ATLANTIS PRESS

Advances in Computer Science Research (ACSR), volume 72

IV International Research Conference "Information Technologies in Science, Management, Social Sphere and Medicine" (ITSMSSM 2017)

Supersymmetry in genomes and chargaff's second rule

Yaroslav Grebnev
Siberian Federal University
Krasnoyarsk, Russia
yaroslav.grebnev@gmail.com

Michail Sadovskiy
Institute of Computational Modeling of Siberian Branch of Russian Academy of sciences
Krasnoyarsk, Russia
msad@icm.krasn.ru

Abstract **— Some preliminary results are provided towards the study of the violation of genomic super-symmetry; that latter is the so called Second Chargaff's rule. The rule stipulates that oligonucleotides that could be read equally in opposite directions with respect to the symbol change according to the complimentary law (complimentary palindromes) should exhibit pretty close frequency. We have checked the genomes of organisms of various taxa ranging from viruses via bacteria, yeasts, animals, plants, etc.; more than 1500 genetic sequences had been studied, totally. The measure for the second rule violation was calculated for a single strand. Both intragenomic, and intergenomic studies have been carried out. It was found that intragenomic variability decays, as the length of string grows up. The intergenomic variability is comparable to the intragenomic one, for considerably short strings.**

Keywords—palindromes, frequency, classification, correlation, taxonomy, evolution

I. INTRODUCTION

The first rule of Chargaff establishes the equality in the DNA molecule of the amount of thymine (T) and the amount of adenine (A), as well as the corresponding equality for guanine (G) and cytosine (C). Later it was established that a similar rule holds for one DNA strand; This equation was called the second rule of Chargaff. The violation of the second Chargaff rule depends on the length of the analyzed region of the genome and can characterize the genome it self. For the whole chromosome of higher eukaryotes, the characteristic error in (approximate) equalities $A \approx T$, $G \approx C$ is. The rules of the Chargaff are universal rules and are subject to the genomes of all organisms from lower plants to higher animals, including extracellular forms of life [1],

At present, not much work has been devoted to investigating the violation of the second rule of Chargaff [2-4], despite the fundamental nature of this fact. The main goal of this work is to assess the degree of violation of the second rule of Chargaff in the genomes of various organisms

II. EASE OF USE

In the present work, a study was made of the behavior of the discrepancy for the genomes of various organisms. Organisms' genomes were decoded texts of nucleotide sequences, which were taken from the EMBL databank (http://www.ebi.ac.uk/genomes/). In the work genomes of various organisms and viruses were used; the following sequences of genomes are analyzed: yeast 81 061 875 nucleotides, fungi 269 875 059 nucleotides, bacteria 36 358 967, mitochondria 243 981, viruses 29142, and also organisms such as Giberella moniliformis 41 104 290, mosquito 230 466 657, bull 2 629 841 282 , 21 human chromosomes 33 216 610, 1 macaque chromosome 232 296 185, Arabidopsis taliana 93 654 490, fruit flies 125 566 102, chimpanzees 106 544 938 and gorillas 9 140.

To determine the residual, the frequency dictionaries of the sequences were compiled. A frequency dictionary is the set of all symbolic subsequences of a given length that occur in the sequence being studied, together with the frequency of their occurrence [5 ± 7]. Within the framework of the present work, frequency dictionaries of thickness from 1 to 8 (i.e., containing words of length 1, 2, ..., 8) were compiled.

Calculation of the discrepancy occurs as follows:

Suppose that an arbitrary hypothetical sequence (5') TTAACCGGGGGGAATGGG (3') is a fragment of the genome text. The default entry also means that the neighboring nucleotides are linked together in a chain by a phosphodiester linkage formed by 5'-phosphate (5'-PO3) and 3'-hydroxyl (3'-OH) groups. To confirm or accentuate the indicated orientation of the text, the above entry may have the following form (3 ') AATTGGCCCCCCTTACC (5').

Next, we compile a frequency dictionary:

The second, complementary chain of nucleotides, the line of which lies under the first when writing, has a relative directionality (3 ') - (5'), so for its generation the existing text of the first chain is rewritten from right to left.

For all manipulations with the second chain, for example, to search for restriction sites (sequences of nucleotides in the DNA molecule that determine the sites of its specific cleavage by the enzyme with a restriction enzyme), the last record should be brought in the appropriate direction (5 ') - (3'), that is is again written in the reverse form: (5 ') CCCATTCCCCCCGGTTAA (3'), thus obtained by a complimentary palindrome.

It was noted above that one of the directions of the analysis of DNA texts is the construction and analysis of dictionaries of nucleotide sequences (subsequences, subsequences). A common algorithm is the algorithm of actions in which a reading device is assigned, which is a mathematical tool that selects with the help of a moving frame a fragment of the nucleotide text that puts it on the list and assigns a special indicator to the unit-the number of copies of the fragment in the genetic text under study. The reader is characterized by the size - the number of nucleotides read in one step, the step with which the frame moves along the studied text, and also the direction of

motion. A reading frame can include any number of nucleotides: two, three, four, or more, and all possible variants of triplets, tetraplets, pentaplets, and other multibyte words present in the DNA string to be studied are obtained. To obtain the frequency, the resulting matches are divided into total number of nucleotides in the investigated genome.

An indicator characterizing the degree of violation of the second Chargaff rule was the magnitude

$$\mu = \frac{2}{4^q}\sqrt{\sum_{\Omega^* \in \omega}\left(f_\omega - f_{\overline{\omega}}\right)^2}\ , \qquad (1)$$

q — length of the words in the context of the words in the dictionary in question, Ω — set of all words that are direct, ω — word, $\overline{\omega}$ — complimentary word, f_ω — direct word frequency, $f_{\overline{\omega}}$ — frequency of a complimentary word.

To analyze the behavior of the magnitude of the residual (1), we estimate its behavior for a random uncorrelated sequence. Suppose that the second Chargaff rule is exactly fulfilled in it:

$$p(\mathrm{A}) = p(\mathrm{T}), \qquad p(\mathrm{C}) = p(\mathrm{G}) \qquad (2)$$

Then $\forall q$ discrepancy (1) is equal to zero and the second rule is fulfilled with absolute accuracy.

Suppose now that relation (2) is not exactly satisfied, but with some error

$$p(\mathrm{A}) = p_W + \varepsilon,\ p(\mathrm{T}) = p_W - \varepsilon$$
$$p(\mathrm{C}) = p_S + \delta,\ p(\mathrm{G}) = p_S - \delta. \qquad (3)$$

$2p_W = p(\mathrm{A}) + p(\mathrm{T})\ \ 2p_S = p(\mathrm{C}) + p(\mathrm{G})$. Then the discrepancy for the thickness dictionary $q = 1$ is given by $\mu = 2(\varepsilon + \delta)$. For frequency dictionaries with $q = 2$ we have three combinations of types of nucleotides for all conceivable words: $\mathrm{WW} \Leftrightarrow \mathrm{WW}$, $\mathrm{SS} \Leftrightarrow \mathrm{SS}$ и $\mathrm{SW} \Leftrightarrow \mathrm{SW}$; here the letters denote weak and strong nucleotides ($\mathrm{W} = \{\mathrm{A}, \mathrm{T}\}$ и $\mathrm{S} = \{\mathrm{C}, \mathrm{G}\}$). For the cases $\mathrm{WW} \Leftrightarrow \mathrm{WW}$ and $\mathrm{SS} \Leftrightarrow \mathrm{SS}$ There are three cases of discrepancy (1) on each side of the palindrome (that is, for each of the terms in parentheses in (1):

$$p_w^2 - \varepsilon^2,\ p_w^2 + 2p_w\varepsilon + \varepsilon^2\ \ p_w^2 - 2p_w\varepsilon + \varepsilon^2 \qquad (4)$$

similarly

$$p_s^2 - \delta^2,\ p_s^2 + 2p_s\delta + \delta^2\ \ p_s^2 - 2p_s\delta + \delta^2 \qquad (5)$$

Discarding the terms of order ε^2, δ^2 and above, we obtain an estimate for each palindrome of the form

$$\frac{\max\{\varepsilon, \delta\}}{2}$$

Since the total number of palindromes in the frequency dictionary is $0{,}5 \times 4^q$, the final expression for estimating the magnitude of the discrepancy (1) is given by

$$\mu \cong \frac{\max\{\varepsilon, \delta\}}{2^q} \qquad (6)$$

q — thickness of the dictionary.

III. RESULTS

Table 1 presents the results of calculating the rates of violation of the second Chargaff rule for various organisms. It is seen that the greatest number of violations of the second Chargaff rule is observed for mitochondria Equus caballus breed Appalosa, further on the degree of violation of the second rule of Chargaff, the mitochondria of artiodactyls follow, which again indicates that in the mitochondria of different genomes, the most violations of the second Chargaff rule occur. Then, the degree of violation of the second rule of the Chargaff is followed by the genomes of higher animals, in particular the gorilla genome, then we can distinguish genomes of extracellular life forms, in particular, the tobacco mosaic virus genome. The next violators of the second rule of the Chargaff are fungal organisms ascomycetes, then - insects. The least violation of the second rule of the Chargaff are observed for plants.

TABLE I. THE MAGNITUDE OF THE DISCREPANCY OF THE SECOND RIGHT CHARGAFF

Organism	Number of organisms studied	Number of nucleotides examined, million pairs	Mean value of the discrepancy	Standard deviation
Aspergilus fumigatus	8	29384958	0,016300	0,155000
Aspergilus niger	19	33975768	0,038000	0,246000
Aspergilus fumigatus	8	29384958	0,016300	0,155000
Aspergilus niger	19	33975768	0,038000	0,246000
Mitochondria	2004	93975768251441	2,1318000	10,40300
Candida albicans	9	12061552	0,032800	0,149000
Candida glabrata	13	12318245	0,043700	0,183616
Equus caballus breed Appalosa	57	8993004	2,228400	2,2284000
Gorilla	10	9140	1,0436000	10,436000
Giberella zeae	4	36358967	0,013676	0,001317
Anofeles	5	230466657	0,010476	0,0035282
Kluweromyces lactis	6	10689156	0,036687	0,159453

IV. Discussion

The data in Table 1 indicate the exponential decrease of the discrepancy (1) with increasing thickness of the dictionary for various taxonomic groups. It should be noted that the variability of the discrepancy for small values of the thickness of the dictionary is very large, but with increasing thickness of the dictionary it falls, which agrees with the estimate (6), carried out above. The greatest number of violations of the second rule of Chargaff among the organisms studied by us was observed in mitochondria and extracellular life forms.

The validity of the estimate (6) is also confirmed by Fig. 1, which shows the course of the values of the ratio of two consecutive values of the residual (1) obtained for one or another group of genomes. It is clearly seen that as the thickness of the dictionary grows (when the thickness k is approached), the ratio of two consecutive values of the residual (1) tends to a value equal to two, which completely agrees with the estimate (6). Apparently, we can expect that the accuracy of the approximation of this ratio to 2 will only increase with increasing thickness of the frequency dictionaries taken into consideration.

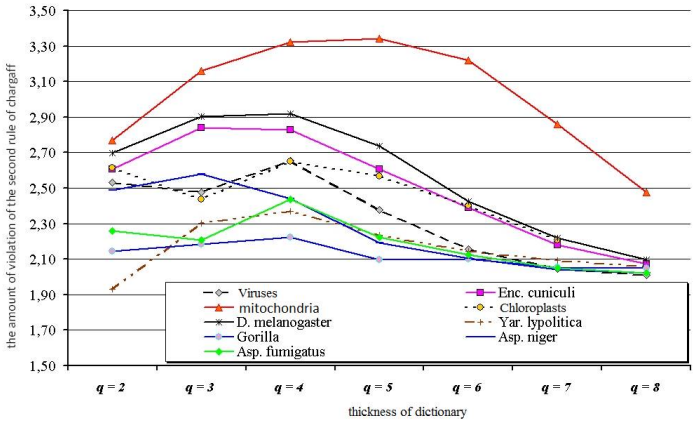

Fig. 1. The ratio of two successive (for the thickness of the dictionary) values of the discrepancy for different taxonomic groups.

Special attention should be paid to a more detailed study of the behavior of the discrepancy itself (1) for relatively small values: If the aspiration of the ratio of two successive values of the discrepancy during growth can be explained, in particular, by the finiteness effects of the character sequence being studied: indeed, the number of different words in the frequency dictionary grows exponentially, which leads to a rapid drop in the number of words that occur in more than one copy, then the behavior of the discrepancy at relatively small word lengths most likely reflects the biological characteristics of the analyzed genetic sequences to the greatest extent.

V. Conclusions

The results presented in the article also indirectly refute one of the hypotheses [1 ± 4] of the origin of the second Chargaff rule, namely the doubling hypothesis. According to this hypothesis, Chargaff's second rule arose as a result of a series of doubling of long and superlong DNA segments. In this case, the sequence itself, which was doubled, was assumed to be close in its properties to random.

However, the estimate (6) shows that for a random sequence - under the condition of an almost exact fulfillment of the second Chargaff rule at the mononucleotide composition level - the second Chargaff rule is also satisfied for words of greater length. Moreover, it can be expected that long and super long repetitions will lead to violations of the second Chargaff rule, at least on average throughout the sequence; It is possible that there is a noticeable heterogeneity of the sequence with respect to the discrepancy index (1), which is determined for different fragments of the original sequence, but this question is beyond the scope of this paper.

References

[1] Albrecht-Bühler G., Inversions and inverted transpositions as the basis for an almost universal “format” of genome sequences, Genomics, 2008, vol.90, pp. 297 – 305.

[2] Nikolaou C, Almirantis Y. Deviations from Chargaff's second parity rule in organellae DNA Insights into the evolution of organellar genomes, Gene, 2006, pp. 34-41.

[3] Mitchell D. GC content and genome length in Chargaff compliant genomes. Biochem. Biophys, Commun, 2007, pp. 207-217.

[4] Rapoport A.E., Trifonov E.N. Compensatory nature of Chargaff's second parity rule, J. Biomol. Struct. Dyn. 2013, 1324-1336.