This article was published and it should be cited as:
Petoukhov S., Petukhova E., Svirin V. (2019) New Symmetries and Fractal-Like Structures in the Genetic Coding System. In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in Computer Science for Engineering and Education, pp. 588-600. ICCSEEA 2018. Advances in Intelligent Systems and Computing, vol 754. Springer, Cham, DOI: https://doi.org/10.1007/978-3-319-91008-6_59.

New Symmetries and Fractal-Like Structures in the Genetic Coding System

Sergey Petoukhov, Elena Petukhova, Vitaliy Svirin

Mechanical Engineering Research Institute, Russian Academy of Sciences, Moscow, M.Kharitonievsky pereulok, 4, Russia
spetoukhov@gmail.com

Abstract. The achievements of molecular genetics and bioinformatics lead to significant changes in technological, medical and many other areas of our lives. This article is devoted to new results of study of structural organization of genetic information in living organisms. A new class of symmetries and fractal-like patterns in long DNA-texts is represented in addition to two Chargaff's parity rules, which played an important role in development of genetics and bioinformatics. Our results provide new approaches for modeling genetic informatics from viewpoints of quantum informatics and theory of dynamic chaos.

Keywords: DNA, Symmetry, Fractal, Probability, Quantum Informatics, Cancer.

1 Tetra-group symmetries in long DNA-texts

The achievements of molecular genetics and biotechnology lead to significant changes in our lives. Genetic engineering and related fields provide not only the diagnostic and therapeutic possibilities of medicine that were unthinkable before, but also the emergence of new materials with surprising properties, new approaches to solving problems of nanotechnology, robotics, artificial intelligence systems, etc. Specialists consider projects for the cultivation of finished bodies of cars from chitin or bones. Several DNA strings connected together form a hinge-type mechanism for nanorobots, capable of bending and unbending by a chemical signal. Of particular importance is the knowledge of the principles of noise immunity of the genetic code in connection with the problem of ensuring noise immunity of information systems of control [1-6].

A road to the knowledge of bioinformational patents of living matter for their use in engineering, medicine and education is inextricably linked with the study of hidden regularities of hereditary information recorded in DNA molecules. The species of living organisms are amazingly diverse, but in all organisms genetic information is recorded in DNA and RNA molecules in the form of long texts of four letters: adenine A, cytosine C, guanine G and thymine T (in RNA uracil U is used instead of thymine). This article represents a new class of symmetries and fractal-like relations in long DNA-texts, the discovery of which shows elements of their fractal grammar. The goal of our research is revealing a participation of fractal structures in long DNA-texts.

DNA molecules are very long. For example, the human genome is a text with several billions of genetic letters A, T, C and G (it is equivalent to a text of thousands of thick books). DNA-texts of different organisms are represented in the GenBank (https://ru.wikipedia.org/wiki/GenBank), which contains hundreds of millions of sequences for more than one hundred thousand organisms. The set of known DNA-texts contains hundreds of billions of letters A, T, C and G.

What rules exist in these basic texts of living organisms? The modern situation is described by the following citation: "*What will we have when these genomic sequences are determined? ... We are in the position of Johann Kepler when he first began looking for patterns in the volumes of data that Tycho Brahe had spent his life accumulating*" [7]. Kepler did not make his own astronomic observations, but he found – in the huge astronomic data from the collection of Tycho Brahe - his Kepler's laws of symmetric planetary movements relative to the Sun. In 100 years after Kepler, thanks to the laws of Kepler, Newton discovered the law of universal gravitation. We have revealed new hidden symmetries in many long texts of single-stranded DNA of several dozen species of organisms, including the complete genomes of some organisms from the GenBank (without exceptions till now).

Below we explain our study but previously we should remind about two Chargaff's parity rules, which are known in genetics long ago. They are important because they point to a kind of "grammar of biology" [8]: a set of hidden rules that govern the structure of DNA. The first Chargaff's parity rule states that in any double-stranded DNA segment, the number of frequencies of adenine A and thymine T are equal, and so are frequencies of cytosine C and guanine G [8, 9]. The rule was an important clue to model the double helix structure of DNA by J.Watson and F.Crick .

The second Chargaff's parity rule states that both %A ≈ %T and %G ≈ %C are approximately valid in single stranded DNA for long nucleotide sequences. Many works of different authors are devoted to confirmations and discussions of this second Chargaff's rule [10 – 25]. Originally, CSPR is meant to be valid only to mononucleotide frequencies in 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 ... This is known in the literature as the Symmetry Principle*" [25, p. 2]. The work [22] 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 all these works, authors concentrate their attention on the comparison of frequencies (or probabilities) of separate fragments of DNA-texts. By contrast to this, we study not individual probabilities of separate fragments but collective (or total) probabilities of special groups of fragments in long DNA-texts.

Let us explain our approach more detailed. Each of long DNA-sequences (for example, the sequence CAGGTATCGAAT…) can be represented not only in the form of the text of 1-letter words (C-A-G-G-T-A-T-C-G-A-A-T-…) but also in the form of the text of 2-letter words (CA-GG-TA-TC-GA-AT-…) or in the form of the text of 3-letter words (CAG-GTA-TCG-AAT-…) or in the form of the text of n-letter words in a general case. We briefly call such representations "n-letter representations" of DNA-texts. In each of such n-letter representations, we study total probabilities of

all members of each of 4 groups of those n-letter words, which have the same letter (A, T, C or G) on the same position $k \leq n$ inside words (we call such groups "tetra-groups"). Each of four DNA-letters A, T, C and G defines its own group of n-letter words inside a complete tetra-group of each n-representation of DNA-text.

By definition, in an n-letter representation of DNA-text, a total (or collective) probability of a group of n-letter words is the ratio: total quantity of all n-letter words of this group divided by the total quantity of all n-letter words. For example, the 2-letter text with 7 words AT-CT-GG-AG-AA-CA-AC contains 4 words with the letter A at their first position. In this text, the total probability $P_2(A_1)$ of such words is equal to 4/7 = 0,571. This text contains also 2 words with the letter A at their second position. Correspondingly their total probability $P_2(A_2)$ is equal to 2/7 = 0,286. In a general case, we denote by the symbol $P_n(A_k)$ a total probability of a group of n-letter words that have the letter A at their position k ($k \leq n$) in a long DNA-text. The similar symbols $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ denote correspondingly total probabilities of all those n-letter words, which have the letters T, C and G at their position k.

It was unexpectedly for us to discover that all these n-letter representations of a long DNA-text are symmetrically interrelated each other on the basis of approximate equalities of total probabilities of all words with the same letter on the same position inside words (n = 1, 2, 3, 4, 5, …. is not too large). These approximate equalities are symmetrical relations, whom we call "tetra-group symmetries in long DNA-texts". For example, Fig. 1 shows results of our calculation of total probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ for DNA-text of the human chromosome № 1 that contains 248956422 letters (here the values of probabilities are rounded to the third decimal place; more detailed results are shown in [4]). One can see in Fig. 1 approximate equalities of high level of accuracy inside the set of $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ under different values n = 1, 2, 3, 4, 5 and $k \leq n$.

$\mathbf{P_1(A_1) = P_2(A_1) = P_2(A_2) = P_3(A_1) = P_3(A_2) = P_3(A_3) = P_4(A_1) = P_4(A_2) = P_4(A_3) = P_4(A_4) = P_5(A_1) = P_5(A_2) = P_5(A_3) = P_5(A_4) = P_5(A_5) = 0{,}291}$.
$\mathbf{P_1(T_1) = P_2(T_1) = P_2(T_2) = P_3(T_1) = P_3(T_2) = P_3(T_3) = P_4(T_1) = P_4(T_2) = P_4(T_3) = P_4(T_4) = P_5(T_1) = P_5(T_2) = P_5(T_3) = P_5(T_4) = P_5(T_5) = 0{,}292.}$
$\mathbf{P_3(C_1) = P_5(C_4) = 0{,}208;\ P_1(C_1) = P_2(C_1) = P_2(C_2) = P_3(C_2) = P_3(C_3) = P_4(C_1) = P_4(C_2) = P_4(C_3) = P_4(C_4) = P_5(C_1) = P_5(C_2) = P_5(C_3) = P_5(C_5) = 0{,}209.}$
$\mathbf{P_1(G_1) = P_2(G_1) = P_2(G_2) = P_3(G_1) = P_3(G_2) = P_3(G_3) = P_4(G_1) = P_4(G_2) = P_4(G_3) = P_4(G_4) = P_5(G_1) = P_5(G_2) = P_5(G_3) = P_5(G_4) = P_5(G_5) = 0{,}209.}$

Fig. 1. Total probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ in the corresponding tetra-groups of n-letter words (n = 1, 2, 3, 4, 5) in the DNA-text of the following sequence, which contains 248956422 letters: Homo sapiens chromosome 1, GRCh38.p7 Primary Assembly. NCBI Reference Sequence: NC_000001.11; https://www.ncbi.nlm.nih.gov/nuccore/NC_000001.11

We have got similar results about an approximate equality of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ (where n = 1, 2, 3, 4, 5) for several dozen long DNA-texts, including the complete genomes of a number of organisms. These results have given

the opportunity to formulate the following rules of the tetra-group symmetries in long DNA-texts [4].

The first rule of tetra-group symmetries in long DNA-texts: if a long sequence of single stranded DNA is represented in different forms of texts of n-letter words (n = 1, 2, 3, 4, 5…. is not too large), then - in these texts - probabilities of words with the letter X (X = A, T, C, G) in their position k ≤ n are approximately equal to each other independently on values n.

The second rule of tetra-group symmetries in long DNA-texts: if a long sequence of single stranded DNA is represented in different forms of texts of n-letter words (n = 1, 2, 3, 4, 5…. is not too large), then - in these texts - probabilities of words with the letter X (X = A, T, C, G) in their position k ≤ n are approximately equal to each other independently on values k.

The third rule of tetra-group symmetries in long DNA-texts: if a long sequence of those single-stranded DNA, that satisfy the second Chargaff's rule, is represented in different forms of texts of n-letter words (n = 1, 2, 3, 4, 5…. is not too large), then - in these texts - probabilities of words with the complementary letters A and T in their position k are approximately equal to each other. The same is true for probabilities of words with the complementary letters C and G in their position k.

These rules are candidacies for the role of universal rules of long DNA-texts in living bodies. Further research is needed to define a degree of universality of these rules and these cooperative genetic symmetries. These phenomenologic rules can be modelled on the basis of a quantum informational approach [4].

2 Tetra-group symmetries in complete sets of chromosomes

Human organisms contain 24 chromosomes: 22 autosomes and 2 sex chromosomes X and Y. These chromosomes are long DNA molecules, the length of texts in which lie in the range from 50 to 250 million letters approximately. Autosomes are numbered from 1 to 22. We have studied tetra-group symmetries of long DNA-texts in each of 24 human chromosomes for cases n = 1, 2, 3, 4, 5. In the result we have obtained not only a confirmation of the described 3 rules of the tetra-group symmetries for all separate chromosomes but also an additional unexpected result concerning the complete set of chromosomes: numeric characteristics of tetra-group symmetries of long DNA-texts of separate chromosomes are approximately equal to each other for all human chromosomes. This result was unexpected since 24 human cromosomes differ greatly by their molecular dimensions, their sequences of letters, kinds and quantities of genes in them, cytogenetic bands (which shows biochemical specifity of different parts of chromosomes), etc. But in relation to values of tetra-group symmetries of their DNA-texts (that is, in relation to their total probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$) 24 human chromosomes are very similar each other [4]. Fig. 2 shows the average values of the probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ for all 24 human chromosomes.

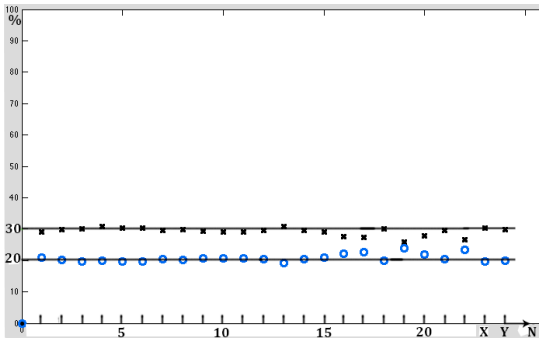

Fig. 2. The graphical representation of the average values of the probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ for all 24 human chromosomes. The abscissa axis contains numberings N of chromosomes, and the ordinate axis contains average values of these probabilities in percent. The symbol "o" corresponds the average values of $P_n(C_k) \approx P_n(G_k)$, and the symbol "x" corresponds the average values of $P_n(A_k) \approx P_n(T_k)$. Direct lines correspond values 20% and 30%.

One should add that fluctuations of values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ for different n and k (n = 1, 2, 3, 4, 5) in relation to their average values were very small. For example, in the case of the human chromosome № 1, fluctuations of the probabilities were ±0,006% (Fig. 3). For other chromosomes, the fluctuations of the probabilities were of the same order of magnitude.

Average value of $P_n(A_k)$ and fuctuations (%)	Average value of $P_n(T_k)$ and fuctuations (%)	Average value of $P_n(C_k)$ and fuctuations (%)	Average value of $P_n(G_k)$ and fuctuations (%)
29,100±0,005	29,176±0,006	20,850±0,006	20,874±0,005

Fig. 3. Fluctuations of the probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ under diferent values n and k in relation to their average values in the case of the DNA-text of human chromosome №1.

It seems that – in the case of all human chromosomes - the average values of the probabilities of $P_n(A_k) \approx P_n(T_k)$ are concentrated around the value of 30% and the average values of the probabilities of $P_n(C_k) \approx P_n(G_k)$ are concentrated around the value of 20%. In theory of musical harmony, the ratio 30/20 = 3/2 is called "quint" (or "fifth").

We also analyzed tetra-group symmetries of the DNA-texts in the complete sets of chromosomes of a few organisms, which are traditionally used as model organisms in the study of genetics, development and disease: a nematode *Caenorhabditis elegans*, a fruit fly *Drosophila melanogaster*, a plant *Arabidopsis thaliana*. All received results show that the represented tetra-group rules are implemented not only for separate long DNA-texts but also for studied complete sets of chromosomes of eukaryotes. These initial results allow putting forward the hypothesis about existence of the following

general rule of tetra-group symmetries of DNA-texts in complete sets of chromosomes of different organisms: in the complete set of chromosomes of each of eukaryot organisms, characteristics of tetra-group symmetries of the DNA-texts of separate chromosomes are approximately equal to each other for all chromosomes [4].

Further researches are needed to check a degree of universality of this rule.

3 Fractal genetic nets, tetra-group symmetries of DNA-texts and a fractal grammar of biology

Our article [26] has introduced the notion of “fractal genetic nets” (FGN) of texts for revealing hidden regularities in long DNA-texts in a connection with Charaff’s thoughts about a "grammar of biology" [8]). Each FGN of texts can contain different fractal genetic trees (FGT). In that work we have represented results testifying in favor of existence of new symmetry principles in long nucleotide sequences in an addition to the known symmetry principle on the basis of the generalized Chargaff’s second parity rule.

Below we represent our results about implementation of the rules of tetra-group symmetries of long DNA-texts at different levels of different fractal genetic trees and nets. In line with our article [26], FGT of various types are constructed by the method of sequential positional convolutions of a long DNA-text into a set of ever-shorter texts. Fig. 4 explains a construction of FGT of various types by means of an example of FGT for a DNA-text, which is represented as a sequence S_0 of 3-letter words (a sequence of triplets). In each triplet, 0, 1 and 2 numbers its three positions correspondingly. At the first level of the text convolution, an initial long sequence S_0 of triplets is transformed by means of a positional convolution into three new sequences of nucleotides $S_{1/0}$, $S_{1/1}$ and $S_{1/2}$, each of which is 3 times shorter in comparison with the initial sequence S_0 (in this notation of sequences, numerator of the index 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 into three new sequences: $S_{1/0}$ is convolved into $S_{2/00}$, $S_{2/01}$, $S_{2/02}$; $S_{1/1}$ – into $S_{2/10}$, $S_{2/11}$, $S_{2/12}$; $S_{1/2}$ – into $S_{2/20}$, $S_{2/21}$, $S_{2/22}$. Similarly, the third level and subsequent levels of the convolution are constructed to form a multi-level tree of sequences of triplets called "the fractal genetic tree for the triplet convolution" or briefly "FGT-3". Texts at lower levels of any FGT can be figuratively called “daughter texts” of the original long DNA-text S_0.

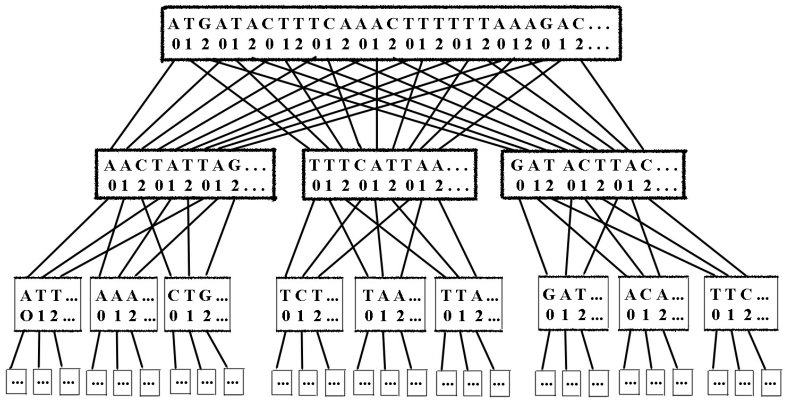

Fig. 4. The scheme of the fractal genetic tree (FGT-3) of a DNA-text, which is represented as a sequence of triplets (from [26]).

This FGT possesses a fractal-like character if the enumeration of positions is only taken into account: each of long sequences of this FGT can be taken as an initial sequence to form a similar genetic net on its basis (Fig. 4). In general case, the FGT can be built not only for triplets, but also for other n-plets (n = 2, 4, 5, ...) 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 build FGT-2, FGT-4, FGT-5, etc. for n = 2, 3, 4, 5,… correspondingly. A set of these FGT-2, FGT-3, FGT-4, FGT-5, … forms a net of separate trees; FGN is a set of such separate trees.

For a long DNA-text of any biological organism, one can study implementation of the described rules of tetra-group symmetries in long texts at different levels of the convolution in cases of FGT-2, FGT-3, FGT-4, etc. Our own results of initial study of enough long DNA-texts of different organisms show implementation of these tetra-group rules in all texts at initial levels of the FGT-2, FGT-3 and FGT-4. Moreover values of the probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ are approximately repeated in all convoluted texts at different initial levels of convolutions in these tested cases of FGT. Figs. 5&6 illustrate this phenomenologic fact for long texts at initial levels of convolutions in FGT-2 and FGT-3 for human sex chromosomes X and Y, whose DNA-texts contain 156040895 and 57227415 letters correspondingly.

One can see from Figs. 5&6 that fluctuation intervals of studied probabilities are very narrow for the set of texts at each of the initial levels of the fractal genetic trees. Moreover, fluctuation intervals for each of separate kinds of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ are approximately equal to each other for the sets of texts at all considered levels of the FGT-2 and the FGT-3. Our results can be considered as evidences in favor of a fractal grammar of genetics in line with the Chargaff's problem about a grammar of biology.

	Level 0	Level 1/0	Level 1/1	Level 2/00	Level 2/01	Level 2/10	Level 2/11
$P_n(A_k)\in$	0.3017÷ 0.3019	0.3017÷ 0.3019	0.3016÷ 0.3020	0.3016÷ 0.3017	0.3017÷ 0.3019	0.3015÷ 0.3015	0.3015÷ 0.3019

$P_n(T_k)\in$	0.3028÷ 0.3029	0.3028÷ 0.3028	0.3027÷ 0.3027	0.3025÷ 0.3029	0.3028÷ 0.3029	0.3027÷ 0.3031	0.3027÷ 0.3027
$P_n(C_k)\in$	0.197÷ 0.1971	0.1969÷ 0.1971	0.1969÷ 0.197	0.1967÷ 0.1971	0.1969÷ 0.1971	0.1968÷ 0.197	0.1971÷ 0.1971
$P_n(G_k)\in$	0.1981÷ 0.1982	0.1981÷ 0.1982	0.1981÷ 0.1982	0.1981÷ 0.1982	0.1981÷ 0.1982	0.198÷ 0.1984	0.1981÷ 0.1983

	Level 0	Level 1/0	Level 1/1	Level 1/2	Level 2/00	Level 2/01
$P_n(A_k)\in$	0.3017÷ 0.3019	0.3016÷ 0.3020	0.3017÷ 0.3017	0.3017÷ 0.3019	0.3015÷ 0.3018	0.3016÷ 0.3019
$P_n(T_k)\in$	0.3028÷ 0.3029	0.3027÷ 0.3029	0.3028÷ 0.3029	0.3027÷ 0.3029	0.3024÷ 0.3029	0.3025÷ 0.3028
$P_n(C_k)\in$	0.1970÷ 0.1971	0.1970÷ 0.1970	0.1970÷ 0.1974	0.1967÷ 0.1971	0.1969÷ 0.1971	0.1967÷ 0.1968
$P_n(G_k)\in$	0.1981÷ 0.1982	0.1981÷ 0.1981	0.1980÷ 0.1980	0.1981÷ 0.1982	0.198÷ 0.1981	0.1979÷ 0.1985

	Level 2/02	Level 2/10	Level 2/11	Level 2/12	Level 2/20	Level 2/21	Level 2/22
$P_n(A_k)\in$	0.3015÷ 0.3019	0.3018÷ 0.3019	0.3015÷ 0.3018	0.3014÷ 0.3014	0.3016÷ 0.3019	0.3016÷ 0.3017	0.3015÷ 0.3018
$P_n(T_k)\in$	0.3025÷ 0.3026	0.3025÷ 0.3025	0.3026÷ 0.3030	0.3024÷ 0.3030	0.3025÷ 0.3029	0.3025÷ 0.3030	0.3027÷ 0.3030
$P_n(C_k)\in$	0.1966÷ 0.1973	0.1967÷ 0.1969	0.1968÷ 0.1969	0.1969÷ 0.1974	0.1968÷ 0.1971	0.1966÷ 0.1968	0.1967÷ 0.1971
$P_n(G_k)\in$	0.1980÷ 0.1981	0.1977÷ 0.1987	0.1979÷ 0.1983	0.1978÷ 0.1982	0.1980÷ 0.1981	0.1980÷ 0.1985	0.1978÷ 0.1981

Fig. 5. Tables of fluctuation intervals of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ for the set of all texts at each of levels of convolutions in the FGT-2 (upper table) and in the FGT-3 (bottom tables) in the case of the human chromosome X (NCBI Reference Sequence: NC_000023.11).

	Level 0	Level 1/0	Level 1/1	Level 2/00	Level 2/01	Level 2/10	Level 2/11
$P_n(A_k)\in$	0.2983÷ 0.2987	0.2982÷ 0.2982	0.2979÷ 0.2987	0.2980÷ 0.2989	0.2981÷ 0.2981	0.2981÷ 0.2987	0.2976÷ 0.2992
$P_n(T_k)\in$	0.3009÷ 0.3012	0.3008÷ 0.3021	0.3007÷ 0.3014	0.3009÷ 0.3013	0.3005÷ 0.3020	0.3005÷ 0.3010	0.3005÷ 0.3009
$P_n(C_k)\in$	0.1998÷ 0.1998	0.1997÷ 0.1997	0.1995÷ 0.1995	0.1996÷ 0.1999	0.1997÷ 0.1997	0.1996÷ 0.2001	0.1995÷ 0.1995
$P_n(G_k)\in$	0.1998÷ 0.2003	0.1994÷ 0.2001	0.1999÷ 0.2004	0.1995÷ 0.1998	0.1994÷ 0.2002	0.1995÷ 0.2002	0.1996÷ 0.2003

	Level 0	Level 1/0	Level 1/1	Level 1/2	Level 2/00	Level 2/01
$P_n(A_k)\in$	0.2983÷ 0.2987	0.2981÷ 0.2988	0.2984÷ 0.2989	0.298÷ 0.2982	0.2981÷ 0.2988	0.2976÷ 0.2984
$P_n(T_k)\in$	0.3009÷ 0.3012	0.3008÷ 0.3012	0.3007÷ 0.3013	0.3008÷ 0.3012	0.300÷ 0.3013	0.3010÷ 0.3019
$P_n(C_k)\in$	0.1998÷ 0.1998	0.1995÷ 0.2004	0.1999÷ 0.2002	0.1999÷ 0.1999	0.199÷ 0.1999	0.1994÷ 0.1994
$P_n(G_k)\in$	0.1998÷ 0.2003	0.1996÷ 0.1996	0.1996÷ 0.1997	0.200÷ 0.2007	0.1996÷ 0.2001	0.1991÷ 0.2003

	Level 2/02	Level 2/10	Level 2/11	Level 2/12	Level 2/20	Level 2/21	Level 2/22
$P_n(A_k)\in$	0.2974÷ 0.2982	0.2978÷ 0.2991	0.2979÷ 0.2979	0.2983÷ 0.2985	0.2975÷ 0.2976	0.2974÷ 0.2990	0.2975÷ 0.2979
$P_n(T_k)\in$	0.3005÷ 0.3023	0.2997÷ 0.3006	0.3002÷ 0.3010	0.3006÷ 0.3008	0.3005÷ 0.3023	0.3005÷ 0.3010	0.3005÷ 0.3006
$P_n(C_k)\in$	0.1994÷ 0.2005	0.1994÷ 0.1996	0.1995÷ 0.2004	0.1996÷ 0.2011	0.1993÷ 0.2003	0.1997÷ 0.2004	0.1993÷ 0.2002
$P_n(G_k)\in$	0.1990÷ 0.1990	0.2000÷ 0.2008	0.1988÷ 0.2007	0.1991÷ 0.1996	0.1997÷ 0.1998	0.1996÷ 0.1996	0.1998÷ 0.2013

Fig. 6. Tables of fluctuation intervals of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ for the set of all texts at each of levels of convolutions in the FGT-2 (upper table) and in the FGT-3 (bottom tables) in the case of the human chromosome Y (NCBI Reference Sequence: NC_000024.10).

These results about a fractal grammar of long DNA-texts are additionally interesting by the following reasons:

- Many biological organisms have fractal-like inherited configurations in their bodies. This phenomenon can be considered as a consequence of the fractal-like organization of long DNA texts with the participation of tetra-group symmetries;
- As known, fractals allow a colossal compression of information (https://en.wikipedia.org/wiki/Fractal_compression). It is obvious that an opportunity of informaton compression is essential for genetic systems. Modern computer science knows a great number of methods of information compression including many methods of fractal compression. Our described results about fractal genetic nets can lead to a discovery of those «genetic» methods of information compression, which are used in genetic systems and in biological bodies in the whole;
- Many authors published their ideas and materials about relations of genomes with fractal structures in different aspects [27-33]. For example, work [33] shown an existence of fractal globule in the three dimensional architecture of whole genomes, where spatial chromosome territories exist and where maximmally dense packing is provided on the basis of a special fractal packing, which provides the ability to easily fold and unfold any genomic locus. One should note that, by contrast to the work [33], in our work we study not spatial packing of whole genomes in a form of fractal globules but the quite another thing: we study the fractal organization of long DNA-texts, in particular, in the form of described fractal genetic nets or trees of different

kinds (FGT-n, where n = 1, 2, 3, 4,…), which are connected with tetra-group symmetries of these texts (these symmetries and fractals were never studyed early).
- Fractals are actively used in study of cancer; some modern data testify that cancer processes are related with fractal patterns and their development [34-38].
- Fractals are connected with theory of dynamic chaos, which has many applications in engineering technologies. We believe that the discovery of fractal-like properties of DNA-texts related with their tetra-group symmetries can lead to new ideas in theoretical and application areas, including problems of artificial intelligence and in-depth study of genetic phenomena for medical and biotechnological tasks [3, 39-44].

Conclusions

A special class of symmetries is implemented in long DNA-texts of different organisms. Long DNA-texts are constructed by Nature with using fractal structures. This can be one of reasons of existence of a great number of inherited fractal configurations in biological bodies in their normal and pathologic states including fractal patterns in cancerous and biorythmic structures. Fractal patterns are connected with the theory of dynamic chaos, which has many applications in sciences and technology. A specifity of fractal genetic nets can provoke a further development of the theory of dynamic chaos and its applications.

References.

1. Petoukhov S. V.: Matrix Genetics, Algebras of the Genetic Code, Noise Immunity. RCD, Moscow, Russia (2008) (in Russian).
2. Petoukhov, S.V., He M.: Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications. IGI Global, Hershey, USA (2010).
3. Petoukhov S., Petukhova E., Hazina L., Stepanyan I., Svirin V., Silova T. The Genetic Coding, United-Hypercomplex Numbers and Artificial Intelligence. In: Hu Z., Petoukhov S., He M. (eds) Advances in Artificial Systems for Medicine and Education. AIMEE 2017. Advances in Intelligent Systems and Computing, vol 658. Springer, Cham; DOI https://doi.org/10.1007/978-3-319-67349-3_1, Print ISBN 978-3-319-67348-6, Online ISBN 978-3-319-67349-3 (online is available from 20 August 2017 – https://link.springer.com/search?query=978-3-319-67348-6).
4. Petoukhov S.V. The rules of long DNA-sequences and tetra-groups of oligonucleotides. (2017) (https://arxiv.org/abs/1709.04943)
5. Hu Z.B., Petoukhov S.V., Petukhova E.S. I-Ching, dyadic groups of binary numbers and the geno-logic coding in living bodies.- *Progress in Biophysics and Molecular Biology*. (2017); in press, available online 18 September 2017, http://authors.elsevier.com/sd/article/S0079610717300949).
6. Hu Z.B., Petoukhov S.V. Generalized crystallography, the genetic system and biochemical esthetics. *Structural Chemistry*, v. 28, №1, pp. 239-247 (2017). doi:10.1007/s11224-016-0880-0 . http://link.springer.com/journal/11224/28/1/page/2
7. Fickett, J., & Burks, Chr. (1989). Development of a database for nucleotide sequences.- In: M.S.Waterman (Ed.), Mathematical Methods in DNA Sequences, pp.1-34. Florida: CRC Press. (1989).

8. Chargaff, E. Preface to a Grammar of Biology: A hundred years of nucleic acid research. - *Science*, 172, pp. 637-642. (1971).
9. Chargaff, E. Structure and function of nucleic acids as cell constituents. - *Fed. Proc*., 10, pp. 654-659. (1951).
10. Albrecht-Buehler G. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc Natl Acad Sci U S A. November 21; 103(47): pp. 17828–17833. (2006).
11. Baisnee P-F., Hampson S., Baldi P. Why are complementary DNA strands symmet ric? – *Bioinformatics*, 18(8),1021-33. (2002).
12. Bell, S. J., Forsdyke, D. R. Deviations from Chargaff's second parity rule correlate with direction of transcription - *J. Theo. Bio*., 197, 63-76. (1999).
13. Chargaff E. A fever of reason. Annu. Rev. Biochem., 44: pp. 1-20. (1975).
14. Dong Q., Cuticchia A. J. Compositional symmetries in complete genomes. *Bioinformatics*, 17, pp. 557-559. (2001).
15. Forsdyke D. R. A stem-loop "kissing" model for the initiation of recombination and the origin of introns. *Molecular Biology and Evolution*, 12, pp. 949–958. (1995).
16. Forsdyke D.R. Symmetry observations in long nucleotide sequences: a commentary on the discovery of Qi and Cuticchia. *Bioinformatics letter*, v.18, № 1, pp. 215-217. (2002).
17. Forsdyke D. R. Evolutionary Bioinformatics. New-York: Springer Verlag. (2006).
18. Forsdyke D. R., Bell S. J. Purine-loading, stem-loops, and Chargaff's second parity rule. *Applied Bioinformatics*, 3, pp. 3–8. (2004).
19. Mitchell, D., Bridge, R. A test of Chargaff's second rule. - *BBRC*, 340, pp. 90-94. (2006).
20. Okamura K., Wei J., Scherer S. Evolutionary implications of inversions that have caused intra-strand parity in DNA. *Bmc Genomics*, 8, 160–166. (2007). http://www.gutenberg.org/files/39713/39713-h/39713-h.htm#Page_264
21. Perez J.-Cl. The "3 Genomic Numbers" Discovery: How Our Genome Single-Stranded DNA Sequence Is "Self-Designed" as a Numerical Whole. - *Applied Mathematics*, 4, 37-53, (2013). http://dx.doi.org/10.4236/am.2013.410A2004
22. Prabhu, V. V. Symmetry observation in long nucleotide sequences. *Nucleic Acids Res*., 21, pp. 2797-2800. (1993).
23. Rapoport A.E., Trifonov E.N. Compensatory nature of Chargaff's second parity rule. *Joик nal of Biomolecular Structure and Dynamics*, November, pp. 1-13, (2012). DOI:10.1080/07391102.2012.736757.
24. Sueoka N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. Journal of Molecular Evolution, 40, pp. 318–325. (1995).
25. Yamagishi, M., Herai, R. Chargaff's "Grammar of Biology": New Fractal-like Rules. http://128.84.158.119/abs/1112.1528v1. (2011).
26. Petoukhov S.V., Svirin V.I. Fractal genetic nets and symmetry principles in long nucleotide sequences. - *Symmetry: Culture and Science*, vol. 23, № 3-4, pp. 303-322. (2012). http://petoukhov.com/PETOUKHOV_SVIRIN_FGN.pdf
27. Jeffrey H.J. Chaos game representation of gene structure. - *Nucleic Acids Research*, Vol. 18, No. 8, pp. 2163-2170. (1990).
28. Peng C.K., Buldyrev S.V., Goldberger A.L., Havlin S., Sclortino F., Simons M., Stanley H.E. Long-range correlations in nucleotide sequences. - Nature 356, pp.168–170. (1992).

29. Peng C.K., Buldyrev S.V., Goldberger A.L., Havlin S., Sclortino F., Simons M., Stanley H.E. Fractal landscape analysis of DNA walks. - Physica A., 191(1-4): 25-9. (1992 Dec 15).
30. Pellionis A.J. The principle of recursive genome function. - Cerebellum 7: pp. 348–359 (2008). DOI 10.1007/s12311-008-0035-y
31. Pellionisz A. J., Graham R., Pellionisz P. A. and Perez J. C. Recursive Genome Function of the Cerebellum: Geometric Unification of Neuroscience and Genomics, In: M. Manto, D. L.Gruol, J. D. Schmahmann, N. Koibuchi and F. Rossi, Eds., Handbook of the Cerebellum and Cerebellar Disorders, pp. 1381-1423 (2012).
32. Perez J.C. Codon populations in single-stranded whole human genome DNA are fractal and fine-tuned by the Golden Ratio 1.618. - *Interdiscip Sci Comput Life Sci* 2: pp. 228–240 (2010). DOI: 10.1007/s12539-010-0022-0
33. Lieberman-Aiden E., van Berkum N. L., Williams L., Imakaev M., Ragoszy T., Telling A., Lajoie B.R., Sabo P.J., Dorschner M.O., Sandstrom R., Bernstein B., Bender M.A., Groudine M., Gnirke A., Stamatoyannopoulos J., Mirny L.A., Lander E.S., Dekker J. Com prehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. - *Science*, Vol. 326, Issue 5950, pp. 289-293, (2009, 09 October). DOI: 10.1126/science.1181369
34. Baish J.W., Jain R.K. Fractals and cancer. – Cancer Research, 60, pp. 3683–3688, (2000).
35. Bizzarri M., Giuliani A., Cucina A., Anselmi F. D., Soto A. M., Sonnenschein C. Fractal analysis in a Systems Biology approach to cancer. - *Semin Cancer Biol* . *June* ; 21(3): pp. 175–182. (2011). doi:10.1016/j.semcancer.2011.04.002.
36. Lennon F.E., Cianci G.C., Cipriani N.A., Hensing T.A., Zhang H.J., Chin-Tu Chen, Murgu S.D., Vokes E.E., Vannier M.W., Salgia R. Lung cancer - a fractal viewpoint. - Nat Rev Clin Oncol, 12(11):_pp. 664–675. (2015_November)._doi:_10.1038/nrclinonc. 2015.108).
37. Dokukin M.E., Guz N.V., Woodworth C.D., Sokolov I. Emergence of fractal geometry on the surface of human cervical epithelial cells during progression towards cancer. - *New J Phys*. 2015 Mar 10; 17(3). pii: 033019 (2015).
38. Perez J.C. Sapiens Mitochondrial DNA Genome Circular Long Range Numerical Meta Structures are Highly Correlated with Cancers and Genetic Diseases mtDNA Mutations. - *J Cancer Sci Ther*, 9:6 (2017). DOI: 10.4172/1948-5956.1000469.
39. Abo-Zahhad M., Ahmed S.M., Abd-Elrahman Sh.A. Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques, *IJITCS*, vol.4, no.8, pp.22-36 (2012). http://www.mecs-press.org/ijitcs/ijitcs-v4-n8/v4n8-3.html
40. Abo-Zahhad M., Ahmed S.M., Abd-Elrahman Sh.A. A Novel Circular Mapping Technique for Spectral Classification of Exons and Introns in Human DNA Sequences. - *IJITCS*, Vol. 6, No. 4, p. 19-29 (March 2014). DOI: 10.5815/ijitcs.2014.04.02
41. Meher J.K., Panigrahi M.R., Dash G.N., Meher P.K. Wavelet Based Lossless DNA Sequence Compression for Faster Detection of Eukaryotic Protein Coding Regions. *IJIGSP*, Vol. 4, No. 7, pp. 47-53 (July 2012). http://www.mecs-press.org/ijigsp/ijigsp-v4-n7/v4n7-5.html
42. Prakash Chandra Srivastava, Anupam Agrawal, Kamta Nath Mishra, P. K. Ojha, R. Garg. Fingerprints, Iris and DNA Features based Multimodal Systems: A Review. - IJITCS, vol.5, no.2, pp.88-111 (2013). DOI:10.5815/ijitcs.2013.02.10.
43. Hamdy M. Mousa. DNA-Genetic Encryption Technique. - *International Journal of Com puter Network and Information Security* (IJCNIS), Vol.8, No.7, pp.1-9 (2016). DOI: 10.5815/ijcnis.2016.07.01
44. Hossein S.M., S.Roy. A Compression & Encryption Algorithm on DNA Sequences Using Dynamic Look up Table and Modified Huffman Techniques. *IJITCS* Vol. 5, No. 10, pp. 39-61 (September 2013). DOI: 10.5815/ijitcs.2013.10.05