The rules of long DNA-sequences and tetra-groups of oligonucleotides

Sergey V. Petoukhov

Head of Laboratory of Biomechanical System, Mechanical Engineering Research Institute of the Russian Academy of Sciences, Moscow
spetoukhov@gmail.com, http://petoukhov.com/

Comment: Some materials of this article were presented by the author in the keynote speech at the congress on energy and information medicine «Energiemedizin 2017» (10-11 June 2017, Bad Soden, Germany, http://dgeim.de/energiemedizin-kongress-2017/ and in the keynote speech at the international conference “Artificial Intelligence, Medical Engineering, Education” (Moscow, Russia, 21-23 August 2017, http://www.ruscnconf.org/aimee2017/index.html).

Abstract. The article represents a new class of hidden symmetries in long sequences of oligonucleotides of single stranded DNA from their representative set. These symmetries are an addition to symmetries described by the second Chargaff’s parity rule (%A ≅ %T and %G ≅ %C). These new symmetries and their rules concern collective probabilities of oligonucleotides from special tetra-groups and their subgroups in long DNA-texts including complete sets of chromosomes of human and some model organisms. These rules of tetra-group probabilities are considered as possible candidacies for the role of universal rules of long DNA-sequences. A quantum-informational model of genetic symmetries of these collective probabilities is proposed on the basis of the known quantum-mechanic statement that quantum state of a multicomponent system is defined by the tensor product of quantum states of its subsystems. In this model, nitrogenous bases C, T, G, A of DNA are represented as computational basis states of 2-qubit quantum CTGA-systems. A model explanation of the second Chargaff’s rule from the standpoint of quantum informatics is proposed. The biological meaning of these new quantum-information symmetries of long DNA texts is associated with the common ability of all living organisms to grow and develop on the basis of incorporation into their body of new and new molecules of nutrients becoming new quantum-mechanic subsystems of the united quantum-mechanic organism. An important role of resonances, photons and photonic crystals in quantum-information genetics is noted.

Key words. Chargaff’s rules, symmetry, long DNA-sequence, chromosome, tetra-group of oligonucleotides, probabilities, tensor product, quantum informatics, qubit, resonance, photon

CONTENTS

1. Introduction
2. Tetra-groups of oligonucleotides and collective probabilities of tetra-groups in long DNA-texts
3. The first, second and third rules of symmetries of collective probabilities of tetra-groups in long DNA-texts
4. DNA-alphabets, genetic binary oppositions and the tensor product of matrices
5. The quantum-information model to explain and predict symmetries of collective probabilities in tetra-groups of long DNA-texts
6. The explanation of the second Chargaff’s parity rule on the basis of the quantum-information approach.
7. About short DNA-texts
8. Resonances, photons and quantum-information genetics
9. The rules of symmetries of collective probabilities in tetra-groups in the complete set of human chromosomes. The fourth rule of symmetries of tetra-group probabilities (for complete sets of chromosomes)
10. Tetra-group rules in complete sets of chromosomes of model organisms: Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Mus musculus
11. Fractal genetic nets and the fifth and sixth rules of symmetries of collective probabilities of tetra-groups. On a fractal grammar of long DNA-texts
12. About letter-ordered representations of long DNA-texts conserving their collective probabilities in tetra-groups.
13. On the biological sense of the symmetries of collective probabilities of tetra-groups in long DNA texts.
14. Some concluding remarks

Appendix 1. Additional data about tetra-group rules and tetra-group symmetries in long DNA-texts
Appendix 2. Symmetries of tetra-group probabilities in the compete set of human chromosomes
Appendix 3. Symmetries of tetra-group probabilities in the complete set of chromosomes of a nematode Caenorhabditis elegans
Appendix 4. Symmetries of tetra-group probabilities in the complete set of chromosomes of Drosophila melanogaster
Appendix 5. Symmetries of tetra-group probabilities in the complete set of chromosomes of Arabidopsis thaliana
Appendix 6. Symmetries of tetra-group probabilities in the complete set of chromosomes of Mus musculus
Appendix 7. Confirmations of model predictions on symmetries of collective probabilities for 4^k-groups of n-plets in long DNA-texts

Acknowledgments
References

1. Introduction.

Two Chargaff’s parity rules are well known in genetics. They are important because they point to a kind of “grammar of biology” (these words were used by E.Chargaff in the title of his article [Chargaf, 1971]): 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 occurrences (or
frequencies) of adenine A and thymine T are equal, and so are frequencies of cytosine C and guanine G [Chargaff, 1951, 1971]. The rule was an important clue that J. Watson and F. Crick used to develop their model of the double helix structure of DNA.

The second Chargaff’s parity rule states that both $\%A \equiv \%T$ and $\%G \equiv \%C$ 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 [Albrecht-Buehler, 2006, 2007; Bainsne, Hampson, Baldi, 2002; Bell, Forsdyke, 1999; Chargaff, 1971, 1975; Dong, Cuticchia, 2001; Forsdyke, 1995, 2002, 2006; Forsdyke, Bell, 2004; Mitchell, Bridge, 2006; Okamura, Wei, Scherer, 2007; Perez, 2013; Prabhu, 1993; Rapoport, Trifonov, 2012; Sueoka, 1995; Yamagishi, Herai, 2011]. Originally, CSPR is meant to be valid only to mononucleotide frequencies (that is quantities of monoplets) 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 [Prabhu, 1993]. This is known in the literature as the Symmetry Principle” [Yamagishi, Herai, 2011, p. 2]. The work [Prabhu, 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 literature, a few synonyms of the term “n-plets” are used: n-tuples, n-words or n-mers).

These parity rules, including generalized Chargaff’s second parity rule for n-plets in long nucleotide sequences, concerns the equality of frequencies of two separate mononucleotides or two separate oligonucleotides, for example: the equality of frequencies of adenine and thymine; the equality of frequencies of the doublet CA and its complement-reverse doublet TG; the equality of the triplets CAT and its complement-reverse triplet ATG, etc. By contrast to this, to study hidden symmetries in long sequences of oligonucleotides of single stranded DNA, we apply a comparative analysis of equalities not for frequencies of separate oligonucleotides but for aggregated frequencies (or collective frequencies) of oligonucleotides from separate subgroups of their certain tetra-groups. Below we explain the notion of these tetra-groups of oligonucleotides and represent new rules of probabilities in tetra-groups of long sequences of oligonucleotides in single stranded DNA of many organisms including complete sets of chromosomes of human and some model organisms.

2. Tetra-groups of oligonucleotides and collective probabilities of tetra-groups in long DNA-texts

Information in DNA strands is written by means of the tetra-group of nitrogenous bases: adenine A, cytosine C, guanine G and thymine T (in RNA the tetra-group of nitrogenous bases contains uracil U instead of thymine T). E. Chargaff has received both his parity rules by comparative analysis of frequencies of each of 4 members of this tetra-group of mononucleotides in DNA. He and his followers studied DNA sequences as sequences of mononucleotides, frequencies of separate fragments of which were compared. In other words, Chargaff studied long DNA-texts in a form of texts of 1-letter words and calculated frequencies (or probabilities) of each of 4 members of the genetic tetra-group of 1-letter words A, T, C and G.
But it is obvious that 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-AT-…) 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-sequences. Any of such long DNA-texts of n-letter words can be considered as a collection of 4 subgroups of possible tetra-groups, each of which is defined by an attribute of one of 4 letters A, T, C, G at a certain position inside n-letter words of the DNA-text. As we know, till now nobody studied systematically frequencies and probabilities of subgroups of such tetra-groups in DNA-texts of n-letter words. The proposed set of representations of long DNA-sequences plays a key role in revealing hidden symmetries in these sequences. This article shows some results of author’s study of hidden symmetries in long DNA-texts of n-letter words.

In other speaking, in contrast to Chargaff, in our approach, firstly, we analyzes DNA sequences not as sequences of mononucleotides but as sequences of oligonucleotides: doublets, triplets, 4-plets, 5-plets, etc. Secondly, we compare not values of separate frequencies of individual oligonucleotides in long nucleotide sequences but values of sums of individual frequencies of all oligonucleotides, which belong to each of 4 subgroups of special tetra-groups of oligonucleotides of an identical length (such sum is called a collective frequency of the subgroup of the tetra-group). At the final stage of the analysis we compare collective probabilities (or percentage) of collective frequencies of different subgroups of such tetra-groups in the considered sequence. The mentioned tetra-groups are formed in each of considered cases by means of certain positional attributes of the letters A, T, C and G inside oligonucleotides of an identical length. Each of 4 subgroups of such tetra-group combines all oligonucleotides of the same length n (n-plets), which possess the identical letter at their certain position. To simplify the explanation, Fig. 1 shows the example of two tetra-groups, which are formed and studied by us for the analysis of long sequences of doublets.

| Subgroups of tetra-groups | Composition of subgroups of doublets with identical letters at their first positions | Composition of subgroups of doublets with identical letters at their second position |
|---------------------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| A-subgroup                | AA, AC, AG, AT                                                                  | AA, CA, GA, TA                                                                  |
| T-subgroup                | TC, TA, TT, TG                                                                  | CT, AT, TT, GT                                                                  |
| C-subgroup                | CC, CA, CT, CG                                                                  | CC, AC, TC, GC                                                                  |
| G-subgroup                | GC, GA, GT, GG                                                                  | CG, AG, TG, GG                                                                  |

Fig. 1. Compositions of two tetra-groups of doublets with 4 doublets in each of their 4 subgroups.

In the first tetra-group in Fig. 1, the complete alphabet of 16 doublets is
divided into 4 subsets with 4 doublets in each by the attribute of an identical letter on the first position in each of doublets. The complect of these 4 subsets is called the tetra-group of doublets on the basis of this attribute; each of the 4 subsets is called a subgroup of the tetra-group; each of 4 subgroup has its individual name with an indication of its characteristic letter (A-subgroup, T-subgroup, C-subgroup and G-subgroup). In the second tetra-group (Fig. 1, right) the complete set of 16 doublets is divided into 4 subsets with 4 doublets in each by the attribute of an identical letter on the second position in each of doublets. The name "tetra-group" is used since 4 subgroups exist here and 4 letters of DNA play a decisive role in the dismemberment of the set of m-plets on the regular subsets in question. It is obvious that corresponding subgroups of both tetra-groups of doublets in Fig. 1 (left and right columns) are interrelated on the basis of cyclic shifts of positions in doublets. For example, the set of doublets AA, AC, AG, AT in the A-subgroup of the first tetra-group (Fig. 1, left column) is transformed into the A-subgroup AA, CA, GA, TA of the second tetra-group (Fig. 1, right column) by the cyclic shift of positions in doublets. Similar mutual transformations of subgroups of corresponding tetra-groups on the basis of cyclic shifts are also valid in the case of alphabets of triplets, 4-plets, 5-plets, etc. Fig. 2 shows three tetra-groups of triplets, which are used for the analysis of long sequences of triplets.

| Subgroups of tetra-groups | Composition of subgroups of triplets with an identical letter at their 1st position | Composition of subgroups of triplets with an identical letter at their 2nd position | Composition of subgroups of triplets with an identical letter at their 3rd position |
|---------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| A-subgroup                | AAA, AAC, AAG, AAT, ATA, ATC, ATG, ATT, ACA, ACC, ACG, ACT, AGA, AGC, AGG, AGT | AAA, CAA, GAA, TAA, AAT, CAT, GAT, TAT, AAC, CAC, GAC, TAC, AGA, CAG, GAG, TAG | AAA, CAA, GAA, TAA, CTA, ATA, TTA, GTA, CCA, CAA, CTA, CGA, CGA, AGA, TGA, GGA |
| T-subgroup                | TAA, TAC, TAG, TAT, TTA, TTC, TTG, TTT, TCA, TCC, TCG, TCT, TGA, TGC, TGG, TGT | ATA, CTA, GTA, TTA, ATT, CTT, GTT, TTT, ATC, CTC, GTC, TTC, ATG, CTG, GTG, TGG | AAT, CAT, GAT, TAT, CTT, ATT, TTT, GTT, CCT, CAT, CTT, CGT, CGT, AGT, TGT, GGT |
| C-subgroup                | CAA, CAC, CAG, CAT, CTA, CTC, CTG, CTT, CCA, CCC, CCG, CCG, CGA, CGC, CGG, CGT | ACA, CCA, GCA, TCA, ACT, CCT, GCT, TCT, ACC, CCC, GCC, TCC, ACG, CCG, GCG, TCG | AAC, CAC, GAC, TAC, CTC, ATC, TTC, GTC, CCC, CAC, CTC, CGC, CGC, AGC, TGC, GGC |
| G-subgroup                | GAA, GAC, GAG, GAT, GTA, GTG, GTT, GCA, GCC, GCG, GCT, GGA, GGC, GGG, GGT | AGA, CGA, GGA, TGA, AGC, CGT, GGT, TGT, AGC, CGC, GGC, TGC, AGG, CCG, GGG, TGG | AAG, CAG, GAG, TAG, CTG, ATG, TTT, GTT, CCG, CAG, CTG, CGG, CCG, AGG, TGG, GGG |

Fig. 2. Compositions of three tetra-groups of triplets with 16 triplets in each of their 4 subgroups.
In a general case of a sequence of n-plets, the complete alphabet of $4^n$ n-plets is divided into 4 subgroups with $4^{n-1}$ n-plets in each by the attribute of an identical letter on the chosen position inside n-plets. In this case n tetra-groups of n-plets are formed:

- The tetra-group on the basis of the attribute of an identical letter on the 1st position of n-plets;
- The tetra-group on the basis of the attribute of an identical letter on the 2nd position of n-plets;
- ...
- The tetra-group on the basis of the attribute of an identical letter on the n-th position of n-plets.

We use the symbol $\Sigma_n$ ($n = 1, 2, 3, \ldots$) to denote the total quantity of n-plets in the considered DNA-sequence of n-plets; for example, the expression $\Sigma_3 = 100000$ means that an analyzed sequence of triplets contains 100000 triplets. Let us define notions and symbols of collective frequencies and collective probabilities of subgroups of tetra-groups in long sequences of n-plets, where each of n-plets has its individual frequency (or number of its occurrences): for example, doublets have their individual frequencies $F(AC), F(CA), \text{ etc.}$

In a long sequence of n-plets, collective frequencies $F_n(A_k), F_n(T_k), F_n(C_k)$ and $F_n(G_k)$ of each of 4 subgroups of a tetra-group of n-plets are defined as the sum of all individual frequencies of n-plets belong to this subgroup (here the index n denotes the length of n-plets; the index $k = 1, 2, 3, \ldots, n$ denotes the position of the letter in n-plets). For example, in the case of a sequence of triplets with the letters A, C, G and T on the second positions of triplets, these collective frequencies are denoted $F_3(A_2), F_3(T_2), F_3(C_2)$ and $F_3(G_2)$; correspondingly the expression $F_3(A_2) = 50000$ means that a considered sequence of triplets contains 50000 triplets with the letter A in their second positions. Fig. 3 shows appropriate definitions of collective frequencies $F_2(A_k), F_2(T_k), F_2(C_k)$ and $F_2(G_k)$ for the case of sequences of doublets (where $k = 1, 2$), which are analyzed from the standpoint of both tetra-groups from Fig. 1. Fig. 3 also shows – for the case of sequences of doublets – collective probabilities $P_n(A_k), P_n(T_k), P_n(C_k)$ and $P_n(G_k)$ of separate subgroups of tetra-groups; in general case these probabilities are defined by expressions $P_n(A_k) = F_n(A_k)/\Sigma_n, P_n(T_k) = F_n(T_k)/\Sigma_n, P_n(C_k) = F_n(C_k)/\Sigma_n$ and $P_n(G_k) = F_n(G_k)/\Sigma_n$. Below we represent the tetra-group rules for these collective probabilities, which are sum of individual probabilities of separate n-plets. More precisely, in the case of a sequence of n-plets, the probability of each of 4 subgroups of a separate tetra-group is a sum of $4^{n-1}$ individual probabilities of such n-plets. For example, in the case of sequence of 5-plets, the probability $P_5(A_1)$ of the A-subgroup, which combines 5-plets with the letter A at their first position, is a sum of $4^4 = 256$ individual probabilities of 5-plets: $P_5(A_1) = P(AAAAA) + P(AAAAT) + P(AAAAC) + \ldots$, etc.
Two subgroups of any tetra-group of n-plets with the complementary letters on the characteristic positions are conditionally called complementary subgroups of the appropriate tetra-group. For example, the A-subgroup and the T-subgroup are complementary subgroups in each of two tetra-groups in Fig. 1. Such complementary subgroups participate in one of the represented tetra-group rules of long sequences of n-plets in single stranded DNA.

3. The first, second and third rules of symmetries of collective probabilities of tetra-groups in long DNA-texts

It is generally accepted that long sequences contain more than 50 thousands or 100 thousands nucleotides [Albrecht-Buehler, 2006; Prahbu, 1993; Rapoport, Trifonov, 2012]. In this Section we show data of analysis of two sequences of Homo sapiens chromosomes, each of which has its length of exactly one million nucleotides (only two these sequences of such length are retrieved from Entrez Search Field of Genbank by the known range operator 1000000:1000001[SLEN], https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html). Fig. 4 shows calculation data of the first of them from the standpoint of the proposed tetra-group approach: Homo sapiens chromosome 7 sequence, ENCODE region ENm012, accession NT_086368, version NT_086368.3, https://www.ncbi.nlm.nih.gov/nuccore/NT_086368.3. These data include collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$, $P_n(G_k)$ of subgroups of appropriate tetra-groups of doublets, triplets, 4-plets and 5-plets of the sequence. We use the data in Fig. 4 to formulate the suppositional general rules of symmetries (or rules of approximate equalities) of these collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$, $P_n(G_k)$ in long texts of single stranded DNA. These rules can be briefly named "tetra-group rules" or "rules of tetra-group symmetries". Below the formulated tetra-group rules will be confirmed by similar analysis of a representative set of other long nucleotide sequences from the Genbank. By analogy with the generalized Chargaff’s second rule, in cases of these new rules it is assumed that the length “n” of considered n-plets is much smaller than the length of the studied sequence. At this stage, we study only long DNA-sequences of doublets, triplets, 4-plets and 5-plets.
| NUCLEOTIDES | DOUBLETS | TRIPLETs | 4-PLETs | 5-PLETs |
|-------------|----------|----------|---------|---------|
| $\Sigma_1 = 1000000$ | $\Sigma_2 = 500000$ | $\Sigma_3 = 333333$ | $\Sigma_4 = 250000$ | $\Sigma_5 = 200000$ |
| $F_i(A_i) = 307519$ | $P_i(A_i) = 0.3075$ | $F_3(A_i) = 102657$ | $P_3(A_i) = 0.3080$ | $F_5(A_i) = 61097$ |
| $P_i(A_i) = 0.3073$ | $F_2(A_i) = 153652$ | $P_2(A_i) = 0.3080$ | $F_4(A_i) = 76990$ | $P_4(A_i) = 0.3080$ |
| $F_i(T_i) = 335023$ | $P_i(T_i) = 0.3350$ | $F_3(T_i) = 111609$ | $P_3(T_i) = 0.3344$ | $F_5(T_i) = 67004$ |
| $P_i(T_i) = 0.3350$ | $F_2(T_i) = 167514$ | $P_2(T_i) = 0.3348$ | $F_4(T_i) = 83590$ | $P_4(T_i) = 0.3350$ |
| $F_i(C_i) = 176692$ | $P_i(C_i) = 0.1767$ | $F_3(C_i) = 58893$ | $P_3(C_i) = 0.1767$ | $F_5(C_i) = 35525$ |
| $P_i(C_i) = 0.1763$ | $F_2(C_i) = 88158$ | $P_2(C_i) = 0.1767$ | $F_4(C_i) = 44181$ | $P_4(C_i) = 0.1776$ |
| $F_i(G_i) = 180766$ | $P_i(G_i) = 0.1808$ | $F_3(G_i) = 60174$ | $P_3(G_i) = 0.1810$ | $F_5(G_i) = 36374$ |
| $P_i(G_i) = 0.1813$ | $F_2(G_i) = 90676$ | $P_2(G_i) = 0.1805$ | $F_4(G_i) = 45239$ | $P_4(G_i) = 0.1819$ |
| $F_3(G_i) = 90676$ | $P_3(G_i) = 0.1808$ | $F_2(G_i) = 60174$ | $P_2(G_i) = 0.1805$ | $F_4(G_i) = 45239$ |
| $P_3(G_i) = 0.1810$ | $F_3(G_i) = 60174$ | $P_2(G_i) = 0.1805$ | $F_4(G_i) = 45239$ | $P_4(G_i) = 0.1819$ |
| $F_4(G_i) = 45239$ | $P_4(G_i) = 0.1819$ | $F_3(G_i) = 60174$ | $P_3(G_i) = 0.1810$ | $F_5(G_i) = 36374$ |
| $P_4(G_i) = 0.1808$ | $F_3(G_i) = 60174$ | $P_3(G_i) = 0.1810$ | $F_5(G_i) = 36374$ | $P_5(G_i) = 0.1819$ |

Fig. 4. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$ and $F_n(G_k)$ and also collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of subgroups of tetra-groups for sequences of n-plets, which have the same letter in their position $k$, in the case of the following sequence: Homo sapiens chromosome 7 sequence, 1000000 bp, encode region ENm012, accession NT_086368, version NT_086368.3, https://www.ncbi.nlm.nih.gov/nuccore/NT_086368.3. Collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ are marked by red for a visual comfort of their comparison each with other.

From data in Fig. 4 one can assume existence of the following tetra-group rules for long sequences of oligonucleotides (these rules are confirmed by similar analizes of a set of other long DNA-sequences represented below).
The first tetra-group rule (the rule of approximate equality of the collective probabilities of n-plets with an identical letter in their fixed position k, regardless of the length n of the considered n-plets):

- In long sequences of n-plets of single stranded DNA, collective probabilities \( P_n(X_k) \) \( (X = A, T, C, G; k \leq n; n = 1, 2, 3, 4, 5... \text{ is not too large}) \) of those subset of n-plets, which have the letter X in their position k, are approximately equal to the individual probability of the nucleotide X regardless on values n.

This first rule can be also formulated in another way:

- 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..., \text{ is not too large}), \text{ then in these texts - probabilities of words with the letter X} \( (X = A, T, C, G) \text{ in their position} \( k \leq n \) \text{ are approximately equal to each other regardless on values n.}\)

For example, Fig. 4 shows the following approximate symmetries:

- \( P_1(A_1) = 0.3075 \approx P_2(A_1) = 0.3073 \approx P_3(A_1) = 0.3080 \approx P_4(A_1) = 0.3080 \approx P_5(A_1) = 0.3055; \)
- \( P_1(T_1) = 0.3350 \approx P_2(T_1) = 0.3350 \approx P_3(T_1) = 0.3348 \approx P_4(T_1) = 0.3344 \approx P_5(T_1) = 0.3350; \)
- \( P_1(C_1) = 0.1767 \approx P_2(C_1) = 0.1763 \approx P_3(C_1) = 0.1767 \approx P_4(C_1) = 0.1767 \approx P_5(C_1) = 0.1776; \)
- \( P_1(G_1) = 0.1808 \approx P_2(G_1) = 0.1813 \approx P_3(G_1) = 0.1805 \approx P_4(G_1) = 0.1810 \approx P_5(G_1) = 0.1819. \)

For a more convenient vision of this rule and corresponding symmetrical relations, Fig. 5 reproduces separately data about collective frequencies \( P_n(X_k) \) (where \( X = A, T, C, G \)) from Fig. 4. Each of tabular rows contains approximately identical values of collective probabilities and – for emphasizing this fact – all its cells are marked by the same color.

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| \( X_1 = 1000000 \) | \( X_2 = 500000 \) | \( X_3 = 333333 \) | \( X_4 = 250000 \) | \( X_5 = 200000 \) |
| \( P(A) = 0.4078 \) | \( P(A) = 0.4072 \) | \( P(A) = 0.4080 \) | \( P(A) = 0.4096 \) | \( P(A) = 0.4055 \) |
| \( P(T) = 0.3350 \) | \( P(T) = 0.3350 \) | \( P(T) = 0.3348 \) | \( P(T) = 0.3344 \) | \( P(T) = 0.3350 \) |
| \( P(C) = 0.1767 \) | \( P(C) = 0.1763 \) | \( P(C) = 0.1767 \) | \( P(C) = 0.1767 \) | \( P(C) = 0.1776 \) |
| \( P(G) = 0.1808 \) | \( P(G) = 0.1813 \) | \( P(G) = 0.1805 \) | \( P(G) = 0.1810 \) | \( P(G) = 0.1819 \) |
| \( P(A) = 0.3075 \) | \( P(A) = 0.3073 \) | \( P(A) = 0.3080 \) | \( P(A) = 0.3080 \) | \( P(A) = 0.3055 \) |
| \( P(T) = 0.3350 \) | \( P(T) = 0.3350 \) | \( P(T) = 0.3348 \) | \( P(T) = 0.3344 \) | \( P(T) = 0.3350 \) |
| \( P(C) = 0.1767 \) | \( P(C) = 0.1763 \) | \( P(C) = 0.1767 \) | \( P(C) = 0.1767 \) | \( P(C) = 0.1776 \) |
| \( P(G) = 0.1808 \) | \( P(G) = 0.1813 \) | \( P(G) = 0.1805 \) | \( P(G) = 0.1810 \) | \( P(G) = 0.1819 \) |

Fig. 5. Collective probabilities \( P_n(A_k), P_n(T_k), P_n(C_k) \) and \( P_n(G_k) \) from Fig. 4.

The first tetra-group rule can be graphically illustrated by a particular example in Fig. 6 for the same long DNA-sequence (Fig. 4).
Fig. 6. The illustration of the dependence of collective probabilities $P_n(A_1), P_n(T_1), P_n(C_1)$ and $P_n(G_1)$ from the length $n = 1, 2, 3, 4, 5$ of n-plets in the case of Homo sapiens chromosome 7 sequence, 1000000 bp. Numerical data are taken from Fig. 4.

The first tetra-group rule and corresponding symmetrical relations can be briefly expressed by the following expression (1) for any of values $n = 1, 2, 3, 4, \ldots$ under the condition of a fixed value of the index $k$:

$$P_1(X_1) \equiv P_n(X_k), \ k \leq n,$$

where $X$ means any of letters $A, T, C$ and $G$; $n = 1, 2, 3, 4, \ldots$ is not too large.

One should remind here that in the expression (1) various collective probabilities $P_n(X_k)$ are sum of individual probabilities of very different quantities of n-plets: the collective probability $P_2(A_1)$ is sum of individual probabilities of 4 doublets; the collective probability $P_3(A_1)$ is sum of individual probabilities of 16 triplets; the collective probability $P_4(A_1)$ is sum of individual probabilities of 64 tetraplets and the collective probability $P_5(A_1)$ is sum of individual probabilities of 256 pentaplets.

The second tetra-group rule (the rule of approximate equalities of collective probabilities of n-plets with an identical letter in their position $k$, regardless of the value of the positional index $k$ in the considered n-plets):

- In long sequences of n-plets of single stranded DNA, collective probabilities $P_n(X_k)$ $(X = A, T, C, G; \ k \leq n; \ n = 1, 2, 3, 4, \ldots$ is not too large) of those subset of n-plets, which have the letter $X$ in their position $k$, are approximately equal to the individual probability of the nucleotide $X$ regardless on values $k$.

This second rule can be also formulated in another way:
- If a long sequence of single stranded DNA is represented in different forms of texts of n-letter words $(n = 1, 2, 3, 4, \ldots$ is not too large), then - in these texts - probabilities of words with the letter $X$ $(X = A, T, C, G)$ in their position $k \leq n$ are approximately equal to each other regardless on values $k$.

Fig. 5 facilitates a vision of this rule and corresponding symmetrical relations in the considered DNA-sequence: each of tabular columns contains approximately...
identical values of collective probabilities and – for emphasizing this fact – all such cells are marked by the same color.

For example, one can see from data in Fig. 4 the following approximate symmetries:

- \( P_5(A_1) = 0.3055 \approx P_5(A_2) = 0.3072 \approx P_5(A_3) = 0.3081 \approx P_5(A_4) = 0.3085 \approx P_5(A_5) = 0.3083 \);
- \( P_5(T_1) = 0.3350 \approx P_5(T_2) = 0.3352 \approx P_5(T_3) = 0.3356 \approx P_5(T_4) = 0.3348 \approx P_5(T_5) = 0.3344 \);
- \( P_5(C_1) = 0.1776 \approx P_5(C_2) = 0.1772 \approx P_5(C_3) = 0.1760 \approx P_5(C_4) = 0.1766 \approx P_5(C_5) = 0.1761 \);
- \( P_5(G_1) = 0.1819 \approx P_5(G_2) = 0.1804 \approx P_5(G_3) = 0.1803 \approx P_5(G_4) = 0.1801 \approx P_5(G_5) = 0.1812 \).

The second tetra-group rule and corresponding approximate symmetries can be graphically illustrated by a particular example in Fig. 7 for the same long DNA-sequence (Fig. 4).

![Graph](image)

**Fig. 7.** The illustration of the dependence of collective probabilities \( P_5(A_k), P_5(T_k), P_5(C_k) \) and \( P_5(G_k) \) of appropriate subgroups of tetra-groups from the index \( k \) of a position of the attributive letter in 5-plets in the case of the sequence of 5-plets of Homo sapiens chromosome 7 sequence, 1000000 bp. Numerical data are taken from Fig. 4.

The second tetra-group rule and corresponding approximate symmetries can be briefly expressed by the following expression (2) for any of values \( k \) under the condition of a fixed value of a length \( n \) of n-plets: 

\[
P_n(X_1) \equiv P_n(X_k), \quad k \leq n,
\]

where \( X \) means any of letters A, T, C and G; \( n = 1, 2, 3, 4,... \) is not too large.

One can see from expressions (1) and (2) that the first rule and the second rule can be jointly expressed in a brief way by the expression (3) without the mentioned conditions of a fixation of values \( n \) and \( k \):

\[
P_1(X_1) \equiv P_n(X_k), \quad k \leq n,
\]

where \( X \) means any of letters A, T, C and G; \( n = 1, 2, 3, 4,... \) is not too large.
The third tetra-group rule for those long DNA-sequences, which correspond to the second Chargaff's rule (the rule of approximate equalities of collective probabilities of complementary subgroups of tetra-groups):

- in tetra-groups of long sequences of n-plets of those single-stranded DNA, that satisfy the second Chargaff's rule, collective probabilities \( P_n(A_k) \) and \( P_n(T_k) \) of the complementary A- and T-subgroups of tetra-groups are approximately equal to each other. The same is true for collective probabilities \( P_n(C_k) \) and \( P_n(G_k) \) of the complementary C- and G-subgroups of tetra-groups.

This third rule can be also formulated in another way:
- 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 \((n = 1, 2, 3, 4, 5\ldots \) 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 \( k \).

This rule is expressed by expressions (4) for any of considered values of \( n \) and \( k \), where \( k \leq n \) and \( n = 1, 2, 3, 4, \ldots \) is not too large:  

\[
P_n(A_k) \cong P_n(T_k) \quad \text{and} \quad P_n(C_k) \cong P_n(G_k).
\]

(4)

For example, one can see from data in Fig. 4 for the case \( k=1 \) the following:

- \( P_1(A_1)=0,3075 \cong P_1(T_1)=0,3350; P_2(A_1)=0,3073 \cong P_2(T_1)=0,3350; \)
- \( P_3(A_1)=0,3080 \cong P_3(T_1)=0,3348; P_4(A_1)=0,3080 \cong P_4(T_1)=0,3344; \)
- \( P_5(A_1)=0,3055 \cong P_5(T_1)=0,3350; \)
- \( P_1(C_1)=0,1767 \cong P_1(G_1)=0,1808; P_2(C_1)=0,1763 \cong P_2(G_1)=0,1813; \)
- \( P_3(C_1)=0,1767 \cong P_3(G_1)=0,1805; P_4(C_1)=0,1767 \cong P_4(G_1)=0,1810; \)
- \( P_5(C_1)=0,1776 \cong P_5(G_1)=0,1819. \)

The similar situation is true for cases \( k = 2, 3, 4, 5 \) in Fig. 4.

We emphasize that approximately equal collective frequencies \( F_n(A_k) \) and \( F_n(T_k) \) of the complementary A- and T-subgroups of tetra-groups (as well as \( F_n(C_k) \) and \( F_n(G_k) \) of the complementary C- and G-subgroups), which are used in expressions (4), can differ significantly by values of individual frequencies of  n-plets in them. For example, for the sequence of doublets in Fig. 4, these collective frequencies are sum of the following individual frequencies of separate doublets:

- \( F_2(A_1) = F(AA)+F(AC)+F(AG)+F(AT) = 52506+23434+31500+46212, \)
- \( F_2(T_1) = F(TT)+F(TG)+F(TC)+F(TA) = 61483+35978+29290+40763, \)
- \( F_2(C_1) = F(CA)+F(CC)+F(CG)+F(CA) = 32721+32721+2800+33533, \)
- \( F_2(G_1) = F(GT)+F(GG)+F(GC)+F(GA) = 26281+19812+16706+27877 \)

(5)

One can see from the expression (5) that, for example, the individual frequency \( F(CG)=2800 \), which is used in the expression of the collective frequency \( F_2(C_1) \) of the C-subgroup, differs by a factor of 6 from the individual
frequency of the complementary doublet \( F(GC) = 16706 \), which is used in the expression of the collective frequency \( P_2(G_1) \) of the complementary G-subgroup. Correspondingly the individual probability \( P(CG) = F(CG)/\Sigma_2 = 2800/500000 = 0.0056 \) differs by a factor of 6 from the individual probability of the complementary doublet \( P(GC) = F(GC)/\Sigma_2 = 16706/500000 = 0.0334 \). This indicates that the described tetra-group rules can’t be reduced to rules of individual n-plets, but they form a special class of rules of a collective organization in oligonucleotide sequences of single stranded DNA.

The third tetra-group rule can be considered as a generalization of the second Chargaff’s parity rule, which states an approximate equality of individual frequencies of complementary letters \( F(A) \approx F(T) \) and \( F(A) \approx F(T) \) (or probabilities \( P(A) \approx P(T) \) and \( P(A) \approx P(T) \)) in long nucleotide sequences of single stranded DNA. In the case of the sequence in Fig. 4, the second Chargaff’s parity rule is expressed by expressions \( P_1(A_1) = 0.3075 \approx P_1(T_1) = 0.3350 \) and \( P_1(C_1) = 0.1767 \approx P_1(G_1) = 0.1808 \). The level of accuracy of the second Chargaff’s rule execution for this sequence is determined by the difference of probabilities: for the probabilities of complementary letters A and T in the mononucleotide sequence, this difference is equal to \( P_1(T_1) - P_1(A_1) = 0.3350 - 0.3075 = 0.0275 \), and for the probabilities of complementary letters C and G it is equal to \( P_1(G_1) - P_1(C_1) = 0.1808 - 0.1767 = 0.0041 \). For the analyzed sequence (Fig. 4), Fig. 8 shows that the same level of accuracy \( P_1(T_1) - P_1(A_1) = 0.0275 \) is approximately executed for all differences \( P_8(T_8) - P_8(A_8) \), and that the same level of accuracy \( P_1(G_1) - P_1(C_1) = 0.0041 \) is approximately executed for all differences \( P_8(G_8) - P_8(C_8) \). It testifies that the second Chargaff’s parity rule can be considered as a particular case of the third tetra-group rule.

| NUCLEOTIDES     | DOUBLETS  | TRIPLETS  | 4-PLETS  | 5-PLETS  |
|-----------------|-----------|-----------|----------|----------|
| \( P_1(T_1) - P_1(A_1) \) = 0.0275 | \( P_2(T_1) - P_2(A_1) \) = 0.0277 | \( P_3(T_1) - P_3(A_1) \) = 0.0268 | \( P_4(T_1) - P_4(A_1) \) = 0.0264 | \( P_5(T_1) - P_5(A_1) \) = 0.0295 |
| \( P_1(G_1) - P_1(C_1) \) = 0.0041 | \( P_2(G_1) - P_2(C_1) \) = 0.0005 | \( P_3(G_1) - P_3(C_1) \) = 0.0038 | \( P_4(G_1) - P_4(C_1) \) = 0.0043 | \( P_5(G_1) - P_5(C_1) \) = 0.0043 |
| \( P_2(T_1) - P_2(A_2) \) = 0.0273 | \( P_3(T_1) - P_3(A_2) \) = 0.0272 | \( P_4(T_1) - P_4(A_2) \) = 0.0261 | \( P_5(T_1) - P_5(A_2) \) = 0.028 |
| \( P_2(G_2) - P_2(C_2) \) = 0.0031 | \( P_3(G_2) - P_3(C_2) \) = 0.0044 | \( P_4(G_2) - P_4(C_2) \) = 0.0027 | \( P_5(G_2) - P_5(C_2) \) = 0.0032 |
| \( P_3(T_3) - P_3(A_3) \) = 0.0285 | \( P_4(T_3) - P_4(A_3) \) = 0.0291 | \( P_5(T_3) - P_5(A_3) \) = 0.0275 |
| \( P_3(G_3) - P_3(C_3) \) = 0.004 | \( P_4(G_3) - P_4(C_3) \) = 0.0058 | \( P_5(G_3) - P_5(C_3) \) = 0.0043 |
| \( P_4(T_4) - P_4(A_4) \) = 0.0284 | \( P_5(T_4) - P_5(A_4) \) = 0.0263 |
| \( P_4(G_4) - P_4(C_4) \) = 0.0034 | \( P_5(G_4) - P_5(C_4) \) = 0.0035 |
| \( P_5(T_5) - P_5(A_5) \) = 0.0261 | \( P_5(G_5) - P_5(C_5) \) = 0.0051 |
The second Chargaff's parity rule is known in the literature as the Symmetry Principle. New kinds of approximate symmetries in long DNA-texts can be conditionally named as tetra-group symmetries. They expand scientific knowledge about Symmetry Principles in organization of long DNA-texts.

For long DNA-sequences of n-plets, the described tetra-group rules have a predictive power: knowing the collective probabilities of only two of 4 subgroups of one of the tetra-groups (for example, $P_1(A_1)$ and $P_1(C_1)$), one can predict approximate values of probabilities of other subgroups of this and other tetra-groups on the basis of the expressions (1-4).

Fig. 9 shows a joint representation of all three tetra-group rules described above: sectors of the same color contain approximately the same values of collective probabilities. Each of its rings corresponds to an appropriate length "n" of n-plets: the smallest ring corresponds to the case of doublets (in this case $k = 1, 2$); the next ring corresponds to the case of triplets (in this case $k = 1, 2, 3$), etc.

Fig. 9. The joint representation of all three tetra-group rules: sectors of the same color contain approximately the same values of collective probabilities.

Now let us turn to the second DNA-sequence with one million nucleotides, which was retrieved from Entrez Search Field of Genbank by the range operator 1000000:1000001[SLEN]: Homo sapiens chromosome 5 sequence, ENCODE region ENm002; accession NT_086358, version NT_086358.1, https://www.ncbi.nlm.nih.gov/nuccore/NT_086358.1. Fig. 10 shows our results of the analysis of this long sequence from the standpoint of tetra-groups of its oligonucleotides (by analogy with data in Fig. 4).

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| $\Sigma_1 = 1000000$ | $\Sigma_2 = 500000$ | $\Sigma_3 = 333333$ | $\Sigma_4 = 250000$ | $\Sigma_5 = 200000$ |
| $F_1(A_1) = 283765$ | $P_2(A_1) = 0.2838$ | $F_2(A_1) = 142054$ | $P_3(A_1) = 0.2841$ | $F_3(A_1) = 94805$ | $P_4(A_1) = 0.2844$ | $F_4(A_1) = 70944$ | $P_5(A_1) = 0.2838$ | $F_5(A_1) = 56865$ | $P_6(A_1) = 0.2843$ |
| $F_1(T_1) = 270465$ | $P_2(T_1) = 0.2705$ | $F_2(T_1) = 135082$ | $P_3(T_1) = 0.2702$ | $F_3(T_1) = 90432$ | $P_4(T_1) = 0.2713$ | $F_4(T_1) = 67590$ | $P_5(T_1) = 0.2704$ | $F_5(T_1) = 53956$ | $P_6(T_1) = 0.2698$ |
| $F_1(C_1) = 226576$ | $P_2(C_1) = 0.2267$ | $F_2(C_1) = 113153$ | $P_3(C_1) = 0.2263$ | $F_3(C_1) = 75409$ | $P_4(C_1) = 0.2262$ | $F_4(C_1) = 56607$ | $P_5(C_1) = 0.2264$ | $F_5(C_1) = 45425$ | $P_6(C_1) = 0.227$ |
| \(F_1(G_1)\)=219194 | \(F_2(G_1)\)=109711 | \(F_3(G_1)\)=72687 | \(F_4(G_1)\)=54859 | \(F_5(G_1)\)=43754 |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| \(P_1(G_1)\)=0.2194 | \(P_2(G_1)\)=0.2181 | \(P_3(G_1)\)=0.2184 | \(P_4(G_1)\)=0.2194 | \(P_5(G_1)\)=0.2188 |
| \(F_2(A_2)\)=141711 | \(F_3(A_2)\)=94612 | \(F_4(A_2)\)=71111 | \(F_5(A_2)\)=56643 |
| \(P_2(A_2)\)=0.2834 | \(P_3(A_2)\)=0.2838 | \(P_4(A_2)\)=0.2844 | \(P_5(A_2)\)=0.2832 |
| \(F_2(T_2)\)=135383 | \(F_3(T_2)\)=89797 | \(F_4(T_2)\)=67600 | \(F_5(T_2)\)=54278 |
| \(P_2(T_2)\)=0.2708 | \(P_3(T_2)\)=0.2694 | \(P_4(T_2)\)=0.2704 | \(P_5(T_2)\)=0.2714 |
| \(F_2(C_2)\)=113423 | \(F_3(C_2)\)=75549 | \(F_4(C_2)\)=56616 | \(F_5(C_2)\)=45153 |
| \(P_2(C_2)\)=0.2268 | \(P_3(C_2)\)=0.2266 | \(P_4(C_2)\)=0.2265 | \(P_5(C_2)\)=0.2258 |
| \(F_2(G_2)\)=109483 | \(F_3(G_2)\)=73375 | \(F_4(G_2)\)=54673 | \(F_5(G_2)\)=43926 |
| \(P_2(G_2)\)=0.2190 | \(P_3(G_2)\)=0.2201 | \(P_4(G_2)\)=0.2187 | \(P_5(G_2)\)=0.2196 |
| \(F_3(A_3)\)=94347 | \(F_4(A_3)\)=71110 | \(F_5(A_3)\)=56493 |
| \(P_3(A_3)\)=0.2830 | \(P_4(A_3)\)=0.2844 | \(P_5(A_3)\)=0.2825 |
| \(F_3(T_3)\)=90236 | \(F_4(T_3)\)=67492 | \(F_5(T_3)\)=54239 |
| \(P_3(T_3)\)=0.2707 | \(P_4(T_3)\)=0.2700 | \(P_5(T_3)\)=0.2712 |
| \(F_3(C_3)\)=75618 | \(F_4(C_3)\)=56546 | \(F_5(C_3)\)=45223 |
| \(P_3(C_3)\)=0.2268 | \(P_4(C_3)\)=0.2262 | \(P_5(C_3)\)=0.2261 |
| \(F_3(G_3)\)=73132 | \(F_4(G_3)\)=54852 | \(F_5(G_3)\)=44045 |
| \(P_3(G_3)\)=0.2194 | \(P_4(G_3)\)=0.2194 | \(P_5(G_3)\)=0.2202 |
| \(F_4(A_4)\)=70600 | \(F_5(A_4)\)=56932 |
| \(P_4(A_4)\)=0.2824 | \(P_5(A_4)\)=0.28467 |
| \(F_4(T_4)\)=67783 | \(F_5(T_4)\)=54087 |
| \(P_4(T_4)\)=0.2711 | \(P_5(T_4)\)=0.2704 |
| \(F_4(C_4)\)=56807 | \(F_5(C_4)\)=45309 |
| \(P_4(C_4)\)=0.2272 | \(P_5(C_4)\)=0.2265 |
| \(F_4(G_4)\)=54810 | \(F_5(G_4)\)=43672 |
| \(P_4(G_4)\)=0.2192 | \(P_5(G_4)\)=0.2184 |
| \(F_5(A_5)\)=56832 | \(F_5(T_5)\)=53905 |
| \(P_5(A_5)\)=0.2842 | \(P_5(T_5)\)=0.2695 |
| \(F_5(C_5)\)=45466 | \(F_5(G_5)\)=43797 |
| \(P_5(C_5)\)=0.2273 | \(P_5(G_5)\)=0.2190 |

Fig. 10. Collective frequencies \(F_n(A_k)\), \(F_n(T_k)\), \(F_n(C_k)\) and \(F_n(G_k)\) and also collective probabilities \(P_n(A_k)\), \(P_n(T_k)\), \(P_n(C_k)\) and \(P_n(G_k)\) \((n = 1, 2, 3, 4, 5\) and \(k \leq n\)) of subgroups of tetra-groups for sequences of \(n\)-plets, which have the same letter in their position \(k\), in the case of the following sequence: Homo sapiens chromosome 5 sequence, 1000000 bp, encode region ENm002; accession NT_086358, version NT_086358.1, [https://www.ncbi.nlm.nih.gov/nuccore/NT_086358.1](https://www.ncbi.nlm.nih.gov/nuccore/NT_086358.1). Collective probabilities \(P_n(A_k)\), \(P_n(T_k)\), \(P_n(C_k)\) and \(P_n(G_k)\) are marked by red for a visual convenient of their comparison each with other.

One can see from data in Fig. 10 that they satisfy the three tetra-group rules by analogy with data in Fig. 4. Fig. 11 shows levels of accuracy between values of collective propabilities (shown in Fig. 10) of complementary subgroups of tetra-groups of \(n\)-plets for this new long DNA-sequence. One can see that these levels of accuracy are approximately equal to the levels of accuracy in Fig. 8 for the previously considered sequence.
Fig. 11. Levels of accuracy between values of collective probabilities (shown in Fig. 6) of complementry subgroups of tetra-groups of n-plets (n = 1, 2, 3, 4, 5) for the sequence Homo sapiens chromosome 5 sequence, 1000000 bp, encode region ENm002; accession NT_086358, version NT_086358.1, https://www.ncbi.nlm.nih.gov/nuccore/NT_086358.1.

And what about the implementation of tetra-group rules for the whole human genome, which contains about 3 billion nucleotides? Fig. 12 shows data, which are taken from [Perez, 2013, Table 3] about individual frequencies of all 64 triplets in the whole human genome.

| triplet frequency | triplet frequency | triplet frequency | triplet frequency | triplet frequency | triplet frequency |
|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| AAA               | 36381293          | CAA               | 17927956          | GAA               | 18678084          | TAA               | 19721149          |
| AAC               | 13794251          | CAC               | 14214421          | GAC               | 8938833           | TAC               | 10755607          |
| AAG               | 18894716          | CAG               | 19176935          | GAG               | 15939419          | TAG               | 12240281          |
| AAT               | 23634011          | CAT               | 17423117          | GAT               | 12658530          | TAT               | 19568343          |
| ACA               | 19073189          | CCA               | 17444649          | GCA               | 13635427          | TCA               | 18565027          |
| ACC               | 11007307          | CCC               | 12428986          | GCC               | 11268094          | TCC               | 14614789          |
| AGG               | 2372235           | CCG               | 2606672           | GCG               | 2247440           | TGG               | 2087422           |
| ACT               | 15251455          | CTT               | 16835177          | GCT               | 13252828          | TCT               | 20990387          |
| AGA               | 20948987          | CGA               | 2085226           | GGA               | 14619310          | TGA               | 18562015          |
| AGC               | 13242724          | CGC               | 2244432           | GGC               | 11258126          | TGC               | 13649076          |
| AGG               | 16810797          | CGG               | 2604253           | GGG               | 12446600          | TGG               | 17480496          |
| ATG               | 15266057          | CGT               | 2379612           | GGT               | 11026602          | TGT               | 19152113          |
| ATA               | 19548709          | CTA               | 12217331          | GTA               | 10766854          | TTA               | 19750578          |
| ATC               | 12650299          | CTC               | 15942742          | GTC               | 8955434           | TTC               | 18708048          |
accuracy in the human genome?

biological evolution, does the accuracy of the realization of tetra symmetries (and of compliance with tetra organims) increase up to the highest level of accuracy in the human genome?

| NUCLEOTIDES | TRIPLETS |
|-------------|----------|
| Σ₁ = 2843411601 | Σ₃ = 947803867 |
| P₁(A) = 0.295359 | P₃(A₁) = 0.295372 |
| P₁(T) = 0.295847 | P₃(T₁) = 0.295821 |
| P₁(C) = 0.204341 | P₃(C₁) = 0.204338 |
| P₁(G) = 0.204453 | P₃(G₁) = 0.204469 |
| P₃(A₂) = 0.295364 | |
| P₃(T₂) = 0.295841 | |
| P₃(C₂) = 0.204347 | |
| P₃(G₂) = 0.204448 | |
| P₃(A₃) = 0.295341 | |
| P₃(T₃) = 0.295879 | |
| P₃(C₃) = 0.204339 | |
| P₃(G₃) = 0.204441 | |

Fig. 12. Frequencies of 64 triplets in the whole human genome in accordance with data from [Perez, 2013, Table 3].

In accordance with these data, the human genome contains 947.803.867 triplets and 2.843.411.601 (=947803867*3) nucleotides. From the data in Fig. 12, one can calculate probabilities $P_1(A)$, $P_1(T)$, $P_1(C)$ and $P_1(G)$ of members of the tetra-group of nucleotides A, T, C, G and also collective probabilities of subgroups of three tetra-groups of triplets: 1) $P_3(A_1)$, $P_3(T_1)$, $P_3(C_1)$, $P_3(G_1)$; 2) $P_3(A_2)$, $P_3(T_2)$, $P_3(C_2)$, $P_3(G_2)$; 3) $P_3(A_3)$, $P_3(T_3)$, $P_3(C_3)$ and $P_3(G_3)$. Results of such calculations are represented in Fig. 13.

![Fig. 12](image-url)

Fig. 13. Probabilities of subgroups of tetra-groups of nucleotides and triplets in the case of the whole human genome. Initial data about frequencies of triplets are shown in Fig. 12 from [Perez, 2013, Table 3].

One can see in Fig. 13 the very high level of accuracy of compliance of the whole human genome with the tetra-group rules of these probabilities (up to the fourth decimal place in each of equalities $P_1(X) = P_3(X_1) = P_3(X_2) = P_3(X_3)$, where X = A, T, C or G):

$P_1(A) = P_3(A_1) = P_3(A_2) = P_3(A_3) \approx 0.2953$

$P_1(T) = P_3(T_1) = P_3(T_2) = P_3(T_3) \approx 0.2958$

$P_1(C) = P_3(C_1) = P_3(C_2) = P_3(C_3) \approx 0.2043$

$P_1(G) = P_3(G_1) = P_3(G_2) = P_3(G_3) \approx 0.2044$

Is it accidental that the highest species of living organisms - the human organism - has a genome with such high accuracy of realization of tetra-group symmetries (and of compliance with tetra-group rules)? In the course of biological evolution, does the accuracy of the realization of tetra-group symmetries in the genomes of biological species increase up to the highest accuracy in the human genome?
Values of tetra-group probabilities and levels of accuracy of the realization of tetra-group symmetries in the genomes of different organisms can be used as a new criteria for formal rankings different species of living organisms. Will these rankings of biological species correspond to the stages of evolutionary development of biological species or to some other biological characteristics of species? These and other questions - arising in connection with the described tetra-group symmetries in long DNA-texts - are open at this stage of our knowledge and subject to study.

The author also analyzed the implementation of tetra-group rules in a representative set of other long DNA-sequences including the complete set of 24 human chromosomes (see below). Many of these sequences were taken from their lists in articles of other authors in order to avoid suspicion of a special choice of DNA-texts. The results of this analysis are presented in Appendixes 1-7. They show that all considered sequences satisfy the three tetra-group rules and demonstrate the existence of symmetries of tetra-group probabilities in these long DNA-texts.

4. DNA-alphabetts, genetic binary oppositions and the tensor product of matrices

It is desirable to create a mathematical model of origin of the described tetra-group symmetries of probabilities of oligonucleotides in long DNA-sequences. A great discovery of twentieth century physics was the probabilistic nature of physical phenomena at atomic scales, described in quantum mechanics. Molecules of heredity DNA and RNA belong to the microworld of quantum mechanics and they should obey to principles of quantum mechanics. In this Section the author proposes one of possible models, which is connected with principles and formalisms of quantum mechanics and quantum informatics.

| №  | Binary Symbols            | C | A | G | T/U |
|-----|---------------------------|---|---|---|-----|
| 1   | 0₁ — pyrimidines 1₁ — purines | 0₁ | 1₁ | 1₁ | 0₁ |
| 2   | 0₂ — amino 1₂ — keto      | 0₂ | 0₂ | 1₂ | 1₂ |
| 3   | 0₃ — three hydrogen bonds; 1₃ — two hydrogen bonds | 0₃ | 1₃ | 0₁ | 1₁ |

Fig. 14. Left: the four nitrogenous bases of DNA: adenine A, guanine G, cytosine C, and thymine T. Right: three binary sub-alphabets of the genetic alphabet on the basis of three pairs of binary-oppositional traits or indicators.

Science does not know why the nucleotide alphabet of DNA has been created by nature from just four letters (A, T, C, G), and why just these very simple molecules were chosen for the DNA-alphabet (out of millions of possible molecules). But science knows [Fimmel, Danielli, Strüngmann, 2013; Petoukhov, 2008; Petoukhov, He, 2009; Stambuk, 1999] that these four molecules are interrelated by means of their symmetrical peculiarities into the united molecular ensemble with its three pairs of binary-oppositional traits or indicators (Fig. 14):
(1) Two letters are purines (A and G), and the other two are pyrimidines (C and T). From the standpoint of these binary-oppositional traits one can denote \( C = T = 0, \ A = G = 1 \). From the standpoint of these traits, any of the DNA-sequences are represented by a corresponding binary sequence. For example, GCATGAAGT is represented by 101011110;
(2) Two letters are amino-molecules (A and C) and the other two are keto-molecules (G and T). From the standpoint of these traits one can designate \( A = C = 0, \ G = T = 1 \). Correspondingly, the same sequence, GCATGAAGT, is represented by another binary sequence, 100110011;
(3) The pairs of complementary letters, A-T and C-G, are linked by 2 and 3 hydrogen bonds, respectively. From the standpoint of these binary traits, one can designate \( C = G = 0, \ A = T = 1 \). Correspondingly, the same sequence, GCATGAAGT, is read as 001101101.

Accordingly, each of the DNA-sequences of nucleotides is the carrier of three parallel messages on three different binary languages. At the same time, these three types of binary representations form a common logic set on the basis of logic operation of modulo-2 addition denoted by the symbol \( \oplus \): modulo-2 addition of any two such binary representations of the DNA-sequence coincides with the third binary representation of the same DNA-sequence: for example, \( 101011110 \oplus 100110011 = 001101101 \). One can be reminded here of the rules of the bitwise modulo-2 addition: \( 0 \oplus 0 = 0; \ 0 \oplus 1 = 1; \ 1 \oplus 0 = 1; \ 1 \oplus 1 = 0 \). (The logic operation of modulo-2 addition is actively used in quantum informatics, which will be considered below in a connection with DNA-texts).

Taking into account the phenomenon of each of DNA-letters C, A, T and G is uniquely defined by any two kinds of mentioned binary-oppositional indicators (Fig. 14), these genetic letters can be represented as corresponding pairs of binary symbols, for example, from the standpoint of two first binary-oppositional indicators. It is convenient for us for further description to show at the first position of each of letters its binary symbol from the second pair of binary-oppositional indicators (the indicator "amino or keto": \( C=A=0, \ T=G=1 \)) and at the second positions of each of letters its binary symbol from the first pair of binary-oppositional indicators (the indicator "pyrimidine or purine": \( C=T=0, \ A=G=1 \)). In this case the letter C is represented by the binary symbol \( 0_20_1 \) (that is as 2-bit binary number), A – by the symbol \( 0_21_1 \), T – by the symbol \( 1_20_1 \), G – by the symbol \( 1_21_1 \). Using these representations of separate letters, each of 16 doublets is represented as concatenation of the binary symbols of its letters (that is as 4-bit binary number): for example, the doublet CC is represented as 4-bit binary number \( 0_20_10_20_1 \), the doublet CA – as 4-bit binary number \( 0_20_10_21_1 \), etc. By analogy, each of 64 triplets is represented as concatenation of the binary symbols of its letters (that is as 6-bit binary number): for example, the triplet CCC is represented as 6-bit binary number \( 0_20_10_20_10_20_1 \), the triplet CCA – as 6-bit binary number \( 0_20_10_20_10_21_1 \), etc. In general, each of n-plets is represented as concatenation of the binary symbols of its letters (below we will not show these indexes 2 and 1 of separate letters in binary representations of n-plets but will remember that each of positions corresponds to its own kind of indicators from the first or from the second set of indicators in Fig. 14).

It is convenient to represent DNA-alphabets of 4 nucleotides, 16 doublets, 64 triplets, ... \( 4^n \) n-plets in a form of appropriate square tables (Fig. 15), which rows
and columns are numerated by binary symbols in line with the following principle. Entries of each column are numerated by binary symbols in line with the first set of binary-oppositional indicators in Fig. 14 (for example, the triplet CAG and all other triplets in the same column are the combination “pyrimidine-purin-purin” and so this column is correspondingly numerated 011). By contrast, entries of each row are numerated by binary numbers in line with the second set of indicators (for example, the same triplet CAG and all other triplets in the same row are the combination “amino-amino-keto” and so this row is correspondingly numerated 001). In such tables (Fig. 15), each of 4 letters, 16 doublets, 64 triplets, ... takes automatically its own individual place and all components of the alphabets are arranged in a strict order.

Fig. 15. The square tables of DNA-alphabets of 4 nucleotides, 16 doublets and 64 triplets with a strict arrangement of all components. Each of tables is constructed in line with the principle of binary numeration of its column and rows (see explanations in text).

Here one can remind a historical fact that the same principle of constructing square tables with quite similar binary numerations of their columns and rows was used else in the Ancient Chinese book «I-Ching», which was written a few thousand years ago and which is the most ancient historical example of systematic usage of binary numbers. The famous table of 64 hexagrams in Fu-Xi’s order exists there, which has many deep analogies with the
genetic matrix of 64 triplets [Hu, Petoukhov, Petukhova, 2017; Petoukhov, 1999, 2008, 2017; Petoukhov, He, 2009]. The ancient Chinese claimed that this table is the universal archetype of nature. They knew nothing about the genetic code, but the genetic code is constructed in line with the “I-Ching”.

For our article about tetra-group symmetries in long DNA-texts it is more important the following: these 3 separate genetic tables form the joint tensor (!) family of matrices since they are interrelated by the known operation of the tensor (or Kronecker) product of matrices. By definition, under tensor multiplication of two matrices, each of entries of the first matrix is multiplied with the whole second matrix [Bellman, 1960]. The second tensor power of the (2*2)-matrix [C, A; T, G] of 4 DNA-letters gives automatically the matrix of 16 doublets; the third tensor power of the matrix of the same matrix of 4 DNA-letters gives the matrix of 64 triplets with the same strict arrangement of entries (Fig. 16).

Fig. 16. The tensor family of genetic matrices [C, A; T, G]^(n) (here tensor power n = 1, 2, 3) of DNA-alphabets of 4 nucleotides, 16 doublets and 64 triplets. The symbol ⊗ means the tensor product.

One can see from a comparison of Fig. 16 with Fig. 15 that the tensor powers of the genetic matrix of 4 letters [C, A; T, G] produce - in a new way - the same square tables of the DNA-alphabets with the same strict arrangements of alphabetic components without any mention about binary-oppositional traits of genetic letters C, A, T and G, which were used to construct the square tables in Fig. 15.

So, the structural organization of the system of DNA-alphabets is connected with the algebraic operation of the tensor product. It is important since the operation of the tensor product is well known in mathematics, physics and informatics, where it gives a way of putting vector spaces together to form larger vector spaces. The following quotation speaks about the meaning of the tensor product: «This construction is crucial to understanding the quantum mechanics of multiparticle systems» [Nielsen, Chuang, 2010, p. 71]. For us the most interesting is that the tensor product is one of basis instruments in quantum informatics. Notions and formalisms of quantum informatics are used
in the next Section to simulate phenomena of the tetra-group symmetries in long DNA-texts and for creation of new quantum-informational approaches in mathematical genetics.

5. The quantum-information model to explain and predict symmetries of collective probabilities in tetra-groups of long DNA-texts

Are there systems in nature or in mathematical natural science, for which the rules of tetra-group symmetry of probabilities are fulfilled, similar to those observed in long DNA texts? Yes. Below we show that, in the field of quantum informatics, 2n-qubit systems in so called separable pure states exist, for which similar rules are fulfilled precisely (in contrast to their approximate implementation in long DNA-texts). Correspondingly the author proposes using these quantum states to model phenomena of tetra-group symmetries of probabilities in long DNA-texts. It should be mentioned that general thoughts and hypotheses about using principles of quantum informatics in living organisms are discussed in interesting publications of many authors [Igamberdiev, 1993, 2012; Matsuno, 1999, 2003, 2015; Matsuno, Paton, 2000; Simeonov, 2013; and others]. For example, the book [Penrose, 1996], concerning the topic of biological quantum computers, appeals to the fact that tubulin proteins exist in two different configurations, and they can switch between these configurations by analogy with triggers to provide bio-computer functions. In contrast to this “protein standpoint”, our model approach testifies in favor that already the genetic level, which is the deepest level of living organisms, is connected with the principles of quantum computers.

DNA- and RNA-molecules belong to the microworld of quantum mechanics and perform the information function. By this reason their informational features should be studied on the basis of quantum informatics. Below the author proposes a possible model approach to DNA-texts from the standpoint of quantum informatics. This Section shows how notions and formalisms of quantum informatics can be introduced into the field of genetic informatics in connection with described tetra-group symmetries and rules of long DNA-texts. Here we use typical notions, denotations and formalisms of quantum informatics from the fundamental book [Nielsen, Chuang, 2010]. In particular we use the notion of quantum bits (or qubits) for model representations of elements of single stranded DNA. We also use ordinary Dirac notations for vectors and operations with them:

- $|\psi\rangle$ means a vector, which is known also as a ket-vector;
- $\langle\psi|$ means a vector, which is dual to $|\psi\rangle$ and known as a bra-vector;
- $\langle\varphi|\psi\rangle$ means scalar (or inner) product between the vectors $\langle\varphi|$ and $|\psi\rangle$;
- $|\varphi\rangle \otimes |\psi\rangle$ means the tensor product of $|\varphi\rangle$ and $|\psi\rangle$;
- $|\varphi\rangle|\psi\rangle$ is abbreviated notation for tensor product of $|\varphi\rangle$ and $|\psi\rangle$;
- $M^\dagger$ – transpose of the M matrix;
- $\langle\varphi|\text{M}|\psi\rangle$ scalar product between $|\varphi\rangle$ and M$|\psi\rangle$.

In quantum informatics, such vector spaces $H$ are considered, which are equipped with the scalar (or inner) product (so called Hilbert spaces). Let $H_1$ and $H_2$ be quantum mechanical state spaces, that is, finite dimensional Hilbert spaces with orthonormal basis states $|\alpha\rangle$ and $|\beta\rangle$, where $i = 1, ..., n$ and $j = 1, ..., m$. By a postulate of quantum mechanics, the state space of the composite system is
given by the tensor product $H_1 \otimes H_2$ with base states $\{ |\alpha_i\rangle \otimes |\beta_j\rangle \}$, or in more compact notation $\{ |\alpha\beta\rangle \}$. “The state space of a composite physical systems is the tensor product of the state spaces of the component physical systems. Moreover, if we have systems numbered 1 through n and system number i is prepared in the state $\rho_i$, then the joint state of the total system is $\rho_1 \otimes \rho_2 \otimes ... \rho_n$” [Nielsen, Chuang, 2010, p. 102]. If a quantum state can be represented as a vector of a Hilbert space, such state is called a pure quantum state. If a pure state $|\psi\rangle \in H_1 \otimes H_2$ can be written in the form $|\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle$, where $|\psi\rangle$ is a pure state of the i-th subsystem, it is said to be separable. Otherwise it is called entangled.

As known, a quantum bit (or qubit) is a unit of quantum information. For two-level quantum systems used as qubits, the state $|0\rangle$ is identified with the vector $(1, 0)$, and similarly the state $|1\rangle$ with the vector $(0, 1)$. Two possible states for a qubit are the states $|0\rangle$ and $|1\rangle$, which correspond to the states of 0 and 1 for a classical bit. The difference between bits and qubits is that a qubit can be in a state other than $|0\rangle$ or $|1\rangle$. It is possible to form linear combinations of states, often called superpositions (6):

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle,$$

The symbol $|\psi\rangle$ means a state of a qubit. The numbers $\alpha$ and $\beta$ can be complex numbers but in our case it is enough to think of them as real numbers. Put another way, the state of a qubit is a vector in a two-dimensional vector space. The standard notation for states in quantum mechanics is the Dirac notation “$| >$". The special states $|0\rangle$ and $|1\rangle$ are known as computational basis states, and form an orthonormal basis for this vector space [Nielsen, Chuang, 2010, p. 13]. As known, we cannot examine a qubit to determine its quantum state, that is, the values of $\alpha$ and $\beta$. Instead, quantum mechanics tells us that we can only acquire much more restricted information about the quantum state. When we measure a qubit we get either the result 0, with probability $|\alpha|^2$, or the result 1, with probability $|\beta|^2$. Naturally, $|\alpha|^2 + |\beta|^2 = 1$, since the probabilities must sum to one. Geometrically, we can interpret this as the condition that the qubit’s state be normalized to length 1. Values $\alpha$ and $\beta$ are called amplitudes of probabilities. Thus, in general a qubit’s state is a unit vector in a two-dimensional complex vector space. Let us emphasize again that when a qubit is measured, it only ever gives “0” or “1” as the measurement result – probabilistically.

In more general case a system of “n” qubits is considered in quantum informatics. The computational basis states of this system are written in the form $|x_1x_2...x_n\rangle$; a quantum state of such a system is specified by $2^n$ amplitudes [Nielsen, Chuang, 2010, p. 17]. In our model approach we interpret DNA-texts as quantum systems of many qubits.

In technical devices of quantum informatics, a qubit can be represented by many ways on the basis of different pairs of binary-oppositional indicators: for example, by two electronic levels in an atom; by two kinds of polarization of a single photon (vertical polarization and horizontal polarization), etc.

In our model approach for genetic informatics, we represent qubits on the basis of different pairs of binary-oppositional indicators of adenine A, guanine G, cytosine C, and thymine T, which were shown above in Fig. 14. As we noted, each of these DNA bases can be uniquely defined by any two kinds of mentioned
By analogy with the previous Section, to characterized each of the DNA-letters C, A, T, G we will use the first kind of indicators («pyrimidine or purine») and the second kind of indicators («three hydrogen bonds or two hydrogen bonds»). On the basis of each of these pairs of binary-oppositional indicators, a corresponding two-level quantum system can be formally introduced with a definition of its appropriate qubit (those qubits, which are introduced in genetic informatics on the basis of binary-oppositional indicators of genetic molecules, can be conditionally called «genetic qubits» or briefly «g-qubits»).

Let us introduce, firstly, the notion of a genetic qubit as a two-level quantum system on the basis of the indicators «pyrimidine or purine»: in this quantum system one level corresponds to the indicator «pyrimidine» and the second level – to the oppositional indicator «purine». In other words, such genetic qubit is represented by these oppositional indicators and the state of such qubit is a vector in its appropriate two-dimensional Hilbert space H₁. One can assume that the state |0> corresponds to the state «pyrimidine», and the state |1> - to the state «purine». By analogy with the expression (6), a state of such genetic qubit can be expressed by the expression (7), where α₀ and α₁ are amplitudes of probabilities of these computational basis states “pyrimidine” and “purine”:

\[ |\psi_1> = \alpha_0 |0> + \alpha_1 |1>, \quad \alpha_0^2 + \alpha_1^2 = 1 \]  (7)

In a particular case, a qubit (7) can represent a pure state of a quantum system in a form of a sequence, which consists of pyrimidines and purines.

Secondly, let us introduce the notion of another genetic qubit as a two-level quantum system on the basis of the indicators «three hydrogen bonds or two hydrogen bonds»: in this quantum system one level corresponds to the indicator «three hydrogen bonds» and the second level – to the indicator «two hydrogen bonds». In other words, such genetic qubit is represented by these two indicators and the state of such qubit is a vector in its appropriate 2-dimensional Hilbert space H₂. One can assume that the state |0> corresponds to the state «three hydrogen bonds», and the state |1> - to the state «two hydrogen bonds». By analogy with the expression (6), a state of such genetic qubit can be expressed by the expression (8), where β₀ and β₁ are amplitudes of probabilities of these computational basis states:

\[ |\psi_2> = \beta_0 |0> + \beta_1 |1>, \quad \beta_0^2 + \beta_1^2 = 1 \]  (8)

In a particular case, a qubit (8) can represent a pure state of a quantum system in a form of a sequence, which consists of elements with three and two hydrogen bonds.

So we have two different 2-dimensional Hilbert spaces H₁ and H₂, to which pure states of genetic qubits (7) and (8) belong correspondingly.

In our genetic case, the tensor product of the two-dimensional Hilbert space H₁ ⊗ H₂ gives one four-dimensional Hilbert space with the following separable pure state of a quantum 2-qubit system:

\[ |\psi_{12}>= |\psi_1> \otimes |\psi_2> = (\alpha_0 |0> + \alpha_1 |1>) \otimes (\beta_0 |0> + \beta_1 |1>) = \]

binary-oppositional indicators (Fig. 14).
\[ \alpha_0 \beta_0 \begin{array}{c} 00 \end{array} + \alpha_0 \beta_1 \begin{array}{c} 01 \end{array} + \alpha_1 \beta_0 \begin{array}{c} 10 \end{array} + \alpha_1 \beta_1 \begin{array}{c} 11 \end{array} \]  

Such 2-qubit system has four computational basis states denoted \( |00\rangle, |01\rangle, |10\rangle, |11\rangle \). Amplitudes \( \alpha_0 \beta_0, \alpha_0 \beta_1, \alpha_1 \beta_0 \) and \( \alpha_1 \beta_1 \) of probabilities satisfy the normalization condition: \( (\alpha_0 \beta_0)^2 + (\alpha_0 \beta_1)^2 + (\alpha_1 \beta_0)^2 + (\alpha_1 \beta_1)^2 = 1 \). One can note that two kinds of indicators “pyrimidine-purine” and “three hydrogen bonds or two hydrogen bonds” for separate nitrogenous bases C, T, G, A define the following correspondence:

- Cytosine C corresponds to the computational basis state \( |00\rangle \) of the 2-qubit system (9) since cytosine C is characterized by the indicator “pyrimidine”, which is the computational basis state \( |0\rangle \) in the first qubit (7), and also by the indicator “three hydrogen bonds”, which is the computational basis state \( |0\rangle \) in the second qubit (8). In the four-dimensional Hilbert space \( \mathbb{H}_1 \otimes \mathbb{H}_2 \), these computational basis states \( |0\rangle \) and \( |0\rangle \) of two genetic qubits (7) and (8) define the computational basis state \( |00\rangle \) of the 2-qubit system (9);
- Thymine T corresponds to the computational basis state \( |01\rangle \) of the 2-qubit system (9) since thymine T is characterized by the indicator “pyrimidine”, which is the computational basis state \( |0\rangle \) in the first qubit (7), and also by the indicator “two hydrogen bonds”, which is the computational basis state \( |1\rangle \) in the second qubit (8). In the four-dimensional Hilbert space \( \mathbb{H}_1 \otimes \mathbb{H}_2 \), these computational basis states \( |0\rangle \) and \( |1\rangle \) of two genetic qubits (7) and (8) define the computational basis state \( |01\rangle \) of the 2-qubit system (9);
- Guanine G corresponds to the computational basis state \( |10\rangle \) of the 2-qubit system (9) since guanine G is characterized by the indicator “purine”, which is the computational basis state \( |1\rangle \) in the first qubit (7), and also by the indicator “three hydrogen bonds”, which is the computational basis state \( |0\rangle \) in the second qubit (8). In the four-dimensional Hilbert space \( \mathbb{H}_1 \otimes \mathbb{H}_2 \), these computational basis states \( |1\rangle \) and \( |0\rangle \) of two genetic qubits (7) and (8) define the computational basis state \( |10\rangle \) of the 2-qubit system (9);
- Adenine A corresponds to the computational basis state \( |11\rangle \) of the 2-qubit system (9) since adenine A is characterized by the indicator “purine”, which is the computational basis state \( |1\rangle \) in the first qubit (7), and also by the indicator “two hydrogen bonds”, which is the computational basis state \( |1\rangle \) in the second qubit (8). In the four-dimensional Hilbert space \( \mathbb{H}_1 \otimes \mathbb{H}_2 \), these computational basis states \( |1\rangle \) and \( |1\rangle \) of two genetic qubits (7) and (8) define the computational basis state \( |11\rangle \) of the 2-qubit system (9).

By this way, members of the tetra-group of nucleotides C, T, G and A get their representations as computational basis states of a 2-qubit system (9) in a four-dimensional Hilbert space \( \mathbb{H}_1 \otimes \mathbb{H}_2 \) with conditional denotations \( |C\rangle = |00\rangle, |T\rangle = |01\rangle, |G\rangle = |10\rangle \) and \( |A\rangle = |11\rangle \). (Below we will show that these computational basis states can be connected with pairs of photons of different frequencies, radiated by molecular elements of these dual indicators). Correspondingly the expression (9) can be rewritten in the following conditional form (10):
$$|\psi_{12}\rangle = \alpha_0 \beta_0 |00\rangle + \alpha_0 \beta_1 |01\rangle + \alpha_1 \beta_0 |10\rangle + \alpha_1 \beta_1 |11\rangle = \alpha_0 \beta_0 |C\rangle + \alpha_0 \beta_1 |T\rangle + \alpha_1 \beta_0 |G\rangle + \alpha_1 \beta_1 |A\rangle$$

(10)

We call the 2-qubit quantum system with its separable pure state (10) as a «monoplet CTGA-system». In a particular case, the state (10) of a 2-qubits monoplet CTGA-system can represent a separable pure state of a quantum system, which is a sequence of elements, where each of elements is pyrimidine or purin and simultaneously has three or two hydrogen bonds by analogy with nucleotides. From (10), the probabilities of computational basis states of separate nucleotides C, T, G and A are equal to the following:

$$P_1(C_1) = (\alpha_0 \beta_0)^2, P_1(T_1) = (\alpha_0 \beta_1)^2, P_1(G_1) = (\alpha_1 \beta_0)^2, P_1(A_1) = (\alpha_1 \beta_1)^2$$

(11)

They should satisfy the normalization condition (12):

$$(\alpha_0 \beta_0)^2 + (\alpha_0 \beta_1)^2 + (\alpha_1 \beta_0)^2 + (\alpha_1 \beta_1)^2 = 1$$

(12)

To consider the case not single monoplets C, T, G, A but doublets CC, CT, CG, CA, TC, TT, TG, TA, GC, GT, GG, GA, AC, AT, AG, AA one should accordingly expand the 4-dimensional Hilbert space \(H_1 \otimes H_2\) to the 16-dimensional Hilbert space \(H_1 \otimes H_2 \otimes H_3 \otimes H_4\). Here the spaces \(H_1\) and \(H_2\) are related with the first letters of doublets and the spaces \(H_3\) and \(H_4\) are related with the second letters of doublets. These spaces \(H_3\) and \(H_4\) are defined in a close analogy with the spaces \(H_1\) and \(H_2\). Let us explain this.

By analogy with the expressions (6, 7), each of four letters C, T, G, A at the second position of doublets is interpreted firstly as a two-level quantum system on the basis of the oppositional indicators «pyrimidine or purine»: in this quantum system one level corresponds to the indicator «pyrimidine» and the second level – to the oppositional indicator «purine». In other words, a new genetic qubit arises on the basis of these oppositional indicators for the second letters inside 16 doublets; the state of such qubit is a vector in its appropriate two-dimensional Hilbert space \(H_3\). One can assume that the state \(|0\rangle\) corresponds to the state «pyrimidine», and the state \(|1\rangle\) - to the state «purine». By analogy with the expression (7), a state of such genetic qubit can be denoted by the expression (13), where \(\alpha_2\) and \(\alpha_3\) are amplitudes of probabilities of these computational basis states «pyrimidine» and «purine» for the second letters in 16 doublets:

$$|\psi_3\rangle = \alpha_2 |0\rangle + \alpha_3 |1\rangle, \quad \alpha_2^2 + \alpha_3^2 = 1$$

(13)

Secondly, for the second letters of 16 doublets, the notion of another genetic qubit as a two-level quantum system on the basis of the oppositional indicators «three hydrogen bonds or two hydrogen bonds» is defined by analogy with the expression (8). In this quantum system, one level corresponds to the indicator «three hydrogen bonds» and the second level – to the indicator «two hydrogen bonds». In other words, such genetic qubit is represented by these two indicators and the state of such qubit is a vector in its appropriate 2-dimensional Hilbert space \(H_4\). One can assume that the state \(|0\rangle\) corresponds to the state
«three hydrogen bonds», and the state |1> - to the state «two hydrogen bonds».
By analogy with the expression (8), a state of such genetic qubit can be expressed by the expression (14), where $\beta_2$ and $\beta_3$ are amplitudes of probabilities of these computational basis states:

$$ |\psi_4\rangle = \beta_2 |0\rangle + \beta_3 |1\rangle, \quad \beta_2^2 + \beta_3^2 = 1 \quad (14) $$

In such way, in the case of long DNA-texts of doublets, we have two additional 2-dimensional Hilbert spaces $H_3$ and $H_4$, to which pure states of genetic qubits (13) and (14) belong correspondingly. The tensor product of the two-dimensional Hilbert space $H_3 \otimes H_4$ gives one four-dimensional Hilbert space with the following separable pure state of a quantum 2-qubit system for the second letters inside 16 doublets:

$$ |\psi_{34}\rangle = |\psi_3\rangle \otimes |\psi_4\rangle = (\alpha_2 |0\rangle + \alpha_3 |1\rangle) \otimes (\beta_2 |0\rangle + \beta_3 |1\rangle) = $$

$$ = \alpha\beta_2 |00\rangle + \alpha\beta_3 |01\rangle + \alpha_3\beta_2 |10\rangle + \alpha_3\beta_3 |11\rangle = $$

$$ = \alpha\beta_2 |C\rangle + \alpha\beta_3 |T\rangle + \alpha_3\beta_2 |G\rangle + \alpha_3\beta_3 |A\rangle \quad (15) $$

Such new 2-qubit system has four computational basis states denoted $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$. By analogy with the case of the Hilbert space space $H_1 \otimes H_2$ for the first letters of doublets with conditional denotations $|C\rangle=|00\rangle$, $|T\rangle=|01\rangle$, $|G\rangle=|10\rangle$ and $|A\rangle=|11\rangle$ (see the expression (10)), in the Hilbert space $H_3 \otimes H_4$ similar denotations are used for 2-qubit systems of the second letters of doublets: $|C\rangle=|00\rangle$, $|T\rangle=|01\rangle$, $|G\rangle=|10\rangle$ and $|A\rangle=|11\rangle$. Amplitudes $\alpha\beta_2$, $\alpha\beta_3$, $\alpha_3\beta_2$ and $\alpha_3\beta_3$ of probabilities in (15) satisfy the normalization condition (16):

$$(\alpha\beta_2)^2 + (\alpha\beta_3)^2 + (\alpha_3\beta_2)^2 + (\alpha_3\beta_3)^2 = 1. \quad (16)$$

In the 16-dimensional Hilbert space $H_1 \otimes H_2 \otimes H_3 \otimes H_4$ for the case of 16 doublets, we have the following 16 computational basis states with their appropriate amplitudes of probabilities:

$$ |\psi_{12}\rangle \otimes |\psi_{34}\rangle = $$

$$ = \alpha_0\beta_0\alpha_2\beta_2 |CC\rangle + \alpha_0\beta_0\alpha_2\beta_2 |CT\rangle + \alpha_0\beta_0\alpha_3\beta_2 |CG\rangle + \alpha_0\beta_0\alpha_3\beta_3 |CA\rangle + $$

$$ + \alpha_0\beta_1\alpha_2\beta_2 |TC\rangle + \alpha_0\beta_1\alpha_2\beta_2 |TT\rangle + \alpha_0\beta_1\alpha_3\beta_2 |TG\rangle + \alpha_0\beta_1\alpha_3\beta_3 |TA\rangle + $$

$$ + \alpha_1\beta_0\alpha_2\beta_2 |GC\rangle + \alpha_1\beta_0\alpha_3\beta_2 |GT\rangle + \alpha_1\beta_0\alpha_3\beta_3 |GG\rangle + \alpha_1\beta_0\beta_3\beta_3 |GA\rangle + $$

$$ + \alpha_1\beta_1\alpha_2\beta_2 |AC\rangle + \alpha_1\beta_1\alpha_3\beta_2 |AT\rangle + \alpha_1\beta_1\alpha_3\beta_3 |AG\rangle + \alpha_1\beta_1\beta_3\beta_3 |AA\rangle \quad (17) $$

In the state (17) of a 4-qubit "doublet CTGA-system", 16 computational basis states are represented by 16 doublets: $|0000\rangle=|CC\rangle$, $|0001\rangle=|CT\rangle$, $|0010\rangle=|CG\rangle$, $|0011\rangle=|CA\rangle$, $|0100\rangle=|TC\rangle$, $|0101\rangle=|TT\rangle$, $|0110\rangle=|TG\rangle$, $|0111\rangle=|TA\rangle$, $|1000\rangle=|GC\rangle$, $|1001\rangle=|GT\rangle$, $|1010\rangle=|GG\rangle$, $|1011\rangle=|GA\rangle$, $|1100\rangle=|AC\rangle$, $|1101\rangle=|AT\rangle$, $|1110\rangle=|AG\rangle$, $|1111\rangle=|AA\rangle$. In our model approach, these 16 computational basis states are interpreted as representations of appropriate 16 genetic doublets.

From the expression (17), we have the following collective probability $P_2(C_1)$ of all 4 doublets with the first letter C in them (they are collected in the first row of this expression):
\[ P_2(C_1) = (\alpha_0\beta_0\alpha_2\beta_2)^2 + (\alpha_0\beta_0\alpha_3\beta_3)^2 + (\alpha_0\beta_0\alpha_3\beta_3)^2 = (\alpha_0\beta_0)^2 \times ((\alpha_2\beta_2)^2 + (\alpha_3\beta_3)^2) = (\alpha_0\beta_0)^2 = P_1(C) \quad (18) \]

Here the sum in curly brackets is equal to 1 according to the normalization condition (16). The expression (18) means that the collective probability \( P_2(C_1) \) of 4 doublets CC, CT, CG and CA is equal to the individual probability \( P_1(C) \) of the letter C in this quantum-informational model of collective and individual probabilities inside long DNA-texts. From the expression (17), similar calculations of collective probabilities \( P_2(T_1), P_2(G_1) \) and \( P_2(A_1) \) of doublets with the first letters T, G, A in them give analogical results (19) of their equality to individual probabilities of letters T, G and A:

\[ P_2(T_1) = P_1(T), \quad P_2(G_1) = P_1(G), \quad P_2(A_1) = P_1(A) \quad (19) \]

These model results correspond to phenomenologic facts reflected in the the first rule of tetra-group symmetries in long DNA-texts (see Section 3 above).

The expression (17) allows a calculation of the collective probability \( P_2(C_2) \) of all 4 doublets with the second letter C in them:

\[ P_2(C_2) = (\alpha_0\beta_0\alpha_2\beta_2)^2 + (\alpha_0\beta_1\alpha_2\beta_2)^2 + (\alpha_1\beta_0\alpha_2\beta_2)^2 + (\alpha_1\beta_1\alpha_2\beta_2)^2 = (\alpha_2\beta_2)^2 \times ((\alpha_0\beta_0)^2 + (\alpha_0\beta_1)^2 + (\alpha_1\beta_0)^2 + (\alpha_1\beta_1)^2) = (\alpha_2\beta_2)^2 = P_1(C) \quad (20) \]

Here the sum in curly brackets is equal to 1 according to the normalization condition (12). The expression (20) means that the collective probability \( P_2(C_2) \) of all doublets with the second letter C (that is, CC, TC, GC, AC) is also equal to the individual probability \( P_1(C) \) of the letter C. From the expression (17), similar calculations of collective probabilities \( P_2(T_2), P_2(G_2) \) and \( P_2(A_2) \) of doublets with the second letters T, G, A in them give analogical results (21) of their equality to individual probabilities of letters T, G and A:

\[ P_2(T_2) = P_1(T), \quad P_2(G_2) = P_1(G), \quad P_2(A_2) = P_1(A) \quad (21) \]

These model results correspond to phenomenologic facts reflected in the the second rule of tetra-group symmetries in long DNA-texts (see Section 3 above).

To consider the case of 64 triplets one should accordingly expand the 16-dimensional Hilbert space \( H_1 \otimes H_2 \otimes H_3 \otimes H_4 \) to the 64-dimensional Hilbert space \( H_1 \otimes H_2 \otimes H_3 \otimes H_4 \otimes H_5 \otimes H_6 \). Here the spaces \( H_1 \) and \( H_2 \) are related with the first letters of doublets, the spaces \( H_3 \) and \( H_4 \) are related with the second letters of doublets and the additional spaces \( H_5 \) and \( H_6 \) are related with the third letters of triplets. These additional spaces \( H_5 \) and \( H_6 \) are defined in a close analogy with the spaces \( H_1, H_2, H_3 \) and \( H_4 \). Repeating for the spaces \( H_5 \) and \( H_6 \) those steps, which were made above to define the Hilbert spaces \( H_3 \) and \( H_4 \) for the second letters of doublets, we get for the third letter of triplets the following expressions:

\[ |\psi_5\rangle = \alpha_4 |0\rangle + \alpha_5 |1\rangle, \quad \alpha_4^2 + \alpha_5^2 = 1 \quad (22) \]
\[ |\psi_6\rangle = \beta_4 |0\rangle + \beta_5 |1\rangle, \quad \beta_4^2 + \beta_5^2 = 1 \quad (23) \]
Here we again use the denotations: $|C| = |00>$, $|T| = |01>$, $|G| = |10>$ and $|A| = |11>$ for the third letters of triplets in the 4-dimensional Hilbert space $H_3 \otimes H_6$.

In the 64-dimensional Hilbert space $H_1 \otimes H_2 \otimes H_3 \otimes H_4 \otimes H_5 \otimes H_6$ for the case of 64 triplets, we have the following 64 computational basis states with their appropriate amplitudes of probabilities:

$$|\Psi_{56}\rangle = |\psi_5\rangle \otimes |\psi_6\rangle = (\alpha_4 |0\rangle + \alpha_5 |1\rangle) \otimes (\beta_4 |0\rangle + \beta_5 |1\rangle) =$$

$$\alpha_4 \beta_4 |00\rangle + \alpha_5 \beta_5 |01\rangle + \alpha_5 \beta_4 |10\rangle + \alpha_5 \beta_5 |11\rangle =$$

$$\alpha_4 \beta_4 |C\rangle + \alpha_5 \beta_5 |T\rangle + \alpha_5 \beta_4 |G\rangle + \alpha_5 \beta_5 |A\rangle$$

(24)

$$(\alpha_4 \beta_4)^2 + (\alpha_5 \beta_5)^2 + (\alpha_5 \beta_4)^2 + (\alpha_5 \beta_5)^2 = 1$$

(25)

In the state (26) of a 6-qubit “triplet CTGA-system”, 64 computational basis states represent 64 triplets in our model approach:

$$|000000\rangle = |CCC\rangle, |000001\rangle = |CTC\rangle, |000010\rangle = |CCG\rangle, |000011\rangle = |CCA\rangle,$$

$$|000100\rangle = |CTC\rangle, |000101\rangle = |CTT\rangle, |000110\rangle = |CTG\rangle, |000111\rangle = |CTA\rangle,$$

$$|001000\rangle = |CCG\rangle, |001001\rangle = |CGT\rangle, |001010\rangle = |CGG\rangle, |001011\rangle = |CGA\rangle,$$

$$|001100\rangle = |CAG\rangle, |001101\rangle = |CAT\rangle, |001110\rangle = |CA\rangle, |001111\rangle = |CAA\rangle,$$

$$|010000\rangle = |TCC\rangle, |010001\rangle = |TCT\rangle, |010010\rangle = |TCG\rangle, |010011\rangle = |TCA\rangle,$$

$$|010100\rangle = |TGT\rangle, |010101\rangle = |TTT\rangle, |010110\rangle = |TTG\rangle, |010111\rangle = |TTA\rangle,$$

$$|011000\rangle = |TGG\rangle, |011001\rangle = |TGT\rangle, |011010\rangle = |TG\rangle, |011011\rangle = |TAG\rangle,$$

$$|011100\rangle = |TAA\rangle, |011101\rangle = |TAT\rangle, |011110\rangle = |T\rangle, |011111\rangle = |A\rangle,$$

$$|100000\rangle = |GCC\rangle, |100001\rangle = |GCT\rangle, |100010\rangle = |GCC\rangle, |100011\rangle = |GCA\rangle,$$

$$|100100\rangle = |GTC\rangle, |100101\rangle = |GT\rangle, |100110\rangle = |GTT\rangle, |100111\rangle = |GTA\rangle,$$

$$|101000\rangle = |GG\rangle, |101001\rangle = |G\rangle, |101010\rangle = |GG\rangle, |101011\rangle = |GA\rangle,$$

$$|101100\rangle = |G\rangle, |101101\rangle = |G\rangle, |101110\rangle = |G\rangle, |101111\rangle = |G\rangle,$$

$$|110000\rangle = |AC\rangle, |110001\rangle = |ACT\rangle, |110010\rangle = |ACG\rangle, |110011\rangle = |ACA\rangle,$$

$$|110100\rangle = |ATC\rangle, |110101\rangle = |ATT\rangle, |110110\rangle = |ATG\rangle, |110111\rangle = |ATA\rangle,$$

$$|111000\rangle = |AG\rangle, |111001\rangle = |AG\rangle, |111010\rangle = |AG\rangle, |111011\rangle = |AG\rangle,$$

$$|111100\rangle = |A\rangle, |111101\rangle = |A\rangle, |111110\rangle = |A\rangle, |111111\rangle = |AAA\rangle.$$
From the expression (26), we have the following collective probability \( P_3(C_1) \) of all 16 triplets with the first letter C in them (they are collected in the first 4 rows of this expression):

\[
P_3(C_1) = (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 = \]

\[
= (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 \times \{\{\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4\}^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2\}.
\]

(28)

In this calculation, the sums in curly brackets are equal to 1 according to the normalization conditions (16, 25).

This mathematical result (28) of the proposed model means that collective probability \( P_3(C_1) \) of all 16 triplets with the first letter C is equal to the individual probability \( P_1(C) \) of the letter C in the same long DNA-text.

From the expression (26), similar calculations of collective probabilities \( P_3(T_1), P_3(G_1) \) and \( P_3(A_1) \) of triplets with the first letters T, G, A in them give analogical results (28) of their equality to individual probabilities of letters T, G and A:

\[
P_3(T_1) = P_1(T), \quad P_3(G_1) = P_1(G), \quad P_3(A_1) = P_1(A)
\]

(29)

These model results (28, 29) correspond to phenomological facts reflected in the first rule of tetra-group symmetries in long DNA-texts (see Section 3 above).

Similar calculations on the basis of the expression (26) give the following results:

- collective probabilities of all triplets with the second letters C, T, G, A in them are also equal to individual probabilities of letters C, T, G and A:
  \( P_3(C_2) = P_1(C), \quad P_3(T_2) = P_1(T), \quad P_3(G_2) = P_1(G) \) and \( P_3(A_2) = P_1(A) \);
- collective probabilities of all triplets with the third letters C, T, G, A in them are also equal to individual probabilities of letters C, T, G and A:
  \( P_3(C_3) = P_1(C), \quad P_3(T_3) = P_1(T), \quad P_3(G_3) = P_1(G) \) and \( P_3(A_3) = P_1(A) \).

These model results correspond to phenomological facts reflected in the second rule of tetra-group symmetries in long DNA-texts (see Section 3 above).

By analogy, in this quantum-informational model approach, one can consider cases of \( 4^n \) n-plents in appropriate \( 4^n \)-dimensional Hilbert space \( (n = 4, 5, ...) \) and get again and again confirmations of correspondence of this model to the described phenomenological rules of approximate tetra-group symmetries of collective and individual probabilities in long DNA-texts. The model describes these symmetries of probabilities as exact symmetries in contrast to phenomenological data, where they exist as approximate symmetries.

An effective model should not only explain known phenomological data but also should predict unknown data to search them in natural systems. Let us show now that the proposed quantum-informational model approach not only
explains the phenomenological facts, but also has predictive power, allowing to open previously unknown properties of long DNA-texts.

For this, consider the expression (26) of separable pure state of a quantum 6-qubit “triplet CTGA-system”. The right part of the expression (26) contains 64 computational basis states |CCC>, |CCT>, ..., |AAA>, which correspond to 64 genetic triplets. The set of 64 triplets contains the following 16 groups of triplets, each of which contains 4 triplets with one of 16 doublets on their identical positions. For example, the doublet CC defines one of 4²-groups of 4 triplets, each of which has this doublet in its beginning: CCC, CCT, CCG and CCA.

One should note that in previous Sections of this article we considered tetra-groups (or 4-groups), 4 subgroups of which were defined by monoplets: C-subgroups, T-subgroups, G-subgroups and A-subgroups (see Figs. 1, 2). Such 4-groups can be called as “4-groups of the first degree”. But now we will pay our attention to the wider sets of the following 4³-groups (that is, tetra-groups of higher degrees s = 2, 3, 4, ...):
- 4²-groups, each of which contains 4 subgroups defined by one of 16 doublets (like the CC-subgroup of triplets CCC, CCT, CCG and CCA);
- 4³-groups, each of which contains 4 subgroups defined by one of 64 triplets (like the CCC-subgroup of tetraplets CCCC, CCCT, CCGG and CCAA);
- etc.

In the right part of the expression (26), each of 16 rows represents one of 4²-groups of triplets with one of 16 doublets in their first positions. Let us take the first row of (26) and calculate the collective probabilities \( P_3(\text{CC}) \) of all 4 triplets CCC, CCT, CCG and CCA with the doublet CC in their beginning, using amplitudes of individual probabilities of the computational basis states |CCC>, |CCT>, |CCG> and |CCG>. The expression (30) shows the result of such calculation:

\[
P_3(\text{CC}) = (\alpha_0\beta_0\alpha_2\beta_3)^2 + (\alpha_0\beta_0\alpha_2\beta_3)^2 + (\alpha_0\beta_0\alpha_2\beta_3)^2 = \left(\alpha_0\beta_0\alpha_2\beta_3\right)^2 = P_2(\text{CC}) \tag{30}
\]

Here the sum in curly brackets is equal to 1 in accordance with the normalization condition (25). The model result (30) shows that the collective probability \( P_3(\text{CC}) \) of the set of all 4 triplets CCC, CCT, CCG, CCA with the doublet CC in their beginning is equal to the individual probability \( P_3(\text{CC}) \) of the doublet CC in the same long DNA-text.

Similar calculations of collective probabilities \( P_3(\text{CT}) \), \( P_3(\text{CG}) \), etc. of other 15 sets of 4 triplets in other 15 rows in (26) give analogical results (31) of their equality to individual probabilities of appropriate doublets:

\[
\begin{align*}
P_3(\text{CT}) &= (\alpha_0\beta_0\alpha_2\beta_3)^2 = P_2(\text{CT}), \\
P_3(\text{CA}) &= (\alpha_0\beta_0\alpha_3\beta_2)^2 = P_2(\text{CA}), \\
P_3(\text{TT}) &= (\alpha_0\beta_1\alpha_2\beta_3)^2 = P_2(\text{TT}), \\
P_3(\text{TA}) &= (\alpha_0\beta_1\alpha_3\beta_2)^2 = P_2(\text{TA}), \\
P_3(\text{GT}) &= (\alpha_1\beta_0\alpha_2\beta_3)^2 = P_2(\text{GT}), \\
P_3(\text{GA}) &= (\alpha_1\beta_0\alpha_3\beta_2)^2 = P_2(\text{GA}), \\
P_3(\text{AT}) &= (\alpha_1\beta_1\alpha_3\beta_2)^2 = P_2(\text{AT}), \\
P_3(\text{AG}) &= (\alpha_1\beta_1\alpha_3\beta_2)^2 = P_2(\text{AG}), \\
\end{align*}
\]

\[
\begin{align*}
P_3(\text{CT}) &= (\alpha_0\beta_0\alpha_2\beta_3)^2 = P_2(\text{CT}), \\
P_3(\text{CG}) &= (\alpha_0\beta_0\alpha_3\beta_2)^2 = P_2(\text{CG}), \\
P_3(\text{CT}) &= (\alpha_0\beta_1\alpha_2\beta_3)^2 = P_2(\text{CT}), \\
P_3(\text{GP}) &= (\alpha_1\beta_0\alpha_2\beta_3)^2 = P_2(\text{GP}), \\
P_3(\text{GT}) &= (\alpha_1\beta_0\alpha_3\beta_2)^2 = P_2(\text{GT}), \\
P_3(\text{GA}) &= (\alpha_1\beta_1\alpha_2\beta_3)^2 = P_2(\text{GA}), \\
P_3(\text{AT}) &= (\alpha_1\beta_1\alpha_3\beta_2)^2 = P_2(\text{AT}), \\
P_3(\text{AG}) &= (\alpha_1\beta_1\alpha_3\beta_2)^2 = P_2(\text{AG}), \\
\end{align*}
\]

(31)
From the expression (26), calculations of collective probabilities $P_3(CC_2)$ of all 4 triplets CCC, CCT, CCG and CCG with the doublet CC in their end, give the following:

$$P_3(CC_2) = (\alpha_0\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_0\beta_1\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_1\beta_0\alpha_2\beta_2\alpha_4\beta_4)^2 + (\alpha_1\beta_1\alpha_2\beta_2\alpha_4\beta_4)^2 =$$
$$= (\alpha_2\beta_2\alpha_4\beta_4)^2 * ((\alpha_0\beta_0)^2 + (\alpha_1\beta_0)^2 + (\alpha_1\beta_1)^2) = (\alpha_2\beta_2\alpha_4\beta_4)^2 = P_2(CC)$$

(32)

Here the sum in curly brackets is equal to 1 in accordance with the normalization condition (12).

From the expression (26), similar calculations of collective probabilities $P_3(CT_2)$, $P_3(CG_2)$, etc. of other 15 sets of 4 triplets with identical doublets in their end give analogical results (33) of their equality to individual probabilities of appropriate doublets:

$$P_3(CT_2) = (\alpha_2\beta_2\alpha_4\beta_5)^2 = P_2(CT),$$
$$P_3(CA_2) = (\alpha_2\beta_3\alpha_5\beta_5)^2 = P_2(CA),$$
$$P_3(TT_2) = (\alpha_2\beta_3\alpha_5\beta_5)^2 = P_2(TT),$$
$$P_3(TA_2) = (\alpha_2\beta_3\alpha_5\beta_5)^2 = P_2(TA),$$
$$P_3(CT_2) = (\alpha_2\beta_3\alpha_5\beta_5)^2 = P_2(CT),$$
$$P_3(TA_2) = (\alpha_2\beta_3\alpha_5\beta_5)^2 = P_2(TA),$$
$$P_3(GA_2) = (\alpha_3\beta_2\alpha_5\beta_3)^2 = P_2(GA),$$
$$P_3(GA_2) = (\alpha_3\beta_2\alpha_5\beta_3)^2 = P_2(GA),$$
$$P_3(AA_2) = (\alpha_3\beta_3\alpha_5\beta_5)^2 = P_2(AA).$$

(33)

One can consider these model results as predictions of a possible existence of additional rules of symmetries of collective probabilities of the following sets of n-plets in long DNA-texts in cases of the mentioned 4-s-groups of oligonucleotides ($s = 2, 3, 4, ...$), for example:

- the CC-subgroup, CT-subgroup, CG-subgroup and CA-subgroup of triplets;
- the CCC-subgroup, CCT-subgroup, CCG-subgroup, CCA-subgroup, CTC-subgroup, CTT-subgroup and other similar subgroups of 4-plets;
- the CCCCC-subgroup, CCCT-subgroup, CCGG-subgroup, CCCA-subgroup and other similar subgroups of 5-plets;
- etc.

The Appendix 7 shows some initial confirmations of these theoretical predictions made on the basis of the described quantum-information model. Of course, these predictions should be checked for much more number of long DNA-texts including genomes and complete sets of chromosomes of different organisms. We hope to publish results of such checking some later.

But can the proposed quantum-information approach to long DNA-texts reveal a reason of implementation of the second Chargaff’s parity rule? The next Section gives the positive answer on this question.

6. The explanation of the second Chargaff’s parity rule on the basis of the quantum-information approach

The second Chargaff’s parity rule says that in long DNA-texts of nucleotides (or in DNA-texts of 1-letter words) the following approximate equalities are true: $P_1(A_1) \equiv P_1(T_1)$ and $P_1(C_1) \equiv P_1(G_1)$. For a possible
explanation of this rule, let us return to the separate pure state $|\psi_{12}\rangle = \alpha_0\beta_0 |C> + \alpha_0\beta_1 |T> + \alpha_1\beta_0 |G> + \alpha_1\beta_1 |A> \quad (10)$ of a quantum 2-qubit CTGA-system, which contains amplitudes $\alpha_0\beta_0, \alpha_0\beta_1, \alpha_1\beta_0, \alpha_1\beta_1$ of probabilities $P_1(C_1), P_1(T_1), P_1(G_1), P_1(A_1)$ of computational basis states $|C>, |T>, |G>, |A>$. In the quantum-informational notations of (10), $P_1(C_1)=(\alpha_0\beta_0)^2$, $P_1(T_1)=(\alpha_0\beta_1)^2$, $P_1(G_1)=(\alpha_1\beta_0)^2$, $P_1(A_1)=(\alpha_1\beta_1)^2$. In these notations, the second Chargaff’s parity rule says the following (34):

$$\ (\alpha_0\beta_1)^2 \equiv (\alpha_1\beta_1)^2 \text{ and } (\alpha_0\beta_0)^2 \equiv (\alpha_1\beta_0)^2 \quad (34)$$

It is obvious that the both equalities (34) are implemented only if the following single equality (35) is implemented:

$$\alpha_0^2 \equiv \alpha_1^2 \quad (35)$$

But in accordance with the expression (7), $\alpha_0$ and $\alpha_1$ denote amplitudes of probabilities of the computational basis states “pyrimidine” and “purine” in a 1-qubit quantum state, which is used for writing separable pure states of 2n-qubit CTGA-systems. The expression (35) shows that the second Chargaff’s rule can be considered as the consequence of the condition of the approximate equality of probabilities of these computational basis states - «pyrimidine» or «purine» - in long DNA-sequences of nucleotides (or in DNA-sequences of 1-letter words) as 2-qubit quantum systems.

It means that the genetic system does not give a quantitative preference to pyrimidines or purines, both of them are in equal situations of their informational usage. One should specially note the following difference between the second Chargaff’s rule and this principle of approximate equality of probabilities of pyrimidines and purines (for brevity, the last can be called «the principle of pyrimidine-purine equality» in long DNA-sequences:

- The second Chargaff’s rule deals with two separate equalities ($P_1(A_1) = P_1(T_1)$ and $P_1(G_1) = P_1(C_1)$), origin of which is difficult to explain in terms of common sense;
- The principle of pyrimidine-purine equality deals only with the single equality (35), origin of which is simply explained by the thought that the genetic system does not give quantitative preference to pyrimidines or purines (the genetic system doesn’t need additional mechanisms to provide a quantitative preference to pyrimidines or purines).

Outside of the Section’s topic about the second Chargaff’s rule, one can make the following small remark of a general type. The described tetra-group rules concern each of two strands of DNA, the complementary nitrogenous bases of which are bound together into pairs (A with T, and C with G). Taking into account that the two strands of DNA run in opposite directions to each other (they are antiparallel), any long DNA-sequence of n-plets ($n = 1, 2, 3, 4, 5, \ldots$) of one of DNA-strand corresponds to an appropriate sequence of complement-reverse n-plets of its second strand. This leads to questions about complement-reverse tetra-groups of oligonucleotides and about a possible existence - in single stranded DNA - additional tetra-group rules connected with such complement-reverse tetra-groups. Fig. 17 shows the complement-reverse tetra-
groups of doublets, which are received from the «direct» tetra-group of doublets in Fig. 1 by the replacement of each of doublets by its complement-reverse doublet. One can see from the comparison of Fig. 1 and Fig. 17 that the initial tetra-groups of doublets in Fig. 1 distinguish from the new complement-reverse tetra-groups of doublets in Fig. 17 only permutations without arising essential new subgroups of tetra-groups: the set of members of the tetra-group of doublets with the same letter at their first (second) position from Fig. 1 is identical to the set of the complement-reverse tetra-group of doublets with the same letter at their second (first) position in Fig. 17. This leads to the conclusion that additional tetra-group rules connected with such complement-reverse tetra-groups are absent since these tetra-groups are simple repetitions - by their essential aspects - of known tetra-groups of doublets. The same is true for tetra-groups in cases of triplets and other n-plets.

| Composition of the complement-reverse tetra-group for doublets with the same letter at their second position | Composition of the complement-reverse tetra-group for doublets with the same letter at their first position |
|---------------------------------------------------------------|---------------------------------------------------------------|
| TT, GT, CT, AT                                               | TT, TG, TC, TA                                               |
| GA, TA, AA, CA                                               | AG, AT, AA, AC                                               |
| GG, TG, AG, GC                                               | GG, GT, GA, GC                                               |
| GC, TC, AC, CC                                               | CG, CT, CA, CC                                               |

Fig. 17. Compositions of the complement-reverse tetra-groups, which are received by the replacement of each of doublets in Fig. 1 by its complement-reverse doublet.

The Chargaff's parity rules are frequently called in the literature as Symmetry Principles [Yamagishi, Herai, 2011, p. 2]. The represented tetra-group rules give new data about hidden symmetries of long sequences in single stranded DNA. This continues the theme of molecular-genetic symmetries described, in particular, in our works [Petoukhov, 2008, 2011, 2012; Petoukhov, He, 2009; Petoukhov, Petukhova, 2017a, b; Petoukhov, Svirin, 2012].

7. About short DNA-texts

The last two Sections have shown a possibility and effectiveness of usage of quantum informatics to model properties of long DNA-texts by means of the formalisms of separable pure states of 2n-qubit quantum CTGA-systems. Now let us pay attention to a possible meaning for genetic systems other types of states of 2n-qubit quantum CTGA-systems. One can see from expression (10), which describes a separable pure state of 2-qubit system, that its amplitudes \(\alpha_0\beta_0, \alpha_0\beta_1, \alpha_1\beta_0, \alpha_1\beta_1\) are closely interrelated with each other. Tetra-group symmetries in long DNA-texts say that these amplitudes in the expressions (10, 13) are interrelated in an appropriate manner. If one of these amplitudes is changed and the corresponding interrelation disappears, then the changed expression (10) will correspond to non-separable pure state.
In a general case, the tetra-group rules are not met for short DNA-texts (for example, for separate genes), and short DNA-texts cannot be represented in a form of separable quantum states $|\Psi> = |\Psi_1> \otimes |\Psi_2>$. In quantum informatics, a state of a composite system, which can’t be written as a tensor product of states of its component systems, is called an entangled state [Nielsen, Chuang, 2010, p. 96; https://en.wikipedia.org/wiki/Separable_state]. Correspondingly, genes and other relative short DNA-texts can be considered as quantum CTGA-systems in entangled states. In quantum informatics, entangled states play very important roles. The study and use of entangled states is one of the main problems of quantum computing: “…entanglement is a key element in effects such as quantum teleportation, fast quantum algorithms, and quantum error-correction. It is, in short, a resource of great utility in quantum computation and quantum information. … entangled states play a crucial role in quantum computation and quantum information” [Nielsen, Chuang, 2010, p. XXIII and p. 96].

The quantum operator, which allows you to entangle two qubits $|x>$ and $|y>$, is called the operator CNOT and is given by the expression (36):

$$P_{12}|x, y> = |x, x \oplus y>$$

(36)

where $x \oplus y$ is the logic operation of modulo-2 addition. Above in the Section 3 we shown that the DNA-alphabet of nucleotides C, T, G, A is connected with the logic operation of modulo-2 addition: any DNA-text is the carrier of three parallel messages on three different binary languages, and these three types of binary representations of the DNA-text form a common logic set on the basis of logic operation of modulo-2 addition. This connection of DNA-texts with the logic operation of modulo-2 addition can be used to represented fragments of long DNA-texts as entangled states of quantum CTGA-systems for study hidden regularities and meanings of genetic messages. Beside this, the connection of DNA-structures with the logic operation of modulo-2 addition testifies additionally in favor of adequacy of the quantum-information approach to analyze DNA-texts. One can add that a possible important meaning of entangled states for functioning biological ensembles of protein molecules was discussed in [Matsuno, Paton, 2000].

One should note here also that, as known, Hadamard matrices («Hadamard gates») play essential role in quantum computers; they are used in in quantum mechanics in the form of unitary operators, etc. But structures of genetic alphabets are naturally connected with Hadamard matrices [Petoukhov, 2008, 2010, 2011, 2016, 2017a; Petoukhov, He, 2009]. It additionally testifies in favor of quantum-informational basises of genetic informatics.

In quantum mechanics, matrix operators play a significant role. One of them is the density operator (or the density matrix) that describes quantum systems in pure and mixed states. All the postulates of quantum mechanics can be reformulated in terms of the density operator language [Nielsen, Chuang, 2010, p. 99]. Quantum informatics possesses a rich set of other useful notions and mathematical formalisms including quantum search algorithms, quantum algorithms for encoding information with error-corrections, the quantum
Fourier transform, the Schmidt decomposition, quantum logic, trace distances, stabilizer codes, von Neumann entropy, quantum circuits, etc. In our opinion, applications of these notions and formalisms in genetic informatics will be very useful for progress of theoretical biology, medicine, systems of artificial intelligence and other fields including the quantum computer science itself. In particular, it will lead to development of quantum-informational genetics as a perspective scientific direction.

8. Resonances, photons and quantum-information genetics

From the point of view of quantum mechanics, the interaction of molecules is based on the emission and absorption of photons with the participation of resonance correspondences.

In modern molecular biology, when considering the interaction of molecules, the idea of the stereochemical correspondence of molecules is usually used, that is, the idea of the correspondence of the spatial configurations of molecules (the model of the "lock-and-key" correspondence). In our quantum-informational approach, quite another idea is in the centre of attention: the idea of exchanges of information among energetical states of interrelated molecules by means of photons on principles of resonances. In our approach, nucleotides and their combinations are represented by not their spatial configurations but, first of all, by their energetical peculiarities, which provide opportunities to exchange intermolecular information due to the emission and absorption of photons of certain energies (wavelengths) with the participation of resonance correspondences.

This Section is devoted to additional explanations of this quantum-informational approach, which is connected with the fundamental notion "resonance" and with photons as carriers of intermolecular information and also as the force carriers for electromagnetic force. Below we represent arguments why one can think about a connection of tetra-group symmetries in long DNA-texts with phenomena of resonances.

Quantum mechanics is closely connected with phenomena of resonances. This science has begun from 1900 year with the pioneer work of M. Planck (1936), who has analyzed a great set of resonant oscillators inside the special cavity to receive his famous law of electromagnetic radiation emitted by a black body in thermal equilibrium. One can say that Planck has represented the matter as a set of vibrating oscillators and set the task to study the equilibrium, which was established in the result of the exchange of energy between the oscillators and radiation. This work has introduced into science the Planck's constant as the quantum of action, central in quantum mechanics. It was originally the proportionality constant between the minimal increment of energy of a hypothetical electrically charged oscillator in the mentioned cavity and the frequency of its associated electromagnetic wave.

Later, after more than 50 years of successful development of quantum mechanics, E. Schrödinger emphasised the basic meaning of resonances: "The one thing which one has to accept and which is the inalienable consequence of the wave-equation as it is used in every problem, under the most various forms, is this: that the interaction between two microscopic physical systems is controlled by a peculiar law of resonance" [Schrödinger, 1952, p. 115]. In considering an exact
balance in nature between bundles of energy, lost by one system and gained by another, he noted: «I maintain that it can in all cases be understood as a resonance phenomenon» (ibid, p. 114). He wrote in his resonance concept of quantum interactions that chemical reactions, including photochemical reactions, can be explained on the base of resonances. One of examples considered in his article was a production of water molecules H2O from a suitable mixture of hydrogen gas H2 and oxygen gas O2 under action of ultraviolet light. In this example, “wave-mechanically the gaseous mixture is represented by a vibration of the combined system, and, by the way, not by one proper vibration since there is anyhow the vast variety of translational and rotational modes; and, of course, the electronic modes. The gaseous compound, H2O, is represented by an entirely different vibration of the same system” (ibid, p. 118).

His book [Schroedinger, 1944] declared that the chromosome is an aperiodic crystal since its atoms are connected each with other by forces of the same nature that atoms in crystals. This standpoint attracts an attention to a very important role of vibrations and resonances in crystals. The presence of the interaction of atoms in the crystal lattice together with the resonance phenomenon leads to the fact that oscillatory motions of lattice elements are combined in a collective oscillation process in a form of a wave propagating in the crystal. In the course of the normal vibrations, all the atoms in the crystal lattice oscillate about their equilibrium positions by harmonic law with the same frequency. As it is known, in a quantum description of small oscillations of a crystal, it is possible to interpreted normal fluctuations of the crystal as special quasiparticles, which are quanta of the field of elastic vibrations of the crystal and which are called phonons. The theory of phonons is one of the bases of physics of crystalls. One can hope that a similar resonance approach can usefully serve in genetics.

The notion “resonance” was introduced into quantum mechanics by W. Heisenberg in 1926 year in connection with analyzes of multi-body systems. He emphasized that in quantum mechanics the phenomenon of resonances has much more general character than in classical physics. In classic theory, two periodic oscillating systems come into their own resonance only in the case when a frequency of a separate sub-system doesn’t depend on energy of the system and when this frequency is approximately equal in both sub-systems. In quantum mechanics, two atomic systems come into their resonance only in the case when a frequency of absorption of one system coincides with a frequency of emitting another system, or vice versa [Heisenberg, 1926, §2]. Quantized electromagnetic field is represented as a set of oscillators.

L.Pauling used ideas of resonances in quantum mechanical systems in his theory of resonance in structural chemistry. His book [Pauling, 1940] about this theory is the most quoted among scientific books of the XX century. The theory was developed to explain the formation of hybrid bonds in molecules. The actual molecule, as Pauling proposed, is a sort of hybrid, a structure that resonates between the two alternative extremes; and whenever there is a resonance between the two forms, the structure is stabilized. His theory uses the fundamental principle of a minimal energy because – in resonant combining of parts into a single unit – each of members of the ensemble requires less energy for performing own work than when working individually. Of course, this fundamental principle can be used in many other cases of resonances in different
systems as the physical base. The principle of energetic minimum in resonance processes has some correlations with the principle of relaxation in morphogenetic processes proposed in [Igamberdiev, 2012].

Concerning problems of bioenergetics, McClare (1974) put forward the hypothesis about long-lived vibrational excitations in proteins, which could play an important role in protein functions. The results of researches on this hypothesis was summarized in the work [Turin, 2009]: “Colin McClare was the first scientist to envision the potential importance of long-lived vibrational excitations in proteins. He did so at a time when there was no experimental evidence for their existence, no plausible mechanism to make them happen, and no interest in the possibility among his peers... Thirty years later (after his death in 1977 – S.P.), we can see that McClare had got one fundamental thing right: vibrational energy can be stored and transmitted in proteins». This hypothesis of McClare is associated with the idea of Bauer (1935) that living systems work in expense of non-equilibrium, and the external energy is used not directly to perform work but to support the stable non-equilibrium state; most of this energy is transformed into the kinetic energy.

For a theme of a general importance of resonances in living matter, interesting materials are represented in works [Ji, 2012, 2015, 2017]. Their author has postulated analogy between enzymic catalysis and blackbody radiation, which was modelled by Planck due to his idea about huge number of resonances. This postulation was made by Ji on the base, first of all, of his observations that some important biological phenomena have their graphical representations analogical to ones of blackbody radiations. This author has proposed a generalization of the Planck equation for modeling different biological phenomena, having long-tailed histograms, from the general standpoint: protein folding, single-molecule enzymology, whole-cell transcriptomics, T-cell receptor variable region diversity, brain neurophysiology, RNA levels in budding yeast, human breast cancer tissues, codon profile in the human genome, etc. By analogy with the principle of quantization of energy in quantum mechanics, Ji postulates a quantization of free energy levels in enzymes. He also proposes an original theory of molecular machines with using Franck-Condon mechanisms concerning vibronic transitions, which are simultaneous changes in electronic and vibrational energy levels of a molecule. His thoughts about musical-wave harmony in organization of living matter are correlated with musical aspects of our concept of systemic-resonance genetics, where an organism is considered as a very complex and developing music synthesizer and some kinds of so called «genetic music» are developed and represented in musical concerts of Moscow P.I.Tchaikovsky Conservatory [Hu, Petoukhov, 2017; Petoukhov, 2015b].

The articles [Petoukhov, 2015a, 2016; Petoukhov, Petukhova, 2017b] describe our concept about the important role of resonances in genetic structures. This concept is based on impressive analogies of some genetic structures and phenomena, including Mendelian laws, with eigenvalues and eigenvectors of tensor families of matrix operators representing oscillatory systems with many degrees of freedom (these operators can be interpreted as quantum operators). In these works we described data to the idea about existence of biological resonant computers inside any organism, which are based on binary-oppositional resonances in genetic systems.
Let us repeat that the interaction of molecules is based on the emission and absorption of photons (particles of light) with the participation of resonance correspondences. Therefore, special attention should be paid to the important role of photons in genetic informatics.

As known, a photon is a type of elementary particle, the quantum of electromagnetic field including electromagnetic radiation such as light, and the force carrier for electromagnetic force (in particle physics, force carriers or messenger particles or intermediate particles are particles that give rise to forces between other particles (https://en.wikipedia.org/wiki/Force_carrier). The photon has zero rest mass and always moves at the speed of light within a vacuum. A photon has two possible polarization states. Like all elementary particles, photons are currently best explained by quantum mechanics and exhibit wave-particle duality, exhibiting properties of both waves and particles. The photon’s wave and quanta qualities are two observable aspects of a single phenomenon, and cannot be described by any mechanical model; a representation of this dual property of light, which assumes certain points on the wavefront to be the seat of the energy, is not possible. The quanta in a light wave cannot be spatially localized (https://en.wikipedia.org/wiki/Photon).

Photon energy is the energy carried by a single photon. The amount of energy is directly proportional to the photon’s electromagnetic frequency and inversely proportional to the wavelength. Photon energy is solely a function of the photon’s wavelength. Other factors, such as the intensity of the radiation, do not affect photon energy. In other words, two photons of light with the same «color» and therefore, same frequency, will have the same photon energy. The equation for photon energy E is the following: \( E = \frac{hc}{\lambda} = hf \), where «h» is the Planck constant, «c» is the speed of light in vacuum, «\( \lambda \)» is the photon’s wavelength and «f» is the photon’s wavefrequency.

Photons, which are radiated by different molecular elements, can differ by their frequencies. In systems of genetic molecules, where each of nitrogenous bases C, T, G, A is a carrier of its individual pair of binary-oppositional molecular indicators (Fig. 14), each of these two molecular indicators can define a radiation of photons with their individual frequencies. Jointly the two molecular indicators can radiate a beam of photons with two kinds of frequencies, which defines its appropriate computational basis state of a 2-qubit system in the expression (10) with the conditional denotations \( |C\rangle=|00\rangle \), \( |T\rangle=|01\rangle \), \( |G\rangle=|10\rangle \) and \( |A\rangle=|11\rangle \).

The existence of DNA-texts is accompanied by a rich set of appropriate beams of multicore photons to provide a cooperative informational functioning of ensembles of genetic elements. From this point of view, nitrogenous bases C, T/U, A and G and their combinations in DNA and RNA are resonance directors of frequencies of biophotonic ensembles within living bodies (or briefly, “biophotons directors”). The reading and transmission of genetic information from DNA and RNA molecules occurs by means of a set of resonance frequencies of biological photons. DNA encodes quantum states of biophoton beams. The photons language is a serious candidacy for the role of a basic language of molecular-genetic information. Figuratively speaking, from this point of view, life in its information aspects is woven from the light.

Photons are actively studied in modern science as elements of quantum computers and devices of quantum cryptography. In models of quantum computers, conventional light polarizers are used to create pure and mixed
states of n-qubit systems of light beams. The idea of ensembles of «multicolor» photons for a creation of n-qubit states, which was noted by us above in the connection with DNA-texts, was independently used in the recent engineering work of Canadian scientists [Caspani et al., 2016]. This work has revealed a new perspective way to create quantum computers. For increasing dimensionality of the photon quantum state, its authors used the ability to generate multiple photon pairs on a frequency comb, corresponding to resonances in specifically designed microcavities. The possibility of fully exploiting the polarization degree of freedom, even for integrated devices, exists for further achievements.

One should note that, as known, living bodies possess inherited opportunities to manage photonic beams using physical principles of photonic crystals with their properties of photon gratings, etc. Many biological phenomena of structural coloration and of animal reflectors are built on this, including a beautiful coloring of butterfly wings, peacock feathers, etc. (see details and lists of references in https://en.wikipedia.org/wiki/Photonic_crystal, https://en.wikipedia.org/wiki/Animal_reflectors, https://en.wikipedia.org/wiki/Structural_coloration). A photonic crystal is a periodic optical nanostructure that affects the motion of photons. Photonic crystals contain regularly repeating regions of high and low dielectric constant. Photons (behaving as waves) either propagate through this structure or not, depending on their wavelength. This gives rise to distinct optical phenomena, such as inhibition of spontaneous emission, high-reflecting omni-directional mirrors, and low-loss-waveguiding. The periodicity of the photonic crystal structure must be around half the wavelength of the electromagnetic waves to be diffracted.

We believe that spatial characteristics of ensembles of genetic and other biological molecules, that form complex diffraction structures, play the managing role of photonic crystals in the problem of controlling biophoton beams that are generated and absorbed by these molecules. In particular, the spatial configuration of genetic molecules as biophotonic crystals is an important factor in controlling the processes of transmission of genetic information from DNA and RNA molecules with using biophoton beams generated by them. Not without reason, Schrödinger called chromosomes aperiodic crystals [Schrödinger, 1944].

In our opinion, the inherited processes of the morphogenesis of living bodies are also determined to a large extent by bio-photon beams, the course of which is not accidental, but is strictly organized by a system of spatial characteristics of ensembles of genetic and other biological molecules as biophotonic crystals. In the course of ontogeny, on the basis of electromagnetic (photonic) interactions, new molecular materials are involved into a naturally growing biological body, which leads to the appropriate growth of the managing system of biophotonic crystals and to the growth of numbers of photon beams. Of course, quantum-mechanic laws of resonances in molecular photonic interactions play a basic role. On this basis, we develop our concept of the "morpho-resonance field" to model morphogenetic phenomena [Petoukhov, 2015cd, 2016].

Classical electrodynamics describes a photon as an electromagnetic wave with its circular right or left polarization. To the theme “life and photons”, one can add many interesting connections of these polarization properties of photons with inherited properties of living bodies, for example, the following:
• One of the biggest mysteries of nature is the asymmetry of biological molecules, accompanied by a preferred direction of the rotation - to the left or to the right - of the polarization plane of light by these molecules (this was discovered by Louis Pasteur). For example, all biological amino acids (except the simplest glycine, which is symmetric), from which the proteins of all living organisms are composed, exist only in one of two possible asymmetric forms - in the left form. Amino acids in this form rotate the plane of polarization of light to the left. Our body doesn’t use amino acids with the opposite right form, rotating the plane of polarization of light to the right. Biological catalysts - enzymes, - being built asymmetrically, act only on one optical antipode, without touching the other. The same asymmetry with respect to the right and left is inherent not only in amino acids, but also in the nucleotides that form DNA and RNA. The reason for this is the asymmetry of the components of the sugar, which is part of the nucleotides and which provides the optical activity of DNA and RNA molecules: they rotate the plane of polarization of light to the right.

• Millions of species of living organisms (insects, mollusks, arthropods, etc.) are endowed with inherited ability to see in polarized light (a human organism does not possess this ability).

At the end of the Section, one can say that our quantum-informational approach for modeling and analysis of the genetic system provides new materials for understanding the role of photons in the organization and functioning of the genetic system.

9. The rules of symmetries of collective probabilities in tetra-groups in the complete set of human chromosomes. The fourth rule of symmetries of tetra-group probabilities (for complete sets of chromosomes)

This Section is devoted to results of author's study of tetra-group symmetries in the complete set of human chromosomes. Fig. A2/1-A2/24 in the Appendix 2 show resulting tables of tetra-group symmetries of all 24 human chromosomes, initial data about which were taken from the GenBank. Let us note some important aspects of the obtained results.

Human organisms contain 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.

First of all, our results demonstrate very high accuracy of implementation of all the rules of tetra-group symmetries for each of human chromosomes. Fig. 18 shows the average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$, expressed in percentage, for each of the 24 chromosomes ($n = 1, 2, 3, 4, 5; k \leq n$) where the numerical values of probabilities in percentage are rounded to the third decimal place. Fluctuations of the probabilities around these averages are also shown. One should note that these fluctuations are very small.
| CHROMOSOME | AVERAGE VALUE OF $P_n(A_k)$ AND FLUCTUATIONS (%) | AVERAGE VALUE OF $P_n(T_k)$ AND FLUCTUATIONS (%) | AVERAGE VALUE OF $P_n(C_k)$ AND FLUCTUATIONS (%) | AVERAGE VALUE OF $P_n(G_k)$ AND FLUCTUATIONS (%) |
|------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|
| 1          | 29,100±0,005                               | 29,176±0,006                               | 20,850±0,006                               | 20,874±0,005                               |
| 2          | 29,845±0,005                               | 29,927±0,006                               | 20,807±0,006                               | 20,142±0,006                               |
| 3          | 30,131±0,006                               | 30,204±0,005                               | 19,805±0,006                               | 19,861±0,006                               |
| 4          | 30,862±0,005                               | 30,895±0,006                               | 19,097±0,006                               | 19,147±0,006                               |
| 5          | 30,176±0,006                               | 30,317±0,006                               | 19,712±0,006                               | 19,794±0,006                               |
| 6          | 30,210±0,006                               | 30,205±0,006                               | 19,794±0,006                               | 19,791±0,006                               |
| 7          | 29,602±0,006                               | 29,701±0,006                               | 20,330±0,006                               | 20,368±0,006                               |
| 8          | 29,939±0,007                               | 29,898±0,007                               | 20,081±0,007                               | 20,081±0,006                               |
| 9          | 29,280±0,007                               | 29,263±0,007                               | 20,736±0,007                               | 20,721±0,007                               |
| 10         | 29,172±0,007                               | 29,286±0,007                               | 20,741±0,007                               | 20,801±0,007                               |
| 11         | 29,202±0,007                               | 29,258±0,007                               | 20,741±0,007                               | 20,799±0,007                               |
| 12         | 29,571±0,007                               | 29,663±0,007                               | 20,349±0,007                               | 20,417±0,007                               |
| 13         | 30,688±0,007                               | 30,788±0,006                               | 19,259±0,007                               | 19,265±0,007                               |
| 14         | 29,452±0,006                               | 29,697±0,006                               | 20,397±0,007                               | 20,453±0,007                               |
| 15         | 28,956±0,007                               | 29,009±0,006                               | 20,974±0,006                               | 21,061±0,007                               |
| 16         | 27,575±0,007                               | 27,840±0,007                               | 22,214±0,007                               | 22,370±0,007                               |
| 17         | 27,303±0,006                               | 27,382±0,006                               | 22,581±0,007                               | 22,734±0,007                               |
| 18         | 30,106±0,006                               | 30,136±0,007                               | 19,866±0,006                               | 19,893±0,007                               |
| 19         | 25,911±0,006                               | 26,151±0,006                               | 23,878±0,007                               | 24,060±0,007                               |
| 20         | 27,780±0,006                               | 28,097±0,006                               | 22,009±0,007                               | 22,113±0,007                               |
| 21         | 29,644±0,006                               | 29,463±0,007                               | 20,465±0,006                               | 20,428±0,006                               |
| 22         | 26,512±0,007                               | 26,483±0,007                               | 23,393±0,006                               | 23,611±0,007                               |
| X          | 30,185±0,005                               | 30,290±0,005                               | 19,706±0,005                               | 19,819±0,006                               |
| Y          | 29,855±0,005                               | 30,120±0,005                               | 20,011±0,006                               | 20,015±0,005                               |

Fig. 18. The table of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ in percentage for each of 24 human chromosomes ($n = 1, 2, 3, 4, 5; k \leq n$). Fluctuations of the probabilities are also shown.

Here let us explain how these average values and fluctuations in relation to them are calculated, for example, in the case of the average value of $P_n(A_k)$ for the human chromosome No. 1. In the Appendix 2, each of tables of tetra-group symmetries (Fig. A2/1-A2/24) contains 15 cells with $P_n(A_k)$. When each of these probabilities $P_n(A_k)$ is calculated up to the seventh decimal place for this DNA-text, the following 15 values arise: $P_1(A_1)=0.2910013$, $P_2(A_2)=0.2910218$, $P_3(A_3)=0.2910489$, $P_4(A_4)=0.2910489$, $P_5(A_5)=0.2910489$, $P_6(A_6)=0.2910234$, $P_7(A_7)=0.2910234$, $P_8(A_8)=0.2910995$, $P_9(A_9)=0.2910202$, $P_{10}(A_{10})=0.2910202$, $P_{11}(A_{11})=0.2910202$, $P_{12}(A_{12})=0.2910202$, $P_{13}(A_{13})=0.2910202$, $P_{14}(A_{14})=0.2910202$, $P_{15}(A_{15})=0.2910202$. The sum of these 15 numbers divided by 15 gives their average value 0.2910013. The probability $P_n(A_k)$=0.2910489 has the strongest deviation from this average value 0.2910013. This deviation is equal to
0.0000476 ≈ 0.00005. Thus, in percentage denotation, we get for the probabilities $P_n(A_k)$ their mean value 29.100% and the deviation (fluctuation) 0.005% from it. The corresponding expression $(29.100±0.005)\%$ is shown in the table in Fig. 18 for the chromosome № 1. By analogy, values in all other cells of the table in Fig. 18 are calculated.

One can see from the table in Fig. 18 that average values of $P_n(A_k)$ and $P_n(T_k)$ are almost equal; the average values of $P_n(C_k)$ and $P_n(G_k)$ are also almost equal. Fig. 19 shows graphically average values of these probabilities.

![Graphical representation of average values of probabilities](image1)

**Fig. 19.** The graphical representation of the average values of the probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ from Fig. 18 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%.

![Diagram of human chromosomes](image2)

**Fig. 20.** Human chromosomes (https://www.ncbi.nlm.nih.gov/genome/51)

These 24 human chromosomes differ greatly in their molecular dimensions, their sequences of letters, kinds and quantities of genes in them, cytogenetic bands (which shows biochemical specificity of different parts of chromosomes), etc. (Fig. 20). Taking into account these great differences among
human chromosomes, it was very unexpected for the author to reveal that the 24 human chromosomes are similar each to other from the point of view of tetra-group symmetries of their DNA-texts. It means that all these chromosomes are not completely individual objects but they are closely interrelated each other in relation to tetra-group symmetries of their DNA-texts. As can be seen from the tables of tetra-group symmetries in Fig. 18, 19 and in the Appendix 2 (Fig. A2/1-A2/24), the maximum similarity in average values of collective probabilities of subgroups of the tetra-groups exists for DNA-texts of chromosomes №№ 1-18, 20, 21, X and Y (that is for 22 chromosomes from 24 chromosomes). Only two chromosomes № 19 and № 22 have their DNA-texts with noticeable deviations of these probabilities in their average values in 15-20% in comparison with the average values of the probabilities in DNA-texts of other 22 chromosomes.

It seems that 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”).

It is obvious that such collective text phenomena in the complete set of human chromosomes cannot be explained on the basis of ideas of stereochemical interrelations among genetic molecules by the principle “lock-and-key”. Here another direction of thoughts is needed. We think that quantum information approach should be used to explain such collective informational phenomena.

In the next Section we represent results of our study of tetra-group symmetries in complete sets of chromosomes of some model organisms, which are traditionally used in the study of genetics, development and disease. All these 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 the validity of the fourth rule of tetra-group symmetries, which concerns complete sets of chromosomes of different organisms:

- In the complete set of chromosomes of each of eukaryot organisms, characteristics of the tetra-group symmetries of separate chromosomes are approximately equal to each other for all chromosomes.

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

10. Tetra-group rules in complete sets of chromosomes of model organisms: Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Mus musculus

This Section is devoted to consideration of perculiarities of tetra-group symmetries in complete sets of chromosomes of a few model organisms, which are used long ago in the study of genetics, development and disease. Tables of tetra-group symmetries of these sets of chromosomes are represented in Appendixes 3-5. Data of these tables confirm the implementation of the tetra-group rules for chromosomes of these organisms and they show also that – for each of the organisms - characteristics of tetra-group symmetries of each of chromosomes are approximately equal for all chromosomes.

We begin with a nematode Caenorhabditis elegans. It is a free-living soil nematode with 959 somatic cells in its body. Caenorhabditis elegans was the first
multicellular eukaryotic genome to be completely sequenced. Its nuclear genome is approximately 100 Mb, distributed among six chromosomes (see tables in the Appendix 3). Fig. 21 and 22 shows the table and the corresponding graphical representation of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$. These average values were calculated from the tables in Fig. A3/1-A3/6 by analogy with the described calculation of average values of the probabilities in cases of human chromosomes (Fig. 18).

| CHROMOSOME | AVERAGE VALUE OF $P_n(A_k)$ (%) | AVERAGE VALUE OF $P_n(T_k)$ (%) | AVERAGE VALUE OF $P_n(C_k)$ (%) | AVERAGE VALUE OF $P_n(G_k)$ (%) |
|------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| 1          | 31,954                          | 32,032                          | 18,018                          | 17,998                          |
| 2          | 31,927                          | 31,871                          | 18,123                          | 18,078                          |
| 3          | 32,246                          | 32,093                          | 17,769                          | 17,892                          |
| 4          | 32,647                          | 32,760                          | 17,347                          | 17,247                          |
| 5          | 32,263                          | 32,309                          | 17,741                          | 17,689                          |
| X          | 32,435                          | 32,361                          | 17,607                          | 17,596                          |

Fig. 21. The table of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ in percentage for each of 6 chromosomes of a nematode *Caenorhabditis elegans* (n = 1, 2, 3, 4, 5; k ≤ n).

Fig. 22. The graphic representation of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ in percentage for each of 6 chromosomes of a nematode *Caenorhabditis elegans* (n = 1, 2, 3, 4, 5; k ≤ n). The abscissa axis contains numberings of chromosomes, and the ordinate axis contains average values of these probabilities in percent. The symbol of a blue diamond corresponds the average values of $P_n(A_k) \approx P_n(T_k)$, and the symbol of a red square corresponds the average values of $P_n(C_k) \approx P_n(G_k)$.

Now let us turn to the second model organism - *Drosophila melanogaster*, which is studied in biology labs for over eighty years. Tables of tetra-group symmetries of its set of 7 chromosomes are represented in the Appendix 4.
Fig. 23 and 24 shows the table and the corresponding graphical representation of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ for all chromosomes of *Drosophila melanogaster*. These average values were calculated from the tables in Fig. A4/1-A4/7 by analogy with the described calculation of average values of the probabilities in cases of human chromosomes (Fig. 18).

| CHROMOSOME | AVERAGE VALUE OF $P_n(A_k)$ (%) | AVERAGE VALUE OF $P_n(T_k)$ (%) | AVERAGE VALUE OF $P_n(C_k)$ (%) | AVERAGE VALUE OF $P_n(G_k)$ (%) |
|------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| X          | 28,679                          | 28,858                          | 21,195                          | 21,267                          |
| 2L         | 29,145                          | 29,074                          | 20,889                          | 20,892                          |
| 2R         | 28,769                          | 28,621                          | 21,342                          | 21,269                          |
| 3L         | 29,092                          | 29,287                          | 20,813                          | 20,807                          |
| 3R         | 28,719                          | 28,693                          | 21,320                          | 21,269                          |
| 4          | 31,945                          | 32,805                          | 17,471                          | 17,779                          |
| Y          | 30,992                          | 29,583                          | 20,023                          | 19,403                          |

Fig. 23. The table of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ in percentage for each of 7 chromosomes of *Drosophila melanogaster* ($n = 1, 2, 3, 4, 5; k \leq n$).

![Graph](image)

Fig. 24. The graphic representation of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ in percentage for each of 7 chromosomes of *Drosophila melanogaster*. The abscissa axis contains numberings of chromosomes, and the ordinate axis contains average values of these probabilities in percent. The symbol of a blue diamond corresponds the average values of $P_n(A_k) \approx P_n(T_k)$, and the symbol of a red square corresponds the average values of $P_n(C_k) \approx P_n(G_k)$.

The third model organism is a plant *Arabidopsis thaliana*. This small flowering plant is used for over fifty years to study plant mutations and for classical genetic
analysis. It became the first plant genome to be fully sequenced. Tables of tetragroup symmetries of its set of 5 chromosomes are represented in the Appendix 5. Fig. 25 and 26 shows the table and the corresponding graphical representation of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ for all chromosomes of Arabidopsis thaliana. These average values were calculated from the tables in Fig. A5/1-A5/5 by analogy with the described calculation of average values of the probabilities in cases of human chromosomes (Fig. 18).

| CHROMOSOME | AVERAGE VALUE OF $P_n(A_k)$ (%) | AVERAGE VALUE OF $P_n(T_k)$ (%) | AVERAGE VALUE OF $P_n(C_k)$ (%) | AVERAGE VALUE OF $P_n(G_k)$ (%) |
|------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| 1          | 32,083                        | 32,043                        | 17,960                        | 17,913                        |
| 2          | 32,066                        | 32,070                        | 17,989                        | 17,876                        |
| 3          | 31,913                        | 31,756                        | 18,156                        | 18,175                        |
| 4          | 31,969                        | 31,827                        | 18,144                        | 18,060                        |
| 5          | 31,976                        | 32,088                        | 17,920                        | 18,018                        |

Fig. 25. The table of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ in percentage for each of 5 chromosomes of Arabidopsis thaliana (n = 1, 2, 3, 4, 5; k ≤ n).

Fig. 26. The graphic representation of average values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ in percentage for each of 7 chromosomes of Arabidopsis thaliana. The abscissa axis contains numberings of chromosomes, and the ordinate axis contains average values of these probabilities in percent. The symbol of a blue diamond corresponds the average values of $P_n(A_k) \approx P_n(T_k)$, and the symbol of a red square corresponds the average values of $P_n(C_k) \approx P_n(G_k)$.

One more analysed organism is the laboratory mouse, which is a major model organism for basic mammalian biology, human disease, and genome evolution. Data about the fulfillment of the described rules of tetra-group symmetries in the complete set of chromosomes of Mus musculus (house mouse) are represented in the Appendix 6.
It seems interesting in the future to study the characteristics of tetra-group symmetries for karyotypes of a variety of different organisms; these characteristics (approximate constants of karyotypes?) can be useful for comparative analysis of organisms, for example, in problems of biological evolution.

11. Fractal genetic nets and the fifth and sixth rules of symmetries of collective probabilities of tetra-groups. On a fractal grammar of long DNA-texts.

The tetra-group symmetries are connected with the fractal-like principle: in n-letter representations of a long DNA-text, the individual probabilities of tetra-group subgroups of n-plets of a lower order are repeated in the collective probabilities of the corresponding subgroups of n-plets of higher orders.

Another connection of long DNA-texts with fractals was described in the article [Petoukhov, Svirin, 2012] where the notion of “fractal genetic nets” (FGN) has been introduced. Each FGN of texts can contain different fractal genetic trees (FGT). In that article we have shown hidden regularities in long DNA-texts in a connection with Chargaff’s thoughts about a “grammar of biology”: our results testified 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. Some new symmetry principles dealing with FGT and FGN have been formulated there.

Below we represent our results that show the implementation of the tetra-group symmetries in convoluted long DNA-texts at different levels of different kinds of fractal genetic nets (or trees). But initially let us remind about FGN.

In line with our article [Petoukhov, Svirin, 2012], 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. 27 explains a construction of FGT of various types by an example of the FGT for a long DNA-text, which is represented as a sequence S_0 of 3-letter words (the 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. This tree is 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 initial long DNA-text $S_0$.

![Diagram of the fractal genetic tree (FGN-3) for a long DNA-text](image)

Fig. 27. The scheme of a fractal genetic tree (FGN-3) of a long DNA-text, which is represented as a sequence of triplets (from [Petoukhov, Svirin, 2012]).

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 tree on its basis (Fig. 4). In general case, the FGT can be built not only for DNA-texts of triplets, but also for DNA-texts of other n-plets ($n = 2, 4, 5, ...$) by means of the repeated positional convolution of each of texts from the previous level into "n" daughter texts 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 of the initial and daughter texts; FGN is a set of such separate trees.

It is obvious that in the general case of FGT- $n$ (where $n = 2, 3, 4, 5,...$), when we move to its next convolution level, the number of texts increases by a factor of n-fold and their lengths are shortened by a factor of $n$. A FGT- $n$ has at its first level of convolution $n$ daughter texts; at its second level of convolution $– n^2$ daughter texts; and at p-th level $– n^p$ daughter texts. The total quantity $N$ of texts in a FGT- $n$, which has p levels of convolution ($p = 1, 2, 3, 4, ...$), is equal to the sum of daughter texts at all p levels:

$$N = 1+n+n^2+n^3+...+n^p$$  \hspace{2cm} (37)

Each of daughter texts at different levels of a FGT- $n$ is very individual and differs from other texts in this FGT- $n$ in the general case.

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 fractal genetic trees FGT-2, FGT-3, FGT-4 for enough long DNA-texts of different organisms show the implementation of these tetra-
group rules in all daughter texts at initial levels of the FGT-2, FGT-3 and FGT-4. Moreover, for each of tested long DNA-tests, 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 daughter texts at different initial levels of these tested cases of FGT-n. In other words, for each daughter text at each level of these fractal genetic trees, values of each of these kinds of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ (where $n = 1, 2, 3, 4, 5; k \leq n$) lie in a narrow numerical interval (it is called an interval of fluctuations of the probability values at an appropriate level of a FGT-n). Fig. 28 and 29 illustrate this phenomenologic fact in a compressive form of tables showing these fluctuation intervals of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ for daughter texts at each of two first levels of the FGT-2 and FGT-3 in cases of human sex chromosomes X and Y (their DNA-texts contain 15604095 and 57227415 letters correspondingly). Let us remind that - in line with the expression (37) – the quantity of different texts at the considered levels (from the level 0 till the level 2) is equal to $1+2+2^2 = 7$ in the case of FGT-2 and $1+3+3^2 = 13$ in the case of FGT-3.

|          | Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|----------|---------|-----------|-----------|------------|------------|------------|------------|
| $P_n(A_k)$ | 0.3017+/0.3019 | 0.3017+/0.3019 | 0.3016+/0.3020 | 0.3016+/0.3017 | 0.3017+/0.3019 | 0.3015+/0.3019 | 0.3015+/0.3019 |
| $P_n(T_k)$ | 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 |
| $P_n(C_k)$ | 0.197+/0.1971 | 0.1969+/0.1971 | 0.1969+/0.197 | 0.1967+/0.1971 | 0.1969+/0.1971 | 0.1968+/0.1971 | 0.1971+/0.1971 |
| $P_n(G_k)$ | 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 2/00 | Level 2/01 | Level 2/10 | Level 2/01 |
|----------|---------|-----------|-----------|------------|------------|------------|------------|
| $P_n(A_k)$ | 0.3017+/0.3019 | 0.3016+/0.3020 | 0.3017+/0.3017 | 0.3015+/0.3019 | 0.3016+/0.3018 | 0.3018+/0.3019 | 0.3019+
| $P_n(T_k)$ | 0.3027+/0.3029 | 0.3027+/0.3027 | 0.3027+/0.3027 | 0.3024+/0.3029 | 0.3024+/0.3029 | 0.3025+/0.3028 | 0.3027+/0.3027 |
| $P_n(C_k)$ | 0.197+/0.1971 | 0.197+/0.197 | 0.197+/0.1974 | 0.1967+/0.1971 | 0.1969+/0.1971 | 0.1967+/0.1968 | 0.197+/0.1968 |
| $P_n(G_k)$ | 0.198+/0.1982 | 0.198+/0.1981 | 0.198+/0.1980 | 0.198+/0.1982 | 0.198+/0.1981 | 0.198+/0.1981 | 0.1979+/0.1985 |

|          | Level 0 | Level 1/0 | Level 1/1 | Level 2/0 | Level 2/1 | Level 2/2 | Level 2/2 |
|----------|---------|-----------|-----------|----------|----------|----------|----------|
| $P_n(A_k)$ | 0.3015+/0.3019 | 0.3018+/0.3019 | 0.3015+/0.3019 | 0.3014+/0.3018 | 0.3016+/0.3019 | 0.3016+/0.3018 | 0.3015+/0.3018 |
| $P_n(T_k)$ | 0.3025+/0.3026 | 0.3025+/0.3025 | 0.3025+/0.3030 | 0.3025+/0.3030 | 0.3025+/0.3030 | 0.3027+/0.3030 | 0.3027+/0.3030 |
| $P_n(C_k)$ | 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 |
| $P_n(G_k)$ | 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. 28. 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 FGT-3 (bottom tables) in the case of the human sex 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 |
|---------|-----------|-----------|------------|------------|------------|------------|
| $P_n(A_k)$ | 0.2983+ 0.2987 | 0.2982+ 0.2986 | 0.2979+ 0.2987 | 0.2980+ 0.2989 | 0.2981+ 0.2987 | 0.2981+ 0.2976+ |
| $P_n(T_k)$ | 0.3009+ 0.3012 | 0.3008+ 0.3021 | 0.3007+ 0.3014 | 0.3009+ 0.3013 | 0.3005+ 0.3020 | 0.3005+ 0.3007+ |
| $P_n(C_k)$ | 0.1998+ 0.2003 | 0.1994+ 0.2001 | 0.1995+ 0.2004 | 0.1995+ 0.2002 | 0.1994+ 0.2002 | 0.1995+ 0.2003 |
| $P_n(G_k)$ | 0.1998+ 0.2003 | 0.1996+ 0.2001 | 0.1996+ 0.2001 | 0.1996+ 0.2001 | 0.1995+ 0.2001 | 0.1995+ 0.2001 |

| Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|------------|------------|------------|------------|------------|------------|------------|
| $P_n(A_k)$ | 0.2974+ 0.2982 | 0.2978+ 0.2991 | 0.2979+ 0.2997 | 0.2983+ 0.2985 | 0.2975+ 0.2976 | 0.2974+ 0.2979 |
| $P_n(T_k)$ | 0.3005+ 0.3023 | 0.3007+ 0.3012 | 0.3006+ 0.3008 | 0.3002+ 0.3010 | 0.3005+ 0.3023 | 0.3005+ 0.3006 |
| $P_n(C_k)$ | 0.1994+ 0.2005 | 0.1994+ 0.2004 | 0.1995+ 0.2004 | 0.1996+ 0.2007 | 0.1993+ 0.2003 | 0.1997+ 0.2002 |
| $P_n(G_k)$ | 0.1990+ 0.1990 | 0.1990+ 0.2008 | 0.1988+ 0.2007 | 0.1991+ 0.1996 | 0.1997+ 0.1998 | 0.1996+ 0.2013 |

Fig. 29. 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 sex chromosome Y (NCBI Reference Sequence: NC_000024.10).

The same is true not only for each of human sex chromosomes X and Y (Fig. 28 and 29) but also for each of 22 human autosomes. Moreover, the same is true for long DNA-texts of all those model organisms and long DNA-texts, which are represented in the Appendixes 3-6.

These our results allow putting forward the hypothesis about the validity of the fifth rule of tetra-group symmetries, which concerns fractal genetic trees of long DNA-texts (including chromosomes) of different organisms:

- For each of fractal genetic trees FGT-n (where n = 2, 3, 4, 5,... is not too large) of a long DNA-text, each of its daughter texts at different levels p of the tree (where p = 1, 2, 3, ... is not too large) has approximately those
values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ like the initial DNA-text.

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

From Fig. 28 and 29 one can see also that fluctuation intervals of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ are approximately equal each to other for all 20 ($=7+13$) texts at all considered levels of the fractal genetic net, which consists of the FGT-2 and FGT-3. It is true not only for each of human sex chromosomes X and Y but also for each of 22 human autosomes. Moreover, the same is true for long DNA-texts of all those model organisms and long DNA-texts, which are represented in the Appendixes 3-6.

These our results allow putting forward the hypothesis about the validity of the sixth rule of tetra-group symmetries, which concerns fractal genetic nets of long DNA-texts (including chromosomes) of different organisms:

- For each of a long DNA-text, in each of its fractal genetic nets (FGN), which consists of a few genetic fractal trees FGT-2, FGT-3, ..., FGT-n (where $n = 2, 3, 4, 5, ...$ is not too large), each of its daughter texts at different levels $p$ of the trees (where $p = 1, 2, 3, ...$ is not too large) has approximately those values of probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ like the initial DNA-text.

Our results for long DNA-texts (about the implementation of the tetra-group symmetries in long convoluted texts at different levels of their FGT-n) testify in favor of existence of a fractal grammar of genetics in line with the Chargaff’s words about a grammar of biology [Chargaff, 1971]. 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](https://en.wikipedia.org/wiki/Fractal_compression)). It is obvious that an opportunity of information 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 [Jeffrey, 1990; Lieberman-Aiden et al., 2009; Pellionisz, 2008; Pellionisz et al., 2012; Pellionisz, Ramanujam, Rajan, 2017; Peng et al., 1992; Perez, 2010]. For example, the work [Lieberman-Aiden et al., 2009] shown an existence of fractal globe in the three dimensional architecture of whole genomes, where spatial chromosome territories exist and where maximally 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 [Lieberman-Aiden et al., 2009], in our work we study not the 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 trees and their nets of different kinds (FGT-n, where \( n = 2, 3, 4, \ldots \)), which are connected with tetra-group symmetries of these texts (these symmetries and fractals were never studied early);

- Fractals are actively used in study of cancer; some modern data testify that cancer processes are related with fractal patterns and their development [Baish, Jain, 2000; Bizzarri et al., 2011; Dokukin et al., 2015; Lennon et al., 2015; Pellionisz, Ramanujam, Rajan, 2017; Perez, 2017].
- 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.

Fig. 27 has shown above the example of a creation of daughter texts at different levels of FGN-n from a long DNA-text by the method of sequental positional convolutions. But one can consider the inverse method, that is the method of the sequential positional assembly those texts, which belong to lower levels of FGN-n, into the extended texts, which belong to its higher levels up to the highest level \( S_0 \). If at a lower level (or at some lower levels) of FGN-n, its texts are permuted in their order, then the method of the sequential positional assembly from these texts leads to new extended texts at higher levels of this fractal tree. Such sequential positional assembly of permutated texts of lower levels of FGN-n can be used in biological evolution of DNA-texts of different organisms. As known, some biological organisms differ each from other by permutations of fragments of their DNA-texts. For example, mouse and human genomes can be viewed as a collection of about 200 fragments which are shuffled (rearranged) in mice as compared to humans; the chromosome 2 in humans is built from fragments that are similar to fragments from mouse DNA residing on chromosomes 1, 2, 6, 8, 11, 12, and 17 [Pevzner, 2000, Fig. 1.4 and 1.5, pp. 15, 16]. Genome rearrangements are a rather common chromosomal abnormality, which are associated with such genetic diseases as Down syndrome.

12. About letter-ordered representations of long DNA-texts conserving their collective probabilities in tetra-groups

The fifth and sixth rules of tetra-group symmetries, which are described in the previous Section about fractal genetic trees, are fulfilled at all not for all long texts of four letters C, G, A and T. To confirm this, let us consider one of many examples of long texts, which don’t satisfy the named rules. For example, take a long text, the first part of which contains only 100000 doublets CG and the second part contains only 100000 doublets AT: CG-CG-CG-...-CG-AT-AT-AT'-...-AT. This text satisfies the second Chargaff’s parity rule (%C=%G and %A=%T). But at the first level of the positional convolution of this text in its dichotomic fractal genetic tree (FGT-2), two texts are generated: the first of which contains only 50000 letters C and 50000 letters A, and the second of which contains only 50000 letters G and 50000 letters T. It is obvious that each of these two texts
doesn’t satisfy the second Chargaff’s rule and other rules of tetra-group symmetries since \( %C \neq %G \) and \( %A \neq %T \), etc.

Is there a variant of ordering rearrangement (shuffling) of all the letters of any long DNA-text, in which the characteristics of tetra-group symmetries of the text do not change? Let’s demonstrate a possible variant of an ordered representation of any long DNA-text, which we call the “letter-ordered representation” (LO-representation). Such representation can be used as a convinient “canonical” form of representations of any long DNA-texts for some tasks of comparison analysis of such texts from the standpoint of their tetra-group symmetries.

Any long DNA-text, which contains \( N_c \) letters C, \( N_G \) letters G, \( N_A \) letters A, \( N_T \) letters T, can be formally represented by a text of the same length with an ordered sequence that contains four parts: its first part consists of \( N_c \) letters C; its second part consists of \( N_G \) letters G; its third part consists of \( N_A \) letters A; its fourth part consists of \( N_T \) letters T. Such letter-ordered sequence can be conditionally denoted \( N_cC-N_GG-N_AA-N_TT \).

We have revealed that such letter-ordered representation of long DNA-texts possesses the same symmetric characteristics of tetra-group probabilities \( P_n(C_k), P_n(G_k), P_n(A_k) \) and \( P_n(T_k) \) as the original DNA-text (here \( n=1,2,3,4,... \) is not too large, \( k \leq n \)).

With a letter-ordered representation of a long DNA-text, a 4-dimensional metric space with a Cartesian coordinate system can be used, along the axes of which frequencies (numbers) of each of the four letters are plotted. In this case, different DNA-texts will be represented by different points of the given space, and distances between them can be calculated for a comparison analysis. (Another 4-dimensional metric space can be used for probabilities \( P_n(C_k), P_n(G_k), P_n(A_k) \) and \( P_n(T_k) \)). Letter-ordered representations of long DNA-texts are also convenient as beginning forms to study those permutations of fragments of long DNA-texts, which conserve characteristics of tetra-group symmetries of the texts.

13. On the biological sense of the symmetries of collective probabilities of tetra-groups in long DNA texts

The world of molecules, including DNA molecules, is subordinate to the principles of quantum mechanics. DNA molecules are carriers of quantum information that modern science analyzes and models on the basis of concepts and methods of quantum information and quantum computing: multi-qubit systems, pure and mixed quantum states, entangled states, edensity matrices, etc. The discovery of the symmetries of the collective probabilities of tetra-groups of oligomers in long DNA texts allowed us to propose a new model approach in which DNA alphabets and long DNA texts are represented by the academic formalisms of quantum informatics in the form of quantum-information multi-qubit systems. From the standpoint of this model approach, the existence of symmetries of these collective probabilities is a quantum-informational phenomenon and part of the general biological quantum-information system for ensuring the transfer of hereditary information in living organisms.

What is the biological sense of the quantum-information phenomenon of the symmetry of the collective probabilities of tetra groups in long DNA texts? A
common property of all living organisms, consisting of quantum-mechanical molecular subsystems, is their ability to grow and develop on the basis of incorporation into their body of new and new molecules of nutrients becoming new parts of the whole organism. In the course of evolution, many species have acquired an even greater ability to unite parts to create single colonial and multicellular organisms. This means that biological evolution endowed organisms with a fundamental way of creating single quantum states of multicomponent quantum systems from the quantum states of their subsystems. But an analogous method in quantum mechanics and quantum informatics is built on the general principles of the mathematical operation of tensor multiplication of vectors and matrices. Figuratively speaking, a whole multicomponent organism is - to a certain extent - a product of tensor multiplication of the quantum states of its subsystems, including its molecular genetic systems (our work [Petoukhov, 2011] shows a connection of the tensor product of matrices with Punnett squares known in Mendelian genetics from 1905 year and included in text-books of genetics about Mendelian crosses of organisms). It is not for nothing that the symmetries of collective probabilities revealed in long DNA texts allow their modeling in the language of separable pure states of multi-qubit quantum systems that unite the quantum states of their subsystems. Our data testify that the biological meaning of symmetries of the collective probabilities of tetra groups in quantum-information long DNA-texts is related to the fundamental problem of the evolutionary quantum-information unification of the set of quantum subsystems into a single organism based on the principles of quantum mechanics. The genetic system, being an integral part of this quantum-information association, ensures the transfer of quantum information to descendants with the possibility of complicating the quantum-genetic messages in the process of biological evolution of organisms. Genetic symmetries of collective probabilities of tetra-groups in long quantum-information DNA texts designate the framework, inside which the evolutionary complication of quantum-genetic messages takes place. From the model standpoint, these symmetries of collective probabilities are a consequence of those quantum-information principles of creating the united states of quantum mechanical systems, which are modeled with the help of tensor products. To this we add that the reverse phenomenon - the appearance of independent "daughter" organisms or individual parts from integral biological objects - can be modeled as a result of the tensor factorization (decomposition) of the quantum states of an integral system upon its disintegration into two or more constituent parts.

The author hopes that the further usage in genetics the concepts and formalisms of quantum informatics, which was undertaken in this article in connection with symmetries of collective probabilities of tetra-groups, will lead to the development of substantial quantum-information genetics. This will promote the inclusion of genetics and all biology in the field of profound mathematical natural science. Consideration of biological phenomena, including the phenomena of inheritance of the intellectual abilities of biological bodies, from the standpoint of the theory of quantum computers, gives many valuable opportunities for their comprehension and also for development of artificial intelligence systems [Petoukhov et al., 2017]. For example, an adult human organism has around 10 trillion \((10^{14})\) human cells and each of cells contains an
identical compлект of DNA, genetic information from which is used for physiological functioning organism as the holistic system of cells. How such huge number of cells can reliably functioning as a cooperative whole? Quantum informatics and associations with quantum computers can help to model and understand such holistic biological systems with their ability to compute complex tasks and to transfer genetic information from one generation to another. The fundamental question - whether the quantum device can solve some tasks more effectively than a classical computer - was first touched upon in the book [Manin, 1980].

14. Some concluding remarks

The modern situation in the theoretic field of genetic informatics, where many millions of nucleotide sequences are described, can be characterized by the following citation: "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. We have the program that runs the cellular machinery, but we know very little about how to read it." [Fickett & Burks, 1989]. Kepler did not make his own astronomic observations, but he found - in the huge astronomic data of Tycho Brahe - his Kepler's laws of symmetric movements of planets relative to the Sun along ellipses. What are the hidden symmetries and rules in long DNA texts, in which the living nature for some reason records hereditary information about the most diverse organisms?

The Chargaff's parity rules played an important role in development of bioinformatics. We hope that the represented class of rules of tetra-group symmetries of collective probabilities in long texts of single stranded DNA will be also useful for further development of bioinformatics and theoretical biology.

The study of these tetra-group rules is now continued with including, in particular, relations of the tetra-groups of triplets with groups of amino acids and stop-codons, which are considered in many publications (see for example, [Dragovich, Dragovich, 2007; Dragovich, Khrennikov, Misic, 2017]).

What is the degree of universality of these genetic rules of collective probabilities? Are these rules (or some of them) true for all long DNA-sequences? Or are there any exceptions or limitations to them? Only further research can provide answers to these questions.

In our opinion, characteristics of tetra-group symmetries of collective probabilities in long DNA-sequences and in sets of chromosomes of different organisms can be useful as new criteria of comparative analysis for taxonomic classification of biological organisms, for problems of evolutionary biology, etc.

The article [Trifonov, 1993] noted that the classical triplet code is not the only code carried by the sequences. DNA is a molecular language consisting of letters, words, and texts, carrying overlapping messages embodying multiple codes (as in our broadband communication channels of the Internet; on this, see additionally [Manin, Manin, 2017]). «The nucleotide sequences are written in an unbroken manner. One way to detect “words” in such a continuous text is to
evaluate the degree of internal correlation by calculating contrast values for the words. This technique allows one to derive vocabularies, which are species- and function-specific. The nucleotide sequences, thus, carry numerous superimposed messages. We do understand only a few of these messages while many more are waiting for their turn to be deciphered» [Trifonov, 1993]. Our discover of symmetries of collective probabilities of tetra-groups of oligonucleotides in long DNA-texts gives new materials to this topic about existence of many genetic codes of quantum-information types, about the geno-logic coding [Petoukhov, 2017a,b; Petoukhov, Petukhova, 2017a] and about the fractal grammar of DNA.

Appendix 1. Additional data about tetra-group rules and tetra-group symmetries in long DNA-texts

The Appendix 1 represents data about the fulfillment of the described tetra-group rules in a representative set of long sequences of oligonucleotides of single stranded DNA from GenBank. Kinds of studied sequences here are taken from articles of other authors to avoid a suspicion about a special choice of sequences. The results, presented below, confirm the fulfillment of the tetra-group rules and the existence of tetra-group symmetries in all the cases considered.

First of all, we represent results of analysis of tetra-group probabilities of 19 sequences from their list in the article about the second Chargaff’s parity rule [Rapoport, Trifonov, 2012, p. 2]: 

“Nucleotide disparities for prokaryotic coding sequences were taken from bacterial genomes of different groups both from Bacteria and Archea. All together 19 genomes were used: Aquifex aeolicus, Acidobacteria bacterium, Bradyrhizobium japonicum, Bacillus subtilis, Chlamydia trachomatis, Chromobacterium violaceum, Dehalococcoides etenogenes, Escherichia coli, Flavobacterium psychrophilum, Gloeobacter violaceus, Helicobacter pilory, Methanosarcina acetivorans, Nanoarchaeum equitans, Syntrophus aciditrophicus, Streptomyces coelicolor, Sulfolobus solfataricus, Treponema denticola, Thermotoga maritima and Thermus thermophiles”. Fig. A1/1-A1/19 show results of analysis of tetra-group probabilities in these 19 DNA-sequences.

Then, Fig. A1/20-A1/35 show results of analysis of tetra-group symmetries in DNA-sequences, the length of which is which is approximately equal to 100000 nucleotides and more, from the article [Prahbu, 1993] devoted also to the second Chargaff’s rule.
### Table: Nucleotides, Doublets, Triplets, 4-Plets, 5-Plets

| Nucleotides | Doublets | Triplets | 4-Plets | 5-Plets |
|-------------|----------|----------|---------|---------|
| P₁(A₁) = 0.2841 | P₂(A₁) = 0.2834 | P₃(A₁) = 0.2820 | P₄(A₁) = 0.2834 | P₅(A₁) = 0.2843 |
| P₁(T₁) = 0.2811 | P₂(T₁) = 0.2818 | P₃(T₁) = 0.2815 | P₄(T₁) = 0.2820 | P₅(T₁) = 0.2821 |
| P₁(C₁) = 0.2168 | P₂(C₁) = 0.2173 | P₃(C₁) = 0.2167 | P₄(C₁) = 0.2170 | P₅(C₁) = 0.2160 |
| P₁(G₁) = 0.2179 | P₂(G₁) = 0.2176 | P₃(G₁) = 0.2198 | P₄(G₁) = 0.2174 | P₅(G₁) = 0.2178 |

### References

Fig. A1/1. Probabilities of subgroups of tetra-groups in the sequence:
Aquifex aeolicus VF5, complete genome
[https://www.ncbi.nlm.nih.gov/nuccore/AE000657.1?report=genbank](https://www.ncbi.nlm.nih.gov/nuccore/AE000657.1?report=genbank);
[https://www.ncbi.nlm.nih.gov/nuccore/AE000657.1?report=fasta](https://www.ncbi.nlm.nih.gov/nuccore/AE000657.1?report=fasta);
LOCUS AE000657, 1551335 bp DNA circular BCT 30-JAN-2014 DEFINITION
Aquifex aeolicus VF5, complete genome. ACCESSION AE000657 AE000669-AE000777 VERSION AE000657.1
**Fig. A1/2.** Probabilities of subgroups of tetra-groups in the sequence: Acidobacteria bacterium KBS 146 M015DRAFT_scf71800000000004_quiver.1_C, whole genome shotgun sequence, GenBank: JHVA01000031.1, -
https://www.ncbi.nlm.nih.gov/nuccore/JHVA01000031.1?report=fasta

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)= 0.1819 | P(T)= 0.1820 | P(A)= 0.1806 | P(G)= 0.1816 | P(A)= 0.1814 |
| P(C)= 0.1815 | P(G)= 0.1814 | P(A)= 0.1797 | P(T)= 0.1817 | P(C)= 0.1818 |
| P(G)= 0.3184 | P(A)= 0.3185 | P(T)= 0.3197 | P(C)= 0.3184 | P(G)= 0.3182 |
| P(A)= 0.3182 | P(T)= 0.3182 | P(G)= 0.3201 | P(G)= 0.3183 | P(G)= 0.3186 |
| P(G)= 0.1818 | P(A)= 0.1818 | P(A)= 0.1816 | P(A)= 0.1816 | P(A)= 0.1819 |
| P(T)= 0.1816 | P(A)= 0.1825 | P(T)= 0.1815 | P(T)= 0.1815 | P(T)= 0.1815 |
| P(C)= 0.3183 | P(T)= 0.3183 | P(C)= 0.3184 | P(C)= 0.3184 | P(C)= 0.3186 |
| P(G)= 0.3182 | P(C)= 0.3173 | P(C)= 0.3185 | P(G)= 0.3179 |
| P(A)= 0.1815 | P(C)= 0.1823 | P(A)= 0.1820 | P(A)= 0.1820 |
| P(A)= 0.1824 | P(T)= 0.1811 | P(T)= 0.1815 | P(A)= 0.1817 |
| P(C)= 0.3189 | P(C)= 0.3185 | P(C)= 0.3186 |
| P(G)= 0.3173 | P(G)= 0.3181 | P(G)= 0.3180 |
| P(A)= 0.1820 | P(A)= 0.1818 | P(T)= 0.1815 |
| P(C)= 0.3182 | P(C)= 0.3183 |
| P(G)= 0.3180 | P(G)= 0.3185 |

**Fig. A1/3.** Probabilities of subgroups of tetra-groups in the sequence: Bradyrhizobium japonicum strain E109, complete genome
GenBank: CP010313.1,
https://www.ncbi.nlm.nih.gov/nuccore/CP010313.1?report=fasta

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)= 0.2149 | P(T)= 0.21694 | P(A)= 0.2859 | P(G)= 0.2823 |
| P(C)= 0.2149 | P(G)= 0.21694 | P(G)= 0.2859 | P(G)= 0.2823 |
| P(T)= 0.2149 | P(A)= 0.2859 | P(G)= 0.2823 | P(G)= 0.2823 |
| P(C)= 0.2149 | P(G)= 0.2823 | P(G)= 0.2823 | P(G)= 0.2823 |
| nucleotides | doublets | triplets | 4-plets | 5-plets |
|-------------|----------|----------|----------|----------|
| P(A)=0.2805 | P(A)=0.2806 | P(A)=0.2807 | P(A)=0.2808 | P(A)=0.2809 |
| P(T)=0.2806 | P(T)=0.2806 | P(T)=0.2807 | P(T)=0.2808 | P(T)=0.2809 |
| P(C)=0.2192 | P(C)=0.2196 | P(C)=0.2195 | P(C)=0.2196 | P(C)=0.2196 |
| P(G)=0.2197 | P(G)=0.2197 | P(G)=0.2197 | P(G)=0.2197 | P(G)=0.2197 |

**Fig. A1/4.** Probabilities of subgroups of tetra-groups in the sequence:

Bacillus subtilis strain UD1022, complete genome, GenBank: CP011534.1, [https://www.ncbi.nlm.nih.gov/nuccore/CP011534.1?report=fasta](https://www.ncbi.nlm.nih.gov/nuccore/CP011534.1?report=fasta) LOCUS CP011534, 4025326 bp, DNA circular BCT 18-FEB-2016 DEFINITION Bacillus subtilis strain UD1022, complete genome. ACCESSION CP011534

| nucleotides | doublets | triplets | 4-plets | 5-plets |
|-------------|----------|----------|----------|----------|
| P(A)=0.2942 | P(A)=0.2939 | P(A)=0.2939 | P(A)=0.2939 | P(A)=0.2939 |
| P(T)=0.2930 | P(T)=0.2931 | P(T)=0.2933 | P(T)=0.2932 | P(T)=0.2932 |
| P(C)=0.2067 | P(C)=0.2066 | P(C)=0.2059 | P(C)=0.2066 | P(C)=0.2066 |
| P(G)=0.2061 | P(G)=0.2063 | P(G)=0.2069 | P(G)=0.2068 | P(G)=0.2068 |

**Fig. A1/5.** Probabilities of subgroups of tetra-groups in the sequence:
Chlamydia trachomatis strain QH111L, complete genome
GenBank: CP018052.1,
https://www.ncbi.nlm.nih.gov/nuccore/CP018052.1?report=genbank
LOCUS CP018052, 1025839 bp, DNA circular BCT 22-NOV-2016
DEFINITION Chlamydia trachomatis strain QH111L, complete genome.
ACCESSION CP018052 VERSION CP018052.1

| NUCLEOTIDES | DOUBLES | TRIPLETS | 4-PLETS | 5-PLETS |
|--------------|---------|----------|---------|---------|
| P_1(A_1)    | 0.1670  | 0.1658   | 0.1758  | 0.1640  | 0.1678  |
| P_1(T_1)    | 0.1819  | 0.1812   | 0.1883  | 0.1809  | 0.1823  |
| P_1(C_1)    | 0.3049  | 0.3044   | 0.3009  | 0.3042  | 0.3055  |
| P_1(G_1)    | 0.3461  | 0.3486   | 0.3350  | 0.3509  | 0.3444  |

Chromobacterium violaceum strain LK30 1, whole genome shotgun sequence
GenBank: LDUX0100000.1,
https://www.ncbi.nlm.nih.gov/nuccore/LDUX0100000.1?report=genbank
LOCUS LDUX01000001, 127377 bp, DNA linear BCT 30-JUN-2015
DEFINITION Chromobacterium violaceum strain LK30 1, whole genome shotgun sequence. ACCESSION LDUX01000001 LDUX01000000 VERSION LDUX01000001.1

| NUCLEOTIDES | DOUBLES | TRIPLETS | 4-PLETS | 5-PLETS |
|--------------|---------|----------|---------|---------|
| P_1(A_1)    | 0.2664  | 0.2665   | 0.2706  | 0.2668  | 0.2673  |
| P_1(T_1)    | 0.2645  | 0.2647   | 0.2653  | 0.2649  | 0.2638  |
| P_1(C_1)    | 0.2338  | 0.2337   | 0.2295  | 0.2337  | 0.2340  |
| P_1(G_1)    | 0.2353  | 0.2351   | 0.2346  | 0.2346  | 0.2349  |

Fig. A1/6. Probabilities of subgroups of tetra-groups in the sequence:
Fig. A1/7. Probabilities of subgroups of tetra-groups in the sequence:

Dehalococcoides mccartyi strain CG3, complete genome
NCBI Reference Sequence: NZ_CP013074.1,
https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP013074.1?report=fasta
https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP013074.1?report=genbank:
LOCUS NZ_CP013074 1521287 bp DNA circular CON 06-APR-2017
DEFINITION Dehalococcoides mccartyi strain CG3, complete genome.
ACCESSION NZ_CP013074 VERSION NZ_CP013074.1

| NUCLEOTIDES | DOUBLETs | TRIPLETs | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)= 0.2480 | P(A)= 0.2481 | P(A)= 0.2479 | P(A)= 0.2479 | P(A)= 0.2483 |
| P(T)= 0.2472 | P(T)= 0.2470 | P(T)= 0.2489 | P(T)= 0.2472 | P(T)= 0.2472 |
| P(C)= 0.2526 | P(C)= 0.2523 | P(C)= 0.2519 | P(C)= 0.2522 | P(C)= 0.2529 |
| P(G)= 0.2522 | P(G)= 0.2526 | P(G)= 0.2514 | P(G)= 0.2526 | P(G)= 0.2516 |
| P(A)= 0.2480 | P(A)= 0.2480 | P(A)= 0.2479 | P(A)= 0.2479 | P(A)= 0.2483 |
| P(T)= 0.2474 | P(T)= 0.2461 | P(T)= 0.2477 | P(T)= 0.2475 | P(T)= 0.2475 |
| P(C)= 0.2528 | P(C)= 0.2544 | P(C)= 0.2530 | P(C)= 0.2522 | P(C)= 0.2522 |
| P(G)= 0.2518 | P(G)= 0.2527 | P(G)= 0.2519 | P(G)= 0.2523 | P(G)= 0.2523 |
| P(A)= 0.2494 | P(A)= 0.2483 | P(A)= 0.2483 | P(A)= 0.2485 | P(A)= 0.2485 |
| P(T)= 0.2466 | P(T)= 0.2467 | P(T)= 0.2470 | P(T)= 0.2470 | P(T)= 0.2470 |
| P(C)= 0.2515 | P(C)= 0.2525 | P(C)= 0.2527 | P(C)= 0.2527 | P(C)= 0.2527 |
| P(G)= 0.2525 | P(G)= 0.2526 | P(G)= 0.2526 | P(G)= 0.2526 | P(G)= 0.2526 |

Fig. A1/8. Probabilities of subgroups of tetra-groups in the sequence:

Escherichia coli CFT073, complete genome, GenBank: AE014075.1,
https://www.ncbi.nlm.nih.gov/nuccore/AE014075.1?report=fasta
https://www.ncbi.nlm.nih.gov/nuccore/AE014075.1?report=genbank:
LOCUS AE014075 5231428 bp DNA circular BCT 31-DEC-2014
DEFINITION Escherichia coli CFT073, complete genome. ACCESSION
AE014075 AE016755 AE016756 AE016757 AE016758 AE016759 AE016760
### Nucleotides

**Definitory: Flavobacterium psychrophilum,**

**Locus:** NC_009613, 2860382 bp, DNA circular CON 03-AUG-2016

**Accession:** NC_009613 VERSION NC_009613.3

| Nucleotides | Doublets | Triplets | 4-Plets | 5-Plets |
|-------------|----------|----------|---------|---------|
| P(A1) = 0.3326 | P(T1) = 0.3420 | P(C1) = 0.1640 | P(G1) = 0.1614 | P(A2) = 0.3326 |
| P(A1) = 0.3326 | P(A2) = 0.3416 | P(C2) = 0.1643 | P(T2) = 0.1615 | P(A1) = 0.3326 |
| P(T1) = 0.3420 | P(T1) = 0.3416 | P(C1) = 0.1640 | P(T2) = 0.1615 | P(T1) = 0.3420 |
| P(C1) = 0.1640 | P(A2) = 0.3406 | P(C1) = 0.1638 | P(C2) = 0.1638 | P(C1) = 0.3406 |

#### Doublets

- **Fig. A1/9.** Probabilities of subgroups of tetra-groups in the sequence: Flavobacterium psychrophilum. 
  
https://www.ncbi.nlm.nih.gov/nuccore/NC_009613.3:

**Locus:** NC_009613, 2860382 bp, DNA circular CON 03-AUG-2016

**Accession:** NC_009613 VERSION NC_009613.3

| Nucleotides | Doublets | Triplets | 4-Plets | 5-Plets |
|-------------|----------|----------|---------|---------|
| P(A1) = 0.1906 | P(T1) = 0.1894 | P(C1) = 0.3101 | P(G1) = 0.3099 | P(A2) = 0.1906 |
| P(T1) = 0.1894 | P(T1) = 0.1899 | P(T1) = 0.1881 | P(C2) = 0.3099 | P(T1) = 0.1894 |
| P(C1) = 0.3101 | P(C1) = 0.3098 | P(C1) = 0.3120 | P(C2) = 0.3074 | P(C1) = 0.3102 |
| P(G1) = 0.3099 | P(G1) = 0.3099 | P(G1) = 0.3107 | P(G2) = 0.3097 | P(G1) = 0.3103 |

#### Triplets

- **Fig. A1/10.** Probabilities of subgroups of tetra-groups in the sequence:
Gloeobacter violaceus PCC 7421 DNA, complete genome, GenBank: BA000045.2, https://www.ncbi.nlm.nih.gov/nuccore/BA000045.2?report=fasta https://www.ncbi.nlm.nih.gov/nuccore/BA000045.2?report=genbank:

LOCUS BA000045 4659019 bp DNA circular BCT 07-OCT-2016
DEFINITION Gloeobacter violaceus PCC 7421 DNA, complete genome.
ACCESSION BA000045 AP006568-AP006583 VERSION BA000045.2

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A1)=0.3033 | P(A)=0.3034 | P4(A)=0.3033 | P4(A)=0.3031 | P5(A)=0.3040 |
| P(T1)=0.3048 | P2(T)=0.3049 | P4(T)=0.3037 | P4(T)=0.3047 | P5(T)=0.3051 |
| P(C1)=0.1970 | P2(C)=0.1969 | P4(C)=0.1989 | P4(C)=0.1971 | P5(C)=0.1963 |
| P(G1)=0.1949 | P2(G)=0.1949 | P4(G)=0.1941 | P4(G)=0.1951 | P5(G)=0.1946 |
| P(A2)=0.3032 | P(A)=0.3042 | P4(A)=0.3036 | P4(A)=0.3029 |
| P(T2)=0.3048 | P2(T)=0.3028 | P4(T)=0.3051 | P4(T)=0.3062 |
| P(C2)=0.1970 | P2(C)=0.1958 | P4(C)=0.1968 | P4(C)=0.1963 |
| P(G2)=0.1950 | P2(G)=0.1972 | P4(G)=0.1945 | P4(G)=0.1946 |
| P(A3)=0.3022 | P(A)=0.3037 | P4(A)=0.3045 |
| P(T3)=0.3080 | P2(T)=0.3050 | P4(T)=0.3028 |
| P(C3)=0.1962 | P2(C)=0.1967 | P4(C)=0.1974 |
| P(G3)=0.1936 | P2(G)=0.1946 | P4(G)=0.1954 |
| P(A4)=0.3028 | P(A)=0.3070 |
| P(T4)=0.3045 | P2(T)=0.3052 |
| P(C4)=0.1973 | P2(C)=0.1974 |
| P(G4)=0.1954 | P2(G)=0.1946 |
| P(A5)=0.3022 |
| P(T5)=0.3049 |
| P(C5)=0.1974 |
| P(G5)=0.1955 |

Fig. A1/11. Probabilities of subgroups of tetra-groups in the sequence: Helicobacter pylori, NCBI Reference Sequence: NC_000921.1
https://www.ncbi.nlm.nih.gov/nuccore/NC_000921.1
LOCUS NC_000921 1643831 bp DNA circular CON 22-MAR-2017
DEFINITION Helicobacter pylori J99, complete genome. ACCESSION
NC_000921 NZ_AE001440-NZ_AE001571 VERSION NC_000921.1

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A1)=0.2848 | P(A)=0.2848 | P4(A)=0.2830 | P4(A)=0.2850 | P5(A)=0.2843 |
| P(T1)=0.2884 | P2(T)=0.2881 | P4(T)=0.2890 | P4(T)=0.2881 | P5(T)=0.2882 |
| P(C1)=0.2136 | P2(C)=0.2139 | P4(C)=0.2148 | P4(C)=0.2134 | P5(C)=0.2137 |
| P(G1)=0.2132 | P2(G)=0.2132 | P4(G)=0.2134 | P4(G)=0.2134 | P5(G)=0.2138 |
| P(A2)=0.2848 | P(A)=0.2851 | P4(A)=0.2846 | P4(A)=0.2850 |
| P(T2)=0.2887 | P2(T)=0.2875 | P4(T)=0.2888 | P4(T)=0.2879 |
| P(C2)=0.2132 | P2(C)=0.2133 | P4(C)=0.2133 | P4(C)=0.2133 |
| P(G2)=0.2133 | P2(G)=0.2141 | P4(G)=0.2134 | P4(G)=0.2137 |
| P(A3)=0.2863 | P4(A)=0.2847 | P5(A)=0.2852 |
| P(T3)=0.2887 | P2(T)=0.2880 | P4(T)=0.2866 |
| P(C3)=0.2126 | P2(C)=0.2144 | P4(C)=0.2137 |
| P(G3)=0.2124 | P2(G)=0.2129 | P4(G)=0.2126 |
| P(A4)=0.2849 | P4(A)=0.2846 |
| P(T4)=0.2887 | P2(T)=0.2889 |
| P(C4)=0.2132 | P4(C)=0.2133 |
Fig. A1/12. Probabilities of subgroups of tetra-groups in the sequence:

Methanosarcina acetivorans str. C2A, complete genome,
GenBank: AE010299.1,
https://www.ncbi.nlm.nih.gov/nuccore/AE010299.1?report=fasta
https://www.ncbi.nlm.nih.gov/nuccore/AE010299

LOCUS AE010299 5751492 bp, DNA circular BCT 01-OCT-2014
DEFINITION Methanosarcina acetivorans str. C2A, complete genome.
ACCESSION AE010299 AE010656-AE011189 VERSION AE010299.1

| NUCLEOTIDES   | DOUBLES    | TRIPLETS | 4-PLETS | 5-PLETS |
|---------------|------------|----------|---------|---------|
| P(A|G)= 0.3422 | P(A|T)= 0.3424 | P(A|G)= 0.3413 | P(A|T)= 0.3422 | P(A|G)= 0.3413 |
| P(T|A)= 0.3422 | P(T|G)= 0.3419 | P(T|A)= 0.3420 | P(T|G)= 0.3422 | P(T|A)= 0.3426 |
| P(C|G)= 0.1576 | P(C|T)= 0.1581 | P(C|G)= 0.1593 | P(C|T)= 0.1572 | P(C|G)= 0.1585 |
| P(G|C)= 0.1580 | P(G|A)= 0.1577 | P(G|C)= 0.1573 | P(G|A)= 0.1584 | P(G|C)= 0.1576 |
| P(G|A)= 0.3420 | P(G|T)= 0.3423 | P(G|A)= 0.3408 | P(G|T)= 0.3433 | P(G|A)= 0.3411 |
| P(G|C)= 0.1571 | P(G|T)= 0.1576 | P(G|C)= 0.1594 | P(G|T)= 0.1596 | P(G|C)= 0.1582 |
| P(G|A)= 0.1583 | P(G|T)= 0.3431 | P(G|A)= 0.3425 | P(G|T)= 0.3430 | P(G|A)= 0.3439 |
| P(G|C)= 0.1563 | P(G|T)= 0.1590 | P(G|C)= 0.1570 | P(G|T)= 0.1545 | P(G|C)= 0.1586 |
| P(G|A)= 0.1573 | P(G|T)= 0.3433 | P(G|A)= 0.3426 | P(G|T)= 0.3432 | P(G|A)= 0.3417 |
| P(G|C)= 0.1571 | P(G|T)= 0.1570 | P(G|C)= 0.1594 | P(G|T)= 0.1574 | P(G|C)= 0.1577 |
| P(G|A)= 0.1570 | P(G|T)= 0.1570 | P(G|A)= 0.1570 | P(G|T)= 0.1570 | P(G|A)= 0.1570 |

Fig. A1/13. Probabilities of subgroups of tetra-groups in the sequence:

Nanoarchaeum equitans Kin4-M, complete genome, GenBank: AE017199.1,
https://www.ncbi.nlm.nih.gov/nuccore/AE017199.1?report=fasta
https://www.ncbi.nlm.nih.gov/nuccore/AE017199

LOCUS AE017199 490885 bp DNA circular BCT 30-JAN-2014
DEFINITION Nanoarchaeum equitans Kin4-M, complete genome. ACCESSION
AE017199 AACL01000000 AACL01000001 VERSION AE017199.1

| NUCLEOTIDES   | DOUBLES    | TRIPLETS | 4-PLETS | 5-PLETS |
|---------------|------------|----------|---------|---------|
| P(A|G)= 0.2431 | P(A|T)= 0.2431 | P(A|G)= 0.2431 | P(A|T)= 0.2435 | P(A|G)= 0.2435 |
| P(T|A)= 0.2423 | P(T|G)= 0.2423 | P(T|A)= 0.2446 | P(T|G)= 0.2423 | P(T|A)= 0.2425 |
| P(C|G)= 0.2556 | P(C|T)= 0.2556 | P(C|G)= 0.2546 | P(C|T)= 0.2557 | P(C|G)= 0.2560 |
| P(G|C)= 0.2590 | P(G|A)= 0.2590 | P(G|C)= 0.2567 | P(G|A)= 0.2585 | P(G|C)= 0.2590 |
**Streptomyces coelicolor**

**DEFINITION**
Syntrophus aciditrophicus SB, complete genome. **ACCESSION**

**LOCUS** CP000252, 3179300 bp, DNA circular BCT 31-JAN-2014 **DEFINITION**
Syntrophus aciditrophicus SB, complete genome. **ACCESSION**

**Fig. A1/14. Probabilities of subgroups of tetra-groups in the sequence:**

| NUCLEOTIDES | DOUBLETs | TRIPLETs | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P₂(A₁) = 0.2431 | P₂(A₂) = 0.2418 | P₂(A₃) = 0.2432 | P₂(A₄) = 0.2435 |
| P₂(T₁) = 0.2424 | P₂(T₂) = 0.2413 | P₂(T₃) = 0.2423 | P₂(T₄) = 0.2418 |
| P₂(G₁) = 0.2557 | P₂(G₂) = 0.2573 | P₂(G₃) = 0.2557 | P₂(G₄) = 0.2554 |
| P₂(C₁) = 0.2589 | P₂(C₂) = 0.2597 | P₂(C₃) = 0.2588 | P₂(C₄) = 0.2592 |

| P₃(A₁) = 0.2434 | P₃(A₂) = 0.2426 | P₃(A₃) = 0.2434 |
| P₃(T₁) = 0.2412 | P₃(T₂) = 0.2422 | P₃(T₃) = 0.2422 |
| P₃(G₁) = 0.2551 | P₃(G₂) = 0.2556 | P₃(G₃) = 0.2557 |
| P₃(C₁) = 0.2604 | P₃(C₂) = 0.2596 | P₃(C₃) = 0.2587 |

**Fig. A1/15. Probabilities of subgroups of tetra-groups in the sequence:**

Streptomyces coelicolor A3(2) complete genome, GenBank: AL645882.2, [https://www.ncbi.nlm.nih.gov/nuccore/AL645882.2?report=genbank](https://www.ncbi.nlm.nih.gov/nuccore/AL645882.2?report=genbank)
**NUCLEOTIDES** | **DOUBLETS** | **TRIPLETS** | **4-PLETS** | **5-PLETS**
---|---|---|---|---
P₃(A₁)= | 0.3184 | P₃(A₂)= | 0.3179 | P₃(A₁)= | 0.3182 | P₃(A₁)= | 0.3177 | P₃(A₁)= | 0.3185 |
P₃(T₁)= | 0.3233 | P₃(T₁)= | 0.3231 | P₃(T₁)= | 0.3225 | P₃(T₁)= | 0.3234 | P₃(T₁)= | 0.3236 |
P₃(C₁)= | 0.1798 | P₃(C₁)= | 0.1797 | P₃(C₁)= | 0.1798 | P₃(C₁)= | 0.1792 | P₃(C₁)= | 0.1793 |
P₃(G₁)= | 0.1788 | P₃(G₁)= | 0.1793 | P₃(G₁)= | 0.1795 | P₃(G₁)= | 0.1798 | P₃(G₁)= | 0.1787 |

Fig. A1/16. Probabilities of subgroups of tetra-groups in the sequence:

Sulfolobus solfataricus strain SULA, complete genome, GenBank: CP011057.1, [https://www.ncbi.nlm.nih.gov/nuccore/CP011057.1?report=fasta](https://www.ncbi.nlm.nih.gov/nuccore/CP011057.1?report=fasta)  
LOCUS CP011057, 2727337 bp, DNA circular BCT 10-JUN-2015  
DEFINITION Sulfolobus solfataricus strain SULA, complete genome. ACCESSION CP011057 VERSION CP011057.1

**NUCLEOTIDES** | **DOUBLETS** | **TRIPLETS** | **4-PLETS** | **5-PLETS**
---|---|---|---|---
P₃(A₁)= | 0.3095 | P₃(A₂)= | 0.3095 | P₃(A₁)= | 0.3079 | P₃(A₁)= | 0.3095 | P₃(A₁)= | 0.3089 |
P₃(T₁)= | 0.3106 | P₃(T₁)= | 0.3104 | P₃(T₁)= | 0.3120 | P₃(T₁)= | 0.3106 | P₃(T₁)= | 0.3106 |
P₃(C₁)= | 0.1880 | P₃(C₁)= | 0.1877 | P₃(C₁)= | 0.1882 | P₃(C₁)= | 0.1871 | P₃(C₁)= | 0.1885 |
P₃(G₁)= | 0.1919 | P₃(G₁)= | 0.1923 | P₃(G₁)= | 0.1918 | P₃(G₁)= | 0.1928 | P₃(G₁)= | 0.1920 |

Sulfolobus solfataricus strain SULA, complete genome, GenBank: CP011057.1, [https://www.ncbi.nlm.nih.gov/nuccore/CP011057.1?report=fasta](https://www.ncbi.nlm.nih.gov/nuccore/CP011057.1?report=fasta)  
LOCUS CP011057, 2727337 bp, DNA circular BCT 10-JUN-2015  
DEFINITION Sulfolobus solfataricus strain SULA, complete genome. ACCESSION CP011057 VERSION CP011057.1
Fig. A1/17. Probabilities of subgroups of tetra-groups in the sequence:

Treponema denticola SP33 supercont1.1, whole genome shotgun sequence
NCBI Reference Sequence: NZ_KB442453.1,
https://www.ncbi.nlm.nih.gov/nuccore/NZ_KB442453.1?report=fasta
https://www.ncbi.nlm.nih.gov/nuccore/NZ_KB442453.1?report=genbank -
LOCUS NZ_KB442453, 1850823 bp, DNA linear CON 02-APR-2017
DEFINITION Treponema denticola SP33 supercont1.1, whole genome shotgun sequence
ACCESSION NZ_KB442453 NZ_AGDZ01000000 VERSION
NZ_KB442453.1

| NUCLEOTIDES | DOUBLETs | TRIPLETs | 4-PLETS | 5-PLETS |
|-------------|----------|----------|----------|----------|
| P(A1)= 0.2695 | P(A1)= 0.2700 | P(A1)= 0.2703 | P(A1)= 0.2700 | P(A1)= 0.2692 |
| P(T1)= 0.2678 | P(T1)= 0.2681 | P(T1)= 0.2689 | P(T1)= 0.2680 | P(T1)= 0.2672 |
| P(C1)= 0.2281 | P(C1)= 0.2277 | P(C1)= 0.2271 | P(C1)= 0.2276 | P(C1)= 0.2291 |
| P(G1)= 0.2347 | P(G1)= 0.2341 | P(G1)= 0.2337 | P(G1)= 0.2345 | P(G1)= 0.2346 |
| P(A2)= 0.2689 | P(A2)= 0.2697 | P(A2)= 0.2688 | P(A2)= 0.2691 |
| P(T2)= 0.2675 | P(T2)= 0.2694 | P(T2)= 0.2671 | P(T2)= 0.2678 |
| P(C2)= 0.2284 | P(C2)= 0.2289 | P(C2)= 0.2295 | P(C2)= 0.2275 |
| P(G2)= 0.2352 | P(G2)= 0.2320 | P(G2)= 0.2345 | P(G2)= 0.2356 |
| P(A3)= 0.2684 | P(A3)= 0.2701 | P(A3)= 0.2703 |
| P(T3)= 0.2651 | P(T3)= 0.2683 | P(T3)= 0.2674 |
| P(C3)= 0.2282 | P(C3)= 0.2279 | P(C3)= 0.2285 |
| P(G3)= 0.2383 | P(G3)= 0.23371 | P(G3)= 0.2339 |
| P(A4)= 0.2690 | P(A4)= 0.2700 |
| P(T4)= 0.2679 | P(T4)= 0.2683 |
| P(C4)= 0.2272 | P(C4)= 0.2272 |
| P(G4)= 0.2359 | P(G4)= 0.2345 |
| P(A5)= 0.2689 |
| P(T5)= 0.2684 |
| P(C5)= 0.2280 |
| P(G5)= 0.2347 |

Fig. A1/18. Probabilities of subgroups of tetra-groups in the sequence:

Thermotoga maritima strain Tma200, complete genome, GenBank: CP010967.1,
https://www.ncbi.nlm.nih.gov/nuccore/CP010967.1?report=fasta
https://www.ncbi.nlm.nih.gov/nuccore/CP010967.1?report=genbank -
LOCUS CP010967, 1859582 bp, DNA circular BCT 10-JUN-2015
DEFINITION Thermotoga maritima strain Tma200, complete genome
ACCESSION CP010967 VERSION CP010967.1

| NUCLEOTIDES | DOUBLETs | TRIPLETs | 4-PLETS | 5-PLETS |
|-------------|----------|----------|----------|----------|
| P(A1)= 0.1543 | P(A1)= 0.1543 | P(A1)= 0.1515 | P(A1)= 0.1541 | P(A1)= 0.1539 |
| P(T1)= 0.1557 | P(T1)= 0.1556 | P(T1)= 0.1556 | P(T1)= 0.1549 | P(T1)= 0.1559 |
| P(C1)= 0.3461 | P(C1)= 0.3458 | P(C1)= 0.3490 | P(C1)= 0.3460 | P(C1)= 0.3461 |
| P(G1)= 0.3439 | P(G1)= 0.3443 | P(G1)= 0.3439 | P(G1)= 0.3450 | P(G1)= 0.3441 |
Thermus thermophilus DNA, complete genome, strain: TMY

GenBank: AP017920.1, -
https://www.ncbi.nlm.nih.gov/nucleotide/AP017920.1?report=fasta
https://www.ncbi.nlm.nih.gov/nucleotide/AP017920.1?report=genbank

LOCUS AP017920, 2121526 bp, DNA circular BCT 26-NOV-2016
DEFINITION Thermus thermophilus DNA, complete genome, strain: TMY.
ACCESSION AP017920 VERSION AP017920.1

### NUCLEOTIDES

| \( \Sigma_1 \) = 229354 | \( \Sigma_2 \) = 114677 | \( \Sigma_3 \) = 76451 | \( \Sigma_4 \) = 57338 | \( \Sigma_5 \) = 45870 |
|---|---|---|---|---|
| \( P_1(A_1) = 0.1542 \) | \( P_2(A_2) = 0.1542 \) | \( P_3(A_3) = 0.1541 \) | \( P_4(A_4) = 0.1547 \) | \( P_5(A_5) = 0.1547 \) |
| \( P_2(T_2) = 0.1558 \) | \( P_2(T_3) = 0.1529 \) | \( P_2(T_4) = 0.1555 \) | \( P_4(T_4) = 0.1561 \) | \( P_5(T_5) = 0.1561 \) |
| \( P_3(C_3) = 0.3464 \) | \( P_4(C_4) = 0.3465 \) | \( P_3(C_3) = 0.3467 \) | \( P_3(C_3) = 0.3447 \) | \( P_3(C_3) = 0.3447 \) |
| \( P_4(G_4) = 0.3435 \) | \( P_4(G_4) = 0.3434 \) | \( P_4(G_4) = 0.3436 \) | \( P_3(G_3) = 0.3445 \) | \( P_3(G_3) = 0.3445 \) |

### DOUBLETs

| \( P_1(A_1) = 0.2157 \) | \( P_2(A_1) = 0.2163 \) | \( P_3(A_1) = 0.2167 \) | \( P_4(A_1) = 0.2176 \) | \( P_5(A_1) = 0.2176 \) |
| \( P_2(T_1) = 0.2127 \) | \( P_3(T_1) = 0.2122 \) | \( P_4(T_1) = 0.2139 \) | \( P_5(T_1) = 0.2106 \) | \( P_5(T_1) = 0.2106 \) |
| \( P_1(C_1) = 0.2830 \) | \( P_2(C_1) = 0.2826 \) | \( P_3(C_1) = 0.2813 \) | \( P_4(C_1) = 0.2832 \) | \( P_5(C_1) = 0.2832 \) |
| \( P_1(G_1) = 0.2886 \) | \( P_2(G_1) = 0.2889 \) | \( P_3(G_1) = 0.2841 \) | \( P_4(G_1) = 0.2881 \) | \( P_5(G_1) = 0.2881 \) |

### TRIPLETS

| \( P_2(A_2) = 0.2152 \) | \( P_3(A_2) = 0.2161 \) | \( P_4(A_2) = 0.3296 \) | \( P_5(A_2) = 0.3296 \) | \( P_5(A_2) = 0.3296 \) |
| \( P_2(T_3) = 0.2131 \) | \( P_3(T_3) = 0.2088 \) | \( P_4(T_3) = 0.2113 \) | \( P_5(T_3) = 0.2116 \) | \( P_5(T_3) = 0.2116 \) |
| \( P_2(C_3) = 0.2834 \) | \( P_3(C_3) = 0.2893 \) | \( P_4(C_3) = 0.2848 \) | \( P_5(C_3) = 0.2860 \) | \( P_5(C_3) = 0.2860 \) |
| \( P_2(G_2) = 0.2883 \) | \( P_3(G_2) = 0.2926 \) | \( P_4(G_2) = 0.2877 \) | \( P_5(G_2) = 0.2877 \) | \( P_5(G_2) = 0.2877 \) |

### 4-PLETS

| \( P_3(A_3) = 0.2159 \) | \( P_4(A_3) = 0.2157 \) | \( P_4(A_3) = 0.2157 \) | \( P_4(A_3) = 0.2157 \) | \( P_4(A_3) = 0.2157 \) |
| \( P_3(T_3) = 0.2120 \) | \( P_4(T_3) = 0.2106 \) | \( P_4(T_3) = 0.2106 \) | \( P_4(T_3) = 0.2106 \) | \( P_4(T_3) = 0.2106 \) |
| \( P_3(C_3) = 0.2811 \) | \( P_4(C_3) = 0.2838 \) | \( P_4(C_3) = 0.2838 \) | \( P_4(C_3) = 0.2838 \) | \( P_4(C_3) = 0.2838 \) |
| \( P_3(G_3) = 0.2891 \) | \( P_4(G_3) = 0.2897 \) | \( P_4(G_3) = 0.2897 \) | \( P_4(G_3) = 0.2897 \) | \( P_4(G_3) = 0.2897 \) |

### 5-PLETS

| \( P_4(A_4) = 0.2279 \) | \( P_5(A_5) = 0.9843 \) | \( P_5(A_5) = 0.9843 \) | \( P_5(A_5) = 0.9843 \) | \( P_5(A_5) = 0.9843 \) |

Fig. A1/19. Probabilities of subgroups of tetra-groups in the sequence:
Fig. A1/20. Collective frequencies \( F_n(A_k) \), \( F_n(T_k) \), \( F_n(C_k) \), \( F_n(G_k) \) and collective probabilities \( P_n(A_k) \), \( P_n(T_k) \), \( P_n(C_k) \) and \( P_n(G_k) \) (\( n = 1, 2, 3, 4, 5 \) and \( k \leq n \)) of tetra-group subgroups in sequences of n-plets in the case of the Human cytomegalovirus strain AD169 complete genome, 229354 bp, GenBank, accession X17403.1. The index \( k \) denotes a position of the letter inside n-plets.
Fig. A1/21. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the VACCG, Vaccinia virus Copenhagen, complete genome, 191737 bp, accession M35027.1, https://www.ncbi.nlm.nih.gov/nucleotide/335317
Fig. A1/22. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the MPOMTCG, Marchantia paleacea isolate A 18 mitochondrion, complete genome, 186609 bp, accession M68929.1, https://www.ncbi.nlm.nih.gov/nuccore/786182

| NUCLEOTIDES | 1-PLETS | 2-PLETS | 3-PLETS | 4-PLETS | 5-PLETS |
|-------------|---------|---------|---------|---------|---------|
| $\Sigma_1 = 184113$ | $\Sigma_2 = 92056$ | $\Sigma_3 = 61371$ | $\Sigma_4 = 46028$ | $\Sigma_5 = 36822$ |
| $F_4(G_3) =$ | 9979 | $P_4(G_3) =$ | 0.2139 | $F_5(G_3) =$ | 7917 |
| $P_4(G_3) =$ | 0.2121 | $F_5(A_3) =$ | 10504 | $P_5(A_3) =$ | 0.2815 |
| $F_5(T_3) =$ | 10860 | $P_5(T_3) =$ | 0.2910 | $F_5(C_3) =$ | 8021 |
| $P_5(C_3) =$ | 0.2149 | $F_5(G_3) =$ | 7936 | $P_5(G_3) =$ | 0.2126 |
| $F_6(G_4) =$ | 92056 | $P_6(G_4) =$ | 0.3022 | $F_6(C_4) =$ | 0.2960 |
| $P_6(C_4) =$ | 0.2974 | $F_6(G_4) =$ | 0.2986 | $P_6(G_4) =$ | 0.2974 |
Fig. A1/23. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ (n = 1, 2, 3, 4, 5 and k ≤ n) of tetra-group subgroups in sequences of n-plets in the case of the HS48958RAJ, Epstein-Barr virus, 184113 bp, accession M80517.1, https://www.ncbi.nlm.nih.gov/nuccore/330330
Fig. A1/24. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the Nicotiana tabacum chloroplast genome DNA,155943 bp, accession Z00044.2, https://www.ncbi.nlm.nih.gov/nuccore/Z00044
Fig. A1/25. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the Equine herpesvirus 1 strain Ab4, complete genome, 150224 bp, accession AY665713.1, https://www.ncbi.nlm.nih.gov/nuccore/AY665713.1
Fig. A1/26. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the Oryza sativa cultivar TN1 chloroplast, complete genome, 134502 bp, accession NC_031333.1, https://www.ncbi.nlm.nih.gov/nuccore/NC_031333.1

| NUCLEOTIDES | DOUBLETs | TRIPLETs | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| $\Sigma_1 = 134226$ | $\Sigma_2 = 67113$ | $\Sigma_3 = 44742$ | $\Sigma_4 = 33556$ | $\Sigma_5 = 26845$ |
| $F_1(A)=28727$ | $F_2(A)=14299$ | $F_3(A)=9855$ | $F_4(A)=7148$ | $F_5(A)=5795$ |
| $P_1(A)=0.2140$ | $P_2(A)=0.2131$ | $P_3(A)=0.2203$ | $P_4(A)=0.2130$ | $P_5(A)=0.2159$ |
| $F_1(T)=30025$ | $F_2(T)=15047$ | $F_3(T)=10633$ | $F_4(T)=7535$ | $F_5(T)=5969$ |
| $P_1(T)=0.2237$ | $P_2(T)=0.2242$ | $P_3(T)=0.2377$ | $P_4(T)=0.2246$ | $P_5(T)=0.2224$ |
| $F_1(C)=37767$ | $F_2(C)=18943$ | $F_3(C)=12066$ | $F_4(C)=9436$ | $F_5(C)=7480$ |
| $P_1(C)=0.2814$ | $P_2(C)=0.2823$ | $P_3(C)=0.2697$ | $P_4(C)=0.2812$ | $P_5(C)=0.2786$ |
| $F_1(G)=37707$ | $F_2(G)=18824$ | $F_3(G)=12188$ | $F_4(G)=9437$ | $F_5(G)=7601$ |
| $P_1(G)=0.2809$ | $P_2(G)=0.2805$ | $P_3(G)=0.2724$ | $P_4(G)=0.2812$ | $P_5(G)=0.2831$ |

$F_5(A)=8197$ $P_5(A)=0.3047$ $F_5(T)=8129$ $P_5(T)=0.3022$ $F_5(C)=5307$ $P_5(C)=0.1973$ $F_5(G)=5267$ $P_5(G)=0.1958$
Fig. A1/27. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the IH1CG, Ictalurid herpesvirus 1 strain Auburn 1, complete genome, 134226 bp, accession M75136.2, https://www.ncbi.nlm.nih.gov/nuccore/519868059
Fig. A1/28. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, and $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of $n$-plets in the case of the Human herpesvirus 3 isolate 667/2005, complete genome, 124884 bp, accession JN704693.1, [https://www.ncbi.nlm.nih.gov/nuccore/JN704693.1](https://www.ncbi.nlm.nih.gov/nuccore/JN704693.1)

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|----------|---------|
| $\Sigma_1 = 121024$ | $\Sigma_2 = 60512$ | $\Sigma_3 =$40341 | $\Sigma_4 =$30256 | $\Sigma_5 =$24204 |
| $F_1(A)=$ 42896 | $F_2(A)=$(21385 | $F_3(A)=$(14349 | $F_4(A)=$(10739 | $F_5(A)=$(8517 |
| $P_1(A)=$0.3544 | $P_2(A)=$(0.3534 | $P_3(A)=$(0.3557 | $P_4(A)=$(0.3549 | $P_5(A)=$(0.3519 |
| $F_1(T)=$(43263 | $F_2(T)=$(21703 | $F_3(T)=$(14637 | $F_4(T)=$(10870 | $F_5(T)=$(8677 |
| $P_1(T)=$(0.3575 | $P_2(T)=$(0.3587 | $P_3(T)=$(0.3628 | $P_4(T)=$(0.3593 | $P_5(T)=$(0.3585 |
| $F_1(C)=$(17309 | $F_2(C)=$(8677 | $F_3(C)=$(5687 | $F_4(C)=$(4340 | $F_5(C)=$(3448 |
| $P_1(C)=$(0.1430 | $P_2(C)=$(0.1434 | $P_3(C)=$(0.1410 | $P_4(C)=$(0.1434 | $P_5(C)=$(0.1425 |
| $F_1(G)=$(17556 | $F_2(G)=$(8747 | $F_3(G)=$(5668 | $F_4(G)=$(4307 | $F_5(G)=$(3562 |
| $P_1(G)=$(0.1451 | $P_2(G)=$(0.1445 | $P_3(G)=$(0.1405 | $P_4(G)=$(0.1424 | $P_5(G)=$(0.1472 |
| $F_1(A)=$(21511 | $F_2(A)=$(14274 | $F_3(A)=$(10782 | $F_4(A)=$(8550 | $F_5(A)=$(8550 |
| $P_1(A)=$(0.3555 | $P_2(A)=$(0.3538 | $P_3(A)=$(0.3564 | $P_4(A)=$(0.3564 | $P_5(A)=$(0.3532 |
| $F_1(T)=$(21560 | $F_2(T)=$(14746 | $F_3(T)=$(10815 | $F_4(T)=$(8658 | $F_5(T)=$(8658 |
| $P_1(T)=$(0.3563 | $P_2(T)=$(0.3655 | $P_3(T)=$(0.3574 | $P_4(T)=$(0.3574 | $P_5(T)=$(0.3577 |
| $F_1(C)=$(8632 | $F_2(C)=$(5657 | $F_3(C)=$(4288 | $F_4(C)=$(3533 | $F_5(C)=$(3533 |
| $P_1(C)=$(0.1426 | $P_2(C)=$(0.1402 | $P_3(C)=$(0.1417 | $P_4(C)=$(0.1417 | $P_5(C)=$(0.1460 |
| $F_1(G)=$(8809 | $F_2(G)=$(5664 | $F_3(G)=$(4371 | $F_4(G)=$(3463 | $F_5(G)=$(3463 |
| $P_1(G)=$(0.1456 | $P_2(G)=$(0.1404 | $P_3(G)=$(0.1445 | $P_4(G)=$(0.1445 | $P_5(G)=$(0.1431 |
| $F_1(A)=$(14273 | $F_2(A)=$(10646 | $F_3(A)=$(8596 | $F_4(A)=$(8596 | $F_5(A)=$(8596 |
| $P_1(A)=$(0.3538 | $P_2(A)=$(0.3519 | $P_3(A)=$(0.3551 | $P_4(A)=$(0.3551 | $P_5(A)=$(0.3551 |
| $F_1(T)=$(13879 | $F_2(T)=$(10833 | $F_3(T)=$(8694 | $F_4(T)=$(8694 | $F_5(T)=$(8694 |
| $P_1(T)=$(0.3440 | $P_2(T)=$(0.3580 | $P_3(T)=$(0.3592 | $P_4(T)=$(0.3592 | $P_5(T)=$(0.3592 |
| $F_1(C)=$(5965 | $F_2(C)=$(4337 | $F_3(C)=$(3416 | $F_4(C)=$(3416 | $F_5(C)=$(3416 |
| $P_1(C)=$(0.1479 | $P_2(C)=$(0.1433 | $P_3(C)=$(0.1411 | $P_4(C)=$(0.1411 | $P_5(C)=$(0.1411 |
| $F_1(G)=$(6224 | $F_2(G)=$(4440 | $F_3(G)=$(3498 | $F_4(G)=$(3498 | $F_5(G)=$(3498 |
| $P_1(G)=$(0.1543 | $P_2(G)=$(0.1467 | $P_3(G)=$(0.1467 | $P_4(G)=$(0.1467 | $P_5(G)=$(0.1467 |
| $F_1(A)=$(10729 | $F_2(A)=$(8680 | $F_3(A)=$(8552 | $F_4(A)=$(8552 | $F_5(A)=$(8552 |
| $P_1(A)=$(0.3546 | $P_2(A)=$(0.3586 | $P_3(A)=$(0.3586 | $P_4(A)=$(0.3586 | $P_5(A)=$(0.3586 |
| $F_1(T)=$(10745 | $F_2(T)=$(8552 | $F_3(T)=$(8552 | $F_4(T)=$(8552 | $F_5(T)=$(8552 |
| $P_1(T)=$(0.3551 | $P_2(T)=$(0.3533 | $P_3(T)=$(0.3533 | $P_4(T)=$(0.3533 | $P_5(T)=$(0.3533 |
| $F_1(C)=$(4344 | $F_2(C)=$(3475 | $F_3(C)=$(3475 | $F_4(C)=$(3475 | $F_5(C)=$(3475 |
| $P_1(C)=$(0.1436 | $P_2(C)=$(0.1436 | $P_3(C)=$(0.1436 | $P_4(C)=$(0.1436 | $P_5(C)=$(0.1436 |
| $F_1(G)=$(4438 | $F_2(G)=$(3497 | $F_3(G)=$(3497 | $F_4(G)=$(3497 | $F_5(G)=$(3497 |
| $P_1(G)=$(0.1467 | $P_2(G)=$(0.1467 | $P_3(G)=$(0.1467 | $P_4(G)=$(0.1467 | $P_5(G)=$(0.1467 |
| $F_1(A)=$(8552 | $F_2(A)=$(8552 | $F_3(A)=$(8552 | $F_4(A)=$(8552 | $F_5(A)=$(8552 |
| $P_1(A)=$(0.3533 | $P_2(A)=$(0.3533 | $P_3(A)=$(0.3533 | $P_4(A)=$(0.3533 | $P_5(A)=$(0.3533 |
| $F_1(T)=$(8680 | $F_2(T)=$(8680 | $F_3(T)=$(8680 | $F_4(T)=$(8680 | $F_5(T)=$(8680 |
| $P_1(T)=$(0.3586 | $P_2(T)=$(0.3586 | $P_3(T)=$(0.3586 | $P_4(T)=$(0.3586 | $P_5(T)=$(0.3586 |
Fig. A1/29. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of $n$-plets in the case of the Marchantia paleacea chloroplast genome DNA, $121024$ bp, accession X04465.1, [https://www.ncbi.nlm.nih.gov/nuccore/X04465](https://www.ncbi.nlm.nih.gov/nuccore/X04465)
Fig. A1/30. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the Escherichia coli strain PSU078 plasmid pPSU078_1, complete sequence, 132464 bp, accession CP012113.1, https://www.ncbi.nlm.nih.gov/nuccore/CP012113.1
Fig. A1/31. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the HSIULR, Human herpesvirus 1 complete genome, 152261 bp, accession X14112.1, [https://www.ncbi.nlm.nih.gov/nuccore/X14112.1](https://www.ncbi.nlm.nih.gov/nuccore/X14112.1)

| NUCLEOTIDES | DOUBLETs | TRIPLETs | 4-PLETs | 5-PLETs |
|-------------|----------|----------|---------|---------|
| $\Sigma_1 = 100849$ | $\Sigma_2 = 50424$ | $\Sigma_3 = 33616$ | $\Sigma_4 = 25212$ | $\Sigma_5 = 20169$ |
| $F_1(A) = 30346$ | $F_1(A) = 15271$ | $F_3(A) = 7636$ | $F_3(A) = 5978$ |
| $P_1(A) = 0,3009$ | $P_1(A) = 0,3029$ | $P_1(A) = 0,3021$ | $P_1(A) = 0,2996$ |
| $F_1(T) = 32481$ | $F_1(T) = 16175$ | $F_3(T) = 8030$ | $F_3(T) = 6570$ |
| $P_1(T) = 0,3221$ | $P_1(T) = 0,3208$ | $P_1(T) = 0,3197$ | $P_1(T) = 0,3185$ |
| $F_1(G) = 18635$ | $F_1(G) = 9390$ | $F_3(G) = 4727$ | $F_3(G) = 3709$ |
| $P_1(G) = 0,1848$ | $P_1(G) = 18,62$ | $P_1(G) = 0,1875$ | $P_1(G) = 0,1875$ |
| $F_1(C) = 19387$ | $F_1(C) = 9588$ | $F_3(C) = 4819$ | $F_3(C) = 3912$ |
| $P_1(C) = 0,1922$ | $P_1(C) = 0,1901$ | $P_1(C) = 0,1911$ | $P_1(C) = 0,1940$ |

$F_5(G_3) = 10503$
$P_5(G_2) = 0,3449$
Fig. A1/32. Collective frequencies $F_n(A_k), F_n(T_k), F_n(C_k), F_n(G_k)$ and collective probabilities $P_n(A_k), P_n(T_k), P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the following case: HUMNEUROF, Human oligodendrocyte myelin glycoprotein (OMG) exons 1-2; neurofibromatosis 1 (NF1) exons 28-49; ecotropic viral integration site 2B (EVI2B) exons 1-2; ecotropic viral integration site 2A (EVI2A) exons 1-2; adenylate kinase (AK3) exons 1-2, 100849 bp, accession L05367.1, https://www.ncbi.nlm.nih.gov/nuccore/189152

| NUCLEOTIDES | DOUBLETS | TRIPLETs | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| $\Sigma_1 = 100314$ | $\Sigma_2 = 50157$ | $\Sigma_3 = 33438$ | $\Sigma_4 = 25078$ | $\Sigma_5 = 20062$ |
| $F_1(A_1) =$ 35804 | $F_2(A_1) =$ 17906 | $F_3(A_1) =$ 11917 | $F_4(A_1) =$ 8814 | $F_5(A_1) =$ 7101 |
| $P_1(A_1) =0,3569$ | $P_2(A_1) =0,3570$ | $P_3(A_1) =0,3564$ | $P_4(A_1) =0,3515$ | $P_5(A_1) =0,3540$ |
| $F_1(T_1) =$ 34358 | $F_2(T_1) =$ 17236 | $F_3(T_1) =$ 11802 | $F_4(T_1) =$ 8693 | $F_5(T_1) =$ 6967 |
| $P_1(T_1) =0,3425$ | $P_2(T_1) =0,3436$ | $P_3(T_1) =0,3530$ | $P_4(T_1) =0,3466$ | $P_5(T_1) =0,3473$ |
| $F_1(C_1) =$ 13428 | $F_2(C_1) =$ 6669 | $F_3(C_1) =$ 4415 | $F_4(C_1) =$ 3413 | $F_5(C_1) =$ 2614 |
| $P_1(C_1) =0,1339$ | $P_2(C_1) =0,1330$ | $P_3(C_1) =0,1320$ | $P_4(C_1) =0,1361$ | $P_5(C_1) =0,1303$ |
| $F_1(G_1) =$ 16724 | $F_2(G_1) =$ 8346 | $F_3(G_1) =$ 5304 | $F_4(G_1) =$ 4158 | $F_5(G_1) =$ 3380 |
| $P_1(G_1) =0,1667$ | $P_2(G_1) =0,1664$ | $P_3(G_1) =0,1586$ | $P_4(G_1) =0,1658$ | $P_5(G_1) =0,1685$ |
| $F_1(A_2) =$ 17898 | $F_2(A_2) =$ 12161 | $F_3(A_2) =$ 8975 | $F_4(A_2) =$ 7134 |
| $P_1(A_2) =0,3568$ | $P_2(A_2) =0,3567$ | $P_3(A_2) =0,3579$ | $P_4(A_2) =0,3556$ |
| $F_1(T_2) =$ 17122 | $F_2(T_2) =$ 11010 | $F_3(T_2) =$ 8550 | $F_4(T_2) =$ 6849 |
| $P_1(T_2) =0,3414$ | $P_2(T_2) =0,3293$ | $P_3(T_2) =0,3409$ | $P_4(T_2) =0,3414$ |
| $F_1(C_2) =$ 6759 | $F_2(C_2) =$ 4448 | $F_3(C_2) =$ 3344 | $F_4(C_2) =$ 2707 |
| $P_1(C_2) =0,1348$ | $P_2(C_2) =0,1330$ | $P_3(C_2) =0,1333$ | $P_4(C_2) =0,1349$ |
| $F_1(G_2) =$ 8378 | $F_2(G_2) =$ 5819 | $F_3(G_2) =$ 4209 | $F_4(G_2) =$ 3372 |
| $P_1(G_2) =0,1670$ | $P_2(G_2) =0,1740$ | $P_3(G_2) =0,1678$ | $P_4(G_2) =0,1681$ |
| $F_1(A_3) =$ 11726 | $F_2(A_3) =$ 9092 | $F_3(A_3) =0,3507$ | $F_4(A_3) =0,3525$ |
| $P_1(A_3) =0,3507$ | $P_2(A_3) =0,3625$ | $P_3(A_3) =0,3554$ |
| $F_3(T_3) =$ 11546 | $F_4(T_3) =$ 8542 | $F_5(T_3) =$ 6833 |
| $P_3(T_3) =0,3453$ | $P_4(T_3) =0,3406$ | $P_5(T_3) =0,3406$ |
| $F_1(C_3) =$ 4565 | $F_2(C_3) =$ 3256 | $F_3(C_3) =$ 2677 |
| $P_1(C_3) =0,1365$ | $P_2(C_3) =0,1298$ | $P_3(C_3) =0,1334$ |
| $F_1(G_3) =$ 5601 | $F_2(G_3) =$ 4188 | $F_3(G_3) =$ 3422 |
| $P_1(G_3) =0,1675$ | $P_2(G_3) =0,1670$ | $P_3(G_3) =0,1706$ |
| $F_2(T_4) =$ 8922 | $F_3(T_4) =$ 7177 | $F_4(T_4) =$ 7177 |
| $P_2(T_4) =0,3558$ | $P_3(T_4) =0,3558$ | $P_4(T_4) =0,3577$ |
| $F_4(C_4) =$ 3415 | $F_5(C_4) =$ 2730 |
| $P_4(C_4) =0,1362$ | $P_5(C_4) =0,1361$ |
| $F_1(G_4) =$ 4169 | $F_2(G_4) =$ 3241 |
| $P_1(G_4) =0,1662$ | $P_2(G_4) =0,1662$ |
| $F_1(A_5) =$ 7261 | $F_2(A_5) =$ 7261 |
| $P_1(A_5) =0,3619$ | $P_2(A_5) =0,3619$ |
| $F_5(T_5) =$ 6792 | $P_5(T_5) =0,3386$ |
| $F_5(C_5) =$ 2700 | $P_5(C_5) =0,1346$ |
| $F_5(G_5) =$ 3309 | $P_5(G_5) =0,1649$ |
Fig. A1/33. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, $F_n(G_k)$ and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the Podospora anserina mitochondrion, complete genome, 100314 bp, accession NC_001329.3, https://www.ncbi.nlm.nih.gov/nuccore/NC_001329.3
Fig. A1/34. Collective frequencies $F_n(A_k)$, $F_n(T_k)$, $F_n(C_k)$, and collective probabilities $P_n(A_k)$, $P_n(T_k)$, $P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the HUMTCRADCV, Human T-cell receptor genes (Human Tcr-c-delta gene, exons 1-4; Tcr-V-delta gene, exons 1-2; T-cell receptor alpha (Tcr-alpha) gene, J1-J61 segments; and Tcr-C-alpha gene, exons 1-4), 97630 bp, accession M94081.1, https://www.ncbi.nlm.nih.gov/nuccore/2627263

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| $\Sigma_1 = 94647$ | $\Sigma_2 = 47323$ | $\Sigma_3 = 31549$ | $\Sigma_4 = 23661$ | $\Sigma_5 = 18929$ |
| $F_1(A_1) = 26359$ | $F_2(A_1) = 13093$ | $F_3(A_1) = 8879$ | $F_4(A_1) = 6527$ | $F_5(A_1) = 5303$ |
| $P_1(A_1) = 0.2785$ | $P_2(A_1) = 0.2767$ | $P_3(A_1) = 0.2814$ | $P_4(A_1) = 0.2759$ | $P_5(A_1) = 0.2802$ |
| $F_1(T_1) = 25769$ | $F_2(T_1) = 13008$ | $F_3(T_1) = 8621$ | $F_4(T_1) = 6476$ | $F_5(T_1) = 5124$ |
| $P_1(T_1) = 0.2723$ | $P_2(T_1) = 0.2749$ | $P_3(T_1) = 0.2733$ | $P_4(T_1) = 0.2737$ | $P_5(T_1) = 0.2707$ |
| $F_1(C_1) = 20790$ | $F_2(C_1) = 10284$ | $F_3(C_1) = 6911$ | $F_4(C_1) = 5166$ | $F_5(C_1) = 4108$ |
| $P_1(C_1) = 0.2197$ | $P_2(C_1) = 0.2173$ | $P_3(C_1) = 0.2191$ | $P_4(C_1) = 0.2183$ | $P_5(C_1) = 0.2170$ |
| $F_1(G_1) = 21729$ | $F_2(G_1) = 10938$ | $F_3(G_1) = 7138$ | $F_4(G_1) = 5492$ | $F_5(G_1) = 4394$ |
| $P_1(G_1) = 0.2296$ | $P_2(G_1) = 0.2311$ | $P_3(G_1) = 0.2263$ | $P_4(G_1) = 0.2321$ | $P_5(G_1) = 0.2321$ |
| $F_2(A_2) = 13265$ | $F_3(A_2) = 8763$ | $F_4(A_2) = 6632$ | $F_5(A_2) = 5317$ |
| $P_2(A_2) = 0.2803$ | $P_3(A_2) = 0.2778$ | $P_4(A_2) = 0.2803$ | $P_5(A_2) = 0.2809$ |
| $F_2(T_2) = 12761$ | $F_3(T_2) = 8621$ | $F_4(T_2) = 6381$ | $F_5(T_2) = 5184$ |
| $P_2(T_2) = 0.2697$ | $P_3(T_2) = 0.2733$ | $P_4(T_2) = 0.2697$ | $P_5(T_2) = 0.2739$ |
| $F_2(C_2) = 10506$ | $F_3(C_2) = 6875$ | $F_4(C_2) = 5247$ | $F_5(C_2) = 4152$ |
| $P_2(C_2) = 0.2220$ | $P_3(C_2) = 0.2179$ | $P_4(C_2) = 0.2218$ | $P_5(C_2) = 0.2193$ |
| $F_2(G_2) = 10791$ | $F_3(G_2) = 7290$ | $F_4(G_2) = 5401$ | $F_5(G_2) = 4276$ |
| $P_2(G_2) = 0.2280$ | $P_3(G_2) = 0.2311$ | $P_4(G_2) = 0.2283$ | $P_5(G_2) = 0.2259$ |
| $F_3(A_3) = 8717$ | $F_4(A_3) = 6566$ | $F_5(A_3) = 5242$ |
| $P_3(A_3) = 0.2763$ | $P_4(A_3) = 0.2775$ | $P_5(A_3) = 0.2769$ |
| $F_3(T_3) = 8527$ | $F_4(T_3) = 6532$ | $F_5(T_3) = 5160$ |
| $P_3(T_3) = 0.2703$ | $P_4(T_3) = 0.2761$ | $P_5(T_3) = 0.2726$ |
| $F_3(C_3) = 7004$ | $F_4(C_3) = 5117$ | $F_5(C_3) = 4163$ |
| $P_3(C_3) = 0.2220$ | $P_4(C_3) = 0.2163$ | $P_5(C_3) = 0.2199$ |
| $F_3(G_3) = 7301$ | $F_4(G_3) = 5446$ | $F_5(G_3) = 4364$ |
| $P_3(G_3) = 0.2314$ | $P_4(G_3) = 0.2302$ | $P_5(G_3) = 0.2305$ |
| $F_4(A_4) = 6633$ | $F_5(A_4) = 5261$ |
| $P_4(A_4) = 0.2803$ | $P_5(A_4) = 0.2779$ |
| $F_4(T_4) = 6380$ | $F_5(T_4) = 5137$ |
| $P_4(T_4) = 0.2696$ | $P_5(T_4) = 0.2714$ |
| $F_4(C_4) = 5259$ | $F_5(C_4) = 4217$ |
| $P_4(C_4) = 0.2223$ | $P_5(C_4) = 0.2228$ |
| $F_4(G_4) = 5389$ | $F_5(G_4) = 4314$ |
| $P_4(G_4) = 0.2278$ | $P_5(G_4) = 0.2279$ |
| $F_5(A_5) = 5235$ | $F_5(T_5) = 5164$ |
| $P_5(A_5) = 0.2766$ | $P_5(T_5) = 0.2728$ |
| $F_5(C_5) = 4150$ | $F_5(G_5) = 4380$ |
| $P_5(C_5) = 0.2192$ | $P_5(G_5) = 0.2314$ |
Fig. A1/35. Collective frequencies $F_n(A_k), F_n(T_k), F_n(C_k), F_n(G_k)$ and collective probabilities $P_n(A_k), P_n(T_k), P_n(C_k)$ and $P_n(G_k)$ ($n = 1, 2, 3, 4, 5$ and $k \leq n$) of tetra-group subgroups in sequences of n-plets in the case of the MUSTCRA, Mouse T-cell receptor alpha/delta chain locus, 94647 bp, accession M64239.1, https://www.ncbi.nlm.nih.gov/nuccore/201744

One can see from the data of the long DNA-sequences in Fig. A1/1-A1/35 that all these sequences satisfy the three tetra-group rules and demonstrate the existence of tetra-group symmetries in them.

Appendix 2. Symmetries of tetra-group probabilities in the compete set of human chromosomes

The Appendix 2 represents data about the fulfillment of the described tetra-group rules in the complete set of 24 chromosomes of human genome. All initial data about chromosomes are taken from the CenBank.

| Nucleotides | Doublets | Triplets | 4-Plets | 5-Plets |
|-------------|----------|----------|---------|---------|
| P(A) = 0.2910 | P(A) = 0.2910 | P(A) = 0.2910 | P(A) = 0.2910 | P(A) = 0.2910 |
| P(T) = 0.2918 | P(T) = 0.2917 | P(T) = 0.2917 | P(T) = 0.2917 | P(T) = 0.2917 |
| P(C) = 0.2085 | P(C) = 0.2085 | P(C) = 0.2084 | P(C) = 0.2085 | P(C) = 0.2085 |
| P(G) = 0.2087 | P(G) = 0.2088 | P(G) = 0.2088 | P(G) = 0.2088 | P(G) = 0.2088 |

Fig. A2/1. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 1, GRCh38.p7 Primary Assembly
NCBI Reference Sequence: NC_000001.11
LOCUS NC_000001 248956422 bp DNA linear CON 06-JUN-2016
DEFINITION Homo sapiens chromosome 1, GRCh38.p7 Primary Assembly.
ACCESSION NC_000001 GPC_000001293 VERSION NC_000001.11
https://www.ncbi.nlm.nih.gov/nuccore/NC_000001.11
https://www.ncbi.nlm.nih.gov/nuccore/NC_000001.11?report=fasta
### Nucleotides

|  | 4-Plots | 5-Plots |
|---|---|---|
| P(A)_1 | 0.3013 | 0.3012 | 0.3013 | 0.3013 | 0.3013 |
| P(A)_2 | 0.3020 | 0.3020 | 0.3020 | 0.3020 | 0.3020 |
| P(A)_3 | 0.1980 | 0.1981 | 0.1979 | 0.1981 | 0.1980 |
| P(G)_1 | 0.1986 | 0.1986 | 0.1986 | 0.1986 | 0.1986 |
| P(G)_2 | 0.1986 | 0.1986 | 0.1986 | 0.1986 | 0.1986 |
| P(G)_3 | 0.1986 | 0.1986 | 0.1986 | 0.1986 | 0.1986 |

---

Fig. A2/2. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 2, GRCh38.p7 Primary Assembly

NCBI Reference Sequence: NC_000002.12

LOCUS NC_000002 242193529 bp DNA linear CON 06-JUN-2016

DEFINITION Homo sapiens chromosome 2, GRCh38.p7 Primary Assembly.

ACCESSION NC_000002 GPC_000001294 VERSION NC_000002.12

https://www.ncbi.nlm.nih.gov/nuccore/NC_000002.12

https://www.ncbi.nlm.nih.gov/nuccore/NC_000002.12?report=fasta

---

### Nucleotides

|  | 4-Plots | 5-Plots |
|---|---|---|
| P(A)_1 | 0.2984 | 0.2985 | 0.2983 | 0.2985 | 0.2986 |
| P(A)_2 | 0.2993 | 0.2992 | 0.2993 | 0.2991 | 0.2992 |
| P(A)_3 | 0.2009 | 0.2009 | 0.2009 | 0.2008 | 0.2008 |
| P(G)_1 | 0.2014 | 0.2014 | 0.2014 | 0.2015 | 0.2014 |
| P(G)_2 | 0.2014 | 0.2014 | 0.2014 | 0.2014 | 0.2014 |
| P(G)_3 | 0.2014 | 0.2014 | 0.2014 | 0.2014 | 0.2014 |

---

Fig. A2/3. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 3, GRCh38,p7 Primary Assembly

NCBI Reference Sequence: NC_000003.12
# NC_000003 Accession

| LOCUS  | NC_000003 | 198295559 bp | DNA | linear | CON 06-JUN-2016 |
|--------|-----------|--------------|-----|--------|-----------------|
| DEFINITION | Homo sapiens chromosome 3, GRCh38.p7 Primary Assembly |
| ACCESSION | NC_000003 GPC_000001295 VERSION | NC_000003.12 |
| NCBI Reference Sequence | NC_000000.12 |

**Fig. A2/4.** The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 4, GRCh38.p7 Primary Assembly

**LOCUS** NC_000004 190214555 bp DNA linear CON 06-JUN-2016

**DEFINITION** Homo sapiens chromosome 4, GRCh38.p7 Primary Assembly.

**ACCESSION** NC_000004 GPC_000001296 VERSION NC_000004.12

**https://www.ncbi.nlm.nih.gov/nuccore/NC_000004.12**

**NCBI Reference Sequence:** NC_000000.12

---

### Nucleotides and Probabilities

| Nucleotides | Doublets | Triplets | 4-Plets | 5-Plets |
|-------------|----------|----------|---------|---------|
| P(A)| 0.3018 | 0.3018 | 0.3018 | 0.3017 | 0.3018 |
| P(G)| 0.1971 | 0.1971 | 0.1971 | 0.1972 | 0.1971 |
| P(C)| 0.1979 | 0.1979 | 0.1989 | 0.1979 | 0.1980 |
| P(T)| 0.3018 | 0.3018 | 0.3017 | 0.3018 | 0.3018 |

---

The table above shows the probabilities of different tetra-groups in the sequence. Each nucleotide (A, C, G, T) has a set of doublet, triplet, 4-plet, and 5-plet probabilities. For example, the probability of finding a specific nucleotide is 0.3018, 0.1971, 0.1979, or 0.3018 depending on the position in the sequence.
# Fig. A2/5. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 5, GRCh38.p7 Primary Assembly
NCBI Reference Sequence: NC_000005.10
LOCUS NC_000005 181538259 bp DNA linear CON 06-JUN-2016
DEFINITION Homo sapiens chromosome 5, GRCh38.p7 Primary Assembly.
ACCESSION NC_000005 GPC_000001297 VERSION NC_000005.10
[https://www.ncbi.nlm.nih.gov/nuccore/NC_000005.10](https://www.ncbi.nlm.nih.gov/nuccore/NC_000005.10)

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|------------|----------|----------|----------|---------|
| P(A)       | 0.3021   | 0.3021   | 0.3021   | 0.3021  |
| P(T)       | 0.3020   | 0.3020   | 0.3021   | 0.3020  |
| P(C)       | 0.1979   | 0.1979   | 0.1979   | 0.1980  |
| P(G)       | 0.1979   | 0.1979   | 0.1980   | 0.1979  |

# Fig. A2/6. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 6, whole genome shotgun sequence
GenBank: CM000257.1
LOCUS CM000257 171718000 bp DNA linear CON 23-MAR-2015
DEFINITION Homo sapiens chromosome 6, whole genome shotgun sequence.
ACCESSION CM000257 AADB02000000 CH003477 VERSION CM000257.1
[https://www.ncbi.nlm.nih.gov/nuccore/CM000257.1](https://www.ncbi.nlm.nih.gov/nuccore/CM000257.1)
[https://www.ncbi.nlm.nih.gov/nuccore/CM000257.1?report=fasta](https://www.ncbi.nlm.nih.gov/nuccore/CM000257.1?report=fasta)

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|------------|----------|----------|----------|---------|
| P(A)       | 0.2960   | 0.2960   | 0.2960   | 0.2960  |
| P(T)       | 0.2970   | 0.2970   | 0.2971   | 0.2970  |
| P(C)       | 0.2033   | 0.2033   | 0.2032   | 0.2032  |
| P(G)       | 0.2037   | 0.2037   | 0.2037   | 0.2037  |

## Nucleotides
- **A**: Adenine
- **T**: Thymine
- **C**: Cytosine
- **G**: Guanine

## Subgroups
- **0**: Adenine
- **1**: Thymine
- **2**: Cytosine
- **3**: Guanine

## Probabilities
- **0.3021**: Probability of subgroup A
- **0.3020**: Probability of subgroup T
- **0.1979**: Probability of subgroup C
- **0.1979**: Probability of subgroup G

## Doublets
- **0.3021**: Probability of doublet AA
- **0.3021**: Probability of doublet AT
- **0.1979**: Probability of doublet AC
- **0.1979**: Probability of doublet AG

## Triplets
- **0.3021**: Probability of triplet AAA
- **0.3020**: Probability of triplet ATA
- **0.1979**: Probability of triplet ACA
- **0.1979**: Probability of triplet AGA

## Tetralets
- **0.2960**: Probability of tetralet AAAA
- **0.2970**: Probability of tetralet ATTT
- **0.2033**: Probability of tetralet CAAA
- **0.2037**: Probability of tetralet CGAA

## Pentalets
- **0.2960**: Probability of pentalet AAAAA
- **0.2970**: Probability of pentalet ATATA
- **0.2033**: Probability of pentalet CAAA 
- **0.2036**: Probability of pentalet CGGA
**Fig. A2/7.** The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 7, GRCh38.p7 Primary Assembly

NCBI Reference Sequence: NC_000007.14

| LOCUS          | 159345973 bp DNA linear CON 06-JUN-2016 | DEFINITION Homo sapiens chromosome 7, GRCh38.p7 Primary Assembly. |
|----------------|----------------------------------------|---------------------------------------------------------------|
| ACCESSION      | NC_000007 GPC_000001299 VERSION       | NC_000007.14                                                 |
|                |                                        | https://www.ncbi.nlm.nih.gov/nuccore/NC_000007.14             |
|                |                                        | https://www.ncbi.nlm.nih.gov/nuccore/NC_000007.14?report=fasta |

| **NUCLEOTIDES** | **DOUBLETs** | **TRIPLETs** | **4-PLETS** | **5-PLETS** |
|-----------------|--------------|--------------|--------------|--------------|
| \( P(A_1) = 0.2994 \) | \( P(A_2) = 0.2994 \) | \( P(A_3) = 0.2994 \) | \( P(A_4) = 0.2993 \) | \( P(A_5) = 0.2993 \) |
| \( P(T_1) = 0.2990 \) | \( P(T_2) = 0.2990 \) | \( P(T_3) = 0.2990 \) | \( P(T_4) = 0.2991 \) | \( P(T_5) = 0.2991 \) |
| \( P(C_1) = 0.2008 \) | \( P(C_2) = 0.2008 \) | \( P(C_3) = 0.2008 \) | \( P(C_4) = 0.2008 \) | \( P(C_5) = 0.2008 \) |
| \( P(G_1) = 0.2008 \) | \( P(G_2) = 0.2008 \) | \( P(G_3) = 0.2008 \) | \( P(G_4) = 0.2008 \) | \( P(G_5) = 0.2008 \) |

**Fig. A2/8.** The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 8, alternate assembly CHM1_1.1, whole genome shotgun sequence

NCBI Reference Sequence: NC_018919.2

| LOCUS          | 146399655 bp DNA linear CON 06-JUN-2016 | DEFINITION Homo sapiens chromosome 8, alternate assembly CHM1_1.1, whole genome shotgun sequence. |
|----------------|----------------------------------------|---------------------------------------------------------------|
| ACCESSION      | NC_018919 GPC_000001167 VERSION        | NC_018919.2                                                   |
|                |                                        | https://www.ncbi.nlm.nih.gov/nuccore/NC_018919.2              |
|                |                                        | https://www.ncbi.nlm.nih.gov/nuccore/NC_018919.2?report=fasta |
| NUCLEOTIDES | DOUBLETs | TRIPLETs | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)=0.2926 | P(A)=0.2928 | P(A)=0.2927 | P(A)=0.2929 | P(A)=0.2927 |
| P(T)=0.2926 | P(T)=0.2928 | P(T)=0.2927 | P(T)=0.2928 | P(T)=0.2926 |
| P(C)=0.2074 | P(C)=0.2073 | P(C)=0.2074 | P(C)=0.2072 | P(C)=0.2074 |
| P(G)=0.2072 | P(G)=0.2073 | P(G)=0.2072 | P(G)=0.2073 | P(G)=0.2073 |

Fig. A2/9. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 9, whole genome shotgun sequence
GenBank: CM000260.1
LOCUS CM000260 111583154 bp DNA linear CON 23-MAR-2015
DEFINITION Homo sapiens chromosome 9, whole genome shotgun sequence.
ACCESSION CM000260 AADB02000000 CH003480 VERSION CM000260.1
https://www.ncbi.nlm.nih.gov/nuccore/CM000260.1
https://www.ncbi.nlm.nih.gov/nuccore/CM000260.1?report=fasta

| NUCLEOTIDES | DOUBLETs | TRIPLETs | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)=0.2917 | P(A)=0.2918 | P(A)=0.2917 | P(A)=0.2918 | P(A)=0.2917 |
| P(T)=0.2928 | P(T)=0.2928 | P(T)=0.2928 | P(T)=0.2928 | P(T)=0.2928 |
| P(C)=0.2074 | P(C)=0.2073 | P(C)=0.2074 | P(C)=0.2074 | P(C)=0.2074 |
| P(G)=0.2080 | P(G)=0.2080 | P(G)=0.2080 | P(G)=0.2080 | P(G)=0.2080 |

Fig. A2/10. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 10, GRCh38.p7 Primary Assembly
NCBI Reference Sequence: NC_000010.11
Fig. A2/11. The table of probabilities of subgroups of tetra-groups in the sequence Homo sapiens chromosome 11, GRCh38.p7 Primary Assembly
NCBI Reference Sequence: NC_000011.10
LOCUS NC_000011 135086622 bp DNA linear CON 06-JUN-2016
DEFINITION Homo sapiens chromosome 11, GRCh38.p7 Primary Assembly.
ACCESSION NC_000011 GPC_000001303 VERSION NC_000011.10
https://www.ncbi.nlm.nih.gov/nuccore/NC_000011.10
https://www.ncbi.nlm.nih.gov/nuccore/NC_000011.10?report=fasta

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P_(A1) | 0.2957 | P_(A1) | 0.2957 | P_(A1) | 0.2958 | P_(A1) | 0.2958 | P_(A1) | 0.2957 |
| P_(T1) | 0.2966 | P_(T1) | 0.2966 | P_(T1) | 0.2966 | P_(T1) | 0.2966 | P_(T1) | 0.2967 |
| P_(C1) | 0.2035 | P_(C1) | 0.2035 | P_(C1) | 0.2035 | P_(C1) | 0.2035 | P_(C1) | 0.2035 |
| P_(G1) | 0.2042 | P_(G1) | 0.2042 | P_(G1) | 0.2041 | P_(G1) | 0.2042 | P_(G1) | 0.2041 |

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P_(A1) | 0.2957 | P_(A1) | 0.2957 | P_(A1) | 0.2958 | P_(A1) | 0.2958 | P_(A1) | 0.2957 |
| P_(T1) | 0.2966 | P_(T1) | 0.2966 | P_(T1) | 0.2966 | P_(T1) | 0.2966 | P_(T1) | 0.2967 |
| P_(C1) | 0.2035 | P_(C1) | 0.2035 | P_(C1) | 0.2035 | P_(C1) | 0.2035 | P_(C1) | 0.2035 |
| P_(G1) | 0.2042 | P_(G1) | 0.2042 | P_(G1) | 0.2041 | P_(G1) | 0.2042 | P_(G1) | 0.2041 |
Fig. A2/12. Probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 12, GRCh38.p7 Primary Assembly
NCBI Reference Sequence: NC_000012.12
LOCUS NC_000012 133275309 bp DNA linear CON 06-JUN-2016
DEFINITION Homo sapiens chromosome 12, GRCh38.p7 Primary Assembly.
ACCESSION NC_000012 GPC_000001304 VERSION NC_000012.12
https://www.ncbi.nlm.nih.gov/nuccore/NC_000012.12
https://www.ncbi.nlm.nih.gov/nuccore/NC_000012.12?report=fasta

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)=0.3069 | P(A)=0.3068 | P(A)=0.3069 | P(A)=0.3066 | P(A)=0.3071 |
| P(T)=0.3079 | P(T)=0.3078 | P(T)=0.3079 | P(T)=0.3077 |         |
| P(C)=0.1926 | P(C)=0.1927 | P(C)=0.1926 | P(C)=0.1925 | P(C)=0.1925 |
| P(G)=0.1926 | P(G)=0.1925 | P(G)=0.1926 | P(G)=0.1925 | P(G)=0.1925 |

Fig. A2/13. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 13, whole genome shotgun sequence
GenBank: CM000264.1
LOCUS CM000264 95789532 bp DNA linear CON 23-MAR-2015
DEFINITION Homo sapiens chromosome 13, whole genome shotgun sequence.
ACCESSION CM000264 AADB02000000 CH003484 VERSION CM000264.1
https://www.ncbi.nlm.nih.gov/nuccore/CM000264.1
https://www.ncbi.nlm.nih.gov/nuccore/CM000264.1?report=fasta

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)=0.2945 | P(A)=0.2944 | P(A)=0.2946 | P(A)=0.2944 | P(A)=0.2945 |
| P(T)=0.2970 | P(T)=0.2970 | P(T)=0.2970 | P(T)=0.2970 | P(T)=0.2970 |
| P(C)=0.2040 | P(C)=0.2041 | P(C)=0.2040 | P(C)=0.2041 | P(C)=0.2039 |
| P(G)=0.2045 | P(G)=0.2045 | P(G)=0.2045 | P(G)=0.2045 | P(G)=0.2045 |
| P(A)=0.2946 | P(A)=0.2946 | P(A)=0.2946 | P(A)=0.2946 | P(A)=0.2946 |
| P(T)=0.2970 | P(T)=0.2970 | P(T)=0.2970 | P(T)=0.2970 | P(T)=0.2970 |
| P(C)=0.2039 | P(C)=0.2039 | P(C)=0.2039 | P(C)=0.2039 | P(C)=0.2039 |
| P(G)=0.2045 | P(G)=0.2045 | P(G)=0.2045 | P(G)=0.2045 | P(G)=0.2045 |
| P(A)=0.2944 | P(A)=0.2945 | P(A)=0.2945 | P(A)=0.2945 | P(A)=0.2945 |
Fig. A2/14. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 14, whole genome shotgun sequence
GenBank: CM000265.1
LOCUS CM000265 87316725 bp DNA linear CON 23-MAR-2015
DEFINITION Homo sapiens chromosome 14, whole genome shotgun sequence.
ACCESSION CM000265 AADB02000000 CH003485 VERSION CM000265.1
https://www.ncbi.nlm.nih.gov/nuccore/CM000265.1
https://www.ncbi.nlm.nih.gov/nuccore/CM000265.1?report=fasta

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(Ai)       | 0.2896   | 0.2896   | 0.2896  | 0.2895  |
| P(Ti)       | 0.2901   | 0.2901   | 0.2900  | 0.2902  |
| P(Gi)       | 0.2097   | 0.2097   | 0.2098  | 0.2098  |
| P(Ci)       | 0.2106   | 0.2106   | 0.2105  | 0.2105  |

Fig. A2/15. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 15, GRCh38.p7 Primary Assembly
NCBI Reference Sequence: NC_000015.10
LOCUS NC_000015 101991189 bp DNA linear CON 06-JUN-2016
DEFINITION Homo sapiens chromosome 15, GRCh38.p7 Primary Assembly.
ACCESSION NC_000015 GPC_000001307 VERSION NC_000015.10
https://www.ncbi.nlm.nih.gov/nuccore/NC_000015.10
https://www.ncbi.nlm.nih.gov/nuccore/NC_000015.10?report=fasta
Fig. A2/16. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 16, GRCh38.p7 Primary Assembly
NCBI Reference Sequence: NC_000016.10
LOCUS NC_000016 90338345 bp DNA linear CON 06-JUN-2016
DEFINITION Homo sapiens chromosome 16, GRCh38.p7 Primary Assembly.
ACCESSION NC_000016 GPC_000001308 VERSION NC_000016.10
https://www.ncbi.nlm.nih.gov/nuccore/NC_000016.10
https://www.ncbi.nlm.nih.gov/nuccore/NC_000016.10?report=fasta

Fig. A2/17. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 17, GRCh38.p7 Primary Assembly
**Fig. A2/18.** The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 18, whole genome shotgun sequence GenBank: CM000269.1

**NCBI Reference Sequence:** NC_000017.11  
**LOCUS** NC_000017  83257441 bp DNA linear CON 06-JUN-2016  
**DEFINITION** Homo sapiens chromosome 17, GRCh38.p7 Primary Assembly.  
**ACCESSION** NC_000017 GPC_000001309 VERSION NC_000017.11  
[https://www.ncbi.nlm.nih.gov/nuccore/NC_000017.11](https://www.ncbi.nlm.nih.gov/nuccore/NC_000017.11)  
[https://www.ncbi.nlm.nih.gov/nuccore/NC_000017.11?report=fasta](https://www.ncbi.nlm.nih.gov/nuccore/NC_000017.11?report=fasta)

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A) = 0.3011 | P(A) = 0.3012 | P(A) = 0.3011 | P(A) = 0.3012 | P(A) = 0.3009 |
| P(T) = 0.3014 | P(T) = 0.3013 | P(T) = 0.3014 | P(T) = 0.3014 | P(T) = 0.3017 |
| P(C) = 0.1987 | P(C) = 0.1985 | P(C) = 0.1987 | P(C) = 0.1986 | P(C) = 0.1986 |
| P(G) = 0.1989 | P(G) = 0.1989 | P(G) = 0.1988 | P(G) = 0.1988 | P(G) = 0.1988 |

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A) = 0.3009 | P(A) = 0.3011 | P(A) = 0.3009 | P(A) = 0.3012 | P(A) = 0.3012 |
| P(T) = 0.3014 | P(T) = 0.3013 | P(T) = 0.3014 | P(T) = 0.3014 | P(T) = 0.3013 |
| P(C) = 0.1988 | P(C) = 0.1987 | P(C) = 0.1987 | P(C) = 0.1986 | P(C) = 0.1986 |
| P(G) = 0.1989 | P(G) = 0.1989 | P(G) = 0.1990 | P(G) = 0.1990 | P(G) = 0.1990 |

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A) = 0.3014 | P(A) = 0.3013 | P(A) = 0.3014 | P(A) = 0.3014 | P(A) = 0.3014 |
| P(T) = 0.3013 | P(T) = 0.3014 | P(T) = 0.3014 | P(T) = 0.3014 | P(T) = 0.3013 |
| P(C) = 0.1986 | P(C) = 0.1985 | P(C) = 0.1985 | P(C) = 0.1986 | P(C) = 0.1986 |
| P(G) = 0.1990 | P(G) = 0.1990 | P(G) = 0.1990 | P(G) = 0.1990 | P(G) = 0.1990 |

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A) = 0.3010 | P(A) = 0.3012 | P(A) = 0.3012 | P(A) = 0.3012 | P(A) = 0.3010 |
| P(T) = 0.3014 | P(T) = 0.3013 | P(T) = 0.3014 | P(T) = 0.3014 | P(T) = 0.3013 |
| P(C) = 0.1988 | P(C) = 0.1987 | P(C) = 0.1987 | P(C) = 0.1986 | P(C) = 0.1986 |
| P(G) = 0.1989 | P(G) = 0.1989 | P(G) = 0.1990 | P(G) = 0.1990 | P(G) = 0.1990 |

**NCBI Reference Sequence:** NC_000017.11  
**LOCUS** NC_000017  74792881 bp DNA linear CON 23-MAR-2015  
**DEFINITION** Homo sapiens chromosome 18, whole genome shotgun sequence.  
**ACCESSION** CM000269 AADB02000000 CH003499 VERSION CM000269.1  
[https://www.ncbi.nlm.nih.gov/nuccore/CM000269.1](https://www.ncbi.nlm.nih.gov/nuccore/CM000269.1)  
[https://www.ncbi.nlm.nih.gov/nuccore/CM000269.1?report=fasta](https://www.ncbi.nlm.nih.gov/nuccore/CM000269.1?report=fasta)
### Fig. A2/19. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 19, GRCh38.p7 Primary Assembly
NCBI Reference Sequence: NC_000019.10
LOCUS NC_000019 58617616 bp DNA linear CON 06-JUN-2016
DEFINITION Homo sapiens chromosome 19, GRCh38.p7 Primary Assembly.
ACCESSION NC_000019 GPC_000001311 VERSION NC_000019.10
https://www.ncbi.nlm.nih.gov/nuccore/NC_000019.10
https://www.ncbi.nlm.nih.gov/nuccore/NC_000019.10?report=fasta

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)=0.2778 | P(A)=0.2779 | P(A)=0.2778 | P(A)=0.2779 | P(A)=0.2777 |
| P(T)=0.2810 | P(T)=0.2808 | P(T)=0.2810 | P(T)=0.2809 | P(T)=0.2811 |
| P(C)=0.2201 | P(C)=0.2202 | P(C)=0.2201 | P(C)=0.2202 | P(C)=0.2200 |
| P(G)=0.2211 | P(G)=0.2212 | P(G)=0.2211 | P(G)=0.2211 | P(G)=0.2212 |

### Fig. A2/20. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 20, whole genome shotgun sequence
GenBank: CM000271.1
LOCUS CM000271 59605541 bp DNA linear CON 23-MAR-2015
DEFINITION Homo sapiens chromosome 20, whole genome shotgun sequence.
ACCESSION CM000271 AADB02000000 CH003491 VERSION CM000271.1
https://www.ncbi.nlm.nih.gov/nuccore/CM000271.1
https://www.ncbi.nlm.nih.gov/nuccore/CM000271.1?report=fasta

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)=0.2964 | P(A)=0.2964 | P(A)=0.2963 | P(A)=0.2966 | P(A)=0.2965 |
| P(T)=0.2946 | P(T)=0.2947 | P(T)=0.2946 | P(T)=0.2946 | P(T)=0.2943 |
| P(C)=0.2047 | P(C)=0.2046 | P(C)=0.2046 | P(C)=0.2046 | P(C)=0.2045 |
| P(G)=0.2043 | P(G)=0.2043 | P(G)=0.2045 | P(G)=0.2042 | P(G)=0.2046 |
| P(A)=0.2965 | P(A)=0.2964 | P(A)=0.2967 | P(A)=0.2966 | P(A)=0.2966 |
| P(T)=0.2945 | P(T)=0.2946 | P(T)=0.2945 | P(T)=0.2945 | P(T)=0.2948 |
Fig. A2/21. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens genomic DNA, chromosome 21q

LOCUS BA000005 3354332 bp DNA linear CON 12-JUL-2008
DEFINITION Homo sapiens genomic DNA, chromosome 21q. ACCESSION BA000005 VERSION BA000005.3
https://www.ncbi.nlm.nih.gov/nuccore/BA000005
https://www.ncbi.nlm.nih.gov/nuccore/BA000005.3?report=fasta

| NUCLEOTIDES | DOUBLETs | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)= 0.2651 | P(T)= 0.2651 | P(A)= 0.2651 | P(A)= 0.2651 | P(A)= 0.2652 |
| P(G)= 0.2648 | P(T)= 0.2648 | P(T)= 0.2649 | P(T)= 0.2647 | P(T)= 0.2651 |
| P(C)= 0.2339 | P(C)= 0.2339 | P(C)= 0.2339 | P(C)= 0.2340 | P(C)= 0.2340 |
| P(G)= 0.2361 | P(G)= 0.2362 | P(G)= 0.2361 | P(G)= 0.2361 | P(G)= 0.2361 |

Fig. A2/22. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome 22, GRCh38.p7 Primary Assembly

NCBI Reference Sequence: NC_000022.11

LOCUS NC_000022 50818468 bp DNA linear CON 06-JUN-2016
DEFINITION Homo sapiens chromosome 22, GRCh38.p7 Primary Assembly. ACCESSION NC_000022 GPC_000001314 VERSION NC_000022.11
https://www.ncbi.nlm.nih.gov/nuccore/NC_000022.11
https://www.ncbi.nlm.nih.gov/nuccore/NC_000022.11?report=fasta
**Fig. A2/23.** The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome X, GRCh38.p7 Primary Assembly

NCBI Reference Sequence: NC_000023.11

LOCUS   NC_000023  156040895 bp DNA linear CON 06-JUN-2016

DEFINITION Homo sapiens chromosome X, GRCh38.p7 Primary Assembly.

ACCESSION  NC_000023 GPC_000001315

VERSION   NC_000023.11

https://www.ncbi.nlm.nih.gov/nuccore/NC_000023.11

https://www.ncbi.nlm.nih.gov/nuccore/NC_000023.11?report=fasta

---

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P(A)= 0.2985 | P(A)= 0.2986 | P(A)= 0.2985 | P(A)= 0.2987 | P(A)= 0.2983 |
| P(T)= 0.3012 | P(T)= 0.3013 | P(T)= 0.3014 | P(T)= 0.3009 | P(T)= 0.3014 |
| P(C)= 0.2001 | P(C)= 0.2001 | P(C)= 0.2000 | P(C)= 0.2004 | P(C)= 0.2004 |
| P(G)= 0.2001 | P(G)= 0.2000 | P(G)= 0.2000 | P(G)= 0.2004 | P(G)= 0.2004 |

---

Fig. A2/24. The table of probabilities of subgroups of tetra-groups in the sequence: Homo sapiens chromosome Y, GRCh38.p7 Primary Assembly
Appendix 3. Symmetries of tetra-group probabilities in the complete set of chromosomes of a nematode *Caenorhabditis elegans*

The appendix 3 represents data about the fulfillment of the described tetra-group rules in the complete set of chromosomes of *Caenorhabditis elegans*. This free-living soil nematode is widely used as a model organism in genetics for a long time. The *Caenorhabditis elegans* nuclear genome is approximately 100 Mb, distributed among six chromosomes. All initial data are taken from the CenBank ([https://www.ncbi.nlm.nih.gov/nuccore/NC_000024.10](https://www.ncbi.nlm.nih.gov/nuccore/NC_000024.10)).

| NUCLEOTIDES | DOUBLETS | TRIPLETs | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P₁(A₁) = 0.3208 | P₂(A₁) = 0.3207 | P₃(A₁) = 0.3205 | P₄(A₁) = 0.3207 | P₅(A₁) = 0.3205 |
| P₁(T₁) = 0.3217 | P₂(T₁) = 0.3218 | P₃(T₁) = 0.3216 | P₄(T₁) = 0.3219 | P₅(T₁) = 0.3219 |
| P₁(C₁) = 0.1789 | P₂(C₁) = 0.1789 | P₃(C₁) = 0.1788 | P₄(C₁) = 0.1789 | P₅(C₁) = 0.1789 |
| P₁(G₁) = 0.1786 | P₂(G₁) = 0.1787 | P₃(G₁) = 0.1791 | P₄(G₁) = 0.1784 | P₅(G₁) = 0.1787 |
| P₂(A₂) = 0.321 | P₃(A₂) = 0.3212 | P₄(A₂) = 0.321 | P₅(A₂) = 0.321 |
| P₂(T₂) = 0.3216 | P₃(T₂) = 0.3216 | P₄(T₂) = 0.3218 | P₅(T₂) = 0.3215 |
| P₂(C₂) = 0.1788 | P₃(C₂) = 0.1793 | P₄(C₂) = 0.1785 | P₅(C₂) = 0.179 |
| P₂(G₂) = 0.1786 | P₃(G₂) = 0.1779 | P₄(G₂) = 0.1788 | P₅(G₂) = 0.1785 |
| P₃(A₃) = 0.3209 | P₄(A₃) = 0.3206 | P₅(A₃) = 0.3208 |
| P₃(T₃) = 0.3218 | P₄(T₃) = 0.3217 | P₅(T₃) = 0.3218 |
| P₃(C₃) = 0.1784 | P₄(C₃) = 0.1788 | P₅(C₃) = 0.1789 |
| P₃(G₃) = 0.1789 | P₄(G₃) = 0.1789 | P₅(G₃) = 0.1785 |
| P₄(A₄) = 0.3211 | P₅(A₄) = 0.321 |
| P₄(T₄) = 0.3213 | P₅(T₄) = 0.3218 |
| P₄(C₄) = 0.1792 | P₅(C₄) = 0.1786 |
| P₄(G₄) = 0.1784 | P₅(G₄) = 0.1787 |
| P₅(A₅) = 0.3211 |
| P₅(T₅) = 0.3213 |
| P₅(C₅) = 0.179 |
| P₅(G₅) = 0.1787 |
Fig. A3/1. The table of probabilities of subgroups of tetra-groups in the sequence: Caenorhabditis elegans chromosome I, NCBI Reference Sequence: NC_003279.8

LOCUS NC_003279 15072434 bp DNA linear CON 11-OCT-2017
DEFINITION Caenorhabditis elegans chromosome I. ACCESSION NC_003279
VERSION NC_003279.8
https://www.ncbi.nlm.nih.gov/nuccore/NC_003279.8

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P₁(A₁)= 0.3193 | P₂(A₁)= 0.3193 | P₃(A₁)= 0.3193 | P₄(A₁)= 0.3193 | P₅(A₁)= 0.3191 |
| P₁(T₁)= 0.3187 | P₂(T₁)= 0.3187 | P₃(T₁)= 0.3189 | P₄(T₁)= 0.3187 | P₅(T₁)= 0.3189 |
| P₁(C₁)= 0.1812 | P₂(C₁)= 0.1813 | P₃(C₁)= 0.1814 | P₄(C₁)= 0.1813 | P₅(C₁)= 0.1814 |
| P₁(G₁)= 0.1808 | P₂(G₁)= 0.1807 | P₃(G₁)= 0.1807 | P₄(G₁)= 0.1806 | P₅(G₁)= 0.1806 |
| P₂(A₂)= 0.3193 | P₂(A₂)= 0.3192 | P₃(A₂)= 0.3193 | P₄(A₂)= 0.3193 | P₅(A₂)= 0.3193 |
| P₂(T₂)= 0.3187 | P₃(T₂)= 0.318 | P₄(T₂)= 0.3187 | P₅(T₂)= 0.319 |
| P₂(C₂)= 0.1812 | P₃(C₂)= 0.1814 | P₄(C₂)= 0.181 | P₅(C₂)= 0.1809 |
| P₂(G₂)= 0.1809 | P₃(G₂)= 0.1814 | P₄(G₂)= 0.181 | P₅(G₂)= 0.1809 |
| P₃(A₃)= 0.3196 | P₄(A₃)= 0.3192 | P₅(A₃)= 0.3193 |
| P₃(T₃)= 0.3192 | P₄(T₃)= 0.3187 | P₅(T₃)= 0.3183 |
| P₃(C₃)= 0.1809 | P₄(C₃)= 0.1813 | P₅(C₃)= 0.1813 |
| P₃(G₃)= 0.1803 | P₄(G₃)= 0.1808 | P₅(G₃)= 0.181 |
| P₄(A₄)= 0.3192 | P₅(A₄)= 0.3194 |
| P₄(T₄)= 0.3187 | P₅(T₄)= 0.3185 |
| P₄(C₄)= 0.1813 | P₅(C₄)= 0.1813 |
| P₄(G₄)= 0.1807 | P₅(G₄)= 0.1808 |
| P₅(A₅)= 0.3193 |
| P₅(T₅)= 0.3189 |
| P₅(C₅)= 0.1813 |
| P₅(G₅)= 0.1805 |

Fig. A3/2. The table of probabilities of subgroups of tetra-groups in the sequence: Caenorhabditis elegans chromosome II,
### NUCLEOTIDES

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P₁(A₁) = 0.3225 | P₂(A₁) = 0.3226 | P₃(A₁) = 0.3223 | P₄(A₁) = 0.3225 | P₅(A₁) = 0.3226 |
| P₁(T₁) = 0.3209 | P₂(T₁) = 0.3209 | P₃(T₁) = 0.3206 | P₄(T₁) = 0.3209 | P₅(T₁) = 0.3211 |
| P₁(C₁) = 0.1777 | P₂(C₁) = 0.1778 | P₃(C₁) = 0.178 | P₄(C₁) = 0.1778 | P₅(C₁) = 0.1777 |
| P₁(G₁) = 0.1789 | P₂(G₁) = 0.1787 | P₃(G₁) = 0.1791 | P₄(G₁) = 0.1788 | P₅(G₁) = 0.1787 |
| P₁(Å) = 0.3223 | P₂(Å) = 0.3227 | P₃(Å) = 0.3223 | P₄(Å) = 0.3223 | P₅(Å) = 0.3223 |
| P₂(T₂) = 0.3209 | P₃(T₂) = 0.3213 | P₄(T₂) = 0.3209 | P₅(T₂) = 0.321 |
| P₂(C₂) = 0.1776 | P₃(C₂) = 0.1776 | P₄(C₂) = 0.1776 | P₅(C₂) = 0.178 |
| P₂(G₂) = 0.1792 | P₃(G₂) = 0.1784 | P₄(G₂) = 0.1792 | P₅(G₂) = 0.1788 |
| P₃(Å) = 0.3223 | P₄(Å) = 0.3227 | P₅(Å) = 0.3222 |
| P₃(T₃) = 0.3209 | P₄(T₃) = 0.321 | P₅(T₃) = 0.3208 |
| P₃(C₃) = 0.1775 | P₄(C₃) = 0.1778 | P₅(C₃) = 0.1779 |
| P₃(G₃) = 0.1792 | P₄(G₃) = 0.1785 | P₅(G₃) = 0.179 |
| P₄(Å) = 0.3224 | P₅(Å) = 0.3225 |
| P₄(T₄) = 0.321 | P₅(T₄) = 0.3209 |
| P₄(C₄) = 0.1775 | P₅(C₄) = 0.1775 |
| P₄(G₄) = 0.1791 | P₅(G₄) = 0.179 |
| P₅(Å) = 0.3227 |
| P₅(T₅) = 0.3208 |
| P₅(C₅) = 0.1773 |
| P₅(G₅) = 0.1792 |

Fig. A3/3. The table of probabilities of subgroups of tetra-groups in the sequence: Caenorhabditis elegans chromosome III, NCBI Reference Sequence: NC_003281.10

LOCUS NC_003281 13783801 bp DNA linear CON 11-OCT-2017
DEFINITION Caenorhabditis elegans chromosome III. ACCESSION NC_003281
VERSION NC_003281.10
https://www.ncbi.nlm.nih.gov/nuccore/NC_003281.10
| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|------------|----------|----------|---------|---------|
| P₁(A₁) = 0.3265 | P₂(A₁) = 0.3263 | P₃(A₁) = 0.3264 | P₄(A₁) = 0.3265 | P₅(A₁) = 0.3263 |
| P₁(T₁) = 0.3276 | P₂(T₁) = 0.3277 | P₃(T₁) = 0.3273 | P₄(T₁) = 0.3276 | P₅(T₁) = 0.328 |
| P₁(C₁) = 0.1735 | P₂(C₁) = 0.1735 | P₃(C₁) = 0.1735 | P₄(C₁) = 0.1736 | P₅(C₁) = 0.1733 |
| P₁(G₁) = 0.1725 | P₂(G₁) = 0.1724 | P₃(G₁) = 0.1728 | P₄(G₁) = 0.1723 | P₅(G₁) = 0.1723 |
| P₂(A₂) = 0.3266 | P₃(A₂) = 0.3265 | P₄(A₂) = 0.3266 | P₅(A₂) = 0.3265 |
| P₂(T₂) = 0.3275 | P₃(T₂) = 0.3275 | P₄(T₂) = 0.3275 | P₅(T₂) = 0.3277 |
| P₂(C₂) = 0.1734 | P₃(C₂) = 0.1737 | P₄(C₂) = 0.1732 | P₅(C₂) = 0.1736 |
| P₂(G₂) = 0.1725 | P₃(G₂) = 0.1724 | P₄(G₂) = 0.1727 | P₅(G₂) = 0.1722 |
| P₃(A₃) = 0.3265 | P₄(A₃) = 0.3262 | P₅(A₃) = 0.3268 |
| P₃(T₃) = 0.328 | P₄(T₃) = 0.3278 | P₅(T₃) = 0.3272 |
| P₃(C₃) = 0.1732 | P₄(C₃) = 0.1735 | P₅(C₃) = 0.1735 |
| P₃(G₃) = 0.1723 | P₄(G₃) = 0.1725 | P₅(G₃) = 0.1725 |
| P₄(A₄) = 0.3266 | P₅(A₄) = 0.3265 |
| P₄(T₄) = 0.3275 | P₅(T₄) = 0.3276 |
| P₄(C₄) = 0.1736 | P₅(C₄) = 0.1733 |
| P₄(G₄) = 0.1724 | P₅(G₄) = 0.1726 |
| P₅(A₅) = 0.3262 |
| P₅(T₅) = 0.3275 |
| P₅(C₅) = 0.1736 |
| P₅(G₅) = 0.1726 |

Fig. A3/4. The table of probabilities of subgroups of tetra-groups in the sequence: Caenorhabditis elegans chromosome IV, NCBI Reference Sequence: NC_003282.8

LOCUS NC_003282 17493829 bp DNA linear CON 11-OCT-2017
DEFINITION Caenorhabditis elegans chromosome IV. ACCESSION NC_003282
VERSION NC_003282.8
https://www.ncbi.nlm.nih.gov/nuccore/NC_003282.8
| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P₁(A₁) = 0.3226 | P₂(A₁) = 0.3225 | P₃(A₁) = 0.3227 | P₄(A₁) = 0.3224 | P₅(A₁) = 0.3227 |
| P₁(T₁) = 0.3231 | P₂(T₁) = 0.3232 | P₃(T₁) = 0.3229 | P₄(T₁) = 0.3233 | P₅(T₁) = 0.3232 |
| P₁(C₁) = 0.1774 | P₂(C₁) = 0.1775 | P₃(C₁) = 0.1772 | P₄(C₁) = 0.1776 | P₅(C₁) = 0.177 |
| P₁(G₁) = 0.1769 | P₂(G₁) = 0.1768 | P₃(G₁) = 0.1772 | P₄(G₁) = 0.1768 | P₅(G₁) = 0.1771 |
| P₂(T₂) = 0.323 | P₃(T₂) = 0.3228 | P₄(T₂) = 0.3231 | P₅(T₂) = 0.3231 |
| P₂(C₂) = 0.1773 | P₃(C₂) = 0.1778 | P₄(C₂) = 0.1772 | P₅(C₂) = 0.1775 |
| P₂(G₂) = 0.177 | P₃(G₂) = 0.1768 | P₄(G₂) = 0.1768 | P₅(G₂) = 0.1765 |
| P₃(A₃) = 0.3225 | P₄(A₃) = 0.3226 | P₅(A₃) = 0.3227 |
| P₃(T₃) = 0.3235 | P₄(T₃) = 0.3231 | P₅(T₃) = 0.3228 |
| P₃(C₃) = 0.1773 | P₄(C₃) = 0.1775 | P₅(C₃) = 0.1777 |
| P₃(G₃) = 0.1767 | P₄(G₃) = 0.1769 | P₅(G₃) = 0.1768 |
| P₄(A₄) = 0.3226 | P₅(A₄) = 0.3222 |
| P₄(T₄) = 0.3229 | P₅(T₄) = 0.3231 |
| P₄(C₄) = 0.1774 | P₅(C₄) = 0.1773 |
| P₄(G₄) = 0.1771 | P₅(G₄) = 0.1773 |
| P₅(A₅) = 0.3227 |
| P₅(T₅) = 0.3232 |
| P₅(C₅) = 0.1774 |
| P₅(G₅) = 0.1767 |

Fig. A3/5. The table of probabilities of subgroups of tetra-groups in the sequence: Caenorhabditis elegans chromosome V, NCBI Reference Sequence: NC_003283.11

LOCUS NC_003283 20924180 bp DNA linear CON 11-OCT-2017
DEFINITION Caenorhabditis elegans chromosome V. ACCESSION NC_003283
VERSION NC_003283.11
https://www.ncbi.nlm.nih.gov/nuccore/NC_003283.11
Fig. A3/6. The table of probabilities of subgroups of tetra-groups in the sequence: Caenorhabditis elegans chromosome X, NCBI Reference Sequence: NC_003284.9

LOCUS NC_003284 17718942 bp DNA linear CON 11-OCT-2017
DEFINITION Caenorhabditis elegans chromosome X. ACCESSION NC_003284
VERSION NC_003284.9

https://www.ncbi.nlm.nih.gov/nuccore/NC_003284.9

Appendix 4. Symmetries of tetra-group probabilities in the complete set of chromosomes of Drosophila melanogaster

The appendix 4 represents data about the fulfillment of the described tetra-group rules in the complete set of chromosomes of Drosophila melanogaster. This fruit fly is used as a model organism in the study of genetics, development and disease for long ago. All initial data are taken from the
| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|------------|----------|----------|----------|---------|
| P₁(A₁) = 0.2868 | P₂(A₁) = 0.2867 | P₃(A₁) = 0.2871 | P₄(A₁) = 0.2868 | P₅(A₁) = 0.2871 |
| P₁(T₁) = 0.2886 | P₂(T₁) = 0.2885 | P₃(T₁) = 0.2885 | P₄(T₁) = 0.2884 | P₅(T₁) = 0.2887 |
| P₁(C₁) = 0.212 | P₂(C₁) = 0.212 | P₃(C₁) = 0.2119 | P₄(C₁) = 0.212 | P₅(C₁) = 0.2119 |
| P₁(G₁) = 0.2127 | P₂(G₁) = 0.2127 | P₃(G₁) = 0.2124 | P₄(G₁) = 0.2128 | P₅(G₁) = 0.2124 |
| P₂(A₂) = 0.2869 | P₃(A₂) = 0.2868 | P₄(A₂) = 0.2868 | P₅(A₂) = 0.2868 |
| P₂(T₂) = 0.2886 | P₃(T₂) = 0.2889 | P₄(T₂) = 0.2887 | P₅(T₂) = 0.2885 |
| P₂(C₂) = 0.2119 | P₃(C₂) = 0.2116 | P₄(C₂) = 0.212 | P₅(C₂) = 0.2118 |
| P₂(G₂) = 0.2126 | P₃(G₂) = 0.2127 | P₄(G₂) = 0.2124 | P₅(G₂) = 0.213 |
| P₃(A₃) = 0.2864 | P₄(A₃) = 0.2867 | P₅(A₃) = 0.2866 |
| P₃(T₃) = 0.2883 | P₄(T₃) = 0.2887 | P₅(T₃) = 0.2887 |
| P₃(C₃) = 0.2123 | P₄(C₃) = 0.212 | P₅(C₃) = 0.212 |
| P₃(G₃) = 0.2129 | P₄(G₃) = 0.2126 | P₅(G₃) = 0.2127 |
| P₄(A₄) = 0.2869 | P₅(A₄) = 0.2865 |
| P₄(T₄) = 0.2885 | P₅(T₄) = 0.2886 |
| P₄(C₄) = 0.2117 | P₅(C₄) = 0.2121 |
| P₄(G₄) = 0.2128 | P₅(G₄) = 0.2129 |
| P₅(A₅) = 0.287 |
| P₅(T₅) = 0.2885 |
| P₅(C₅) = 0.212 |
| P₅(G₅) = 0.2125 |

Fig. A4/1. The table of probabilities of subgroups of tetra-groups in the sequence: Drosophila melanogaster chromosome X, NCBI Reference Sequence: NC_004354.4

LOCUS NC_004354 23542271 bp DNA linear CON 14-MAR-2017
DEFINITION Drosophila melanogaster chromosome X.ACESSION NC_004354 NW_001844882 NW_001845036 NW_001848859 VERSION NC_004354.4
https://www.ncbi.nlm.nih.gov/nuccore/NC_004354.4
| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P₁(A₁)     | 0.2915   | 0.2914   | 0.2915  | 0.2915  |
| P₁(T₁)     | 0.2907   | 0.2908   | 0.2911  | 0.2909  |
| P₁(C₁)     | 0.2089   | 0.2089   | 0.2089  | 0.2087  |
| P₁(G₁)     | 0.2089   | 0.209    | 0.2085  | 0.2089  |
| P₂(A₂)     | 0.2915   | 0.2915   | 0.2915  | 0.2915  |
| P₂(T₂)     | 0.2907   | 0.2912   | 0.2909  | 0.2903  |
| P₂(C₂)     | 0.2089   | 0.2085   | 0.209   | 0.2089  |
| P₂(G₂)     | 0.2089   | 0.2087   | 0.2086  | 0.2093  |
| P₃(A₃)     | 0.2913   | 0.2916   | 0.2914  |         |
| P₃(T₃)     | 0.2899   | 0.2907   | 0.291   |         |
| P₃(C₃)     | 0.2092   | 0.2087   | 0.2088  |         |
| P₃(G₃)     | 0.2096   | 0.209    | 0.2088  |         |
| P₄(A₄)     | 0.2915   | 0.2915   | 0.2916  |         |
| P₄(T₄)     | 0.2905   | 0.2907   | 0.2907  |         |
| P₄(C₄)     | 0.2089   | 0.2091   | 0.2091  |         |
| P₄(G₄)     | 0.2091   | 0.2089   | 0.2089  |         |
| P₅(A₅)     | 0.2916   | 0.2916   | 0.2916  |         |
| P₅(T₅)     | 0.2908   | 0.2908   |         |         |
| P₅(C₅)     |          | 0.2090   |         |         |
| P₅(G₅)     |          | 0.2086   |         |         |

Fig. A4/2. The table of probabilities of subgroups of tetra-groups in the sequence: *Drosophila melanogaster* chromosome 2L, NCBI Reference Sequence: NT_033779.5

**LOCUS**  NT_033779   23513712 bp   DNA   linear   CON 14-MAR-2017
**DEFINITION** Drosophila melanogaster chromosome 2L. ACCESSION NT_033779 NW_001848855 VERSION NT_033779.5
[https://www.ncbi.nlm.nih.gov/nuccore/NT_033779.5](https://www.ncbi.nlm.nih.gov/nuccore/NT_033779.5)
| NUCLEOTIDES | DOUBLETS  | TRIPLETS | 4-PLETS | 5-PLETS |
|------------|-----------|----------|---------|---------|
| P₁(A₁)    | 0.2877    | P₂(A₁)   | 0.2877  | P₃(A₁)  | 0.2874  | P₄(A₁)  | 0.2878  |
| P₁(T₁)    | 0.2862    | P₂(T₁)   | 0.2861  | P₃(T₁)  | 0.2859  | P₄(T₁)  | 0.2863  | P₅(T₁)  | 0.2866  |
| P₁(C₁)    | 0.2134    | P₂(C₁)   | 0.2134  | P₃(C₁)  | 0.2136  | P₄(C₁)  | 0.2133  | P₅(C₁)  | 0.2133  |
| P₁(G₁)    | 0.2127    | P₂(G₁)   | 0.2129  | P₃(G₁)  | 0.2128  | P₄(G₁)  | 0.213   | P₅(G₁)  | 0.2129  |
| P₂(A₂)    | 0.2877    | P₃(A₂)   | 0.2879  | P₄(A₂)  | 0.2878  | P₅(A₂)  | 0.2878  |
| P₂(T₂)    | 0.2863    | P₃(T₂)   | 0.2862  | P₄(T₂)  | 0.2864  | P₅(T₂)  | 0.2864  |
| P₂(C₂)    | 0.2135    | P₃(C₂)   | 0.2134  | P₄(C₂)  | 0.2133  | P₅(C₂)  | 0.2133  |
| P₂(G₂)    | 0.2125    | P₃(G₂)   | 0.2126  | P₄(G₂)  | 0.2126  | P₅(G₂)  | 0.2125  |
| P₃(A₃)    | 0.2875    | P₄(A₃)   | 0.2879  | P₅(A₃)  | 0.2878  |
| P₃(T₃)    | 0.2865    | P₄(T₃)   | 0.2859  | P₅(T₃)  | 0.2862  |
| P₃(C₃)    | 0.2133    | P₄(C₃)   | 0.2134  | P₅(C₃)  | 0.2136  |
| P₃(G₃)    | 0.2127    | P₄(G₃)   | 0.2128  | P₅(G₃)  | 0.2124  |
| P₄(A₄)    | 0.2876    | P₅(A₄)   | 0.2877  |
| P₄(T₄)    | 0.2863    | P₅(T₄)   | 0.2863  |
| P₄(C₄)    | 0.2137    | P₅(C₄)   | 0.2133  |
| P₄(G₄)    | 0.2124    | P₅(G₄)   | 0.2128  |
| P₅(A₅)    | 0.2874    | P₅(T₅)   | 0.2862  |
| P₅(C₅)    | 0.2135    |
| P₅(G₅)    | 0.2128    |

Fig. A4/3. The table of probabilities of subgroups of tetra-groups in the sequence: Drosophila melanogaster chromosome 2R,
NCBI Reference Sequence: NT_033778.4

LOCUS NT_033778  25286936 bp DNA linear CON 14-MAR-2017
DEFINITION Drosophila melanogaster chromosome 2R. ACCESSION NT_033778 NW_001844732 NW_001844738 NW_001848856 VERSION NT_033778.4
https://www.ncbi.nlm.nih.gov/nuccore/NT_033778.4
| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|------------|----------|----------|---------|---------|
| $P_1(A_1)$= 0.2909 | $P_2(A_1)$= 0.2909 | $P_3(A_1)$= 0.2909 | $P_4(A_1)$= 0.2908 | $P_5(A_1)$= 0.2909 |
| $P_1(T_1)$= 0.2929 | $P_2(T_1)$= 0.2928 | $P_3(T_1)$= 0.2924 | $P_4(T_1)$= 0.2929 | $P_5(T_1)$= 0.2929 |
| $P_1(C_1)$= 0.2081 | $P_2(C_1)$= 0.2082 | $P_3(C_1)$= 0.2084 | $P_4(C_1)$= 0.2083 | $P_5(C_1)$= 0.2079 |
| $P_1(G_1)$= 0.2081 | $P_2(G_1)$= 0.2080 | $P_3(G_1)$= 0.2083 | $P_4(G_1)$= 0.208 | $P_5(G_1)$= 0.2082 |
| $P_2(A_2)$= 0.2909 | $P_3(A_2)$= 0.2913 | $P_4(A_2)$= 0.2909 | $P_5(A_2)$= 0.291 |
| $P_2(T_2)$= 0.2929 | $P_3(T_2)$= 0.2933 | $P_4(T_2)$= 0.2931 | $P_5(T_2)$= 0.2928 |
| $P_2(C_2)$= 0.208 | $P_3(C_2)$= 0.208 | $P_4(C_2)$= 0.2078 | $P_5(C_2)$= 0.2082 |
| $P_2(G_2)$= 0.2081 | $P_3(G_2)$= 0.2074 | $P_4(G_2)$= 0.2082 | $P_5(G_2)$= 0.208 |
| $P_3(A_3)$= 0.2906 | $P_4(A_3)$= 0.291 | $P_5(A_3)$= 0.291 |
| $P_3(T_3)$= 0.2929 | $P_4(T_3)$= 0.2927 | $P_5(T_3)$= 0.2927 |
| $P_3(C_3)$= 0.208 | $P_4(C_3)$= 0.2082 | $P_5(C_3)$= 0.2082 |
| $P_3(G_3)$= 0.2085 | $P_4(G_3)$= 0.2081 | $P_5(G_3)$= 0.2081 |
| $P_4(A_4)$= 0.291 | $P_5(A_4)$= 0.2911 |
| $P_4(T_4)$= 0.2927 | $P_5(T_4)$= 0.2928 |
| $P_4(C_4)$= 0.2083 | $P_5(C_4)$= 0.2081 |
| $P_4(G_4)$= 0.208 | $P_5(G_4)$= 0.208 |
| $P_5(A_5)$= 0.2906 |
| $P_5(T_5)$= 0.2932 |
| $P_5(C_5)$= 0.2082 |
| $P_5(G_5)$= 0.208 |

Fig. A4/4. The table of probabilities of subgroups of tetra-groups in the sequence: Drosophila melanogaster chromosome 3L, NCBI Reference Sequence: NT_037436.4

LOCUS NT_037436 28110227 bp DNA linear CON 14-MAR-2017
DEFINITION Drosophila melanogaster chromosome 3L. ACCESSION NT_037436 NW_001844849 NW_001848857 VERSION NT_037436.4
https://www.ncbi.nlm.nih.gov/nuccore/NT_037436.4
| NUCLEOTIDES | DOUBLETS | TRIPLET S | 4-PLETS | 5-PLETS |
|-------------|-----------|-----------|----------|---------|
| P₁(A₁) = 0.2872 | P₂(A₁) = 0.2873 | P₃(A₁) = 0.2874 | P₄(A₁) = 0.2872 | P₅(A₁) = 0.2872 |
| P₁(T₁) = 0.2869 | P₂(T₁) = 0.2869 | P₃(T₁) = 0.2871 | P₄(T₁) = 0.2869 | P₅(T₁) = 0.287 |
| P₁(C₁) = 0.2132 | P₂(C₁) = 0.2131 | P₃(C₁) = 0.2131 | P₄(C₁) = 0.2132 | P₅(C₁) = 0.2133 |
| P₁(G₁) = 0.2127 | P₂(G₁) = 0.2127 | P₃(G₁) = 0.2125 | P₄(G₁) = 0.2127 | P₅(G₁) = 0.2125 |
| P₂(A₂) = 0.2871 | P₃(A₂) = 0.2871 | P₄(A₂) = 0.2872 | P₅(A₂) = 0.2872 |
| P₂(T₂) = 0.287 | P₃(T₂) = 0.287 | P₄(T₂) = 0.2868 | P₅(T₂) = 0.2868 |
| P₂(C₂) = 0.2132 | P₃(C₂) = 0.2131 | P₄(C₂) = 0.2133 | P₅(C₂) = 0.213 |
| P₂(G₂) = 0.2127 | P₃(G₂) = 0.2129 | P₄(G₂) = 0.2127 | P₅(G₂) = 0.2129 |
| P₃(A₃) = 0.2871 | P₄(A₃) = 0.2873 | P₅(A₃) = 0.2874 |
| P₃(T₃) = 0.2867 | P₄(T₃) = 0.2869 | P₅(T₃) = 0.287 |
| P₃(C₃) = 0.2135 | P₄(C₃) = 0.2131 | P₅(C₃) = 0.2132 |
| P₃(G₃) = 0.2127 | P₄(G₃) = 0.2127 | P₅(G₃) = 0.2124 |
| P₄(A₄) = 0.287 | P₅(A₄) = 0.2869 |
| P₄(T₄) = 0.2871 | P₅(T₄) = 0.287 |
| P₄(C₄) = 0.2132 | P₅(C₄) = 0.2133 |
| P₄(G₄) = 0.2127 | P₅(G₄) = 0.2128 |
| P₅(A₅) = 0.2872 |
| P₅(T₅) = 0.2869 |
| P₅(C₅) = 0.2132 |
| P₅(G₅) = 0.2128 |

Fig. A4/5. The table of probabilities of subgroups of tetra-groups in the sequence: Drosophila melanogaster chromosome 3R, NCBI Reference Sequence: NT_033777.3

LOCUS NT_033777 32079331 bp DNA linear CON 14-MAR-2017
DEFINITION Drosophila melanogaster chromosome 3R. ACCESSION NT_033777 NW_001844733 NW_001844734 NW_001844736 NW_001844852 NW_001844855 NW_001844895 NW_001844858 VERSION NT_033777.3
https://www.ncbi.nlm.nih.gov/nuccore/NT_033777.3
| NUCLEOTIDES | DOUBLES | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|---------|----------|---------|---------|
| P₁(A₁) = 0.3195 | P₂(A₁) = 0.3191 | P₃(A₁) = 0.3199 | P₄(A₁) = 0.3184 | P₅(A₁) = 0.3218 |
| P₁(T₁) = 0.328  | P₂(T₁) = 0.3285 | P₃(T₁) = 0.3290 | P₄(T₁) = 0.3288 | P₅(T₁) = 0.3262 |
| P₁(C₁) = 0.1747 | P₂(C₁) = 0.175  | P₃(C₁) = 0.1734 | P₄(C₁) = 0.1752 | P₅(C₁) = 0.1747 |
| P₁(G₁) = 0.1778 | P₂(G₁) = 0.1774 | P₃(G₁) = 0.1777 | P₄(G₁) = 0.1775 | P₅(G₁) = 0.1773 |
|  | P₂(A₂) = 0.3198 | P₃(A₂) = 0.3197 | P₄(A₂) = 0.3199 | P₅(A₂) = 0.3193 |
|  | P₂(T₂) = 0.3276 | P₃(T₂) = 0.3273 | P₄(T₂) = 0.3286 | P₅(T₂) = 0.3292 |
|  | P₂(C₂) = 0.1744 | P₃(C₂) = 0.1747 | P₄(C₂) = 0.1739 | P₅(C₂) = 0.1744 |
|  | P₂(G₂) = 0.1781 | P₃(G₂) = 0.1784 | P₄(G₂) = 0.1776 | P₅(G₂) = 0.1771 |
|  |  | P₃(A₃) = 0.3188 | P₄(A₃) = 0.3198 | P₅(A₃) = 0.3174 |
|  |  | P₃(T₃) = 0.3279 | P₄(T₃) = 0.3281 | P₅(T₃) = 0.3296 |
|  |  | P₃(C₃) = 0.1760 | P₄(C₃) = 0.1747 | P₅(C₃) = 0.1744 |
|  |  | P₃(G₃) = 0.1773 | P₄(G₃) = 0.1774 | P₅(G₃) = 0.1786 |
|  |  | P₄(A₄) = 0.3197 | P₅(A₄) = 0.3194 |
|  |  | P₄(T₄) = 0.3266 | P₅(T₄) = 0.3275 |
|  |  | P₄(C₄) = 0.175  | P₅(C₄) = 0.1755 |
|  |  | P₄(G₄) = 0.1787 | P₅(G₄) = 0.1776 |
|  |  | P₅(A₅) = 0.3193 |
|  |  | P₅(T₅) = 0.3278 |
|  |  | P₅(C₅) = 0.1746 |
|  |  | P₅(G₅) = 0.1783 |

Fig. A4/6. The table of probabilities of subgroups of tetra-groups in the sequence: Drosophila melanogaster chromosome 4, NCBI Reference Sequence: NC_004353.4

LOCUS NC_004353 1348131 bp DNA linear CON 14-MAR-2017
DEFINITION Drosophila melanogaster chromosome 4. ACCESSION NC_004353
NW_001845041 VERSION NC_004353.4
https://www.ncbi.nlm.nih.gov/nuccore/NC_004353.4
### Nucleotides

| Nucleotides | Doublets | Triplets | 4-Plets | 5-Plets |
|-------------|----------|----------|---------|---------|
| P₁(A₁) = 0.3099 | P₂(A₁) = 0.3098 | P₃(A₁) = 0.3104 | P₄(A₁) = 0.3102 | P₅(A₁) = 0.3105 |
| P₁(T₁) = 0.2958 | P₂(T₁) = 0.2960 | P₃(T₁) = 0.2955 | P₄(T₁) = 0.2959 | P₅(T₁) = 0.2956 |
| P₁(C₁) = 0.2002 | P₂(C₁) = 0.2003 | P₃(C₁) = 0.2002 | P₄(C₁) = 0.2001 | P₅(C₁) = 0.2003 |
| P₁(G₁) = 0.1940 | P₂(G₁) = 0.1939 | P₃(G₁) = 0.1938 | P₄(G₁) = 0.1938 | P₅(G₁) = 0.1936 |
| P₂(A₂) = 0.3100 | P₃(A₂) = 0.3098 | P₄(A₂) = 0.3095 | P₅(A₂) = 0.31 |
| P₂(T₂) = 0.2957 | P₃(T₂) = 0.2954 | P₄(T₂) = 0.2961 | P₅(T₂) = 0.2957 |
| P₂(C₂) = 0.2002 | P₃(C₂) = 0.2003 | P₄(C₂) = 0.2000 | P₅(C₂) = 0.2004 |
| P₂(G₂) = 0.1941 | P₃(G₂) = 0.1945 | P₄(G₂) = 0.1944 | P₅(G₂) = 0.1939 |
| P₃(A₃) = 0.3096 | P₄(A₃) = 0.3094 | P₅(A₃) = 0.3099 |
| P₃(T₃) = 0.2965 | P₄(T₃) = 0.2961 | P₅(T₃) = 0.2956 |
| P₃(C₃) = 0.2001 | P₄(C₃) = 0.2004 | P₅(C₃) = 0.2006 |
| P₃(G₃) = 0.1938 | P₄(G₃) = 0.1941 | P₅(G₃) = 0.194 |
| P₄(A₄) = 0.3105 | P₅(A₄) = 0.31 |
| P₄(T₄) = 0.2953 | P₅(T₄) = 0.2962 |
| P₄(C₄) = 0.2004 | P₅(C₄) = 0.1992 |
| P₄(G₄) = 0.1938 | P₅(G₄) = 0.1946 |
| P₅(A₅) = 0.3093 |
| P₅(T₅) = 0.2960 |
| P₅(C₅) = 0.2007 |
| P₅(G₅) = 0.1941 |

Fig. A4/7. The table of probabilities of subgroups of tetra-groups in the sequence: Drosophila melanogaster chromosome Y, NCBI Reference Sequence: NC_024512.1

**LOCUS** NC_024512 3667352 bp DNA linear CON 14-MAR-2017
**DEFINITION** Drosophila melanogaster chromosome Y. ACCESSION NC_024512
NW_001844735 NW_001844843 NW_001844868 NW_001844869 NW_001844874 NW_001844917 NW_001845004 NW_001845447 NW_001845711 NW_001845825 NW_001846080 NW_001846110 NW_001847094 NW_001848860 VERSION NC_024512.1
[https://www.ncbi.nlm.nih.gov/nuccore/NC_024512.1](https://www.ncbi.nlm.nih.gov/nuccore/NC_024512.1)
Appendix 5. Symmetries of tetra-group probabilities in the complete set of chromosomes of *Arabidopsis Thaliana*

The appendix 5 represents data about the fulfillment of the described tetra-group rules in the complete set of chromosomes of *Arabidopsis Thaliana*. This small flowering plant has been used for over fifty years to study plant mutations and for classical genetic analysis. It became the first plant genome to be fully sequenced; it has a small genome of ~120 Mb. All initial data about chromosomes are taken from the CenBank (https://www.ncbi.nlm.nih.gov/)

| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P₄(A₁) = 0.3208 | P₂(A₂) = 0.3207 | P₃(A₂) = 0.3212 | P₄(A₃) = 0.3206 | P₅(A₁) = 0.3209 |
| P₃(T₁) = 0.3204 | P₃(T₁) = 0.3204 | P₃(T₁) = 0.3204 | P₄(T₁) = 0.3207 | P₅(T₁) = 0.3201 |
| P₃(C₁) = 0.1796 | P₂(C₁) = 0.1796 | P₃(C₁) = 0.1794 | P₄(C₁) = 0.1797 | P₅(C₁) = 0.1797 |
| P₃(G₁) = 0.1791 | P₂(G₁) = 0.1792 | P₃(G₁) = 0.1790 | P₄(G₁) = 0.179 | P₅(G₁) = 0.1793 |
| P₂(A₂) = 0.3209 | P₃(A₂) = 0.3206 | P₄(A₂) = 0.3209 | P₅(A₂) = 0.3209 |
| P₃(T₂) = 0.3204 | P₃(T₂) = 0.3207 | P₄(T₂) = 0.3207 | P₅(T₂) = 0.3207 |
| P₂(C₂) = 0.1796 | P₃(C₂) = 0.1798 | P₄(C₂) = 0.1797 | P₅(C₂) = 0.1795 |
| P₂(G₂) = 0.1791 | P₃(G₂) = 0.1789 | P₄(G₂) = 0.1791 | P₅(G₂) = 0.179 |
| P₃(A₃) = 0.3207 | P₄(A₃) = 0.3208 | P₅(A₃) = 0.3207 |
| P₃(T₃) = 0.3202 | P₄(T₃) = 0.3202 | P₅(T₃) = 0.3206 |
| P₃(C₃) = 0.1796 | P₄(C₃) = 0.1796 | P₅(C₃) = 0.1795 |
| P₃(G₃) = 0.1795 | P₄(G₃) = 0.1793 | P₅(G₃) = 0.1792 |
| P₄(A₄) = 0.321 | P₅(A₄) = 0.3211 |
| P₄(T₄) = 0.3205 | P₅(T₄) = 0.3204 |
| P₄(C₄) = 0.1794 | P₅(C₄) = 0.1794 |
| P₄(G₄) = 0.1791 | P₅(G₄) = 0.1791 |
| P₅(A₅) = 0.3206 |
| P₅(T₅) = 0.3204 |
| P₅(C₅) = 0.1799 |

Fig. A5/1. The table of probabilities of subgroups of tetra-groups in the sequence: *Arabidopsis thaliana* chromosome 1 sequence, NCBI Reference Sequence: NC_003070.9
| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P₁(A₁) = 0.3207 | P₂(A₁) = 0.3205 | P₃(A₁) = 0.3205 | P₄(A₁) = 0.3207 | P₅(A₁) = 0.3205 |
| P₁(T₁) = 0.3207 | P₂(T₁) = 0.3208 | P₃(T₁) = 0.321 | P₄(T₁) = 0.3208 | P₅(T₁) = 0.3205 |
| P₁(C₁) = 0.1799 | P₂(C₁) = 0.18 | P₃(C₁) = 0.1797 | P₄(C₁) = 0.1797 | P₅(C₁) = 0.1805 |
| P₁(G₁) = 0.1788 | P₂(G₁) = 0.1787 | P₃(G₁) = 0.1788 | P₄(G₁) = 0.1788 | P₅(G₁) = 0.1785 |

Fig. A5/2. The table of probabilities of subgroups of tetra-groups in the sequence: Arabidopsis thaliana chromosome 2 sequence, NCBI Reference Sequence: NC_003071.7

LOCUS NC_003070 30427671 bp DNA linear CON 20-MAR-2017
DEFINITION Arabidopsis thaliana chromosome 1 sequence. ACCESSION NC_003070 VERSION NC_003070.9
https://www.ncbi.nlm.nih.gov/nuccore/NC_003070.9

LOCUS NC_003071 19698289 bp DNA linear CON 20-MAR-2017
DEFINITION Arabidopsis thaliana chromosome 2 sequence. ACCESSION NC_003071 VERSION NC_003071.7
https://www.ncbi.nlm.nih.gov/nuccore/NC_003071.7
| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|------------|----------|----------|---------|---------|
| P_1(A)=   | 0.3191   | P_2(A)=  | 0.3194  | P_3(A)= | 0.3195  | P_4(A)= | 0.3192  | P_5(A)= | 0.3191  |
| P_1(T)=   | 0.3176   | P_2(T)=  | 0.3175  | P_3(T)= | 0.3171  | P_4(T)= | 0.3176  | P_5(T)= | 0.3175  |
| P_1(C)=   | 0.1816   | P_2(C)=  | 0.1815  | P_3(C)= | 0.1816  | P_4(C)= | 0.1815  | P_5(C)= | 0.1816  |
| P_1(G)=   | 0.1817   | P_2(G)=  | 0.1817  | P_3(G)= | 0.1818  | P_4(G)= | 0.1817  | P_5(G)= | 0.1818  |
| P_2(A)=   | 0.3189   | P_3(A)=  | 0.3192  | P_4(A)= | 0.319   | P_5(A)= | 0.3194  |
| P_2(T)=   | 0.3177   | P_3(T)=  | 0.3179  | P_4(T)= | 0.3175  | P_5(T)= | 0.3172  |
| P_2(C)=   | 0.1816   | P_3(C)=  | 0.1813  | P_4(C)= | 0.1816  | P_5(C)= | 0.1813  |
| P_2(G)=   | 0.1818   | P_3(G)=  | 0.1816  | P_4(G)= | 0.1819  | P_5(G)= | 0.182   |
| P_3(A)=   | 0.3187   | P_4(A)=  | 0.3195  | P_5(A)= | 0.3189  |
| P_3(T)=   | 0.3176   | P_4(T)=  | 0.3173  | P_5(T)= | 0.3178  |
| P_3(C)=   | 0.1818   | P_4(C)=  | 0.1816  | P_5(C)= | 0.1818  |
| P_3(G)=   | 0.1818   | P_4(G)=  | 0.1816  | P_5(G)= | 0.1816  |
| P_4(A)=   | 0.3189   | P_5(A)=  | 0.319   |
| P_4(T)=   | 0.3178   | P_5(T)=  | 0.3178  |
| P_4(C)=   | 0.1816   | P_5(C)=  | 0.1815  |
| P_4(G)=   | 0.1818   | P_5(G)=  | 0.1816  |
| P_5(A)=   | 0.3192   |
| P_5(T)=   | 0.3175   |
| P_5(C)=   | 0.1815   |
| P_5(G)=   | 0.1818   |

Fig. A5/3. The table of probabilities of subgroups of tetra-groups in the sequence: Arabidopsis thaliana chromosome 3 sequence, NCBI Reference Sequence: NC_003074.8
LOCUS NC_003074 23459830 bp DNA linear CON 20-MAR-2017
DEFINITION Arabidopsis thaliana chromosome 3 sequence. ACCESSION NC_003074 VERSION NC_003074.8
https://www.ncbi.nlm.nih.gov/nuccore/NC_003074.8?report=fasta
| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| $P_1(A_1)$  | 0.3197   | $P_2(A_1)$= 0.3197 | $P_3(A_1)$= 0.3198 | $P_4(A_1)$= 0.3195 | $P_5(A_1)$= 0.3199 |
| $P_1(T_1)$  | 0.3183   | $P_2(T_1)$= 0.3182 | $P_3(T_1)$= 0.318 | $P_4(T_1)$= 0.3183 | $P_5(T_1)$= 0.3183 |
| $P_1(C_1)$  | 0.1814 | $P_2(C_1)$= 0.1814 | $P_3(C_1)$= 0.1812 | $P_4(C_1)$= 0.1815 | $P_5(C_1)$= 0.1813 |
| $P_1(G_1)$  | 0.1806 | $P_2(G_1)$= 0.1807 | $P_3(G_1)$= 0.1810 | $P_4(G_1)$= 0.1807 | $P_5(G_1)$= 0.1806 |
| $P_2(A_2)$  | 0.3196 | $P_3(A_2)$= 0.3195 | $P_4(A_2)$= 0.3197 | $P_5(A_2)$= 0.3199 |
| $P_2(T_2)$  | 0.3183 | $P_3(T_2)$= 0.3183 | $P_4(T_2)$= 0.3183 | $P_5(T_2)$= 0.3184 |
| $P_2(C_2)$  | 0.1815 | $P_3(C_2)$= 0.1817 | $P_4(C_2)$= 0.1815 | $P_5(C_2)$= 0.1813 |
| $P_2(G_2)$  | 0.1805 | $P_3(G_2)$= 0.1805 | $P_4(G_2)$= 0.1805 | $P_5(G_2)$= 0.1804 |
| $P_3(A_3)$  | 0.3197 | $P_4(A_3)$= 0.32 | $P_5(A_3)$= 0.3195 |
| $P_3(T_3)$  | 0.3185 | $P_4(T_3)$= 0.3181 | $P_5(T_3)$= 0.3182 |
| $P_3(C_3)$  | 0.1814 | $P_4(C_3)$= 0.1813 | $P_5(C_3)$= 0.1817 |
| $P_3(G_3)$  | 0.1803 | $P_4(G_3)$= 0.1807 | $P_5(G_3)$= 0.1806 |
| $P_4(A_4)$  | 0.3196 | $P_5(A_4)$= 0.3197 |
| $P_4(T_4)$  | 0.3184 | $P_5(T_4)$= 0.3182 |
| $P_4(C_4)$  | 0.1815 | $P_5(C_4)$= 0.1813 |
| $P_4(G_4)$  | 0.1805 | $P_5(G_4)$= 0.1808 |
| $P_5(A_5)$  | 0.3196 |
| $P_5(T_5)$  | 0.3183 |
| $P_5(C_5)$  | 0.1816 |
| $P_5(G_5)$  | 0.1806 |

Fig. A5/4. The table of probabilities of subgroups of tetra-groups in the sequence: Arabidopsis thaliana chromosome 4 sequence, NCBI Reference Sequence: NC_003075.7

LOCUS NC_003075 18585056 bp DNA linear CON 20-MAR-2017
DEFINITION Arabidopsis thaliana chromosome 4 sequence. ACCESSION NC_003075 VERSION NC_003075.7

https://www.ncbi.nlm.nih.gov/nuccore/NC_003075.7
| NUCLEOTIDES | DOUBLETS | TRIPLETS | 4-PLETS | 5-PLETS |
|-------------|----------|----------|---------|---------|
| P₁(A₁) = 0.3197 | P₂(A₁) = 0.3197 | P₃(A₁) = 0.3196 | P₄(A₁) = 0.3195 | P₅(A₁) = 0.3195 |
| P₁(T₁) = 0.3209 | P₂(T₁) = 0.3208 | P₃(T₁) = 0.321 | P₄(T₁) = 0.3209 | P₅(T₁) = 0.3208 |
| P₁(C₁) = 0.1792 | P₂(C₁) = 0.1793 | P₃(C₁) = 0.1796 | P₄(C₁) = 0.1794 | P₅(C₁) = 0.1796 |
| P₁(G₁) = 0.1802 | P₂(G₁) = 0.1802 | P₃(G₁) = 0.1797 | P₄(G₁) = 0.1802 | P₅(G₁) = 0.1802 |
| P₂(A₂) = 0.3198 | P₃(A₂) = 0.3199 | P₄(A₂) = 0.3198 | P₅(A₂) = 0.3198 |
| P₂(T₂) = 0.3209 | P₃(T₂) = 0.3206 | P₄(T₂) = 0.3207 | P₅(T₂) = 0.3208 |
| P₂(C₂) = 0.1791 | P₃(C₂) = 0.179 | P₄(C₂) = 0.1793 | P₅(C₂) = 0.1792 |
| P₂(G₂) = 0.1802 | P₃(G₂) = 0.1805 | P₄(G₂) = 0.1802 | P₅(G₂) = 0.1802 |
| P₃(A₃) = 0.3198 | P₄(A₃) = 0.3199 | P₅(A₃) = 0.3201 |
| P₃(T₃) = 0.321 | P₄(T₃) = 0.3208 | P₅(T₃) = 0.3208 |
| P₃(C₃) = 0.179 | P₄(C₃) = 0.1792 | P₅(C₃) = 0.179 |
| P₃(G₃) = 0.1803 | P₄(G₃) = 0.1802 | P₅(G₃) = 0.1801 |
| P₄(A₄) = 0.3199 | P₅(A₄) = 0.3198 |
| P₄(T₄) = 0.3211 | P₅(T₄) = 0.321 |
| P₄(C₄) = 0.1789 | P₅(C₄) = 0.1791 |
| P₄(G₄) = 0.1801 | P₅(G₄) = 0.1802 |
| P₅(A₅) = 0.3196 |
| P₅(T₅) = 0.3211 |
| P₅(C₅) = 0.1791 |
| P₅(G₅) = 0.1802 |

Fig. A5/5. The table of probabilities of subgroups of tetra-groups in the sequence: Arabidopsis thaliana chromosome 5 sequence, NCBI Reference Sequence: NC_003076.8

LOCUS NC_003076 26975502 bp DNA  linear  CON 20-MAR-2017
DEFINITION Arabidopsis thaliana chromosome 5 sequence. ACCESSION NC_003076 VERSION NC_003076.8
https://www.ncbi.nlm.nih.gov/nuccore/NC_003076.8
Appendix 6. Symmetries of tetra-group probabilities in the complete set of chromosomes of Mus musculus

The appendix 6 represents data about the fulfillment of the described tetra-group rules in the complete set of chromosomes of Mus musculus (house mouse). The laboratory mouse is a major model organism for basic mammalian biology, human disease, and genome evolution. All initial data are taken from the CenBank [https://www.ncbi.nlm.nih.gov/genome?term=mus%20musculus]. To show the fulfillment of the tetra-group rules of symmetries in texts at different initial levels of positional convolutions in the fractal genetic trees FGT-2, FGT-3, etc., tables of fluctuation intervals for values of probabilities for the set of all texts at each of levels for all chromosomes are given below (see the Section 11 about fractal genetic trees FGT-2, FGT-3, etc. and also about tables of fluctuation intervals). The fluctuation intervals of the probabilities $P_n(A_k), P_n(T_k), P_n(C_k)$ and $P_n(G_k)$ for each initial long DNA-text are shown in the columns «Level 0».

| Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|---------|-----------|-----------|------------|------------|------------|------------|
| $P_n(A_k)$ | $0.2945 \div 0.2946$ | $0.2945 \div 0.2948$ | $0.2944 \div 0.2945$ | $0.2944 \div 0.2945$ | $0.2944 \div 0.2948$ | $0.2944 \div 0.2945$ |
| $P_n(T_k)$ | $0.2939 \div 0.294$ | $0.2939 \div 0.2939$ | $0.2938 \div 0.2941$ | $0.2937 \div 0.294$ | $0.2937 \div 0.2938$ | $0.2937 \div 0.2942$ |
| $P_n(C_k)$ | $0.2057 \div 0.2058$ | $0.2057 \div 0.2057$ | $0.2056 \div 0.2058$ | $0.2056 \div 0.2058$ | $0.2056 \div 0.2057$ | $0.2056 \div 0.2058$ |
| $P_n(G_k)$ | $0.2055 \div 0.2057$ | $0.2055 \div 0.2057$ | $0.2056 \div 0.2056$ | $0.2056 \div 0.2054$ | $0.2056 \div 0.2057$ | $0.2056 \div 0.2056$ |

| Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 |
|---------|-----------|-----------|------------|------------|
| $P_n(A_k)$ | $0.2945 \div 0.2946$ | $0.2943 \div 0.2945$ | $0.2945 \div 0.2946$ | $0.2942 \div 0.2942$ |
| $P_n(T_k)$ | $0.2938 \div 0.294$ | $0.2938 \div 0.2938$ | $0.2938 \div 0.2938$ | $0.2937 \div 0.2942$ |
| $P_n(C_k)$ | $0.2057 \div 0.2058$ | $0.2057 \div 0.2057$ | $0.2057 \div 0.2058$ | $0.2057 \div 0.2059$ |
| $P_n(G_k)$ | $0.2055 \div 0.2057$ | $0.2055 \div 0.2057$ | $0.2055 \div 0.2058$ | $0.2054 \div 0.2054$ |
| Level 0            | Level 1/0   | Level 1/1   | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|-------------------|-------------|-------------|------------|------------|------------|------------|
| $P_n(A_k) \in$    | 0.2892÷0.2945 | 0.2893÷0.2949 | 0.2894÷0.2947 | 0.2894÷0.2946 | 0.2943÷0.2946 | 0.2943÷0.2944 |
| $P_n(T_k) \in$    | 0.2897÷0.2943 | 0.2897÷0.2943 | 0.2896÷0.2946 | 0.2896÷0.2947 | 0.2896÷0.2948 | 0.2897÷0.2949 |
| $P_n(C_k) \in$    | 0.2102÷0.2103 | 0.2102÷0.2103 | 0.2102÷0.2103 | 0.2102÷0.2103 | 0.2102÷0.2103 | 0.2102÷0.2103 |
| $P_n(G_k) \in$    | 0.2104÷0.2105 | 0.2104÷0.2105 | 0.2104÷0.2104 | 0.2104÷0.2105 | 0.2104÷0.2105 | 0.2104÷0.2105 |

Fig. A6/1. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 1, GRChm38.p4 C57BL/6J. NCBI Reference Sequence: NC_000067.6. 195471971 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000067.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000067.6). Here $n = 1, 2, 3, 4, 5$, $k \leq n$. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
| Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| $P_n(A_k)\in$ | $0.289/\pm0.2894$ | $0.2892/\pm0.2896$ | $0.2889/\pm0.2896$ | $0.2889/\pm0.2896$ | $0.2889/\pm0.2896$ | $0.2889/\pm0.2896$ |
| $P_n(T_k)\in$ | $0.2896/\pm0.2898$ | $0.2894/\pm0.2894$ | $0.2894/\pm0.2896$ | $0.2894/\pm0.2896$ | $0.2894/\pm0.2896$ | $0.2894/\pm0.2896$ |
| $P_n(C_k)\in$ | $0.2099/\pm0.2102$ | $0.21/\pm0.2105$ | $0.2101/\pm0.2105$ | $0.2101/\pm0.2105$ | $0.2101/\pm0.2105$ | $0.2101/\pm0.2105$ |
| $P_n(G_k)\in$ | $0.2104/\pm0.2106$ | $0.2102/\pm0.2105$ | $0.2105/\pm0.2105$ | $0.2105/\pm0.2105$ | $0.2105/\pm0.2105$ | $0.2101/\pm0.2105$ |

Fig. A6/2. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6j chromosome 2, GRCm38.p4 C57BL/6j, NCBI Reference Sequence: NC_000068.7. 182113224 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000068.7](https://www.ncbi.nlm.nih.gov/nuccore/NC_000068.7). Here $n = 1, 2, 3, 4, 5, k \leq n$. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
Level 2/02  Level 2/10  Level 2/11  Level 2/12  Level 2/20  Level 2/21  Level 2/22
\[ P_n(A_k) \in [0.2971, 0.2973] \]

Level 2/02  Level 2/01  Level 2/10  Level 2/11  Level 2/20  Level 2/21  Level 2/22
\[ P_n(T_k) \in [0.2978, 0.2985] \]

Level 2/02  Level 2/10  Level 2/11  Level 2/12  Level 2/20  Level 2/21  Level 2/22
\[ P_n(C_k) \in [0.202, 0.2023] \]

Level 2/02  Level 2/10  Level 2/11  Level 2/12  Level 2/20  Level 2/21  Level 2/22
\[ P_n(G_k) \in [0.2021, 0.2028] \]

Fig. A6/3. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 3, GRChm38.p4 C57BL/6J. NCBI Reference Sequence: NC_000069.6. 160039680 bp. https://www.ncbi.nlm.nih.gov/nuccore/NC_000069.6. Here \( n = 1, 2, 3, 4, 5 \), \( k \leq n \). Level 0 corresponds to the initial DNA-text with its different \( n \)-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
| Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| \( P_n(A_k) \in [0.2872, 0.2873] \) | \( 0.2872 \) | \( 0.2873 \) | \( 0.2872 \) | \( 0.2873 \) | \( 0.2872 \) | \( 0.2873 \) |
| \( P_n(T_k) \in [0.2873, 0.2874] \) | \( 0.2873 \) | \( 0.2874 \) | \( 0.2873 \) | \( 0.2874 \) | \( 0.2873 \) | \( 0.2874 \) |
| \( P_n(C_k) \in [0.2126, 0.2127] \) | \( 0.2126 \) | \( 0.2127 \) | \( 0.2126 \) | \( 0.2127 \) | \( 0.2126 \) | \( 0.2127 \) |
| \( P_n(G_k) \in [0.2125, 0.2126] \) | \( 0.2125 \) | \( 0.2126 \) | \( 0.2125 \) | \( 0.2126 \) | \( 0.2125 \) | \( 0.2126 \) |

Fig. A6/4. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 4, GRCh38.p4 C57BL/6J. NCBI Reference Sequence: NC_000070.6. 156508116 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000070.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000070.6). Here \( n = 1, 2, 3, 4, 5 \), \( k \leq n \). Level 0 corresponds to the initial DNA-text with its different \( n \)-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
Fig. A6/5. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 5, GRCh38.p4 C57BL/6J. NCBI Reference Sequence: NC_000071.6. 151834684 bp. https://www.ncbi.nlm.nih.gov/nuccore/NC_000071.6. Here $n = 1, 2, 3, 4, 5, k \leq n$. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.

| Level  | Level  | Level  | Level  | Level  | Level  | Level  |
|-------|-------|-------|-------|-------|-------|-------|
|       | 2/02  | 2/10  | 2/11  | 2/12  | 2/20  | 2/21  | 2/22  |
| $P_n(A_k)$ | \(0.2868\)÷\(0.2871\) | \(0.287\)÷\(0.2876\) | \(0.2867\)÷\(0.2872\) | \(0.287\)÷\(0.2873\) | \(0.287\)÷\(0.2871\) | \(0.2869\)÷\(0.2872\) | \(0.287\)÷\(0.2871\) |
| $P_n(T_k)$ | \(0.2868\)÷\(0.2874\) | \(0.2869\)÷\(0.2875\) | \(0.2871\)÷\(0.2879\) | \(0.287\)÷\(0.2872\) | \(0.2873\)÷\(0.2876\) | \(0.287\)÷\(0.2872\) | \(0.287\)÷\(0.288\) |
| $P_n(C_k)$ | \(0.2125\)÷\(0.2131\) | \(0.2123\)÷\(0.2124\) | \(0.2124\)÷\(0.2124\) | \(0.2124\)÷\(0.2128\) | \(0.2124\)÷\(0.2127\) | \(0.2123\)÷\(0.2127\) | \(0.2123\)÷\(0.2126\) |
| $P_n(G_k)$ | \(0.2121\)÷\(0.2123\) | \(0.2125\)÷\(0.2125\) | \(0.2125\)÷\(0.2125\) | \(0.2125\)÷\(0.2127\) | \(0.2123\)÷\(0.2127\) | \(0.2124\)÷\(0.2127\) | \(0.2121\)÷\(0.2123\) |

| Level  | Level  | Level  | Level  | Level  | Level  | Level  |
|-------|-------|-------|-------|-------|-------|-------|
|       | 0/0   | 1/0   | 1/1   | 2/00  | 2/01  | 2/10  |
| $P_n(A_k)$ | \(0.2927\)÷\(0.2928\) | \(0.2927\)÷\(0.2927\) | \(0.2926\)÷\(0.2929\) | \(0.2927\)÷\(0.2928\) | \(0.2926\)÷\(0.2927\) | \(0.2925\)÷\(0.2928\) |
| $P_n(T_k)$ | \(0.293\)÷\(0.293\) | \(0.293\)÷\(0.293\) | \(0.2929\)÷\(0.2929\) | \(0.2929\)÷\(0.2929\) | \(0.2929\)÷\(0.2929\) | \(0.2928\)÷\(0.2932\) |
| $P_n(C_k)$ | \(0.2071\)÷\(0.2072\) | \(0.207\)÷\(0.2072\) | \(0.2069\)÷\(0.2069\) | \(0.2066\)÷\(0.2071\) | \(0.207\)÷\(0.2072\) | \(0.207\)÷\(0.2071\) |
| $P_n(G_k)$ | \(0.2069\)÷\(0.207\) | \(0.2068\)÷\(0.207\) | \(0.2069\)÷\(0.2071\) | \(0.2067\)÷\(0.2072\) | \(0.2069\)÷\(0.2072\) | \(0.2069\)÷\(0.2072\) |
Fig. A6/6. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 6, GRCm38.p4 C57BL/6J. NCBI Reference Sequence: NC_000072.6. 149736546 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000072.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000072.6). Here \( n = 1, 2, 3, 4, 5, k \leq n \). Level 0 corresponds to the initial DNA-text with its different \( n \)-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.

| Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|------------|------------|------------|------------|------------|------------|------------|
| \( P_n(A_k) \) ∈ | 0.2925+ 0.2926 | 0.2925+ 0.2929 | 0.2925+ 0.2927 | 0.2925+ 0.2925 | 0.2927+ 0.2928 | 0.2925+ 0.293 | 0.2923+ 0.2927 |
| \( P_n(T_k) \) ∈ | 0.293+ 0.293 | 0.2926+ 0.2929 | 0.2926+ 0.2927 | 0.2928+ 0.2929 | 0.2928+ 0.2933 | 0.2929+ 0.2933 | 0.2926+ 0.2929 |
| \( P_n(C_k) \) ∈ | 0.2069+ 0.2071 | 0.2069+ 0.2073 | 0.2071+ 0.2072 | 0.207+ 0.2073 | 0.207+ 0.2073 | 0.2069+ 0.2069 | 0.2068+ 0.2073 |
| \( P_n(G_k) \) ∈ | 0.2068+ 0.2074 | 0.2067+ 0.2069 | 0.2069+ 0.2074 | 0.2069+ 0.2073 | 0.2064+ 0.2067 | 0.2066+ 0.2068 | 0.2068+ 0.207 |

| Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|----------|----------|----------|----------|----------|----------|----------|
| \( P_n(A_k) \) ∈ | 0.2837+ 0.284 | 0.2839+ 0.284 | 0.2836+ 0.2839 | 0.2838+ 0.2838 | 0.2839+ 0.2841 | 0.2835+ 0.2836 | 0.2837+ 0.2838 |
| \( P_n(T_k) \) ∈ | 0.2855+ 0.2855 | 0.2854+ 0.2854 | 0.2855+ 0.2855 | 0.2853+ 0.2855 | 0.2853+ 0.2854 | 0.2854+ 0.2857 | 0.2854+ 0.2857 |
| \( P_n(C_k) \) ∈ | 0.2152+ 0.2153 | 0.2152+ 0.2155 | 0.2153+ 0.2153 | 0.2153+ 0.2157 | 0.215+ 0.2154 | 0.2153+ 0.2155 | 0.2152+ 0.2153 |
| \( P_n(G_k) \) ∈ | 0.215+ 0.2151 | 0.215+ 0.2151 | 0.215+ 0.2153 | 0.2149+ 0.215 | 0.215+ 0.2152 | 0.2148+ 0.2153 | 0.2151+ 0.2151 |

| Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 |
|----------|----------|----------|----------|----------|
| \( P_n(A_k) \) ∈ | 0.2837+ 0.284 | 0.2838+ 0.2839 | 0.2836+ 0.2836 | 0.2836+ 0.284 | 0.2837+ 0.2837 | 0.2837+ 0.2839 |
| \( P_n(T_k) \) ∈ | 0.2855+ 0.2855 | 0.2853+ 0.2855 | 0.2853+ 0.2856 | 0.2855+ 0.2856 | 0.2855+ 0.2856 | 0.2853+ 0.2857 |
| \( P_n(C_k) \) ∈ | 0.2152+ 0.2153 | 0.2152+ 0.2154 | 0.2153+ 0.2156 | 0.215+ 0.2152 | 0.215+ 0.2156 | 0.215+ 0.2152 |
| \( P_n(G_k) \) ∈ | 0.215+ 0.2151 | 0.215+ 0.2152 | 0.215+ 0.2151 | 0.2149+ 0.2152 | 0.2149+ 0.2152 | 0.2148+ 0.2151 |
Fig. A6/7. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6j chromosome 7, GRCm38.p4 C57BL/6j. NCBI Reference Sequence: NC_000073.6. 145441459 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000073.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000073.6). Here $n = 1, 2, 3, 4, 5, k \leq n$. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.

| Level  | Level  | Level  | Level  | Level  | Level  | Level  |
|--------|--------|--------|--------|--------|--------|--------|
|        | 2/02   | 2/10   | 2/11   | 2/12   | 2/20   | 2/21   | 2/22   |
| $P_n(A_k)$ | 0.2834÷0.2834 | 0.2836÷0.2837 | 0.2833÷0.284 | 0.2834÷0.2834 | 0.2834÷0.2839 | 0.2835÷0.2836 | 0.2836÷0.2839 |
| $P_n(T_k)$ | 0.2852÷0.2859 | 0.2853÷0.2858 | 0.285÷0.2852 | 0.2853÷0.2858 | 0.2856÷0.2856 | 0.2855÷0.2857 | 0.2853÷0.2856 |
| $P_n(C_k)$ | 0.2149÷0.2155 | 0.2151÷0.2156 | 0.2151÷0.2156 | 0.2149÷0.2157 | 0.2151÷0.2153 | 0.2149÷0.2154 |
| $P_n(G_k)$ | 0.2149÷0.2152 | 0.2148÷0.2153 | 0.2148÷0.2153 | 0.2147÷0.2154 | 0.2148÷0.2154 | 0.2147÷0.2151 |

| Level  | Level  | Level  | Level  | Level  | Level  | Level  |
|--------|--------|--------|--------|--------|--------|--------|
|        | 1/0    | 1/1    | 1/2    | 2/00   | 2/01   | 2/10   | 2/11   |
| $P_n(A_k)$ | 0.2883÷0.2884 | 0.2882÷0.2884 | 0.2882÷0.2885 | 0.2882÷0.2883 | 0.288÷0.288 | 0.2882÷0.2885 | 0.2882÷0.2885 |
| $P_n(T_k)$ | 0.2878÷0.2879 | 0.2877÷0.2879 | 0.2877÷0.288 | 0.2878÷0.2878 | 0.2874÷0.288 | 0.2877÷0.2881 | 0.2876÷0.288 |
| $P_n(C_k)$ | 0.2118÷0.2119 | 0.2118÷0.2119 | 0.2117÷0.2117 | 0.2117÷0.2122 | 0.2119÷0.212 | 0.2116÷0.212 | 0.2116÷0.2118 |
| $P_n(G_k)$ | 0.2117÷0.2119 | 0.2117÷0.2118 | 0.2115÷0.2118 | 0.2115÷0.2117 | 0.2116÷0.212 | 0.2114÷0.2115 | 0.2116÷0.2117 |
| Level  | Level  | Level  | Level  | Level  | Level  | Level  |
|--------|--------|--------|--------|--------|--------|--------|
| 2/02   | 2/10   | 2/11   | 2/12   | 2/20   | 2/21   | 2/22   |
| \(P_n(A_k)\) ∈ | 0.2876±0.2882 | 0.2883±0.2897 | 0.2882±0.2885 | 0.2879±0.2882 | 0.2881±0.2893 | 0.288±0.2895 | 0.2881±0.2885 |
| \(P_n(T_k)\) ∈ | 0.2877±0.2881 | 0.2877±0.2888 | 0.2876±0.2879 | 0.2873±0.2874 | 0.2878±0.2884 | 0.2878±0.2878 | 0.2876±0.2877 |
| \(P_n(C_k)\) ∈ | 0.2117±0.2117 | 0.2116±0.2118 | 0.2113±0.2119 | 0.2115±0.2118 | 0.2118±0.2118 | 0.2115±0.2118 |
| \(P_n(G_k)\) ∈ | 0.2116±0.2116 | 0.2115±0.2116 | 0.2113±0.2116 | 0.2115±0.2116 | 0.2114±0.2114 | 0.2111±0.2112 |

Fig. A6/8. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J, chromosome 8, GRCh38.p4 C57BL/6J, NCBI Reference Sequence: NC_000074.6. 129401213 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000074.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000074.6). Here \(n = 1, 2, 3, 4, 5\), \(k \leq n\). Level 0 corresponds to the initial DNA-text with its different \(n\)-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
Fig. A6/9. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 9, GRChm38.p4 C57BL/6J. NCBI Reference Sequence: NC_000075.6. 124595110 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000075.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000075.6). Here $n = 1, 2, 3, 4, 5$, $k \leq n$. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.

| Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| $P_n(A_k) \in$ | 0.2865+ 0.2867 | 0.286+ 0.286 | 0.2861+ 0.2867 | 0.2863+ 0.2867 | 0.286+ 0.2863 | 0.2863+ 0.2868 | 0.2861+ 0.2862 |
| $P_n(T_k) \in$ | 0.286+ 0.2862 | 0.286+ 0.2866 | 0.2861+ 0.2862 | 0.2858+ 0.2864 | 0.2861+ 0.2868 | 0.2857+ 0.2864 | 0.286+ 0.2862 |
| $P_n(C_k) \in$ | 0.2133+ 0.2137 | 0.2134+ 0.2139 | 0.2133+ 0.2134 | 0.2132+ 0.2134 | 0.2132+ 0.2134 | 0.2132+ 0.2134 | 0.2132+ 0.2134 |
| $P_n(G_k) \in$ | 0.2132+ 0.2134 | 0.2131+ 0.2135 | 0.2131+ 0.2137 | 0.2132+ 0.2135 | 0.2129+ 0.2136 | 0.2129+ 0.2136 | 0.213+ 0.2134 |

| Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|---------|----------|----------|-----------|-----------|-----------|-----------|
| $P_n(A_k) \in$ | 0.2925+ 0.2926 | 0.2924+ 0.2924 | 0.2925+ 0.2925 | 0.2922+ 0.2926 | 0.2925+ 0.2926 | 0.2926+ 0.2926 | 0.2925+ 0.2925 |
| $P_n(T_k) \in$ | 0.2932+ 0.2933 | 0.2931+ 0.2933 | 0.2933+ 0.2934 | 0.293+ 0.2934 | 0.2933+ 0.2934 | 0.2932+ 0.2934 | 0.2932+ 0.2933 |
| $P_n(C_k) \in$ | 0.2067+ 0.2069 | 0.2067+ 0.2069 | 0.2067+ 0.2069 | 0.2067+ 0.2069 | 0.2067+ 0.2069 | 0.2066+ 0.2068 | 0.2066+ 0.2068 |
| $P_n(G_k) \in$ | 0.2071+ 0.2072 | 0.207+ 0.2074 | 0.2071+ 0.2072 | 0.2072+ 0.2073 | 0.2069+ 0.2073 | 0.207+ 0.2071 | 0.2071+ 0.2073 |

| Level 0 | Level 1/0 | Level 1/1 | Level 1/2 | Level 2/00 | Level 2/01 |
|---------|----------|----------|----------|-----------|-----------|
| $P_n(A_k) \in$ | 0.2925+ 0.2926 | 0.2925+ 0.2927 | 0.2923+ 0.2924 | 0.2924+ 0.2925 | 0.2924+ 0.2926 | 0.2925+ 0.2928 |
| $P_n(T_k) \in$ | 0.2932+ 0.2933 | 0.293+ 0.2933 | 0.2933+ 0.2937 | 0.293+ 0.2935 | 0.2932+ 0.2932 | 0.2928+ 0.2932 | 0.2929+ 0.2933 |
| $P_n(C_k) \in$ | 0.2067+ 0.2069 | 0.2067+ 0.2067 | 0.2066+ 0.2068 | 0.2066+ 0.2068 | 0.2065+ 0.2069 | 0.2063+ 0.2065 | 0.2063+ 0.2065 |
| $P_n(G_k) \in$ | 0.2071+ 0.2072 | 0.207+ 0.2072 | 0.2069+ 0.2071 | 0.2071+ 0.2072 | 0.2069+ 0.2072 | 0.2069+ 0.2072 | 0.2069+ 0.2074 |
|        | Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|--------|------------|------------|------------|------------|------------|------------|------------|
| $P_n(A_k)$ | $0.2921\div 0.2923$ | $0.2922\div 0.2926$ | $0.2921\div 0.2929$ | $0.2924\div 0.2924$ | $0.2921\div 0.293$ | $0.2925\div 0.2929$ | $0.2922\div 0.2926$ |
| $P_n(T_k)$ | $0.293\div 0.2936$ | $0.2934\div 0.2938$ | $0.293\div 0.2932$ | $0.2931\div 0.2933$ | $0.2928\div 0.2933$ | $0.293\div 0.2936$ | $0.293\div 0.2936$ |
| $P_n(C_k)$ | $0.2065\div 0.2065$ | $0.2066\div 0.2066$ | $0.2065\div 0.2069$ | $0.2064\div 0.2064$ | $0.2064\div 0.2065$ | $0.2066\div 0.2065$ | $0.2066\div 0.207$ |
| $P_n(G_k)$ | $0.2072\div 0.2075$ | $0.2067\div 0.207$ | $0.2068\div 0.2073$ | $0.2067\div 0.2072$ | $0.2068\div 0.2069$ | $0.2068\div 0.2069$ | $0.2068\div 0.2069$ |

Fig. A6/10. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6j chromosome 10, GRCM38.p4 C57BL/6j. NCBI Reference Sequence: NC_000076.6. 130694993 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000076.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000076.6). Here $n = 1, 2, 3, 4, 5$, $k \leq n$. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
| Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|------------|------------|------------|------------|------------|------------|------------|
| \( P_n(A_k) \) ∈ 0.2809÷0.2809 | 0.2808÷0.2814 | 0.2811÷0.2813 | 0.2808÷0.2808 | 0.2811÷0.2811 | 0.2811÷0.2813 | 0.281÷0.2817 |
| \( P_n(T_k) \) ∈ 0.2803÷0.2808 | 0.2801÷0.2807 | 0.2803÷0.2808 | 0.28÷0.2802 | 0.2803÷0.2812 | 0.2805÷0.2807 | 0.2803÷0.2804 |
| \( P_n(C_k) \) ∈ 0.2192÷0.2192 | 0.2187÷0.2191 | 0.2187÷0.2193 | 0.2186÷0.2187 | 0.219÷0.2192 | 0.219÷0.2189 | 0.2189÷0.2189 |
| \( P_n(G_k) \) ∈ 0.2185÷0.2191 | 0.219÷0.219 | 0.2186÷0.2188 | 0.219÷0.2196 | 0.2186÷0.2188 | 0.2188÷0.2188 | 0.219 |

Fig. A6/11. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J, chromosome 11, GRCh38.p4 C57BL/6J. NCBI Reference Sequence: NC_000077.6 122082543 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000077.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000077.6). Here \( n = 1, 2, 3, 4, 5 \), \( k \leq n \). Level 0 corresponds to the initial DNA-text with its different \( n \)-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
Fig. A6/12. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 12, GRCm38.p4 C57BL/6J. NCBI Reference Sequence: NC_000078.6. 120129022 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000078.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000078.6). Here $n = 1, 2, 3, 4, 5$, $k \leq n$. Level 0 corresponds to the initial DNA-text with its different $n$-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
|        | Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|--------|------------|------------|------------|------------|------------|------------|------------|
| $P_n(A_k)\in$ | 0.2921+ 0.2923 | 0.2919+ 0.2925 | 0.2923+ 0.2927 | 0.2921+ 0.2921 | 0.2922+ 0.2931 | 0.2922+ 0.2926 | 0.2924+ 0.2931 |
| $P_n(T_k)\in$ | 0.2912+ 0.2914 | 0.2909+ 0.2914 | 0.2911+ 0.2911 | 0.2906+ 0.2906 | 0.2908+ 0.2912 | 0.2909+ 0.2913 |
| $P_n(C_k)\in$ | 0.2074+ 0.2078 | 0.208+ 0.208 | 0.2078+ 0.2082 | 0.208+ 0.208 | 0.2078+ 0.2083 | 0.2077+ 0.2079 | 0.2077+ 0.2078 |
| $P_n(G_k)\in$ | 0.2079+ 0.2085 | 0.2077+ 0.2081 | 0.2079+ 0.208 | 0.2076+ 0.2082 | 0.2078+ 0.2079 | 0.2079+ 0.2083 | 0.2078+ 0.2078 |

Fig. A6/13. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain 57BL/6 chromosome 13, GRCh38.p4 57BL/6, NCBI Reference Sequence: NC_000079.6, 120421639 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000079.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000079.6) . Here $n = 1, 2, 3, 4, 5$, $k \leq n$. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.

|        | Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|--------|---------|-----------|-----------|-----------|-----------|-----------|-----------|
| $P_n(A_k)\in$ | 0.2939+ 0.2939 | 0.2938+ 0.294 | 0.2939+ 0.2939 | 0.2937+ 0.2938 | 0.2937+ 0.294 | 0.2937+ 0.294 | 0.2938+ 0.2938 |
| $P_n(T_k)\in$ | 0.2943+ 0.2944 | 0.2943+ 0.2944 | 0.2944+ 0.2945 | 0.2942+ 0.2944 | 0.2941+ 0.2944 | 0.2942+ 0.2944 | 0.2942+ 0.2942 |
| $P_n(C_k)\in$ | 0.2055+ 0.2057 | 0.2055+ 0.2057 | 0.2055+ 0.2058 | 0.2054+ 0.2058 | 0.2055+ 0.2059 | 0.2054+ 0.2059 | 0.2053+ 0.2058 |
| $P_n(G_k)\in$ | 0.2058+ 0.206 | 0.2058+ 0.2059 | 0.2058+ 0.206 | 0.2059+ 0.206 | 0.2057+ 0.2059 | 0.2058+ 0.206 | 0.2056+ 0.2062 |

|        | Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 |
|--------|---------|-----------|-----------|-----------|-----------|
| $P_n(A_k)\in$ | 0.2939+ 0.2939 | 0.2939+ 0.2939 | 0.2936+ 0.2937 | 0.2937+ 0.2937 | 0.2937+ 0.2942 | 0.2938+ 0.294 |
| $P_n(T_k)\in$ | 0.2943+ 0.2944 | 0.2941+ 0.2946 | 0.2943+ 0.2945 | 0.2943+ 0.2946 | 0.2939+ 0.2939 | 0.2942+ 0.2943 |
| $P_n(C_k)\in$ | 0.2055+ 0.2057 | 0.2053+ 0.2053 | 0.2056+ 0.2057 | 0.2056+ 0.2057 | 0.2053+ 0.2056 | 0.2051+ 0.2057 |
| $P_n(G_k)\in$ | 0.2058+ 0.206 | 0.2059+ 0.2061 | 0.2057+ 0.2061 | 0.2057+ 0.206 | 0.2057+ 0.2062 | 0.2058+ 0.2059 |
Fig. A6/14. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus C57BL/6j chromosome 14, GRChm38.p4 C57BL/6j. NCBI Reference Sequence: NC_000080.6. 12490244 bp. [https://www.ncbi.nlm.nih.gov/nuccore/NC_000080.6](https://www.ncbi.nlm.nih.gov/nuccore/NC_000080.6). Here $n = 1, 2, 3, 4, 5$, $k \leq n$. Level 0 corresponds to the initial DNA-text with its different $n$-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.

| Level 0  | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|---------|-----------|-----------|------------|------------|------------|------------|
| $P_n(A_k)$ | $0.2898 \div 0.2898$ | $0.2897 \div 0.2898$ | $0.2895 \div 0.2895$ | $0.2896 \div 0.2896$ | $0.2897 \div 0.2897$ | $0.2896 \div 0.2896$ |
| $P_n(T_k)$ | $0.2904 \div 0.2907$ | $0.2903 \div 0.2904$ | $0.2905 \div 0.2905$ | $0.2902 \div 0.2902$ | $0.2902 \div 0.2902$ | $0.2904 \div 0.2904$ |
| $P_n(C_k)$ | $0.2098 \div 0.2098$ | $0.2097 \div 0.2097$ | $0.2098 \div 0.2098$ | $0.2096 \div 0.2096$ | $0.2097 \div 0.2097$ | $0.2096 \div 0.2096$ |
| $P_n(G_k)$ | $0.2096 \div 0.2096$ | $0.2097 \div 0.2097$ | $0.2095 \div 0.2095$ | $0.2096 \div 0.2096$ | $0.2095 \div 0.2095$ | $0.2094 \div 0.2094$ |

| Level 0  | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 |
|---------|-----------|-----------|------------|------------|
| $P_n(A_k)$ | $0.2898 \div 0.2898$ | $0.2898 \div 0.2898$ | $0.2897 \div 0.2897$ | $0.2895 \div 0.2895$ |
| $P_n(T_k)$ | $0.2904 \div 0.2907$ | $0.2903 \div 0.2906$ | $0.2903 \div 0.2906$ | $0.2904 \div 0.2904$ |
| $P_n(C_k)$ | $0.2098 \div 0.2098$ | $0.2097 \div 0.2097$ | $0.2096 \div 0.2096$ | $0.2097 \div 0.2097$ |
| $P_n(G_k)$ | $0.2096 \div 0.2096$ | $0.2097 \div 0.2097$ | $0.2095 \div 0.2095$ | $0.2096 \div 0.2096$ |
Fig. A6/15. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 15, GRCm38.p4 C57BL/6J NCBI Reference Sequence: NC_000081.6, LOCUS NC_000081 104043685 bp DNA linear CON 22-JUN-2016 DEFINITION Mus musculus strain C57BL/6J chromosome 15, GRCm38.p4 C57BL/6J. ACCESSION NC_000081 GPC_000000788 VERSION NC_000081.6 https://www.ncbi.nlm.nih.gov/nuccore/NC_000081.6. Here $n = 1, 2, 3, 4, 5, k \leq n$. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
|         | Level 0 | Level 1/0 | Level 1/1 | Level 1/2 | Level 2/00 | Level 2/01 |
|---------|---------|-----------|-----------|-----------|------------|------------|
| $P_n(A_k)$ | 0.2949±0.2951 | 0.2947±0.2947 | 0.2948±0.295 | 0.295±0.2954 | 0.2945±0.2951 | 0.2947±0.295 |
| $P_n(T_k)$ | 0.295±0.2956 | 0.2953±0.2957 | 0.2955±0.2958 | 0.2953±0.2956 | 0.2953±0.2956 | 0.2952±0.2956 |
| $P_n(C_k)$ | 0.2045±0.2045 | 0.2044±0.2046 | 0.2044±0.2046 | 0.2043±0.2044 | 0.2044±0.2044 | 0.2041±0.2047 |
| $P_n(G_k)$ | 0.2047±0.2047 | 0.2046±0.2049 | 0.2046±0.2047 | 0.2047±0.2049 | 0.2045±0.2049 | 0.2047±0.2047 |

|         | Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|---------|------------|------------|------------|------------|------------|------------|------------|
| $P_n(A_k)$ | 0.2947±0.2948 | 0.2944±0.2948 | 0.2946±0.2948 | 0.2947±0.2956 | 0.2944±0.2944 | 0.2948±0.2953 | 0.2949±0.2951 |
| $P_n(T_k)$ | 0.2949±0.2958 | 0.2954±0.2959 | 0.2954±0.2963 | 0.2951±0.2958 | 0.2948±0.2956 | 0.2947±0.2957 | 0.2951±0.2953 |
| $P_n(C_k)$ | 0.2043±0.2047 | 0.2039±0.2043 | 0.2043±0.2045 | 0.2043±0.2044 | 0.2045±0.2044 | 0.2039±0.2044 | 0.2045±0.2047 |
| $P_n(G_k)$ | 0.2046±0.2047 | 0.2043±0.2049 | 0.2043±0.2045 | 0.2042±0.2045 | 0.2045±0.2045 | 0.2046±0.2046 | 0.2044±0.2049 |

Fig. A6/16. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 16, GRCm38.p4 C57BL/6J, NCBI Reference Sequence: NC_000082.6, LOCUS NC_000082 98207768 bp DNA linear CON 22-JUN-2016 DEFINITION Mus musculus strain C57BL/6J chromosome 16, GRCm38.p4 C57BL/6J. ACCESSION NC_000082 GPC_000000789 VERSION NC_000082.6, https://www.ncbi.nlm.nih.gov/nuccore/NC_000082.6. Here n = 1, 2, 3, 4, 5, k ≤ n. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
| Level 0 | Level 1/0 | Level 1/1 | Level 1/2 | Level 2/00 | Level 2/01 |
|--------|----------|----------|----------|-----------|-----------|
| \( P_n(A_k) \) | \( 0.286/0.2864 \) | \( 0.2859/0.2864 \) | \( 0.2861/0.2864 \) | \( 0.2859/0.2864 \) | \( 0.2857/0.2864 \) | \( 0.2858/0.2862 \) |
| \( P_n(T_k) \) | \( 0.2867/0.2867 \) | \( 0.2868/0.2868 \) | \( 0.2863/0.2867 \) | \( 0.2866/0.2869 \) | \( 0.2867/0.2869 \) | \( 0.2868/0.287 \) |
| \( P_n(C_k) \) | \( 0.2134/0.2136 \) | \( 0.2133/0.2139 \) | \( 0.2134/0.2138 \) | \( 0.2134/0.2137 \) | \( 0.2129/0.2132 \) | \( 0.2133/0.2135 \) |
| \( P_n(G_k) \) | \( 0.2133/0.2133 \) | \( 0.2133/0.2132 \) | \( 0.2133/0.2131 \) | \( 0.2133/0.2135 \) | \( 0.2133/0.2132 \) | \( 0.2133/0.2132 \) |

| Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| \( P_n(A_k) \) | \( 0.2858/0.2859 \) | \( 0.286/0.2865 \) | \( 0.2859/0.2863 \) | \( 0.2861/0.2863 \) | \( 0.2859/0.2863 \) | \( 0.2861/0.2867 \) | \( 0.2851/0.2862 \) |
| \( P_n(T_k) \) | \( 0.2864/0.2867 \) | \( 0.2863/0.2869 \) | \( 0.2862/0.2863 \) | \( 0.2861/0.2867 \) | \( 0.2863/0.2867 \) | \( 0.2864/0.2869 \) | \( 0.2867/0.2869 \) |
| \( P_n(C_k) \) | \( 0.2132/0.2142 \) | \( 0.2133/0.2135 \) | \( 0.2132/0.2134 \) | \( 0.2131/0.2135 \) | \( 0.2133/0.2133 \) | \( 0.2133/0.2133 \) | \( 0.2132/0.214 \) |
| \( P_n(G_k) \) | \( 0.2131/0.2132 \) | \( 0.2126/0.2131 \) | \( 0.2134/0.2134 \) | \( 0.2131/0.2137 \) | \( 0.2131/0.2137 \) | \( 0.2131/0.2137 \) | \( 0.2129/0.2129 \) |

Fig. A6/17. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 17, GRCm38.p4 C57BL/6J, CBI Reference Sequence: NC_000083.6LOCUS NC_000083, https://www.ncbi.nlm.nih.gov/nuccore/NC_000083.6. Here \( n = 1, 2, 3, 4, 5, k \leq n \). Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
|          | Level 0 | Level 1/0 | Level 1/1 | Level 1/2 | Level 2/00 | Level 2/01 |
|----------|---------|-----------|-----------|-----------|------------|------------|
| $P_n(A_k)$ | $0.2928 \div 0.2928$ | $0.2928 \div 0.2928$ | $0.2925 \div 0.2928$ | $0.2928 \div 0.2928$ | $0.2926 \div 0.2927$ | $0.2927 \div 0.2927$ |
| $P_n(T_k)$ | $0.2925 \div 0.2926$ | $0.2923 \div 0.2924$ | $0.2923 \div 0.2925$ | $0.2925 \div 0.2927$ | $0.2923 \div 0.2927$ | $0.2923 \div 0.2923$ |
| $P_n(C_k)$ | $0.2068 \div 0.2076$ | $0.2067 \div 0.2069$ | $0.2068 \div 0.2070$ | $0.2068 \div 0.2071$ | $0.2065 \div 0.2069$ | $0.2066 \div 0.2071$ |
| $P_n(G_k)$ | $0.2074 \div 0.2076$ | $0.2072 \div 0.2076$ | $0.2074 \div 0.2074$ | $0.2073 \div 0.2073$ | $0.2073 \div 0.2073$ | $0.2073 \div 0.2073$ |

|          | Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|----------|------------|------------|------------|------------|------------|------------|------------|
| $P_n(A_k)$ | $0.2928 \div 0.293$ | $0.2924 \div 0.2924$ | $0.2924 \div 0.293$ | $0.2924 \div 0.293$ | $0.2924 \div 0.293$ | $0.2924 \div 0.293$ | $0.2924 \div 0.293$ |
| $P_n(T_k)$ | $0.2923 \div 0.2925$ | $0.2922 \div 0.2923$ | $0.2922 \div 0.2929$ | $0.2924 \div 0.2931$ | $0.2921 \div 0.2931$ | $0.2924 \div 0.2924$ | $0.2922 \div 0.2922$ |
| $P_n(C_k)$ | $0.2064 \div 0.207$ | $0.2066 \div 0.2067$ | $0.2067 \div 0.2067$ | $0.2068 \div 0.2067$ | $0.2066 \div 0.2067$ | $0.2066 \div 0.2067$ | $0.2068 \div 0.2068$ |
| $P_n(G_k)$ | $0.207 \div 0.2075$ | $0.2074 \div 0.2075$ | $0.2072 \div 0.2075$ | $0.2072 \div 0.2074$ | $0.2071 \div 0.2071$ | $0.2071 \div 0.2071$ | $0.2068 \div 0.2068$ |

Fig. A6/18. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 18, GRCm38.p4 C57BL/6J, NCBI Reference Sequence: NC_000084.6, LOCUS NC_000084 90702639 bp DNA linear CON 22-JUN-2016 DEFINITION Mus musculus strain C57BL/6J chromosome 18, GRCm38.p4 C57BL/6J. ACCESSION NC_000084 GPC_000000791 VERSION NC_000084.6, https://www.ncbi.nlm.nih.gov/nuccore/NC_000084.6. Here $n = 1, 2, 3, 4, 5$, $k \leq n$. Level 0 corresponds to the initial DNA-text with its different n-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.
| Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|--------|-----------|-----------|------------|------------|------------|------------|
| $P_n(A_k)$ | 0.2873÷0.2876 | 0.2871÷0.2878 | 0.2873÷0.2876 | 0.2867÷0.2879 | 0.2871÷0.2876 | 0.2873÷0.2876 |
| $P_n(T_k)$ | 0.285÷0.2852 | 0.2849÷0.2853 | 0.2849÷0.2853 | 0.2848÷0.2848 | 0.2847÷0.2855 | 0.2851÷0.2854 |
| $P_n(C_k)$ | 0.2137÷0.2139 | 0.2138÷0.2138 | 0.2138÷0.2138 | 0.2138÷0.2138 | 0.2138÷0.2138 | 0.2134÷0.2138 |
| $P_n(G_k)$ | 0.2132÷0.2133 | 0.2131÷0.2134 | 0.2131÷0.2136 | 0.2131÷0.2136 | 0.2131÷0.2136 | 0.2131÷0.2136 |

| Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|------------|------------|------------|------------|------------|------------|------------|
| $P_n(A_k)$ | 0.2873÷0.2874 | 0.2873÷0.2874 | 0.287÷0.2873 | 0.2869÷0.2869 | 0.2869÷0.2869 | 0.2869÷0.2869 |
| $P_n(T_k)$ | 0.2846÷0.2851 | 0.2847÷0.2853 | 0.2846÷0.2853 | 0.2847÷0.2854 | 0.2849÷0.2856 | 0.2843÷0.2854 |
| $P_n(C_k)$ | 0.2135÷0.2144 | 0.2134÷0.2134 | 0.2134÷0.2134 | 0.2132÷0.2141 | 0.2133÷0.2141 | 0.2135÷0.2137 |
| $P_n(G_k)$ | 0.2127÷0.2131 | 0.2128÷0.2134 | 0.2129÷0.2136 | 0.2125÷0.2134 | 0.2134÷0.2134 | 0.2128÷0.2135 |

Fig. A6/19. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome 19, GRCm38.p4 C57BL/6J, NCBI Reference Sequence: NC_000085.6, LOCUS NC_000085 61431566 bp DNA linear CON 22-JUN-2016 DEFINITION Mus musculus strain C57BL/6J chromosome 19, GRCm38.p4 C57BL/6J. ACCESSION NC_000085 GPC_000000792 VERSION NC_000085.6, https://www.ncbi.nlm.nih.gov/nuccore/NC_000085.6 . Here n = 1, 2, 3, 4, 5, k ≤ n. Level 0 corresponds to the initial DNA-text with its different n-letter
representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.

| Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|---------|-----------|-----------|------------|------------|------------|------------|
| $P_n(A_k) \in$ | 0.3037±0.0309 | 0.3037±0.0309 | 0.3036±0.0309 | 0.3035±0.0307 | 0.3036±0.0308 | 0.3034±0.0307 | 0.3036±0.0307 |
| $P_n(T_k) \in$ | 0.3036±0.0306 | 0.3035±0.0303 | 0.3035±0.0309 | 0.3035±0.0307 | 0.3035±0.0305 | 0.3033±0.0307 | 0.3035±0.0307 |
| $P_n(C_k) \in$ | 0.1961±0.1963 | 0.1961±0.1962 | 0.1961±0.1962 | 0.1959±0.1961 | 0.196±0.1962 | 0.196±0.1962 | 0.196±0.1962 |
| $P_n(G_k) \in$ | 0.1962±0.1963 | 0.1962±0.1964 | 0.1962±0.1961 | 0.1961±0.1966 | 0.1963±0.1963 | 0.1963±0.1963 | 0.1963±0.1963 |

| Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|------------|------------|------------|------------|------------|------------|------------|
| $P_n(A_k) \in$ | 0.3034±0.0304 | 0.3033±0.0303 | 0.3035±0.0304 | 0.3034±0.0303 | 0.3035±0.0303 | 0.3033±0.0303 | 0.3034±0.0303 |
| $P_n(T_k) \in$ | 0.3037±0.0307 | 0.3034±0.0304 | 0.3035±0.0305 | 0.3034±0.0304 | 0.3035±0.0305 | 0.3033±0.0305 | 0.3032±0.0305 |
| $P_n(C_k) \in$ | 0.1959±0.1958 | 0.1958±0.1958 | 0.196±0.1961 | 0.196±0.1962 | 0.196±0.1962 | 0.196±0.1963 | 0.196±0.1963 |
| $P_n(G_k) \in$ | 0.1959±0.1959 | 0.1962±0.1963 | 0.1958±0.1963 | 0.196±0.1963 | 0.196±0.1963 | 0.196±0.1964 | 0.1959±0.1965 |

Fig. A6/20. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome X, GRCm38.p4 C57BL/6J, NCBI Reference Sequence: NC_000086.7, LOCUS NC_000086 171031299 bp DNA linear CON 22-JUN-2016 DEFINITION Mus musculus strain C57BL/6J chromosome X, GRCm38.p4 C57BL/6J. ACCESSION NC_000086 GPC_000000793
Here \( n = 1, 2, 3, 4, 5 \), \( k \leq n \). Level 0 corresponds to the initial DNA-text with its different \( n \)-letter representations. Other levels in this table correspond to sets of all daughter texts at appropriate levels of positional convolutions in the FGT-2 and the FGT-3.

| Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|---------|----------|----------|-----------|-----------|-----------|-----------|
| \( P_n(A_k) \) | 0.3045±0.0305 | 0.3043±0.0306 | 0.3043±0.0306 | 0.3039±0.0304 | 0.3044±0.0304 | 0.3042±0.0305 | 0.3044±0.0304 |
| \( P_n(T_k) \) | 0.3064±0.0306 | 0.3063±0.0306 | 0.3064±0.0306 | 0.3062±0.0306 | 0.3063±0.0306 | 0.3062±0.0306 | 0.3065±0.0306 |
| \( P_n(C_k) \) | 0.1948±0.1948 | 0.1948±0.1948 | 0.1947±0.1951 | 0.1946±0.1948 | 0.1946±0.1946 | 0.1947±0.1947 |
| \( P_n(G_k) \) | 0.1937±0.1942 | 0.1937±0.1939 | 0.1938±0.1942 | 0.1936±0.1941 | 0.1936±0.1941 | 0.1938±0.1938 | 0.1937±0.1944 |

| Level 0 | Level 1/0 | Level 1/1 | Level 2/00 | Level 2/01 | Level 2/10 | Level 2/11 |
|---------|----------|----------|-----------|-----------|-----------|-----------|
| \( P_n(A_k) \) | 0.3045±0.0305 | 0.3041±0.0306 | 0.3043±0.0306 | 0.3045±0.0304 | 0.3039±0.0304 | 0.3043±0.0304 |
| \( P_n(T_k) \) | 0.3064±0.0306 | 0.3062±0.0306 | 0.3065±0.0306 | 0.3064±0.0306 | 0.3061±0.0306 | 0.306±0.0306 |
| \( P_n(C_k) \) | 0.1948±0.1948 | 0.1947±0.1947 | 0.1946±0.1948 | 0.1946±0.1946 | 0.1945±0.1953 | 0.1946±0.1948 |
| \( P_n(G_k) \) | 0.1937±0.1942 | 0.194±0.194 | 0.1934±0.1942 | 0.1937±0.194 | 0.1937±0.194 | 0.1937±0.1943 |

| Level 0 | Level 2/02 | Level 2/10 | Level 2/11 | Level 2/12 | Level 2/20 | Level 2/21 | Level 2/22 |
|---------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| \( P_n(A_k) \) | 0.304±0.0302 | 0.3042±0.0307 | 0.3044±0.0305 | 0.3042±0.0304 | 0.3041±0.0304 | 0.3045±0.0302 | 0.3042±0.0304 |
| \( P_n(T_k) \) | 0.3061±0.0306 | 0.3062±0.0306 | 0.3067±0.0306 | 0.3065±0.0306 | 0.3064±0.0306 | 0.3059±0.0306 | 0.3064±0.0306 |
| \( P_n(C_k) \) | 0.1945±0.1951 | 0.1947±0.1953 | 0.1942±0.1948 | 0.1944±0.1948 | 0.1946±0.1953 | 0.1946±0.195 | 0.1945±0.1945 |
| \( P_n(G_k) \) | 0.1939±0.1944 | 0.1936±0.1936 | 0.1932±0.1939 | 0.1934±0.1939 | 0.1935±0.1935 | 0.1935±0.1937 | 0.1937±0.1945 |

Fig. A6/21. 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 FGT-3 (bottom tables) in the case of the sequence: Mus musculus strain C57BL/6J chromosome Y, GRCm38.p4 C57BL/6J.
Appendix 7. Confirmations of model predictions on symmetries of collective probabilities for 4*-groups of n-plets in long DNA-texts

This Appendix represents a few initial confirmations of our model predictions made above in the end of the Section 5. These confirmations were received in the result of analysis of the single DNA-sequence with one million letters in it (this sequence was already analysed above for the table in Fig. 4). For this sequence, the table in Fig. A7/1 shows collective probabilities of 4*-groups, each of which contains subgroups defined by one of 16 doublets. Denotations in the table are similar for denotations used above. For example, $P_2(CC_1)$ means the individual probability of the doublet CC in this sequence; $P_3(CC_1)$ means the collective probability of the CC-subgroup of 4 triplets CCC, CCT, CCG and CCA with the doublet CC in their beginnings; $P_4(CC_3)$ means the collective probability of the CC-subgroup of 16 tetraplets (like AACC, ATCC, TGCC, etc.) with the same doublet CC in their ends.

One can see from the table in Fig. A7/1 that, for example, $P_3(AA_1)\approx P_3(AA_1)\approx P_3(AA_2)\approx P_4(AA_1)\approx P_4(AA_2)\approx P_4(AA_3)$. These approximate equalities correspond to the first and the second rules of tetra-group symmetries, which were described in the Section 3 for another case, that is for “4*-groups of the first degree”: C-subgroups, T-subgroups, G-subgroups and A-subgroups (see Figs. 1, 2). Data of the table in Fig. A7/1 testify in favor that – in line with our model prediction - similar rules of symmetries of collective probabilities are also fulfilled for 4*-groups («tetra-groups of the second degree»), subgroups of which are defined by 16 doublets: CC-subgroups, TG-subgroups, etc.
| $P_2(AA)$ | $P_3(AA_1)$ | $P_4(AA_2)$ | $P_4(AA_3)$ | $P_4(AA_4)$ |
|----------|-------------|-------------|-------------|-------------|
| 0.10501  | 0.10541     | 0.10442     | 0.10458     | 0.10495     |
| $P_2(AC)$ | $P_3(AC_1)$ | $P_4(AC_2)$ | $P_4(AC_3)$ | $P_4(AC_4)$ |
| 0.04687  | 0.04675     | 0.04714     | 0.04667     | 0.04676     |
| $P_2(AG)$ | $P_3(AG_1)$ | $P_4(AG_2)$ | $P_4(AG_3)$ | $P_4(AG_4)$ |
| 0.06300  | 0.06348     | 0.06382     | 0.06417     | 0.06273     |
| $P_2(AT)$ | $P_3(AT_1)$ | $P_4(AT_2)$ | $P_4(AT_3)$ | $P_4(AT_4)$ |
| 0.09242  | 0.09233     | 0.09241     | 0.09344     | 0.09220     |
| $P_2(CA)$ | $P_3(CA_1)$ | $P_4(CA_2)$ | $P_4(CA_3)$ | $P_4(CA_4)$ |
| 0.065442 | 0.06527     | 0.06551     | 0.06568     | 0.06520     |
| $P_2(CC)$ | $P_3(CC_1)$ | $P_4(CC_2)$ | $P_4(CC_3)$ | $P_4(CC_4)$ |
| 0.03821  | 0.03852     | 0.03852     | 0.03834     | 0.03827     |
| $P_2(CG)$ | $P_3(CG_1)$ | $P_4(CG_2)$ | $P_4(CG_3)$ | $P_4(CG_4)$ |
| 0.00560  | 0.00584     | 0.00542     | 0.00536     | 0.00551     |
| $P_2(CT)$ | $P_3(CT_1)$ | $P_4(CT_2)$ | $P_4(CT_3)$ | $P_4(CT_4)$ |
| 0.067066 | 0.067041    | 0.06762     | 0.067244    | 0.06692     |
| $P_2(GA)$ | $P_3(GA_1)$ | $P_4(GA_2)$ | $P_4(GA_3)$ | $P_4(GA_4)$ |
| 0.05575  | 0.05562     | 0.05548     | 0.05500     | 0.05550     |
| $P_2(GC)$ | $P_3(GC_1)$ | $P_4(GC_2)$ | $P_4(GC_3)$ | $P_4(GC_4)$ |
| 0.03341  | 0.03282     | 0.03356     | 0.03279     | 0.03340     |
| $P_2(GG)$ | $P_3(GG_1)$ | $P_4(GG_2)$ | $P_4(GG_3)$ | $P_4(GG_4)$ |
| 0.03962  | 0.03932     | 0.03951     | 0.03924     | 0.04034     |
| $P_2(GT)$ | $P_3(GT_1)$ | $P_4(GT_2)$ | $P_4(GT_3)$ | $P_4(GT_4)$ |
| 0.05256  | 0.05277     | 0.05224     | 0.05184     | 0.05250     |
| $P_2(TA)$ | $P_3(TA_1)$ | $P_4(TA_2)$ | $P_4(TA_3)$ | $P_4(TA_4)$ |
| 0.08153  | 0.08149     | 0.08140     | 0.08144     | 0.08094     |
| $P_2(TC)$ | $P_3(TC_1)$ | $P_4(TC_2)$ | $P_4(TC_3)$ | $P_4(TC_4)$ |
| 0.05858  | 0.05834     | 0.05836     | 0.05810     | 0.05902     |
| $P_2(TG)$ | $P_3(TG_1)$ | $P_4(TG_2)$ | $P_4(TG_3)$ | $P_4(TG_4)$ |
| 0.07196  | 0.07215     | 0.07222     | 0.07298     | 0.07234     |
| $P_2(TT)$ | $P_3(TT_1)$ | $P_4(TT_2)$ | $P_4(TT_3)$ | $P_4(TT_4)$ |
| 0.12297  | 0.12284     | 0.12299     | 0.12250     | 0.12339     |
Fig. A7/1. Analysis of the collective probabilities of 4²-groups of n-plets in the Homo sapiens chromosome 5 sequence, 1000000 bp, accession NT_086358, version NT_086358.1, https://www.ncbi.nlm.nih.gov/nuccore/NT_086358.1. Each of 4²-groups contains subgroups defined by one of 16 doublets. There are shown the individual probabilities of each of 16 doublets (like P₂(AA), etc.) and also collective probabilities for those sets of triplets and 4-plets (like P₃(AA₁), P₄(AA₂), etc.), where each of 16 doublets plays the decisive attributive role, taking its position in the beginning or in the middle or in the end of these n-plets (n=3, 4). See additional explanations in the text.

The table in Fig. A7/2 shows collective probabilities for the case of 4³-groups, each of which contains subgroups defined by one of 64 triplets. In this table the denotation P₃(CCC) means the individual probability of the triplet CCC in the considered sequence; P₄(CCC₁) means the collective probability of the CCC-subgroup of 4 tetraplets CCCA, CCCC, CCCG and CCCCG with the triplet CCC in their beginnings; P₄(CCC₂) means the collective probability of the CCC-subgroup of 16 tetraplets with the same triplet CCC in their ends (like ACCC, TCCC, GCCC, CCCG), etc. One can see from the table in Fig. A7/2 that, for example, P₃(AAA)≈P₄(AA₁). Data of the table in Fig. A7/1 testify in favor that – in line with the prediction of our model in the Section 5 - similar rules of symmetries of collective probabilities are also fulfilled for 4³-groups («tetra-groups of the third degree»), subgroups of which are defined by 64 triplets: CCC-subgroups, TGA-subgroups, etc.
Fig. A7/2. Analysis of the collective probabilities of $4^3$-groups of n-plets in the Homo sapiens chromosome 5 sequence, 1000000 bp, accession NT_086358, version NT_086358.1, https://www.ncbi.nlm.nih.gov/nuccore/NT_086358.1. Each of $4^3$-groups contains subgroups defined by one of 64 triplets. There are shown the individual probabilities of each of 64 triplets (like $P_3$(AAA), etc.) and also collective probabilities for those sets of 4-plets (like $P_4$(AAA)1, etc.), where each of 64 triplets plays the decisive attributive role, taking its position in the beginning of these 4-plets. See additional explanations in the text.

**Acknowledgments:** Some results of this paper have been possible due to a long-term cooperation between Russian and Hungarian Academies of Sciences on the topic “Non-linear models and symmetrologic analysis in biomechanics, bioinformatics, and the theory of self-organizing systems”, where S.V. Petoukhov was a scientific chief from the Russian Academy of Sciences. The author is grateful to G. Darvas and I. Stepanyan for their collaboration. A special thanks is to V. Svirin for his computer programs, which were made on the basis of the author’s technical tasks and descriptions, to study automatically long DNA-sequences (initial author’s results were obtained by a half-manually manner). The author notes with gratitude the support from the side of Z.B. Hu in carrying out this study. The author is very grateful to E. Fimmel, M. Gumbel, L. Strüngmann and A. Karpuzoglu for their selective verification and the confirmation of some of described results about tetra-groups symmetries by means of their independently created computer program and also for their fruitful discussion of the described results; it was done during the author's internship in autumn 2017 at the Institute of Mathematical Biology of the Mannheim University of Applied Sciences (Germany) on the basis of a scholarship provided by the German Academic Exchange Service (DAAD).
REFERENCES

Albrecht-Buehler G. (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. *Proc Natl Acad Sci USA*. November 21; 103(47): 17828–17833.

Albrecht-Buehler G. (2007) Inversions and inverted transpositions as the basis for an almost universal “format” of genome sequences. *Genomics*, 90, 297–305.

Baish J.W., Jain R.K. (2000). Fractals and cancer. *Cancer Research*, 60, pp. 3683–3688, July 15 2000.

Baisnee P-F., Hampson S., Baldi P. (2002) Why are complementary DNA strands symmetric? – *Bioinformatics*, 18(8),1021-33.

Bauer, E.S. (1935). Theoretical Biology. VIEM, Moscow and Leningrad (in Russian Bauer, E.S. (1982). Theoretical Biology, with English translation. Akadémiai Kiadó, Budapest, 1982).

Bell, S. J., Forsdyke, D. R. (1999) Deviations from Chargaff's second parity rule correlate with direction of transcription - *J. Theo. Bio.*, 197, 63-76.

Bellman, R. (1960) *Introduction to Matrix Analysis.* New-York: Mcgraw-Hill Book Company, Inc, 351 pp.

Bizzarri M., Giuliani A., Cucina A., Anselmi F. D., Soto A. M., Sonnenschein C. (2011). Fractal analysis in a Systems Biology approach to cancer. - *Semin Cancer Biol.* June ; 21(3): pp. 175–182. doi:10.1016/j.semcancer.2011.04.002.

Caspani L., Reimer Ch., Kues M., Roztock P., Clerici M., Wetzel B., Jestin Y., Ferrera M., Peccianti M., Pasquazi M., Razzari L., Little B.E., Sai Chu T., Moss D.J., Morandotti R. (2016). Multifrequency sources of quantum correlated photon pairs on-chip: a path toward integrated Quantum Frequency Combs. – *Nanophotonics*, 5, p. 351–362 (doi:10.1515/nanoph-2016-0029), [http://eprints.gla.ac.uk/118888/](http://eprints.gla.ac.uk/118888/).

Chargaff, E. (1951). Structure and function of nucleic acids as cell constituents. - *Fed. Proc.*, 10, 654-659.

Chargaff, E. (1971). Preface to a Grammar of Biology: A hundred years of nucleic acid research. - *Science*, 172, 637-642

Chargaff E. (1975). A fever of reason. *Annu. Rev. Biochem.*, 44:1-20.

Dokukin M.E., Guz N.V., Woodworth C.D., Sokolov I. (2015). 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.

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

Dragovich B., Dragovich A. (2007). p-Adic Degeneracy of the Genetic Code. - [https://arxiv.org/abs/0707.0764](https://arxiv.org/abs/0707.0764).

Dragovich B., Khrennikov A., Misic N.Z. (2017). Ultrametrics in the genetic code and the genome. - *Applied Mathematics and Computation* 309 (2017) 350–358.

Fickett J.W., Burks C. (1989). Development of a database for nucleotide sequences. – In “Mathematical Methods for DNA Sequences” (Ed. Waterman M.S.), p. 1-34. Florida: CRC Press, Inc.

Fimmel, E., Danielli A., Strüngmann L. (2013). On dichotomic classes and bijections of the genetic code. *J. Theor. Biol.*, 336, 221–230.

Fimmel, E., Giannerini, S., Gonzalez, D., Strüngmann, L. (2014). Circular codes, symmetries and transformations. - *J. Math. Biol.*, 70, 1623–1644.
Fimmel E., Strüngmann L. (2016). Yury Borisovich Rumer and his ‘biological papers’ on the genetic code. Phil. Trans. R. Soc. A 374: 20150228.

Forsdyke D. R. (1995). A stem-loop “kissing” model for the initiation of recombination and the origin of introns. Molecular Biology and Evolution, 12, 949–958.

Forsdyke D.R. (2002). Symmetry observations in long nucleotide sequences: a commentary on the discovery of Qi and Cuticchia. Bioinformatics letter, v.18, № 1, p.215-217.

Forsdyke D. R. (2006). Evolutionary Bioinformatics. New-York: Springer Verlag.

Forsdyke D. R., Bell S. J. (2004). Purine-loading, stem-loops, and Chargaff’s second parity rule. Applied Bioinformatics, 3, 3–8.

Heisenberg W. (1926). Mehrkorperproblem und Resonanz in Quantenmechanik. – Zs. Phys., 38, 411-426.

Holevo A. S. (2013). Quantum systems, channels, information. De Gruyter Studies in Mathematical Physics 16, 349 p.

Hu Z.B., Petoukhov S.V. (2017). Generalized crystallography, the genetic system and biochemical esthetics. - Structural Chemistry, v. 28, №1, pp. 239-247. doi:10.1007/s11224-016-0880-0, http://link.springer.com/journal/11224/28/1/page/2

Hu Z.B., Petoukhov S.V., Petukhova E.S. (2017). I-Ching, dyadic groups of binary numbers and the geno-logic coding in living bodies. - Progress in Biophysics and Molecular Biology (in print, https://doi.org/10.1016/j.pbiomolbio.2017.08.018).

Igamberdiev A.U. (1993). Quantum mechanical properties of biosystems: a framework for complexity, structural stability, and transformations.- Biosystems, 31 (1), pp. 65–73.

Igamberdiev, A.U. (2012). Biomechanical and coherent phenomena in morphogenetic relaxation processes. Biosystems, 109, p. 336–345.

Jeffrey, H.J. (1990). Chaos game representation of gene structure. - Nucleic Acids Research, Vol. 18, No. 8, pp. 2163-2170.

Ji, S. (2012). Molecular Theory of the Living Cell: Concepts, Molecular Mechanisms, and Biomedical Applications, New York: Springer.

Ji, S. (2015). Planckian distributions in molecular machines, living cells, and brains: The wave-particle duality in biomedical sciences, - Proc. Intern. Conf. on Biology and Biomedical Engineering. Vienna, March 15-17, pp. 115-137.

Ji S. (2017). The Cell Language Theory: Connecting Mind and Matter. - World Scientific Publishing Europe Ltd., 450 p. ISBN-13: 978-1848166608.

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. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. - Science, Vol. 326, Issue 5950, pp. 289-293, (2009, 09 October).

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. (2015 November). Lung cancer - a fractal viewpoint. - Nat Rev Clin Oncol, 12(11):pp. 664–675. doi:10.1038/nrclinonc. 2015.108.

Manin Yu. I. (1980). Computable and Uncomputable, Moscow Sovetskoie Radio (in Russian).
Manin D.Yu., Manin Yu.I. (2017). Cognitive networks: brains, internet, and civilizations. – Presentation at the workshop of the Competence Center for Algorithmic and Mathematical methods in Biology, Biotechnology and Medicine of the Mannheim University of Applied Sciences, 27 November 2017. https://www.cammbio.hs-mannheim.de/fileadmin/user_upload/projekte/cammbio/events/20171127-kickoff/EP-Manin.pdf

Matsuno K. (1999). Cell motility as an entangled quantum coherence. – BioSystems, 51, pp. 15–19.

Matsuno K. (2003). Quantum mechanics in first, second and third person descriptions. – BioSystems, 68, pp. 107-118.

Matsuno K. (2015). On the physics of the emergence of sensorimotor control in the absence of the brain. – Progress in Biophysics and Molecular Biology, 119, pp. 313-323.

Matsuno K., Paton R.C. (2000). Is there a biology of quantum information? – BioSystems, 55, 39–46

McClare, C.W.F. (1974). Resonance in bioenergetics. - Ann. N. Y. Acad. Sci., 227, 74–97.

Mitchell, D., Bridge, R. (2006). A test of Chargaff’s second rule, BBRC, 340, 90-94.

Nielsen M.A., Chuang I.L. (2010). Quantum Computation and Quantum Information. New York: Cambridge University Press.

Okamura K., Wei J., Scherer S. (2007). Evolutionary implications of inversions that have caused intra-strand parity in DNA. Bmc Genomics, 8, 160–166.

Pauling, L. (1940). The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry. Oxford University Press, London, 2nd edn, 664 p.

Pellionis A.J. (2008). The principle of recursive genome function. – Cerebellum, 7: pp. 348–359. DOI 10.1007/s12311-008-0035-y

Pellionisz A. J., Graham R., Pellionisz P. A. and Perez J. C. (2012). 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 Cerebrum and Cerebellar Disorders, pp. 1381-1423.

Pellionisz, A.J., Ramanujam, M.V., Rajan, E.G. (2017). Genome Editing – A Novel Business Opportunity for India as a BRICS Country to Excel in Global Genomics Enterprise. In: Proceedings of ICSCI 2017 Hyderabad, India Conference, pp. 1-3. http://www.usa-siliconvalley.com/2017_genome_editing.pdf

Peng C.K., Buldyrev S.V., Goldberger A.L., Havlin S., Sclortino F., Simons M., Stanley H.E. (1992). Long-range correlations in nucleotide sequences. - Nature, 356, pp.168–170.

Peng C.K., Buldyrev S.V., Goldberger A.L., Havlin S., Sclortino F., Simons M., Stanley H.E. (1992). Fractal landscape analysis of DNA walks. - Physica A., 191(1-4): 25-9.

Penrose R. (1996). Shadows of the Mind: A Search for the Missing Science of Consciousness. - Oxford University Press, USA, 480 p.
Perez J.C. (2010). 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. DOI: 10.1007/s12539-010-0022-0

Perez J.-Cl. (2013). The “3 Genomic Numbers” Discovery: How Our Genome Single-Stranded DNA Sequence Is “Self-Designed” as a Numerical Whole. - *Applied Mathematics*, 4, 37-53, http://dx.doi.org/10.4236/am.2013.410A2004

Perez J.C. (2017). 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. DOI: 10.4172/1948-5956.1000469.

Petoukhov S.V. (1999). Genetic Code and the Ancient Chinese “Book of Changes”. - *Symmetry: Culture & Science*, vol. 10, №3-4, 1999, p.211-226.

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

Petoukhov S.V. (2010). The degeneracy of the genetic code and Hadamard matrices. - https://arxiv.org/abs/0802.3366 from 26 December 2010.

Petoukhov S.V. (2011). Matrix genetics and algebraic properties of the multilevel system of genetic alphabets. - *Neuroquantology*, 2011, Vol 9, No 4, p. 60-81, http://www.neuroquantology.com/index.php/journal/article/view/501

Petoukhov S.V. (2012) Symmetries of the genetic code, hypercomplex numbers and genetic matrices with internal complementarities. - “*Symmetry: Culture and Science*, vol. 23, № 3-4, p. 275-301. http://symmetry.hu/scs_online/SCS_23_3-4.pdf

Petoukhov S.V. (2015a). Resonances and genetic bioinformatics. - *Symmetry: Culture and Science*, t. 26, №3, p. 379-397, http://petoukhov.com/PETOUKHOV_IN_SCS_2015.pdf

Petoukhov S.V. (2015b). Music and the modeling approach to genetic systems of biological resonances. – Lecture at the International conference «ISIS Summit Vienna 2015. The information society at the crossroads», 3-7 June, 2015, Vienna, Austria, http://sciforum.net/conference/70/paper/2812.

Petoukhov S.V. (2015c). Resonances and genetic bioinformatics. - *Symmetry: Culture and Science*, t. 26, №3, 2015, p. 379-397, http://petoukhov.com/PETOUKHOV_IN_SCS_2015.pdf

Petoukhov S.V. (2015d). The concept of resonances in genetics ("resonant bioinformatics"). - Selected Works of the VII International Congress "Weak and Superweak Fields and Radiation in Biology and Medicine", September 07-11, 2015, St. Petersburg, pp. 1-41, http://www.biophys.ru/lib/sci/resonance/460-resonance-00001 (in Russian).

Petoukhov S.V. (2016). The system-resonance approach in modeling genetic structures. – *Biosystems*, January 2016, v. 139, p. 1-11, http://petoukhov.com/PETOUKHOV_ARTICLE_IN_BIOSYSTEMS.pdf

Petoukhov S.V. (2017a). The genetic code, algebra of projection operators and problems of inherited biological ensembles. - https://arxiv.org/abs/1307.7882, 9th version from 8 August 2017.

Petoukhov S.V. (2017b). I-Ching, dyadic groups of binary numbers and the geno-logic coding in living bodies. - *Progress in Biophysics and Molecular Biology*, v. 131, December 2017, pp. 354-368.
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., Petukhova E.S. (2017a). Symmetries in genetic systems and the concept of geno-logical coding. - Information, 2017, 8(1), 2; doi:10.3390/info8010002, http://www.mdpi.com/2078-2489/8/1/2/htm

Petoukhov S.V., Petukhova E.S. (2017b). Resonances and the quest for transdisciplinarity. - Information Studies and the Quest for Transdisciplinarity. Editors M. Burgin, W.Hofkirchner, World Scientific, p. 467-487.

Petoukhov S.V., Svirin V.I. (2012). Fractal genetic nets and symmetry principles in long nucleotide sequences. - Symmetry: Culture and Science, vol. 23, № 3-4, p. 303-322. http://petoukhov.com/PETOUKHOV_SVIRIN_FGN.pdf

Petoukhov S.V., Petukhova E.S., Hazina L.V., Svirin V.I., Stepanyan I.V., Silova T.V. (2017). Geno-logical coding, united-hypercomplex numbers and systems of artificial intelligence. – In the book “Advances in Artificial Systems for Medicine and Education”, Editors Z.B. Hu, S.V.Petoukhov and M.He (the conference proceedings “Artificial Intelligence, Medical Engineering, Education”, Moscow, Russia, 21-23 August 2017). Part of the “Advances in Intelligent Systems and Computing” book series (AISAC, volume 658), pages 2-13, Springer, https://link.springer.com/search?query=978-3-319-67348-6.

Poincare H. (2009). The foundations of science. - (In the Chapter «Science and hypothesis»), Cornell University Library, 580 p. http://www.gutenberg.org/files/39713/39713-h/39713-h.htm#Page_140

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

Rapoport A.E., Trifonov E.N. (2012): Compensatory nature of Chargaff's second parity rule. Journal of Biomolecular Structure and Dynamics, November, 1-13, DOI:10.1080/07391102.2012.736757

Schrödinger E. (1944). What is Life? – United Kingdom: Cambridge University Press, 194 p.

Schrödinger E. (1952). Are there quantum jumps? Part I. - The British Journal for the Philosophy of Science, Vol. 3, No. 10, pp. 109-123. Published by: Oxford University Press on behalf of The British Society for the Philosophy of Science Stable URL: http://www.jstor.org/stable/685552.

Simeonov P.L. (2013). On Some Recent Insights in Integral Biomathics. - J. Progress in Biophysics and Molecular Biology. Special Theme Issue on Integral Biomathics: Can Biology Create a Profoundly New Mathematics and Computation? Elsevier. ISSN: 0079-6107. DOI: 10.1016/j.pbiomolbio.2013.06.001. http://arxiv.org/abs/1306.2843.

Stambuk N. (1999). Circular coding properties of gene and protein sequences. Croat. Chem. Acta, 72, 999–1008.

Sueoka N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. Journal of Molecular Evolution, 40, 318–325.

Trifonov E.N. (1993). DNA as a language. - Bioinformatics, Supercomputing and Complex Genome Analysis: pp. 103-110. http://www.worldscientific.com/doi/abs/10.1142/9789814503655_0009

Turin, L. (2009). Colin McClare (1937–1977): a tribute. - Journal of Biological Physics, vol. 35, Issue 1, 9-15.
Yamagishi, M., Herai, R (2011). Chargaff's "Grammar of Biology": New Fractal-like Rules. http://128.84.158.119/abs/1112.1528v1