BioSystems 180 (2019) 101–108

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

On comparing composition principles of long DNA sequences with those of random ones

Elena Fimmel[a], Markus Gumbel[a], Ali Karpuzoglu[a], Sergey Petoukhov[b]

[a] Competence Center for Mathematical and Algorithmical Methods in Biology, Biotechnology and Medicine, Mannheim University of Applied Sciences, 68163 Mannheim, Germany

[b] Laboratory of biomechanical systems, Russian Academy of Sciences, Moscow, Russia

ARTICLE INFO

Keywords:
DNA
Chargaff's parity rules
Genetic information

ABSTRACT

The revelation of compositional principles of the organization of long DNA sequences is one of the crucial tasks in the study of biosystems. This paper is devoted to the analysis of compositional differences between real DNA sequences and Markov-like randomly generated similar sequences. We formulate, among other things, a generalization of Chargaff's second rule and verify it empirically on DNA sequences of five model organisms taken from Genbank. Moreover, we apply the same frequency analysis to simulated sequences. When comparing the afore mentioned – real and random – sequences, significant similarities, on the one hand, as well as essential differences between them, on the other hand, are revealed and described. The significance and possible origin of these differences, including those from the viewpoint of maximum informativeness of genetic texts, is discussed. Besides, the paper discusses the question of what is a "long" DNA sequence and quantifies the choice of length. More precisely, the standard deviations of relative frequencies of bases stabilize from the length of approximately 100 000 bases, whereas the deviations are about three times as large at the length of approximately 25 000 bases.

1. Introduction

> It is the harmony of the diverse parts, their symmetry, their happy balance; in a word it is all that introduces order, all that gives unity, that permits us to see clearly and to comprehend at once both the ensemble and the details. Henri Poincaré

In the last few years, huge amounts of genetic data from different organisms have become available. In view of this, the importance to reveal hidden construction principles and symmetries in long DNA sequences (for comparison, the DNA text of the human genome consists of approximately 3 billions of nucleotide bases) is increasing permanently. The following quotation, although quite old, describes the current situation to the point (Fickett and Burks, 1989): "What will we have when these genomic sequences are determined? … We are in the position of Johann Kepler when he first began looking for patterns in the volumes of data that Tycho Brahe had spent his life accumulating. We have the program that runs the cellular machinery, but we know very little about how to read it."

Since the middle of last century, the so-called Chargaff's rules have been known (Chargaff et al., 1952; Chargaff, 1971). The first rule holds that a double-stranded DNA molecule globally contains equally pyrimidine and purine bases and, more specifically, that the amount of guanine (purine) should be equal to cytosine (pyrimidine) and the amount of adenine (purine) should be equal to thymine (pyrimidine). This rule was theoretically confirmed by the double-helix model of the DNA by Watson and Crick (1953). The second rule states that the same parity is approximately valid for each of the two long DNA strands alone. According to Albrecht-Buehler (2006), this rule applies to the eukaryotic chromosomes, the bacterial chromosomes, the double stranded DNA viral genomes, and the archaeal chromosomes provided they are long enough. However, within the scientific community there is no generally accepted explanation for this rule yet (compare Shporer et al. (2016) and Rapoport and Trifonov (2012)). Perez extended in Perez (2010) Chargaff's second rule to codons: more precisely, for the sequences divided into triplets (codons) the parity of complementary bases takes place in each of these three positions. Petoukhov suggested in Petoukhov (2017) a generalization of Chargaff's second rule and an extension to the approach of Perez for all sizes n of n-plets at each of the n positions (see definition in Section 2.1). This was empirically proven for $n \leq 5$. Shporer et al. (2016) investigated, in following Prabhu's ideas (Prabhu, 1993), the parity of k-nucleotides and their inverted

E-mail addresses: e.fimmel@hs-mannheim.de (E. Fimmel), m.gumbel@hs-mannheim.de (M. Gumbel), ali.karpuzoglu@gmail.com (A. Karpuzoglu), spetoukhov@gmail.com (S. Petoukhov).

https://doi.org/10.1016/j.biosystems.2019.04.003
Received 11 February 2019; Received in revised form 5 April 2019; Accepted 6 April 2019
Available online 09 April 2019
0303-2647/

complementary k-nucleotides in long DNA-sequences: "Inversion Symmetry (IS): the counts of a k-mer of nucleotides on a chromosomal strand are almost equal to those of its inverse (reverse-complement) string". For $k = 1$ it is equivalent to the validity of Chargaff's second rule. Yamagishi (2017) investigated the inversion symmetry very deeply and some explanative hypotheses for it are formulated.

A useful review of publications from different authors on Chargaff's second parity rule (CSPR) and its possible origin is given by Rosandic et al. (2016). In particular, the work of Rapoport and Trifonov (2012) emphasizes that this rule may be maintained in nature by alternating sequence segments with different signs of deviation from parity. Alternatively, it was suggested that CSPR would probably exist from the very beginning of genome evolution. Concerning the fundamental question on evolution of genomes and possible "grammar" rules in them, one should note the following. "Genomes are not static collections of DNA materials. Various biochemical and cellular processes – including point mutation, recombination, gene conversion, replication slippage, DNA repair, translocation, imprinting, and horizontal transfer – constantly act on genomes and drive the genomes to evolve dynamically. … Present genomes can be viewed as a snapshot of an ongoing genome evolution process …." (Zhou and Mishra, 2004). Assuming symmetries arising from primitive genomes could shed light on the origin of genomes, and even on the origin of life (Zhou and Mishra, 2004; Zhang and Huang, 2010; Sobottka and Hart, 2011).

Different chromosomes of a biological species can greatly differ in their length, characteristics and quantities of genes within them, the cytogenetic bands (which show the biochemical specificity of the different parts of chromosomes), etc. However, Chargaff's second rule and its generalizations are performed almost identically for different chromosomes (see for example Okamura et al. (2007)). It was noted that the CSPR can reveal general properties common to all species and have remarkable implications of some unknown mechanism that seems to be present (Albrecht-Buehler, 2006; Rapoport and Trifonov, 2012). The work by Rosandic et al. (2016) proposes that DNA growth might be viewed as being programmed from start by non-local natural symmetry laws of DNA creation; it considers "interplay of DNA language and symmetry forcing – as a possible simple but magnificent aspect for the code of life".

Mascher et al. (2013) revalidated Chargaff's second rule using the genomes of 16 different organisms including mammals, plants and invertebrates. A so-called skew-measure (relative difference of base's frequencies) was proposed in order to quantify the accuracy of this rule. In this paper, we generalize this skew-measure by considering it for subsequences of a sequence of nucleotide bases which are uniformly distributed, i.e. have fixed distances $n = 1, 2, 3, \ldots$ between their bases within the entire sequence. This generalization of Chargaff's second rule is empirically validated for up to about $n = 50$ following the approach from Petoukhov (2017). Then the results are compared with randomly generated chromosomes similar to those of the analyzed organisms in order to understand the construction principles of long DNA sequences. Moreover, the question is addressed whether long DNA sequences are able to convey maximal information due to minimal correlations amongst parts. As known, the base composition in various parts of DNA-sequences is very different, in particular in regions with genes and here in introns and exons. The findings by Nikolaou and Almirantis (2006) indicate that the replication process plays a major role in the shaping of the genome structure. The useful method of the skew as introduced by Mascher et al. (2013) reveals variations of frequencies of DNA nucleotides in different parts of long DNA-sequences. This enables a comparison of the validity of Chargaff's second rule in each of these parts. This is also consistent with the results of the work of Arqués and Michel (1990a,b) and Michel (1986). This method will be applied in the following for a new analytical approach to study long DNA-sequences.

The n-plet skew analysis is performed by means of the R package `abcd` (Analysis of Base Composition of long DNA sequences). This package is part of the software suite *Genetic Analysis Toolkit (GCAT)* (Fimmel et al., 2018; Kraljic et al., 2018) and will be available as open source on Github at https://github.com/informatik-mannheim/abcd-R-package.

2. Methods

2.1. Notations and definitions

Let us denote the nucleotide bases *alphabet* by

$$\mathscr{B} := \{A, C, G, T(U)\}$$

whose *letters* A, C, G and $T(U)$ stand for Adenine, Cytosine, Guanine and Thymine (Uracil), respectively. Thus the alphabet consists of four letters and its powers $\mathcal{B}^2 = \{N_1N_2 : N_i \in \mathcal{B}\}$ and $\mathcal{B}^3 = \{N_1N_2N_3 : N_i \in \mathcal{B}\}$ contain all *dinucleotides* and all *trinucleotides*, the latter being the codons. For an arbitrary $n \in \mathbb{N}$ we will call elements of $\mathcal{B}^n = \{N_1N_2...N_n : N_i \in \mathcal{B}\}$ n-nucleotides or, equivalently, n-plets.

Definition 1. Let $N, M \in \mathcal{B}$, $n \in \mathbb{N}$ and $X \in \mathcal{B}^*$ a sequence of nucleotide bases. We will denote as

1 X_n the sequence X divided into n-plets, $|X_n|$ the number of n-plets in X_n;

2 $F(X; N)$

the absolute frequency (the number of occurrences) of the nucleotide base N in X,

$F_n(X; N, k)$

the absolute frequency of the nucleotide base N in the k-th position ($1 \le k \le n$) of each n-plet in X_n;

3 $$P(X; N) = \frac{F(X; N)}{|X|}$$

the relative frequency (probability) of the nucleotide base N in X,

$$P_n(X; N; k) = \frac{F_n(X; N; k)}{|X_n|}$$

the relative frequency (probability) of the nucleotide base N in the k-th position ($1 \le k \le n$) of each n-plet in X_n;

4 $$S_{N\sim M}(X, n, k) = \frac{F_n(X; N; k) - F_n(X; M; k)}{F_n(X; N; k) + F_n(X; M; k)}$$

the mononucleotide skews (compare Mascher et al. (2013));

Remark 2. Obviously, we have for all $X \in \mathcal{B}^*$ and all $N, M \in \mathcal{B}$

1 $X_1 = X$, $F_1(N, 1) = F(N)$, $P_1(N; 1) = P(N)$

2 For all $n \in \mathbb{N}$ and $1 \le k \le n$ the following inequality takes place:

$$-1 \le S_{N\sim M}(n, k) \le 1.$$

Moreover, the equality

$$S_{N\sim M}(n, k) = 0$$

means that the frequencies of the nucleotide bases N and M in the k-th positions of X_n are exactly equal, the equality

$$|S_{N\sim M}(n, k)| = 1$$

means that one of the bases N or M lacks in the k-th positions of X_n.

3 $S_{N\sim M}(1, 1)$ is the skew as defined in Mascher et al. (2013).

2.2. Sequences being analyzed

DNA sequences taken from Genbank were used for the empirical analysis of the following model organisms (compare Mascher et al.

(2013) and Petoukhov (2017)):

- *Homo sapiens* (human). All 22 autosomes and the two allosomes were analyzed. The genome build is GRCh38.p7 Primary Assembly. E.g. fasta data for chromosome 1 was downloaded form https://www.ncbi.nlm.nih.gov/nuccore/NC_000001.11
- *Caenorhabditis elegans* (a transparent nematode). Chromosomes I to V and X. E.g. chromosome I https://www.ncbi.nlm.nih.gov/nuccore/BX284601.5
- *Arabidopsis thaliana* (thale cress). Chromosomes 1 to 5. E.g. chromosome 1 https://www.ncbi.nlm.nih.gov/nuccore/NC_003070.9
- *Oryza sativa Japonica* (Japanese rice). Chromosomes 1 to 12. E.g. chromosome https://www.ncbi.nlm.nih.gov/nuccore/AP008207.2
- *Chlamydomonas reinhardtii* (single cell green alga). Chromosomes 1 to 12. E.g. chromosome 1 https://www.ncbi.nlm.nih.gov/nuccore/CM008962.1

2.3. Random generated sequences

We like to compare the n-plet analysis on biological sequences with random sequences to verify the significance of our findings. On the one hand, this comparison will help us to understand whether the real genetic sequences behave the same way as random ones referring to the distribution of single nucleotide bases. On the other hand, the comparison should help to assess the importance of the length of a sequence examined.

As we have very long DNA sequences, local properties in the DNA like CpG islands, genes or introns and exons cannot be considered. For simplicity, a Markovian-like random model was used where only transition probabilities $P(N \to M)$ are considered. Here, $P(N \to M)$ indicates the probability for a base N to have the base M as a successor. As an example, the transition probabilities were calculated for the first chromosome of *Homo sapiens* and listed in Table 1). The stationary distribution for this chromosome can be derived from the transition probabilities. Let $p^{(1)} = (p_A^{(1)}, p_T^{(1)}, p_C^{(1)}, p_G^{(1)})^T$ be a vector of probabilities for the occurrence of the four bases A, T, C, G in the first position of the sequence ($p_A^{(1)} + p_T^{(1)} + p_C^{(1)} + p_G^{(1)} = 1$). The second base has then the probabilities

$$p^{(2)} = \mathcal{A} \cdot p^{(1)},$$

where $\mathcal{A}^T$ is the matrix with the transitional probabilities (see Table 1). Using this Markov-like approach we get

$$p^{(n)} = \mathcal{A} \cdot \ldots \cdot \mathcal{A} \cdot (\mathcal{A} \cdot p^{(1)}) = \mathcal{A}^n p^{(1)}$$

for n matrix multiplications. Table 2 shows the stationary results approximated by 100 iterations ($\mathcal{A}^{100}$). These probabilities are identical to the relative frequencies for each base counted in human chromosome 1.

A random chromosome has the same number of bases like the biological one and starts with a random base. Then the missing bases are added step by step according to the transition probabilities.

2.4. Sample-size normalized skews

It is likely that the deviation of the skew increases with the size n of the n-plets. This is because for larger n the number of n-plets decreases and accordingly the sample size. Here we elaborate how the skews for different n-plet sizes can be compared amongst each other.

Table 1
Transition probabilities as measured in human chromosome 1. N is a base and M the next base in sequence.

$P(N \to M)$	A	T	C	G
A	0.327	0.255	0.173	0.245
T	0.216	0.328	0.206	0.250
C	0.349	0.342	0.259	0.049
G	0.288	0.242	0.211	0.260

Table 2
Derived stationary distribution for the random model based on transition probabilities as defined in Table 1. Chargaff's second parity rule is confirmed here as the frequencies of Adenine (A) and Thymine (T) are almost identical as well as the frequencies for Cytosine (C) and Guanine (G).

Base	Probability
A	0.291
T	0.292
C	0.208
G	0.209

A sample-size normalized skew analysis is where the sample size is the same for all n-plet sizes ($n = 1, 2, \ldots, n^*$) where n^* is the largest n-plet size. $|X_{n^*}|$ is the minimal number of n-plets that every data series has (compare Fig. 1). We will consider subsequences $Y^{(n,n^*)}$ of successive bases of X with length:

$$|Y^{(n,n^*)}| = l(n, n^*) = |X_{n^*}| \cdot n$$

The position p in X of the first nucleotide in $Y^{(n,n^*)}$ starts at a multiple of n, i.e. $p = (R \cdot n) + 1$ with $0 \le R < |X_n|$. A preliminary version of a sample-size normalized skew is then defined as:

$$S_{N \sim M}(Y^{(n,n^*)}, n, k)$$

However, this definition has some drawbacks. Let us point out that $l(n, n^*)$ is smaller than the entire sequence length for $n < n^*$ and, thus, a sequence of such a length covers then only a fraction of the entire sequence. For instance, human chromosome 1 in the version used in this paper has 248,956,422 bases and for $n^* = 50$ the sample size is $|X_{50}| = 4,979,128$. The equivalent sequence length for n-plets of size $n = 2$ is then $|X_{50}| \cdot 2 = 9,958,256$ bases which is $n/n^* = 2/50 = 4\%$ of chromosome 1. To balance this situation, we consider several subsequences $Y_i^{(n,n^*)}$ with $1 \le i \le r$ that start at random positions in X but whose start index is still a multiple of n. The number of subsequences r is set to $r = \lfloor 10 \cdot n^*/n \rfloor$ in order to cover the entire sequence in average 10 times and to have a good chance to cover the entire sequence.

As we will see in the result Section 3.3 the skews will vary significantly in different parts of a chromosome. For this reason, the skews are averaged. This is defined as the sample-size normalized skew:

$$\bar{S}_{N \sim M}(X, n, k) = \frac{1}{r} \sum_{i=1}^{r} S_{N \sim M}(Y_i^{(n,n^*)}, n, k)$$

In particular, the standard deviation of all k positions ($1 \le k \le n$) of $\bar{S}_{N \sim M}(X, n, k)$ is later applied in Section 3.1 and Fig. 3.

2.5. Splitting a sequence into equal-sized partitions

While our approach with the n-plets was splitting a sequence bottom-up we will proceed contrary now: The sequence analyzed will be cut into smaller equal-sized partitions and each partition will be examined separately. The question we are trying to answer is whether different regions of the DNA have the same characteristics, i.e. does the first half of a sequence have the same structure as the second half and so on. The entire sequence is split into $P \in \mathbb{N}$ ($P \ge 1$) subsequences (or partitions) of equal length $|X|/P$. In this paper P is a power of two in order to have always the same partition boundaries.

3. Results and discussion

3.1. Generalized Chargaff's second rule

The first question we will address is (compare also Petoukhov

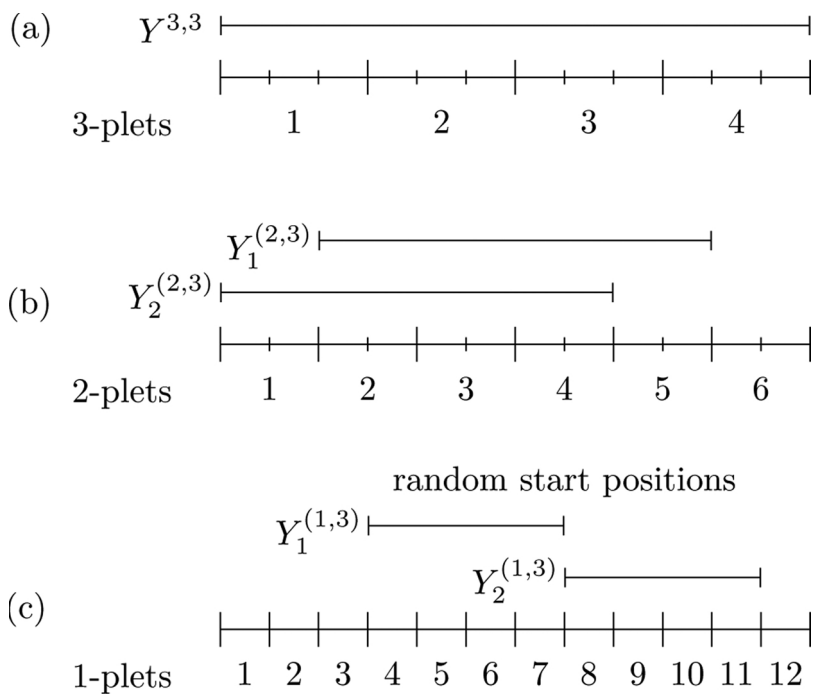

Fig. 1. Example for the computation of sample-size normalized skews with $n^* = 3$. This sequence has 12 bases (separated by minor ticks). (a) For an n-plet size of 3 we get four 3-plets (major ticks). The only subsequence $Y^{(3,3)}$ is the entire sequence. (b) For an n-plet size of 2 we get six 2-plets. Any subsequence $Y_i^{(2,3)}$ has the length of eight bases. As these subsequences do not cover the entire sequence, many subsequences with a random start position are chosen. In this example two subsequences ($r = 2$) are shown. (c) For an n-plet size of 1 we get 12 1-plets, i.e. the bases themselves. Any subsequence $Y_i^{(1,3)}$ has the length of four bases and also a random start position.

(2017)): For all genetic sequences X which are long enough and all $n \in \mathbb{N}$, $n \leq 50$, $1 \leq k \leq n$ the following parity rules apply:

1 $S_{A\sim T}(n, k) \approx 0$,
2 $S_{G\sim C}(n, k) \approx 0$.

This observation was confirmed in the analysis of all five organisms. The complete data for all species is shown in the appendix. As an example we will give the results for human chromosome 1 in Fig. 2 (a). For every n-plet size there are n values for the position (positional skews). The mean values for the positional skews are always close to 0 no matter what the n-plet size is. Next the same analysis was performed on a random sequence for human chromosome 1 (see methods in Section 2.3). The results are shown in Fig. 2 (b). Clearly, the box and whisker plots look very similar. This is true for all organisms and their chromosomes analyzed in the appendix.

However, the data also reveals that the distributions of the positional skews spread more for larger n-plet sizes. As mentioned in Section 2.4 this could be explained with the less sample size in the skew calculation. Fig. 3 shows the standard deviation of normalized skews for the biological and a random sequence of human chromosome 1. Apparently, the deviation of the skews still depends on the n-plet size when normalized skews $\bar{S}_{N\sim M}(X, n, k)$ are considered. The same holds for the random sequence of chromosome 1.

3.2. Long sequences

Another question is: What does it mean – "long DNA sequences"? In the literature, we find similar but not exactly superposable views on this matter:

- The article by Prabhu (1993) says "all sequences longer than 50,000 nucleotides";
- The article by Albrecht-Buehler (2006) states: "… a sufficiently long (> 100 kilobases) strand of genomic DNA that contains N copies of a mono- or oligonucleotide, also contains N copies of its reverse complementary mono- or oligonucleotide on the same strand";
- The article by Rapoport and Trifonov (2012) says: "Although the initial definition of this rule referred to mononucleotides, further works (Albrecht-Buehler, 2006; Prabhu, 1993; Qi and Cuticchia, 2001) demonstrated that it can be formulated more generally: '… a sufficiently long (> 100 kilobases) strand of genomic DNA that contains N copies of a mono- or oligonucleotide, also contains N copies of its reverse complementary mono- or oligonucleotideon the same strand" (Albrecht-Buehler (2006));
- A good review on Chargaff's second rule at the website http://www.epigenetics.com.ua/?p=165 says in Russian[1] : "the accuracy of the equality holds on lengths of up to 70–100 thousand base pairs – independently, coding regions there or not – and then begins to subside".

In this article, we quantify the choice of the minimal length of a sequence using the standard deviations of their corresponding skews as a measure. Fig. 4 shows the results for "shorter" sequences, i.e. sequences with a length of maximal 500,000 bases and Fig. 5 shows the results for "longer" sequences up to 10 mio. bases. Again the biological data in human chromosome 1 is compared with a random human chromosome 1 sequence. A shortened sequence in the real chromosome is a fragment that starts at a position equal to 5% of the total chromosome length and has the intended length, e.g. 25,000 bases. The offset of 5% ensures that bases classified as `N` (unknown) at the beginning of the fasta file are ignored. Those unclassified bases are often found at the beginning of a chromosome.

The findings in Figs. 4 and 5 reveal that the standard deviation of the skews in the real and random human chromosome 1 do not differ significantly. The n-plet size here was set to 20 but the effect is also visible for other values of n. However, there are two sizes where the

[1] Translated by the authors.

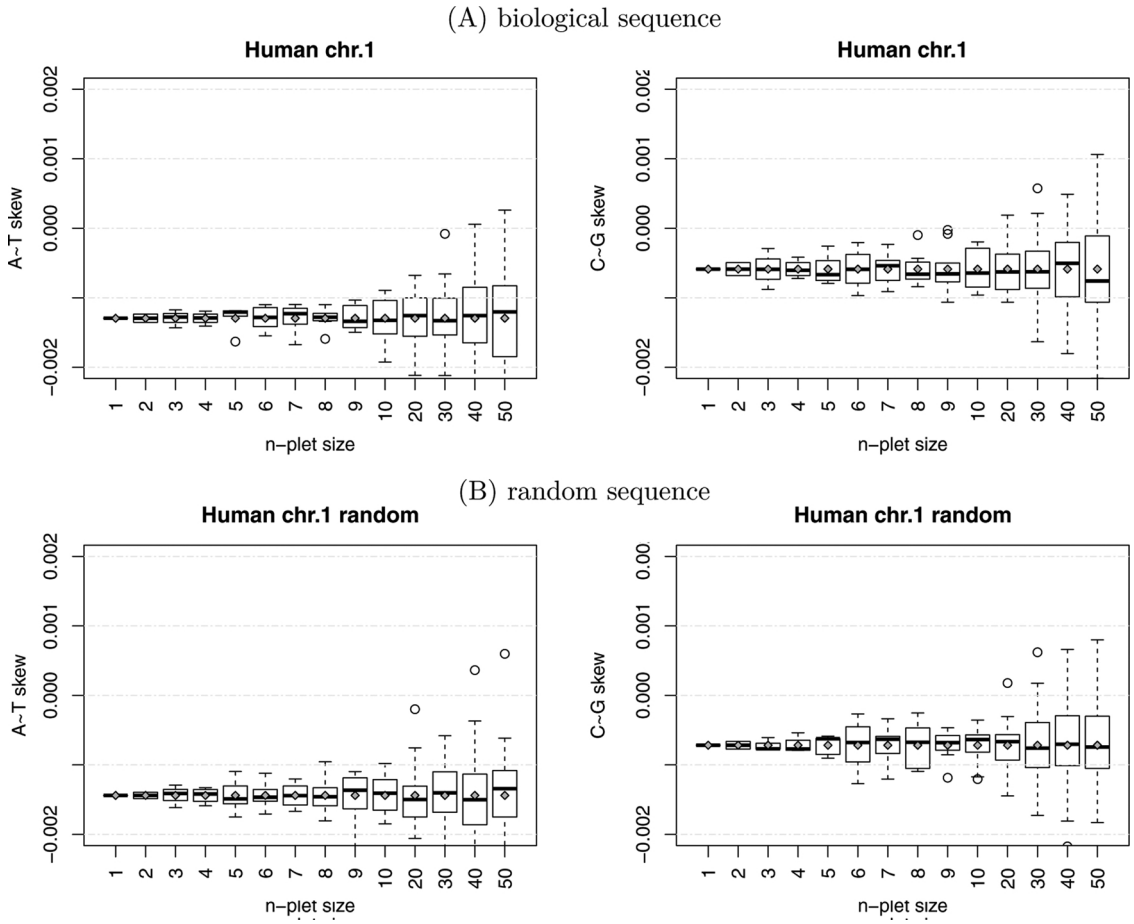

Fig. 2. A ~ T (left column) and C ~ G (right column) skews of *n*-plets for (a) human chromosome 1 and (b) a random human chromosome 1. The skews look very similar for the biological and random sequence. The gray diamond shows the mean value. The height of the boxes represents the interquartile range (IQR). The whiskers are $1.5 \cdot IQR$.

shape of the bar diagram changes: The standard deviation decreases quickly until a size of about 100,000 bases is reached (cf Fig. 4). Then the standard deviations remain quite stable and decreases slowly. In a similar way the same happens at a size of about 2 mio. bases (cf Fig. 5). Overall the gradient reminds on a negative exponential function (exp $(-x)$). It is striking that the standard deviations of skews are distributed in a fractal-like manner.

The landmark of 100,000 bases confirms the statements listed above, e.g. from Prabhu ("all sequences longer than 50,000 nucleotides") and Albrecht-Buehler "(> 100 kilobases)". The possible 2 mio. border and the exponential gradient, however, needs further investigation.

3.3. Partitions

This section shows the findings when a sequence is split into partitions as introduced in Section 2.5. As shown in Fig. 6 (a) the A ~ T and C ~ G skews are not uniformly distributed over the chromosome (again human chr. 1).

In contrast, the skews of the randomly generated DNA stayed very constant at a value very close to the average, barely changing in any part of the chromosome (cf Fig. 6(b)). The biological chromosomes, however, had many regions with a larger absolute skew. While the values for the A ~ T skew in chromosome 1 are in the range from -0.08 to 0.08, the value in a randomly generated chromosome 1 are between $8 \cdot 10^{-3}$ and $10 \cdot 10^{-3}$. This shows that the natural chromosome does not have a constant pattern, at least when using the genomic skew as a measure. Another interesting observation is that the A ~ T and C ~ G genomic skews do not change in the same pattern. The A ~ T skew has a higher variation than the C ~ G skew (cf Fig. 6(a)).

4. Conclusions

In the present work, we are providing a statistical analysis of DNA-sequences of five model organisms taken from GenBank in order to better understand their construction principles and are comparing their compliance with Chargaff's second rule with the behavior of randomly generated sequences. Chargaff's second rule and questions about the grammar of biology attract great attention of many authors in the field of theoretical biology (compare, for instance, Albrecht-Buehler, 2006; Nikolaou and Almirantis, 2006; Okamura et al., 2007; Patel, 2001; Perez, 2010; Prabhu, 1993; Rapoport and Trifonov, 2012; Rosandic et al., 2016; Shporer et al., 2016; Sobottka and Hart, 2011; Yamagishi, 2017; Zhang and Huang, 2010; Zhou and Mishra, 2004). As a consequence of this rule, every long DNA sequence has a certain mathematical characteristic, namely transition probabilities between its bases. Such transition probabilities can be used to construct a great number of appropriate random sequences, each of them can be interpreted as one of possible models of DNA-texts that also satisfies Chargaff's second rule. But it is obvious that real DNA sequences are not random sequences at all since they, for instance, contain genes where the order of bases *A*, *C*, *T* and *G* is very definite. The useful method of the A ~ T- and C ~ G-skews (Mascher et al., 2013), which we developed further in this paper, allows revealing significant similarities, on the one hand, as well as differences, on the other hand, between a long DNA sequence and its model random sequence.

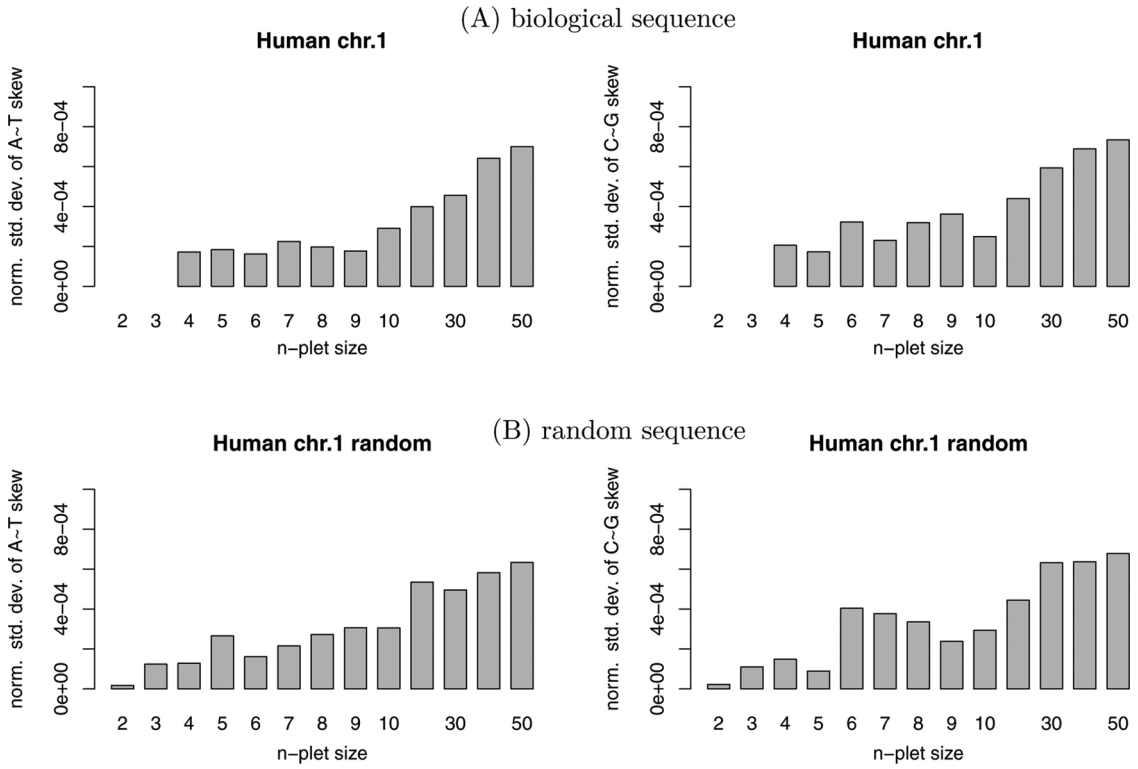

Fig. 3. Standard deviation of all k positions ($1 \le k \le n$) of the sample-size normalized skew $\bar{S}_{N\sim M}(X, n, k)$ for different n-plet sizes (compare Section 2.4). Left column: A ~ T skews and right column: C ~ G skews. Top row (a): human chromosome 1 and bottom row (b): a random human chromosome 1.

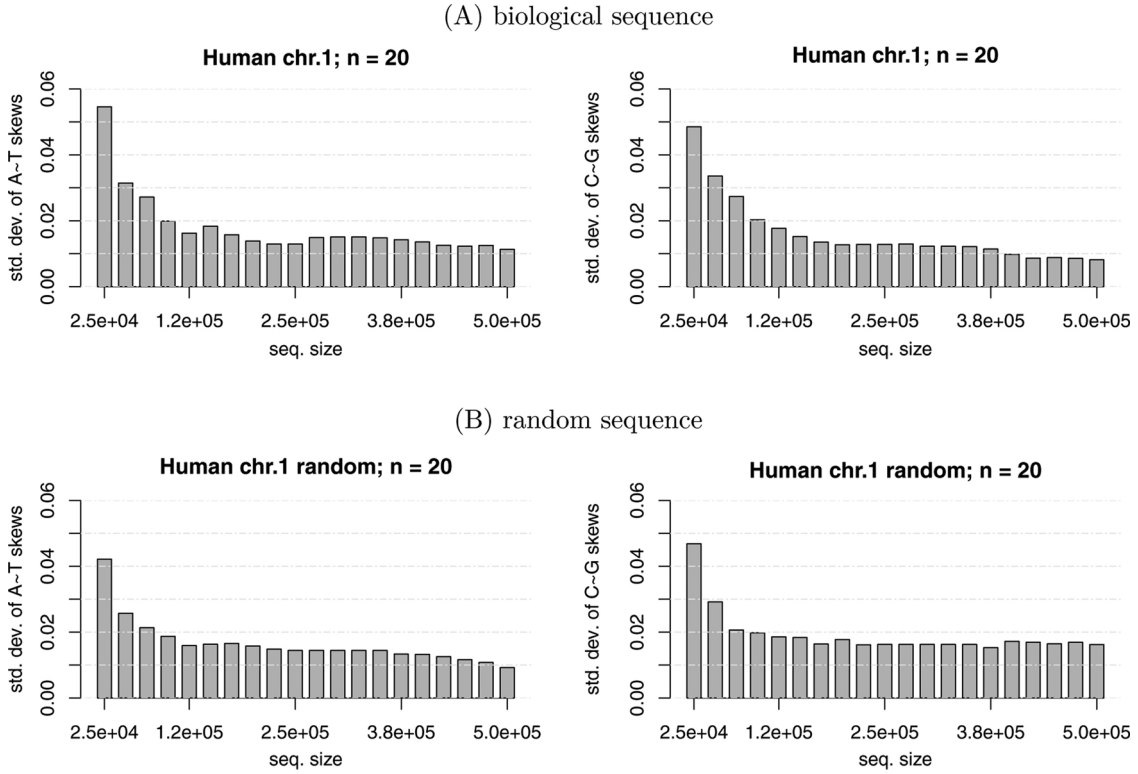

Fig. 4. Standard deviation of all 20 positions ($1 \le k \le 20$) of the skew $S_{N\sim M}(X, n = 20, k)$ for subsequences X of different lengths in the range from 25,000 to 500,000 bases (step wide of 25,000 bases). Left column: A ~ T skews and right column: C ~ G skews. Top row (a) human chromosome and bottom row (b) random human chromosome 1.

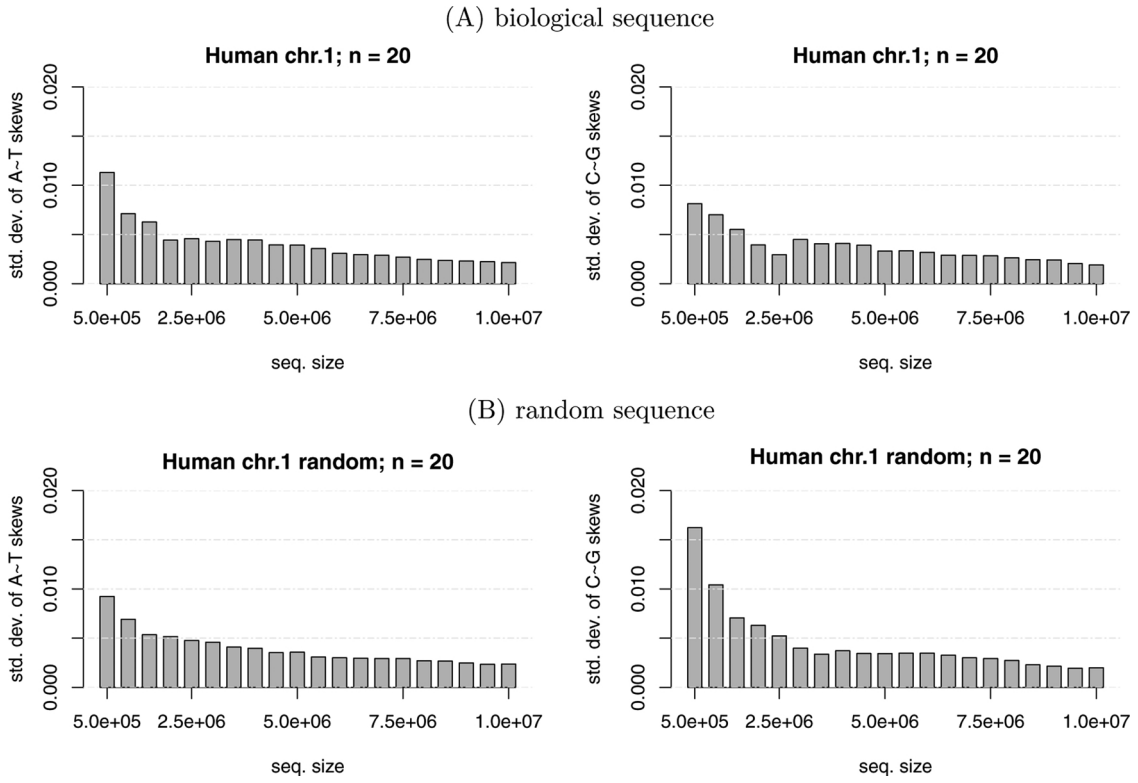

Fig. 5. Standard deviation of all 20 positions ($1 \leq k \leq 20$) of the skew $S_{N\sim M}(X, n = 20, k)$ for subsequences X of different lengths in the range from 500,000 to 10^7 bases (step wide of 500,000 bases). Left column: A ~ T skews and right column: C ~ G skews. Top row (a) human chromosome and bottom row (b) random human chromosome 1.

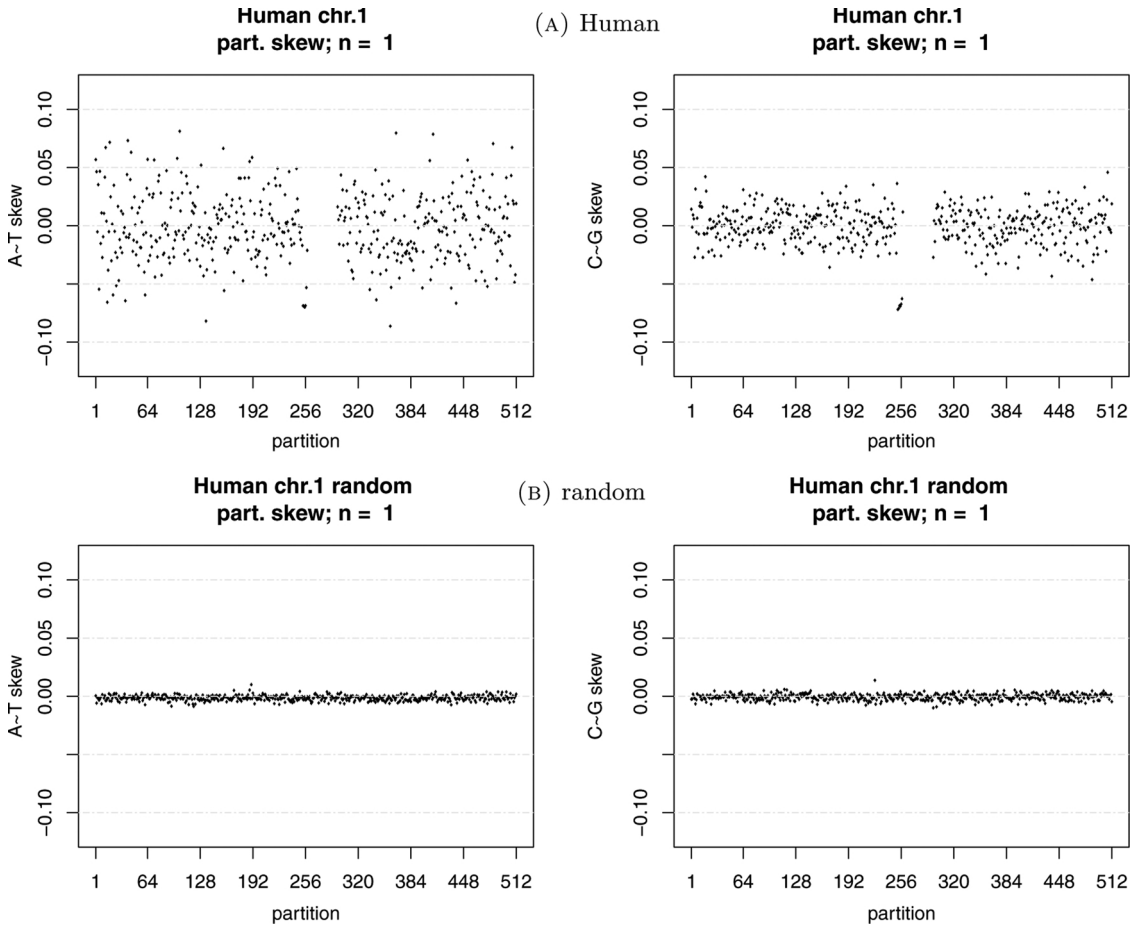

Fig. 6. A ~ T and C ~ G skew in human chromosome 1 compared to the randomly generated DNA divided into 512 partitions. The empty space in the middle of the chromosome 1 (a) is caused by bases classified as N (unknown) during the sequence assembly.

The main findings of the study can be summarized in the following way:

(1) In our work, we represent every DNA sequence as a sequence of n-plets. We show that each of these sequences of n-plets (at least up to $n = 50$) satisfies not only Chargaff's second rule but also its generalization in the following form:
In long sequences of n-plets of single stranded DNA, probabilities of the nucleotide N in their position k are approximately equal to the probability of this nucleotide N and at the same time to the probability of its complementary nucleotide ($A \leftrightarrow T$ and $C \leftrightarrow G$, correspondingly) in the DNA sequence of monoplets regardless of values n and k.
The generalized Chargaff's second rule could be empirically verified for all subsequences which are uniformly distributed with the firmly chosen distance $n = 1, 2, \ldots, 50$ between their bases in the sequences considered.

(2) In cases of DNA sequences of n-plets, one can also construct their model random sequences using knowledge about transition probabilities of arrangements of nucleotides A, T, C and G in each of original long DNA sequences. We reveal that in such random sequences the mentioned generalization of the second Chargaff's rule also holds true. And we can state:
The behavior of uniformly distributed subsequences of a long "real" DNA-sequence does not show noticeable differences in comparison with the randomly generated sequences of the same length (compare Fig. 2). This interesting fact can be explained in the following way: "When there are repetitive structures or correlations amongst different sections of a message, that reduces its capacity to convey new information – part of the variables are wasted in repeating what is already conveyed. Claude Shannon showed that the information content of a fixed length message is maximized when all the correlations are eliminated and each of the variables is made as random as possible." (Patel, 2001)

(3) Dividing long sequences in equal-sized parts and calculating relative differences of the complementary nucleotide bases (skews) for each of the parts, we have shown that the behavior of these skews is significantly different when compared with the corresponding values in randomly generated sequences in all organisms considered. Moreover, the variation of the A ~ T-skews is significantly higher than that of the C ~ G-skews (compare Fig. 6). An explanation of this fact can be connected to the hypothesis that various regions of long DNA sequences play different roles in the whole genetic informatics and for this reason they possess different distributions of nucleotides A, T, C and G.

(4) We quantified the notion of a "long DNA-sequence" using the accuracy measure suggested in Mascher et al. (2013). More precisely, we present how the standard deviations of relative differences (skews) of the bases A and T and C and G, correspondingly, depend on the length of the sequence chosen. To be precise, it has been shown that standard deviations of skews of the complementary bases stabilize and stay under the mark of about 0.02 if a sequence contains more than 10^5 nucleotide bases while the standard deviations are about three times bigger when considering sequences of length, for instance, about 25,000 bases.

With the present study, we hope to shed some light on the darkness surrounding the construction principles of the DNA. In our opinion, the findings of this work deserve comprehensive further investigation and contribute to understanding the "program that runs the cellular machinery" (Fickett and Burks, 1989).

Acknowledgments

We would like to thank Wiebke Werft, Lutz Strüngmann and Philipp Wiedemann for stimulating discussions and Thomas Ihme for providing the CUDA-server.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.biosystems.2019.04.003.

The supplementary material found in the attachment lists the full results for all five organisms.

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 103 (47), 17828–17833. https://doi.org/10.1073/pnas.0605553103.

Arqués, D.G., Michel, C.J., 1990a. Periodicities in coding and noncoding regions of the genes. J. Theor. Biol. 143, 307–318.

Arqués, D.G., Michel, C.J., 1990b. A model of DNA sequence evolution, part 1: statistical features and classification of gene populations, 743–753. Part 2: simulation model, 753–766. Part 3: return of the model to the reality, 766–770. Bull. Math. Biol. 52, 741–772.

Chargaff, E., Lipshitz, R., Green, C., 1952. Composition of the deoxypentose nucleic acids of four genera of sea-urchin. J. Biol. Chem. 195 (1), 155–160.

Chargaff, E., 1971. Preface to a grammar of biology: a hundred years of nucleic acid research. Science 172, 637–642.

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

Fimmel, E., Gumbel, M., Strüngmann, L., 2018. Exploring structure and evolution of the genetic code with the software tool GCAT. AIMEE 2017: Advances in Artificial Systems for Medicine and Education, 658 14–22. https://doi.org/10.1007/978-3-319-67349-3-2.

Kraljic, K., Strüngmann, L., Fimmel, E., Gumbel, M., 2018. Genetic code analysis toolkit: a novel tool to explore the coding properties of the genetic code and DNA sequences. SoftwareX 7, 12–14. https://doi.org/10.1016/j.softx.2017.10.008.

Mascher, M., Schubert, I., Scholz, U., Friedel, S., 2013. Patterns of nucleotide asymmetries in plant and animal genomes. BioSystems 111, 181–189.

Michel, C.J., 1986. New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation. J. Theor. Biol. 120, 223–236.

Nikolaou, C., Almirantis, Y., 2006. Deviations from Chargaff's second parity rule in organellar DNA. Insights into the evolution of organellar genomes. Gene 381, 34–41.

Okamura, K., Wei, J., Scherer, S.W., 2007. Evolutionary implications of inversions that have caused intra-strand parity in DNA. BMC Genomics 8, 160. https://doi.org/10.1186/1471-2164-8-160.

Patel, A., 2001. Quantum Algorithms and the Genetic Code. arXiv:quant-ph/0002037.

Perez, J.-C., 2010. Codon populations in single-stranded whole human genome DNA are fractal and fine-tuned by the Golden Ratio 1. 618. In: Interdisciplinary Sciences: Computational Life Science. 2, Nr. 3, September 2010 228–240. https://doi.org/10.1007/s12539-010-0022-0.

Petoukhov, S., 2017. The rules of long DNA-sequences and tetra-groups of oligonucleotides. arXiv:1709.04943v4 [q-bio.OT] b.

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

Qi, D., Cuticchia, A.J., 2001. Compositional symmetries in complete genomes. Bioinformatics 17, 557–559.

Rapoport, A.R., Trifonov, E.N., 2012. Compensatory nature of Chargaff's second parity rule. J. Biomol. Struct. Dynam. https://doi.org/10.1080/07391102.2012.736757.

Rosandic, M., Vlahovic, I., Gluncic, M., Paar, V., 2016. Trinucleotide's quadruplet symmetries and natural symmetry law of DNA creation ensuing Chargaff's second parity rule. J. Biomol. Struct. Dynam. 34 (7), 1383–1394. https://doi.org/10.1080/07391102.2015.1080628.

Shporer, S., Chor, B., Rosset, S., Horn, D., 2016. Inversion symmetry of DNA k-mer counts: validity and deviations. BMC Genomics 17 (1), 696.

Sobottka, M., Hart, A.G., 2011. A model capturing novel strand symmetries in bacterial DNA. Biochem. Biophys. Res. Commun. 410 (4), 823–828.

Watson, J.D., Crick, F.H.C., 1953. A structure for deoxyribose nucleic acid. Nature 171, 737–738.

Yamagishi, M.E.B., 2017. Mathematical Grammar of Biology. Springer. http://www.springer.com/us/book/9783319626888.

Zhang, S.-H., Huang, Y.-Z., 2010. Limited contribution of stemloop potential to symmetry of single-stranded genomic DNA. Bioinformatics 26, 478–485. https://doi.org/10.1093/bioinformatics/btp703.

Zhou, Y., Mishra, B., 2004. Models of genome evolution. In: Ciobanu, G., Rozenberg, G. (Eds.), Modeling in Molecular Biology. Springer, Berlin, pp. 287–304. https://cs.nyu.edu/mishra/PUBLICATIONS/03.evbydup.pdf.