	50	49	50	50	49	49	49	50	49
F_{BLACK} %	51	50	51	50	50	51	51	51	50	51

Figure 9: percentage of frequencies F_{WHITE} and F_{BLACK} of white and black triplets (from Figure 7) in long sequences S_0, $S_{1/0}$, ..., $S_{3/222}$ in the first four levels of the FGN-3 for the Mycoplasma crocodyli MP145 chromosome, complete genome (NCBI Reference Sequence: NC_014014.1 (NCBI, 2012b)). The sequence S_0 contains contains 934379 nucleotides.

From Figure 9, one can observe the fact of approximate equality of total quantities of white and black triplets in all these sequences S_0, $S_{1/0}$, ..., $S_{3/222}$.

It appears that the described FGN-3 and fractal-like properties of long genetic sequences that are related to the invariance of these Symmetry Principles, have a biological value (a biological sense) associated with mutational changes of such sequences and with evolutionary creation of new types of DNA-sequences. The authors presume that mechanisms of biological evolution use these permutational and other described properties of long nucleotide sequences in producing new biological organisms and organelles. For instance, new DNA sequences can be constructed in the course of biological evolution of organisms by means of combinatorics of nucleotide sequences from different levels of FGN (including genetic crossing among long nucleotide sequences from different levels of FGN by analogy with well-known examples of genetic crossing). One should note here that the question about permutation properties of DNA-sequences is very important because some biological organisms differ each from other only by permutations in their DNA sequences (see for example the book (Pevzner, 2000)). The proposed method of the FGN is the new effective and useful approach in the field of bioinformatics, molecular genetics, and evolutionary biology. It generates new data in the field of symmetrology (Darvas, 2007; Cristea, 2005, etc.)

In addition, one can mention here about fractal images in genetic systems. A number of publications are devoted to fractal features of genetic texts (Gusev et al, 2009; Jeffry, 1990; Pellionisz et al, 2012a; Petoukhov, 2008b; Petoukhov, He, 2009; Yam, 1995, etc). Interesting data about fractal approaches in genetics, including materials about an important connection of fractal defects with cancer, are presented at the website

(Pellionisz, 2012b). Research in this direction continues all over the world. In this article, the authors propose Fractal Genetics Nets (FGN) as a new tool to study fractal-like properties of long DNA sequences that also describes new fractal-like properties of such nucleotide sequences. We believe that these FGN and fractal-like properties of long nucleotide sequences can lead to new principles and systems in the field of signal processing, recognition of images and artificial intellect. The list of these scientific tools includes also genetic algorithms developed intensively in scientific world during last decades (for example see (Goldberg, Korb, Deb, 1989; Forrest, Mitchell, 1991)). Our findings described here contribute to the evidences of the idea about algebraic essence of the genetic coding system (Petoukhov, 2008a-d, 2011, 2012a,b,c; Petoukhov, He, 2009).

We plan to publish in the nearest future other results of our studies toward FGN and the Symmetry Principles related to a wide list of long DNA sequences of different organells and organisms from different taxonomical classes. These results would require a large volume for their publication and, therefore, are not included in the limited volume of this article.

3. DISCUSSION

The genetic coding system possesses impressive noise-immunity properties. Modern technology of noise-immunity coding is based on matrix presentations of discrete signals. This technology allows noise-immunity transferring, for example, photos of a surface of Mars through millions kilometers of spaces with noises to provide a receiving the high-quality photos on Earth. The authors are studying hidden regularities of the genetic coding system by means of known matrix methods from this communication technology. In the result, a special scientific direction called “matrix genetics” is developing during last year (Petoukhov, 2008a-d, 2011, 2012a,b,c; Petoukhov, He, 2009; Petoukhov, Svirin, 2012). The results described in our article are closely connected with many other results of this effective direction of researches where many connections have been revealed between the genetic coding system and mathematics of discrete signals processing including noise-immunity coding. In particular, the list of relevant mathematical formalisms includes Hadamard matrices, orthogonal systems of Walsh functions and Rademacher functions, Kronecker families of matrices, dyadic-shift matrices and dyadic-shift decompositions of matrices, hyper-complex numbers (including Hamilton quaternions and bi-quaternions), new matrix presentations of complex numbers and split-complex numbers, algebras of projective

operators, etc. Without these algebraic results, we couldn't offer fractal genetic nets as a new tool for genetic analyses and we couldn't receive the described phenomenological data about the proposed symmetry principles in long nucleotide sequences. Our matrix approach to the genetic system gives opportunity to receive data in favor of existence of not only symmetry principles proposed above but also some other symmetry principles that will be published in the nearest future.

Modern science knows that deep knowledge about phenomenological relations of symmetry among separate parts of a complex natural system can tell many important things about the evolution and mechanisms of these systems. It should be noted that fantastic successes of molecular genetics were defined in particular by a disclosure of phenomenological facts of symmetry in molecular constructions of genetic code and by skilful using of these facts in theoretical modeling. A bright example is a disclosure of a symmetrological fact, reflected in the first rule by E. Chargaff, of an equality of quantities of nitrogenous bases in their appropriate pairs (adenine-thymine and cytosine-guanine) in molecules of DNA in different organisms. This phenomenological rule was used skilfully in a theoretic modeling of a double helix of DNA by F. Crick and J. Watson with using of additional symmetrological principles.

Biological organisms belong to a category of very complex natural systems, which correspond to a huge number of biological species with inherited properties. But surprisingly, molecular genetics has discovered that all organisms are identical to each other by their basic molecular-genetic structures. Due to this revolutionary discovery, a great unification of all biological organisms has happened in the science. The information-genetic line of investigations has become one of the most prospective lines not only in biology, but also in science as a whole. The more science studies living matter, the more facts of unification in other physiological systems (metabolic biosystems, energy biosystems, etc.) are discovered. The searching of unification principles in living matter is an important direction of developing modern science. Materials of our article belong to this direction.

Modern science recognizes a key meaning of information principles for inherited self-organization of living matter. Modern informatics is an independent branch of science, which possesses its own language and mathematical formalisms and exists together with physics, chemistry and other scientific branches. A problem of information evolution of living matter has been investigated intensively in the last decades in addition to studies of the classical problem of biochemical evolution. Not only physics and chemistry deal with principles and methods of symmetry, informatics and digital

signal processing also pay great attention to them. How is theory of signal processing connected to geometry and geometrical symmetries? Signals are represented there in a form of a sequence of the numeric values of their amplitude in reference points. The theory of signal processing is based on the interpretation of discrete signals as a form of vector in multi-dimensional spaces. In every tact of time, a signal value is interpreted as the corresponding value of a coordinate in a multi-dimensional vector space of signals. In this way, the theory of discrete signals turns out to be the science of geometries of multi-dimensional spaces where different multidimensional numeric systems can be useful. The number of dimensions of such a space is equal to the quantity of reference points for the signal. Metric notions and all other necessary things are introduced in these multi-dimensional vector spaces for those or other problems of maintenance of reliability, speed and economy of the signal information. On this geometrical basis, many methods and algorithms of recognition of signals and images, coding information, detection and correction of information mistakes, artificial intellect and training of robots are constructed. One can add here the importance of symmetries in permutations of components for coding signals, in spectral analysis of signals, in orthogonal and other transformations of signals, and so on. Investigation of symmetrical and structural analogies between computer informatics and genetic informatics is also needed for the creation of DNA-computers, DNA-robotics, for so called “genetic algorithms” that is widely used in modern engineering, etc. The authors of the article hope that the proposed symmetry principles described in the article will be useful not only for fundamental knowledge but also for technologic applications.

Thoughts and dreams of Chargaff about a disclosure of a grammar of biology on the basis of symmetrologic analysis of hidden regularities of DNA are still valid and they determine the important area of researches that are additionally supported by this article.

Acknowledgments: The described research was conducted in a framework of a long-term cooperation between Russian and Hungarian Academies of Sciences. The authors are grateful to G. Darvas, M. He, A. Pellionisz and I. Stepanyan for their support. Some results of this paper have been possible due to the Russian State scientific contract P377 from July 30, 2009.

REFERENCES

Albrecht-Buehler, G. (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. *Proceedings of the National Academy of Sciences*, November 21, 103(47), 17828–17833.

Bell, S. J., Forsdyke, D. R. (1999) Deviations from Chargaff's Second Parity Rule Correlate with Direction of Transcription, *Journal of Theoretical Biology*, 197, 63-76

Chargaff, E. (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experimentia*, 6, 201

Chargaff, E. (1971) Preface to a Grammar of Biology: A hundred years of nucleic acid research, Science, 172, 637-642, http://www.sciencemag.org/content/172/3984/637.full.pdf?ijkey=99298aa2ffc516d64de947c301cfa5f6a56d3c08&keytype2=tf_ipsecsha

Chargaff, E. (1975) A fever of reason, *Annual Review of Biochemistry*, 44, 1-20

Cristea, P.D. (2005) Representation and analysis of DNA sequences. Genomic Signal Processing and Statistics, Chapter 1, E. Daugherty et al. Eds., Hindawi Publishing Corp., pp. 15–65.

Darvas, G. (2007) Symmetry. Basel: Birkhauser, xi + 508 pp.

Dong, Q., Cuticchia, A.J. (2001) Compositional symmetries in complete genomes. *Bioinformatics*, 17, 557-559.

Forrest, S., Mitchell, M. (1991) The performance of genetic algorithms on Walsh polynomials: Some anomalous results and their explanation. – In R.K.Belew and L.B.Booker, editors, *Proceedings of the Fourth International Conference on Genetic Algorithms*, pp.182-189. Morgan Kaufmann, San Mateo, CA.

Forsdyke, D. R. (2002) Symmetry observations in long nucleotide sequences: a commentary on the Discovery Note of Qi and Citicchia. *Bioinformatics letter*, v. 18, 1, 215-217.

Forsdyke, D. R., Bell, S. J. (2004) A discussion of the application of elementary principles to early chemical observations. *Applied Bioinformatics*, 3, 3-8.

Goldberg, D.E., Korb B., Deb K. (1989) Messy genetic algorithms: Motivation, analysis, and first results. *Complex systems,* 1989, 3(5), 493-530.

Gusev, V., Miroshnichenko, L., Chuzhanova, N. (2009). Detection of fractal structures in the DNA sequences. - International Book Series “Information Science and Omputing”, Book 8, *Classification, Forecasting, Data Mining*, p.117-124, in Russian (Supplement to International Journal "*Information Technologies and Knowledge*", v. 3) http://www.foibg.com/ibs_isc/ibs-08/ibs-08-p17.pdf

Jeffrey, H.J. (1990) Chaos game representation of gene structure. *Nucleic Acids Research*, v.18, 8, 2163-2170

Kong, S-G, Fan W-L, Chen, H-D, Hsu, Z-T, Zhou, N, et al. (2009) *Inverse Symmetry in Complete Genomes and Whole-Genome Inverse Duplication*, PLoS ONE 4(11): e7553. doi:10.1371/journal.pone.0007553

Mitchell, D., Bridge, R. (2006) A test of Chargaff's second rule. *Biochemical and Biophysical Research Communications*, 340(1): 90-94, http://www.ncbi.nlm.nih.gov/pubmed/16364245 .

NCBI. (2012a). http://www.ncbi.nlm.nih.gov/.

NCBI. (2012b). http://www.ncbi.nlm.nih.gov/nuccore/294155300.

Pellionisz, A.J, Graham, R., Pellionisz, P.A., Perez, J.C. (2012a) Recursive Genome Function of the Cerebellum: Geometric Unification of Neuroscience and Genomics. In: Springer Handbook "*The Cerebellum*" pp. 1381-1423 M. Manto, D.L. Gruol, J.D. Schmahmann, N. Koibuchi, F. Rossi (eds.), Handbook of the Cerebellum and Cerebellar Disorders, Submitted October 20, Accepted November 1, 2011.DOI 10.1007/978-94-007-1333-8_61, #Springer Science+Business Media Dordrecht 2012 (full text in http://fr.scribd.com/doc/111439455/BOOK-Unification- of-Neuroscience-and-Genomics-Pellionisz-Et-Al-in-Section-4-Springer-the-Cerebellum-Handbook-2012).

Pellionisz, A.J. (2012b). http://www.junkdna.com/the_genome_is_fractal.html.

Perez, J.-C. (2010). Codon populations in single-stranded whole human genome DNA are fractal and fine-tuned by the golden ratio 1.618. Interdisciplinary Sciences Computational Life Sciences, 2, 1–13. http://www.ncbi.nlm.nih.gov/pubmed/20658335 , full text in: (http://fr.scribd.com/doc/95641538/Codon-Populations-in-Single-stranded-Whole-Human-Genome-DNA-Are-Fractal-and-Fine-tuned-by-the- Golden-Ratio-1-618).

Petoukhov, S.V. (2008a) The degeneracy of the genetic code and Hadamard matrices. arXiv:0802.3366 [q-bio.QM].

Petoukhov, S.V. (2008b) *Matrix genetics, algebras of the genetic code, noise immunity*. Moscow: RCD, 316 p. (in Russian).

Petoukhov, S.V. (2008c) Matrix genetics, part 1: Permutations of positions in triplets and symmetries of genetic matrices, http://arxiv.org/abs/0803.0888, 6th version, 1-34.

Petoukhov, S.V. (2008d) Matrix genetics, part 3: the evolution of the genetic code from the viewpoint of the genetic octave Yin-Yang-algebra. arXiv:0805.4692[q-bio.QM].

Petoukhov, S.V. (2011) Hypercomplex numbers and the algebraic system of genetic alphabets. Elements of algebraic biology. *Hypercomplex numbers in geometry and physics*, v. 8, 2(16), 118-139 (Gipercompleksnyie chisla v geometrii i fizike, in Russian)

Petoukhov, S.V. (2012a) The genetic code, 8-dimensional hypercomplex numbers and dyadic shifts. (7th version from January, 30, 2012), http://arxiv.org/abs/1102.3596

Petoukhov, S.V. (2012b) On fractal structure of long nucleotide sequences. *Joint scientific journal (Ob'edinennyi nauchnyi journal)*, # 6-7, 50 (in Russian)

Petoukhov, S.V. (2012c) Symmetries of the genetic code, hypercomplex numbers and genetic matrices with internal complementarities. *Symmetry: Culture and Science*, in this issue

Petoukhov, S.V., He, M. (2009) Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications. Hershey, USA: IGI Global. 271 p.

Petoukhov, S.V., Svirin, V.I. (2012) Fractal genetic nets and the rules of long genetic sequences. *Joint scientific journal (Ob'edinennyi nauchnyi journal)*, # 8-9, 50-52 (in Russian)

Pevzner, P.A. (2000) *Computational molecular biology. An algorithmic approach*. – Cambridge, Massachusetts: MIT Press.

Prabhu, V. V. (1993) Symmetry observation in long nucleotide sequences. *Nucleic Acids Research*, 21, 2797-2800.

Sueoka, N. (1999) Two aspects of DNA base composition: G + C content and translation-coupled deviation from intra-strand rule of A = T and G = C. – *Journal of Molecular Evolution*, 49, 49–62

Yam, Ph. (1995). Talking trash (Linguistic patterns show up in junk DNA). – *Scientific America,* 272(3), 12-15.

Yamagishi, M.E.B., Herai, R.H. (2011) Chargaff's "Grammar of Biology": New Fractal-like Rules. arXiv:1112.1528v1 from 07.12.2011

Watson, J. D., Crick, F. H. C. (1953) Molecular Structure of Nucleic Acids. *Nature*, 4356, 737.