Journal of Theoretical Biology 483 (2019) 109985

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Novel look at DNA and life—Symmetry as evolutionary forcing

Marija Rosandić[a,d], Ines Vlahović[b,c,*], Vladimir Paar[a,b]

[a] Croatian Academy of Sciences and Arts, 10000 Zagreb, Croatia
[b] Department of Physics, Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia
[c] Algebra University College, 10000 Zagreb, Croatia
[d] University hospital centre Zagreb (ret.), Zagreb, Croatia

ARTICLE INFO

Article history:
Received 8 February 2018
Revised 21 June 2018
Accepted 22 August 2019
Available online 27 August 2019

Keywords:
Chargaff's second parity rule
Strand symmetry
Trinucleotide classification
Genetic code table
Entropy

ABSTRACT

After explanation of the Chargaffś first parity rule in terms of the Watson-Crick base-pairing between the two DNA strands, the Chargaffś second parity rule for each strand of DNA (also named strand symmetry), which cannot be explained by Watson-Crick base-pairing only, is still a challenging issue already fifty years. We show that during evolution DNA preserves its identity in the form of quadruplet A+T and C+G rich matrices based on purine-pyrimidine mirror symmetries of trinucleotides. Identical symmetries are present in our classification of trinucleotides and the genetic code table. All eukaryotes and almost all prokaryotes (bacteria and archaea) have quadruplet mirror symmetries in structural form and frequencies following the principle of Chargaff's second parity rule and Natural symmetry law of DNA creation and conservation. Some rare symbionts have mirror symmetry only in their structural form within each DNA strand. Based on our matrix analysis of closely related species, humans and Neanderthals, we find that the circular cycle of inverse proportionality between trinucleotides preserves identical relative frequencies of trinucleotides in each quadruplet and in the whole genome. According to our calculations, a change in frequencies in quadruplet matrices could lead to the creation of new species. Violation of quadruplet symmetries is practically inconsistent with life. DNA symmetries provide a key for understanding the restriction of disorder (entropy) due to mutations in the evolution of DNA.

1. Introduction

Comparative genomic analyses revealed some universals of genome evolution in the form of nucleotide distributions or their specific relationships. The question is whether they reflect some fundamental "laws" of genome evolution or whether they are a kind of statistical patterns (Koonin, 2011). The idea that natural laws are associated with symmetry is present in science, but the symbiosis of mathematics and natural laws is still not fully understood (Wigner, 1969a). Emmy Noether achieved such spectacular result in 1918 when she proved her famous theorem, relating symmetry in time and the energy conservation law (König and Wiss, 1918; Gross, 1996; Kosmann-Schwarzbach, 2010). The use of the concept of symmetry has been spreading throughout science, for example Wigner (1969b), Muller (2003), Mainzer (2005), Zee (2007), Bindi et al. (2015), including biology (Monod, 1978; Bashford et al., 1998; Nikolajewa et al., 2005; Kong, 2009; Ramos et al., 2010; Yamagishi and Herai, 2011; Glazebrook and Wallace, 2012; Rosandić et al., 2013a; Rosandić and Paar, 2014; Afreixo et al., 2015). D.J. Gross expressed a general remark regarding the symmetry principle as the primary feature of nature (Gross, 1996): "We are embarked on a new stage of exploration of fundamental laws of nature, a voyage guided largely by the search for the discovery of new symmetries."

Jacques Monod attached great significance to symmetry and function in biological systems. He pointed out that the word symmetry, here, must not be understood in its purely geometrical connotation, but in the much wider sense (Monod, 1978): "The concept of symmetry becomes almost identical with that of order within a structure, whether in space or time, or purely in abstracto. The difficulties stem precisely from the extreme complexity of biological order, even though it often does express itself, partially, in some very simple and very obvious symmetry elements."

Chargaff's second parity rule stating a marked similarity between the frequencies of nucleotides and oligonucleotides and those of their respective reverse complements within each strand of sufficiently long (>100kb) double stranded DNA, is an interesting empirical pattern (Rudner et al., 1968; Prabhu, 1993; Qi and Cuticchia, 2001; Kong, 2009; Baisnee et al., 2002; Zhang and

* Corresponding author.
E-mail addresses: rosandic@hazu.hr (M. Rosandić), ines@phy.hr, ines.vlahovic@algebra.hr (I. Vlahović), paar@hazu.hr (V. Paar).

https://doi.org/10.1016/j.jtbi.2019.08.016

Huang, 2008; Perez, 2010; Sobottka and Hart, 2011; Mascher et al., 2013; Rapoport and Trifonov, 2013; Rosandić et al., 2013b; Zhang et al., 2013). This rule generally holds for double stranded DNA genomes, with the exception of some rare DNA symbionts, single-stranded genomes and organelles (Mitchell and Bridge, 2006; Nikolau and Almirantis, 2006).

Chargaff's first parity rule for pairs of A-T and C-G nucleotides between strands (Chargaff, 1951) was fully explained by the Watson-Crick pairing in the DNA double helix (Watson and Crick, 1953).

On the other hand, in spite of several proposals, a definitive explanation of Chargaff's second parity rule has not been fully accepted yet. Its fundamental cause is still somewhat controversial (Baisnee et al., 2002; Zhang and Huang, 2008; Mascher et al., 2013; Rapoport and Trifonov, 2013; Forsdyke and Bell, 2004; Chen and Zhao, 2005; Albrecht-Buehler, 2006, 2007; Okamura et al., 2007; Kong, 2009; Rosandić et al., 2016; Afreixo et al., 2016). It was suggested that Chargaff's second parity rule could probably exist from the very beginning of genome evolution. Information emerging from modern genome structures, in terms of small oligonucleotide frequencies could be helpful for the reconstruction of the primordial genome as well as for further understanding of the pattern of genome evolution. This information could shed light on the origin of genomes, and even on the origin of life (Zhang et al., 2013). Thus, it was noted that Chargaff's second parity rule could reveal general species-independent properties and have implications of some unknown mechanism that is likely to be present (Albrecht-Buehler, 2007; Rapoport and Trifonov, 2013).

As a basic feature of genomic sequences, oligonucleotide (in particular trinucleotide) frequency profiles have been used in studying genome evolution (Sobottka and Hart, 2011; Albrecht-Buehler, 2006, 2007; Rosandić et al., 2016; Afreixo et al., 2016; Zhang and Wang, 2011, 2012) . Regarding the importance of trinucleotides in the whole genomic sequences, new evidence was recently provided for their fundamental role in evolution (Frenkel and Trifonov, 2012; Frenkel et al., 2013).

In earlier investigations, the individual frequencies of all possible 64 different trinucleotides were determined in alphabetical ordering, and near equality of frequencies were found for reverse complement pairs within each strand as a characteristic of this binary system. We showed that in the trinucleotide quadruplet approach, a given genomic sequence is mapped into an array of 20 symbolic Q-diboxes, characterized by combined frequencies (Rosandić et al., 2016). Thus, the genomic sequence of concatenated nucleotides is mapped into a schematic presentation of weighted trinucleotides (characterized by frequency of occurrence), organized into twenty trinucleotide quadruplets, characterized by quadruplet symmetries (Table 1). These quadruplets embody the information of nucleotide content in a genomic sequence, but are disentangled from the information about the way in which these nucleotides are distributed throughout the genomic sequence. As we have recently shown, a consequence of quadruplet symmetries between both strands of DNA is Chargaff's second parity rule within the same strand (Rosandić et al., 2016).

A large number of sequenced genomes of different species open a wide framework for broader investigations of symmetries. Here we investigate the following related questions: a) Is the structure of DNA characterized by a random arrangement of nucleotides or was it created in the first species *ab ovo* according to some strict rules? b) Is a genome an "open book" where point mutations and indels appear randomly during life or are they subject to some strict rules? c) Is the entrance of nucleotides into DNA a random and unrestricted process or are there some limitations like self-protection of genome, i.e. how does DNA preserve its integrity or growth? d) In which way restricts the quadruplet symmetry an increase of disorder, which could arise from random mutations in genome? The Watson-Crick rule, binding purine from one strand and pyrimidine from the other, A↔T, G↔C, does not provide an answer to these questions because Watson-Crick pairing does not occur within the same strand of DNA.

2. Quadruplet classification of trinucleotides

The basic role of DNA is related to the genetic code. Its constituents in coding DNA are codons and, therefore, in noncoding sequences, we consider trinucleotides as basic entities. Thus, the four different nucleotides (A, T, C and G) provide 64 possible trinucleotide combinations, which are usually classified alphabetically. Because the alphabetical ordering is purely artificial, without any biological background, it cannot reveal any biochemical correlations.

Our quadruplet classification of trinucleotides is based on the following: each trinucleotide (denoted *D*) belongs to its quadruplet consisting of direct (D) - reverse complement (RC) - complement (C) - reverse (R) trinucleotides (Table 1). For example, the quadruplet corresponding to the ATG trinucleotide is: ATG (D), TAC (C), GTA (R), and CAT (RC). If the TAC trinucleotide within the same quadruplet is chosen as direct, then the other three members of the quadruplet are ATG (C), CAT (R) and GTA (RC). Thus each quadruplet consists of the same four trinucleotides, regardless of which of the four is taken as direct, and none of them belong to any of the other quadruplets.

We constructed the classification of trinucleotides from two quadruplet groups: ten A+T rich and ten C+G rich (Rosandić et al., 2013b, 2016), which encompass all 64 trinucleotides (Table 1). To each member of A+T rich quadruplet belongs a member from C+G rich quadruplet due to purine-pyrimidine transformation within the A, C and T, G amino-keto pairs: A→C, C→A and T→G, G→T. In this way, both A+T rich and C+G rich groups are segmented into three subgroups: 1) nonsymmetrical trinucleotides, each consisting of three different nucleotides, 2) nonsymmetrical trinucleotides, each consisting of two different nucleotides, and 3) symmetrical trinucleotides.

Symmetrical trinucleotides (subgroups Ic and IIc) contain duplicated trinucleotides because in each quadruplet the reverse is equal to the direct, and the complement is equal to the reverse complement. Therefore, each quadruplet frequency of direct and complement symmetrical trinucleotides should be divided by 2, because it simultaneously contains frequencies for both direct and reverse (for example, AAA(D)↔AAA(R) or TGT(D)↔TGT(R)), and also for their complement and reverse complement (for example, for the same case TTT(C)↔TTT(RC) or ACA(C)↔ACA(RC)).

The first four quadruplets in the A+T rich group are generated by codons like start/stop trinucleotides ATG, TGA, TAG, TAA. We made this choice because the corresponding codons have the well-known biological function as start/stop signals AUG, UGA, UAG, UAA. Each start/stop signal like trinucleotide belongs to its quadruplet in the A+T rich group.

We note that in the real DNA sequences the trinucleotides are not ordered in compact quadruplets, but organized nonlocally, and the computation of relative frequencies does not depend on location of trinucleotides. Therefore, the same results will characterize the shuffled genomes too. Also, it is important to note that each of 10 A+T rich and 10 C+G rich quadruplets of trinucleotides is specific and unique, consisting always of D, C, RC, R forms of the same trinucleotides that are mutually related by the Watson-Crick rule (A and C in one strand are coupled to T and G in the other strand, and vice versa, respectively.) Randomly combining any four mono/oligonucleotides cannot create quadruplet mirror symmetries (see later).

Our classification of trinucleotides enables the recognition of their quadruplet symmetry organization and purine-pyrimidine

Table 1
Quadruplet classification of 64 possible trinucleotides. Each quadruplet consists of trinucleotides denoted as direct (D) and its reverse complement (denoted RC(D) or shorter RC), complement (denoted C(D) or shorter C), and reverse (denoted R(D) or shorter R). Ten A+T rich and ten C+G rich quadruplets are organized in three subgroups. Ia (blue), consisting of nonsymmetrical trinucleotides containing four different nucleotides in D and RC. Ib (violet), consisting of nonsymmetrical trinucleotides containing two different nucleotides in D and RC. Ic (green), symmetrical trinucleotides which contain duplicated trinucleotides labeled with an asterisk (D=RC, C=*R*). First four A+T rich quadruplets are generated with start/stop signals like trinucleotides: ATG, TGA, TAG and TAA. The C+G rich trinucleotides correspond to purine-pyrimidine transformation of A+T rich trinucleotides within A (purine) in C (pyrimidine), and T (pyrimidine) in G (purine) amino-keto pairs. Three symmetries are present in our trinucleotides classification: 1) mirror symmetry between direct – reverse and complement – reverse complement in the same quadruplet; 2) purine-pyrimidine symmetries in each quadruplet, 3) purine (0) -pyrimidine (1) symmetries within and between A+T rich and C+G rich quadruplets in the same row.

A+T rich group (I)					**C+G rich group (II)**				
D	**RC(D)**	**C(D)**	**R(D)**	**Subgroup**	***D***	***RC(D)***	***C(D)***	***R(D)***	**Subgroup**
ATG 010	CAT 101	TAC 101	GTA 010	Ia	**CGT** 101	ACG 010	GCA 010	TGC 101	IIa
TGA 100	TCA 110	ACT 011	AGT 001		**GTC** 011	GAC 001	CAG 100	CTG 110	
TAG 100	CTA 110	ATC 011	GAT 001		**GCT** 011	AGC 001	CGA 100	TCG 110	
TAA 100	TTA 110	ATT 011	AAT 001	Ib	**GCC** 011	GGC 001	CGG 100	CCG 110	IIb
AAC 001	GTT 011	TTG 110	CAA 100		**CCA** 110	TGG 100	GGT 001	ACC 011	
AAG 000	CTT 111	TTC 111	GAA 000		**CCT** 111	AGG 000	GGA 000	TCC 111	
ATA 010	TAT 101	TAT* 101	ATA* 010	Ic	**CGC** 101	GCG 010	GCG* 010	CGC* 101	IIc
ACA 010	TGT 101	TGT* 101	ACA* 010		**CAC** 101	GTG 010	GTG* 010	CAC* 101	
AGA 000	TCT 111	TCT* 111	AGA* 000		**CTC** 111	GAG 000	GAG* 000	CTC* 111	
AAA 000	TTT 111	TTT* 111	AAA* 000		**CCC** 111	GGG 000	GGG* 000	CCC* 111	

symmetries (Table 1), which is present not only between the two DNA strands due to the Watson-Crick rule of nucleotide pairing, but also within the same strand. Here, this classification is modified with respect to the classification from Rosandić et al. (2013b). An important guideline for understanding DNA symmetries is provided by dividing trinucleotides into A+T rich and C+G rich, as will be shown later.

The role of symmetry was related to the concept of genetic code since the appearance of degeneracy can generally be a consequence of some kind of symmetry that acts as an organizing principle. Possible symmetries of genetic code and its broader scope have been considered (Findley et al., 1982; Forger et al., 1997; Kozirev and Khrennikov, 2010; Ramos et al., 2010; Barbieri, 2012; Michel and Pirillo, 2013; Rosandić and Paar, 2014; Shu, 2016). Barbieri investigated the general framework of genetic coding and copying. A step in the direction of a partially more symmetrical genetic code table was made by Shu (2016). A novel highly symmetric genetic code table was introduced (Rosandić and Paar, 2014), referred to as “ideal” with respect to the symmetry principle, with topologically connected trinucleotides, and based on three types of symmetries: purine-pyrimidine, A+T rich–C+G rich and direct-complement (Table 2). This “ideal” genetic code table differs from

Table 2
"Ideal" classification scheme of the genetic code. This genetic code table is created by trinucleotide sextets based on exact purine/pyrimidine symmetries, A+T rich / C+G rich symmetries and direct/complement symmetries. This genetic code table is modified with respect to Rosandić and Paar (2014). Italics: A+T rich codons; bold: C+G rich codons; leading group of codons: columns I and II; non-leading group of codons: columns III and IV; 0 purine, 1 pyrimidine. The "ideal" genetic code is determined by biochemical properties of sextets with serine as the "leader" in contrast to the classification of the standard genetic code where the alphabet key is used for nucleotides within codons (Rosandić and Paar, 2014). In our code, three symmetries are also present. First, A+T rich (italics) and C+G rich (bold) codons alternate between pairs of codon columns. Second, purine-pyrimidine structure of all four codon columns within the same row is identical and the consecutive pairs of codon rows also have an identical purine-pyrimidine structure. Third, boxes 5–8 and 13–16 (white) are complements of boxes 1–4 and 9–12 (light gray), respectively *and* vice versa. Amino acids are characterized by polarity, acid-base property, and an aromatic ring, approximately equally distributed between leading and non-leading groups of amino acids.

nonpolar	polar	basic	acidic	aromatic

Box	Leading group						Non-leading group					
	Amino acid	I. Codons	Pu/Py	Pu/Py	II. Codons	Amino acid	Amino acid	III. Codons	Pu/Py	Pu/Py	IV. Codons	Amino acid
Direct boxes 1-4	Start-Met	*AUG*	010	010	**GCA**	Ala	Val	**GUG**	010	010	*ACA*	Thr
	Ile	*AUA*	010	010	**GCG**			*GUA*	010	010	**ACG**	
		AUC	011	011	**GCU**			**GUC**	011	011	*ACU*	
		AUU	011	011	**GCC**			*GUU*	011	011	**ACC**	
Complement boxes 5-8	Tyr	*UAC*	101	101	**CGU**	Arg	His	**CAC**	101	101	*UGU*	Cis
		UAU	101	101	**CGC**			*CAU*	101	101	**UGC**	
	Stop	*UAG*	100	100	**CGA**		Gln	**CAG**	100	100	*UGA*	Stop
	Stop	*UAA*	100	100	**CGG**			*CAA*	100	100	**UGG**	Trp
Direct boxes 9-12	Glu	**GAG**	000	000	*AGA*		Lys	*AAG*	000	000	**GGA**	Gly
		GAA	000	000	**AGG**			*AAA*	000	000	**GGG**	
	Asp	**GAC**	001	001	*AGU*	Ser	Asn	*AAC*	001	001	**GGU**	
		GAU	001	001	***AGC***			*AAU*	001	001	**GGC**	
Complement boxes 13-16	Leu	**CUC**	111	111	*UCU*		Phe	*UUC*	111	111	**CCU**	Pro
		CUU	111	111	**UCC**			*UUU*	111	111	**CCC**	
		CUG	110	110	*UCA*		Leu	*UUG*	110	110	**CCA**	
		CUA	110	110	**UCG**			*UUA*	110	110	**CCG**	

the well-known "universal" genetic code table, where codons are organized alphabetically into the table which does not reveal complete symmetries.

3. Results

3.1. "Butterfly" quadruplet symmetries

It is empirically known that for a number of species the frequency of each direct trinucleotide in each DNA strand is very nearly identical to the frequency of its reverse complement (Albrecht-Buehler, 2006, 2007; Sobottka and Hart, 2011; Zhang and Wang, 2011; 2012; Frenkel et al., 2013; Rapoport and Trifonov, 2013; Rosandić et al., 2016; Afreixo et al., 2016). We have attributed this feature to the quadruplet symmetries of DNA (Rosandić et al., 2016) (Fig. 1).

All trinucleotides from A+T rich and C+G rich quadruplet groups, containing all four different types of nucleotides (A, T, C, G) in D and RC form of trinucleotides have the following numbers of nucleotides in each strand of the corresponding quadruplet: in

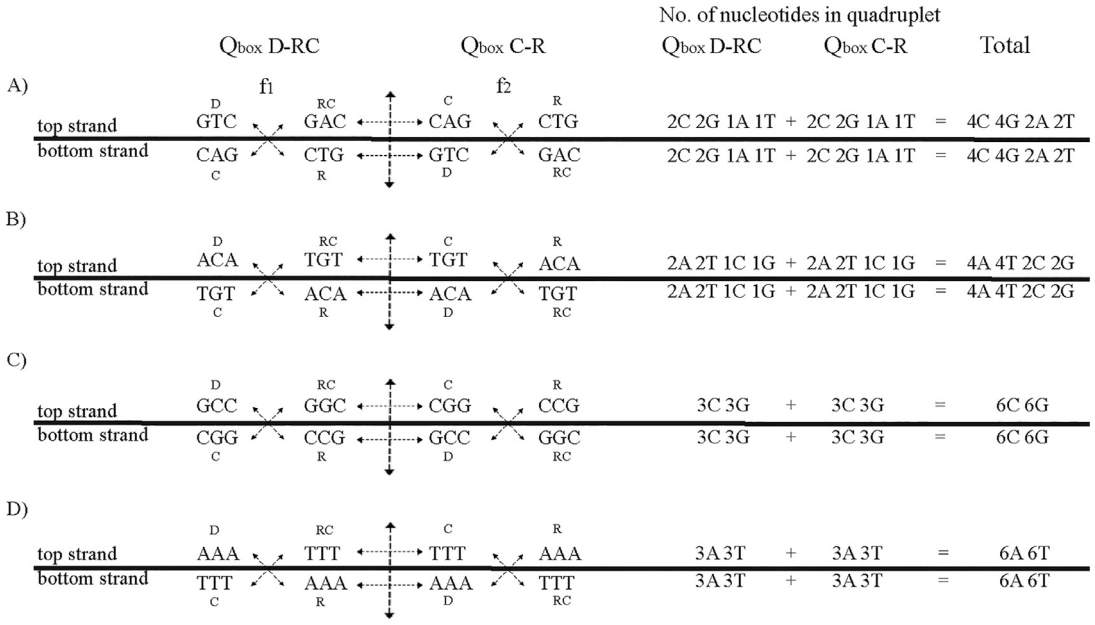


Fig. 1. Illustrations of "butterfly" symmetries of quadruplets leading to Chargaff's second parity rule. ↔, mirror symmetry within quadruplets. Reading unidirectional, the frequencies of all four members of the same quadruplet are identical in real genomes, while f_1(Qbox$_{D\text{-}RC}$) and f_2(Qbox$_{C\text{-}R}$) for asymmetric trinucleotides differ one from the other in genomes. Due to mirror symmetry, the frequencies D↔R between the two strands in the same quadruplets are identical. Therefore, the frequency between D↔RC members within the same quadruplet is also identical. In quadruplets where D and RC of trinucleotides contain all four nucleotides (A, T, C, G), the total number of nucleotides is the same in both strands, as illustrated by examples A) and B). Analogously, in quadruplets where D and RC contain only two different nucleotides, the total number of nucleotides is the same in both strands, as illustrated by examples C) and D). Identical frequencies of nucleotides A=T and C=G in both strands of the same quadruplet and their symmetries lead to Chargaff's second parity rule: the number of every nucleotide is equal to the number of its reverse complement in the same strand.

A+T rich quadruplets - 4A, 4T, 2C, 2G; and in C+G rich quadruplets - 4C, 4G, 2A, 2T.

All trinucleotides from A+T rich and C+G rich quadruplet groups containing nucleotides of only two different types, or of only one type in D and RC form of trinucleotides have the following numbers of nucleotides in their quadruplets in each strand: in A+T rich quadruplets - 6A, 6T; and in C+G rich quadruplets - 6C, 6G.

We see that in each quadruplet the number of nucleotides is A=T and C=G. Thus, the quadruplets act as basic building units. The multiplication of quadruplets in natural genomes does not change this relationship. Thus, the quadruplets appear as a non-local pattern dependent only on the frequency of occurrence of trinucleotide frequencies in each genome, but independent of their ordering within the DNA sequence. One could argue that in this way they resemble a kind of long-range interaction pattern. We show that the quadruplet organization of DNA with its symmetries solves the 50-years old problem of etiology of Chargaff's second parity rule. Accordingly, this rule would be a secondary consequence to quadruplet symmetries. We note that Afreixo et al. have recently investigated it using a sophisticated mathematical approach (Afreixo et al., 2015, 2016). They concluded that: "the exceptional symmetry is a local phenomenon in genome sequences of each chromosome".

The DNA genome has a complete quadruplet mirror symmetrical structure if in each strand it has identical frequencies of trinucleotides in D↔RC and C↔R pairs. Fig. 1 displays the symmetrical structure of quadruplets. Each quadruplet in a single strand of DNA creates two quadruplet *Q*-boxes simultaneously in both strands: Qbox$_{D\text{-}RC}$ and Qbox$_{C\text{-}R}$. Within each *Q*-box, there is a crossing mirror symmetry of trinucleotides between two strands and simultaneously there is also a mirror symmetry between both *Q*-boxes. Symmetries within each quadruplet correspond not only to trinucleotides, but also to their individual nucleotides (Fig. 2). In this way, each quadruplet in a strand creates a pattern of six identical quadruplets simultaneously in both strands: in each of the two strands one quadruplet, in each of the two *Q*-boxes one quadruplet, one laterally and one medially from both *Q*-boxes (Fig. 1). Simultaneously, both Q-boxes of quadruplets with their eight trinucleotides are built in fact of only two types of trinucleotides: one direct and its complement. They change their position inside the quadruplet *Q*-box construction from top to bottom strand or their configuration from direct to reverse. Inspired by mathematical and chemical purine-pyrimidine symmetries and their symmetrical beauty, we named this DNA pattern as quadruplet "butterfly" symmetries (Rosandić et al., 2016).

We note that when we apply unidirectional counting to both strands within the same *Q*-box, the relative frequencies of all four members of trinucleotides are identical (Fig. 3c). However, the relative frequencies differ in two different *Q*-boxes (Qbox$_{D\text{-}RC}$ and Qbox$_{C\text{-}R}$) belonging to the same quadruplet of nonsymmetrical trinucleotides. For example, in the top strand of GTC-generating quadruplet, the frequency f(GTC(D)) = f(GAC(RC)) (denoted by f_1) is equal, but it differs from f(CAG(C)) = f(CTG(R)) (denoted by f_2) in the same quadruplet. Accordingly, all trinucleotides in each quadruplet-box (Qbox$_{D\text{-}RC}$ or Qbox$_{C\text{-}R}$) have not only mirror symmetry in form, but also symmetry in frequency between the two strands. This rule applies to eukaryotes, prokaryotes like archaea and free-living bacteria and even some symbionts (see later).

The ten A+T rich and ten C+G rich quadruplets represent A+T rich and C+G rich matrices of relative trinucleotide frequencies, specific to each chromosome of eukaryotes or to the whole genome of prokaryotes. We note that identical frequencies of all four members of the same quadruplet in both strands of DNA can be obtained for unidirectional reading. This characteristic of quadruplets is used in our histograms. On the other hand, for bidirectional reading the quadruplet structure is reduced to a binary system, and each of the two *Q*-boxes is independent from each other and their mutual quadruplet connection, and quadruplet purine-pyrimidine symmetry is not recognizable. So, there is no D = RC, C = R frequency symmetry, and thus DNA in both strands is reduced to D = RC frequency relation only. In this way, if

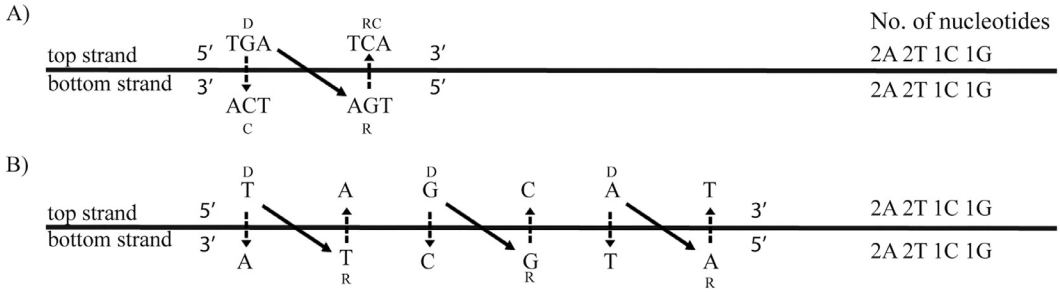
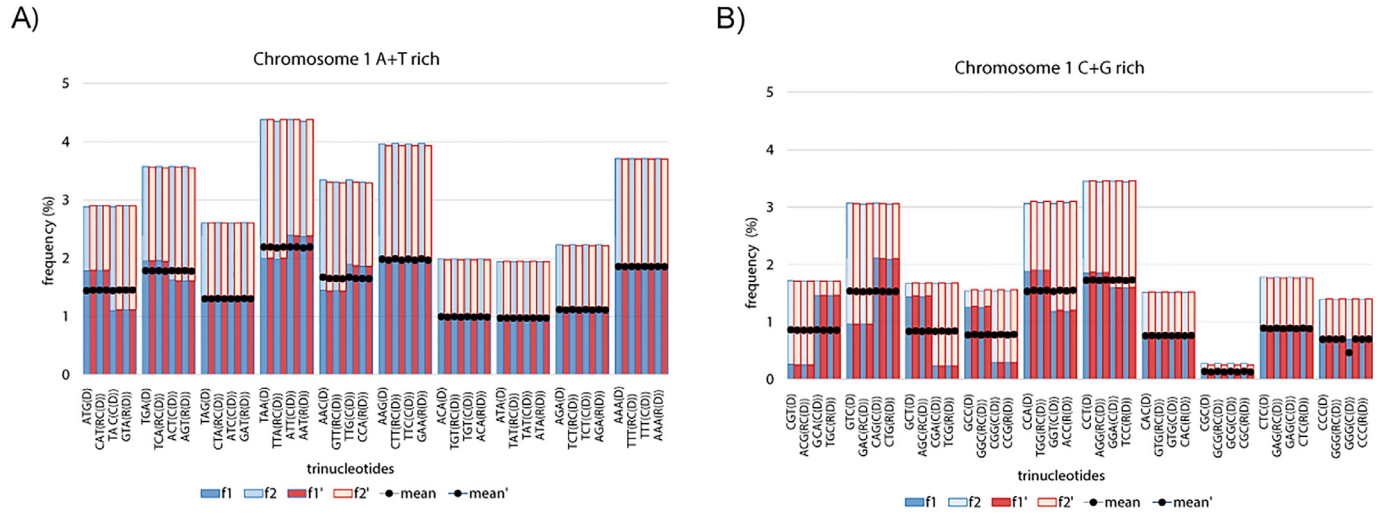


Fig. 2. Examples for Natural symmetry law of DNA creation and conservation. Inclined arrow ↘, direction of D↔R mirror symmetry with the entry of the same nucleotide into both strands of DNA according to the Natural symmetry law of the creation and conservation of DNA; Vertical arrows ↑↓, direction of the creation of complementary purine-pyrimidine base pairs in both strands of DNA according to the Watson-Crick rule. Regardless of whether A) trinucleotides or B) single nucleotide enters DNA, irrespectively of the location of the entry, the total number of inserted nucleotides in both examples is identical due to quadruplet symmetry. In this way, each mono/oligonucleotide mutation is also inserted into DNA like a "four-wheeled" system to keep quadruplet symmetries. Simultaneously, due to the Natural law of DNA creation and conservation bidirectional 5′→3′ orientation is established in both strands.



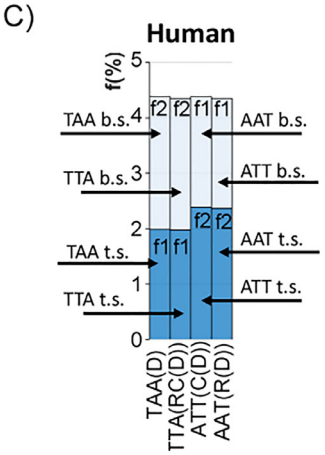


Fig. 3. Comparison of quadruplet matrices for chromosome 1 of *H. sapiens sapiens* (blue) and *H. sapiens neanderthaliensis* (red). A) Histogram of ten A+T rich quadruplets (A+T rich matrix, supplementary Table 1). B) Histogram of ten C+G rich quadruplets (C+G rich matrix, supplementary Table 1). Blue: *H. sapiens sapiens*; red: *H. sapiens neanderthalensis*; dark blue and dark red: top strand; light blue and light red: bottom strand. In each quadruplet we marked the mean values of relative frequencies of each trinucleotide to show its inverse proportionality between the top and bottom strands (i.e., between Qbox $_{D\text{-}RC}$ and Qbox $_{C\text{-}R}$). Relative frequencies in both strands of DNA for all four members of each quadruplet of A+T rich matrix and C+G rich matrix are identical in accordance with quadruplet symmetries. Beside inverse proportionality of the frequency of trinucleotide frequencies between $Qbox_{D\text{-}RC}$ and $Qbox_{C\text{-}R}$ in each quadruplet, inverse proportionality is shown also for A+T rich and C+G rich quadruplet matrices. In this way, the sum of trinucleotide relative frequency is conserved in every single quadruplet and in A+T rich and C+G rich quadruplet matrices. In summary, the frequency of trinucleotides in each matrix and the frequency sums in the whole genome are conserved. From the histogram, it can be seen that the A+T rich matrices as well as C+G rich matrices from chromosome 1 in both species are nearly the same. C) A detailed presentation of a segment of A+T rich TAA-quadruplet from the frequency histogram for human chromosome 1 (blue columns from A). Frequencies are displayed for the top strand (t.s.) (lower section of four columns), the bottom strand (b.s.) (upper section of four columns) and their sum for four trinucleotides in TAA-quadruplet: TAA(D), TTA(RC(D) (reverse complement of direct), ATT(C(D) (complement of direct)), AAT(R(D) (reverse of direct). The lower section of each column corresponds to the top strand (t.s.), and the corresponding frequency is denoted by f_1 for D, RC (TAA, TTA) columns and by f_2 for C, R (ATT, AAT) columns. The upper section (b.s.) in TAA and TTA columns corresponds to the frequency equal to f_2, and in the ATT, AAT columns to the frequency equal to f_1. Thus the combined two-strand frequency for each of the four columns of TAA-quadruplet is the same $f_1 + f_2$ sum. In each quadruplet there is an inverse proportionality between $f_1(Qbox_{D\text{-}RC})$ and $f_2(Qbox_{C\text{-}R})$.

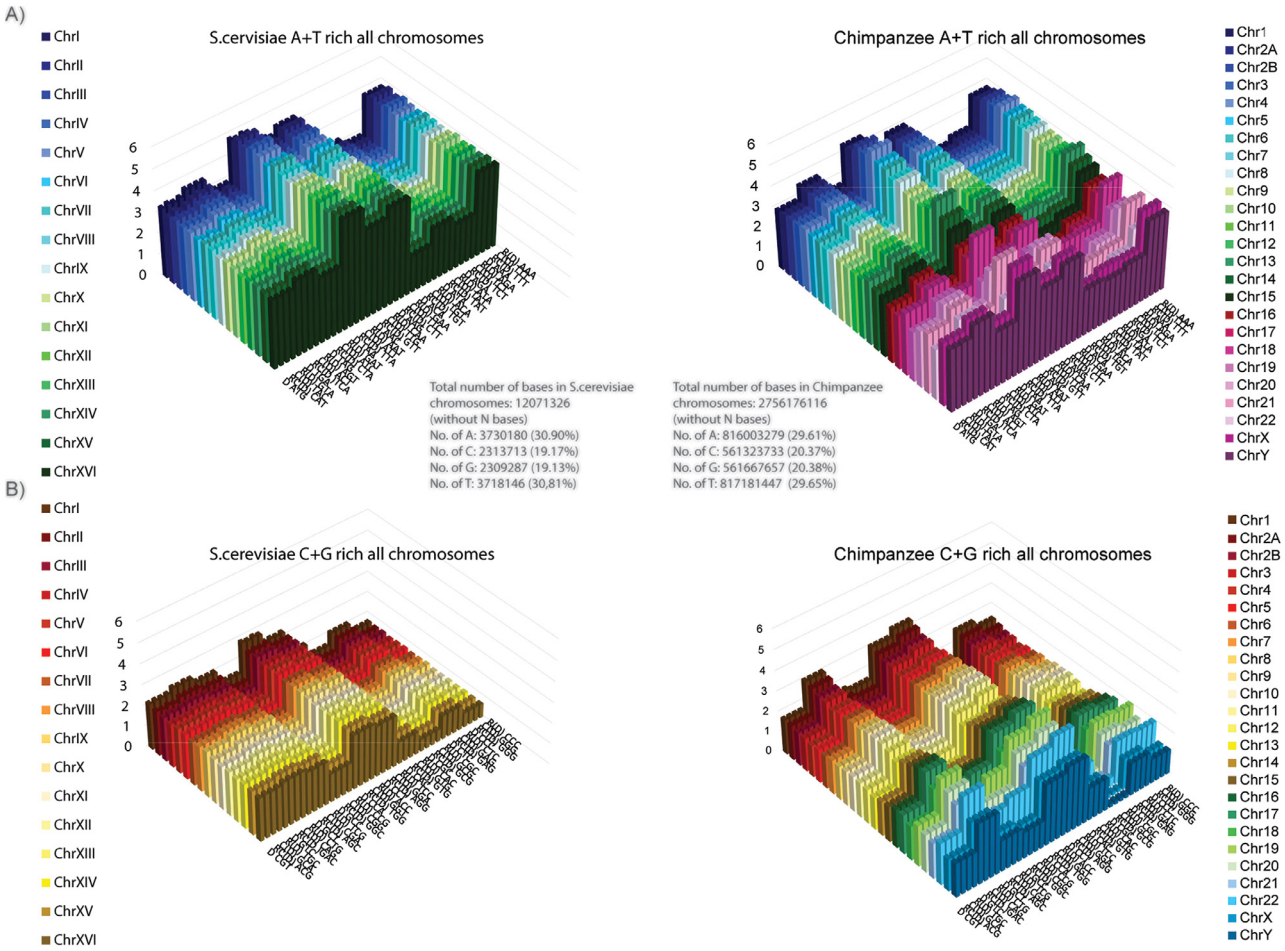


Fig. 4. 3D-diagrams from A+T rich and C+G rich matrices of all chromosomes in *Saccharomyces cerevisiae* and *Pan troglodytes*. *S. cerevisiae*: yeast; *Pan troglodytes*: chimpanzee. A) A+T rich quadruplet matrices, B) C+G rich quadruplet matrices. *S. cerevisiae* (as one of the first surviving eukaryotes) and *P. troglodytes* (as high primate) have very different number of nucleotides in whole genome and differ in number of chromosomes. Although they are evolutionary mutually distant by about 500 million years, their A+T rich and C+G rich matrices are mutually very similar with very small differences in relative frequencies (supplementary Table 3). It should be pointed out that this is pronounced for the first sixteen chromosomes in the chimpanzee and all yeast chromosomes; they may be considered "archaic" and responsible for the basic living processes. The other chimpanzee chromosomes are more specific. An analogous result was also obtained for *H.sapiens sapiens* (Rosandić et al., 2016).

we analyze DNA as a genetic code or replication and transcription, it is useful to employ the bidirectional 5′3→ 3′5 reading. However, if we analyze symmetries and numerical values within both strands of DNA, it is necessary to use unidirectional reading, because it involves the quartic system.

Relative frequencies of trinucleotides in quadruplets composed of all chromosomes in whole genome sequences of eukaryotes do not differ significantly from each other. They give rise to two basic chromosome matrices, A+T rich and C+G rich. However, even such small differences are specific for each individual species (Fig. 4). These differences can be easily detected and recognized with the trinucleotide quadruplet classification, but this cannot be done with the alphabetical ordering of trinucleotides.

The present study of quadruplet symmetries was extended to prokaryotes using combined mean values of relative frequencies of dinucleotides from 1309 bacteria and 133 archaea genomes (Zhang and Huang, 2008; Zhang et al., 2013) compared to our present calculated results for mean values of relative dinucleotide frequencies for all human chromosomes (Fig. 5). These results are in accordance with strand symmetry and with quadruplet symmetry.

A significant difference between human and bacteria/archaea genomes appears in dinucleotide quadruplet [GC(D)-GC(RC), CG(C)-CG(R)]: for human genome the relative frequencies are much smaller than for bacteria and archaea. As compensation for lower values of this quadruplet in human DNA, the inverse proportionality of quadruplet [AA(D)-TT(RC), TT(C)-AA(R)] takes place. Namely, in human genome the ratio of A+T to C+G rich nucleotides is 60:40, because of contribution from A+T rich noncoding DNA, while in bacteria and archaea the ratio is about 50:50 or C+G prevailing because of abundant C+G rich coding DNA.

Analyzing the frequencies of higher-order oligonucleotides in human genome, it was concluded that strand symmetry holds for oligonucleotides up to 6 nucleotides and is no longer statistically significant for oligonucleotides of higher orders (Afreixo et al., 2013). Zhang (2015) pointed out that further analysis shows that strand symmetry would persist for higher order oligonucleotides up to 9 nucleotides in the human genome and that symmetry would break gradually, but not abruptly. Afreixo et al. (2015, 2016) have shown that human genome exhibits high local exceptional symmetry.

We show that, in addition to mononucleotides and trinucleotides, the quadruplet structure is also present for oligonucleotides with a larger number of nucleotides (2, 3, 4, 5, 6, 7, …) but possible combinations of nucleotides become increasingly numerous (4^2 = 16, 4^3 = 64, 4^4 = 256, 4^5 = 1024, 4^6 = 4096, 4^7 = 16,384,… 4^{10} = 1,048,576). Therefore, for the identification of quadruplet symmetries and Chargaff's second parity rule with an increasing number of nucleotides, much larger DNA sequences

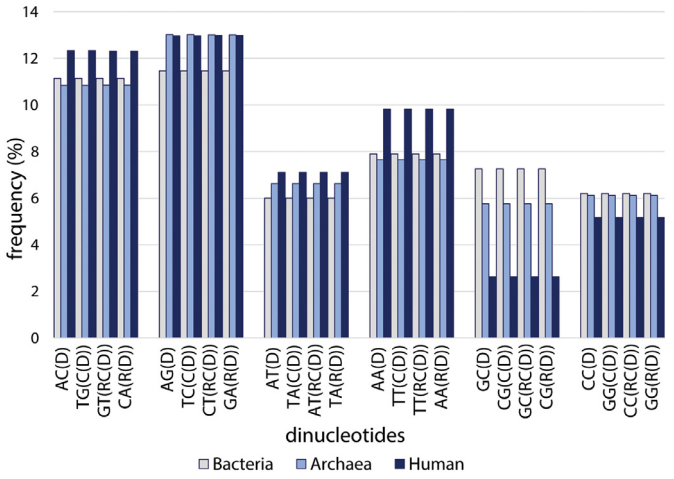


Fig. 5. Comparison of dinucleotide quadruplets in bacteria, archaea and human genomes. For each quadruplet relative frequencies in both strands of DNA are determined using the mean values of dinucleotide frequencies of 1309 bacteria and 133 archaea (Zhang and Huang, 2008) and mean dinucleotide frequencies of all human chromosomes calculated in this work. In two quadruplets, [AT(D)-AT(RC), TA(C)-TA(R)] and [GC(D)-GC(RC), CG(C)-CG(R)] the same dinucleotides appear for direct and reverse complement, and also for complement and reverse dinucleotides. Therefore, the frequencies obtained for AT, TA and GC, CG are divided by 2. Analogous case appears for quadruplets [AA(D)-TT(RC), TT(C)-AA(R)] and [CC(D)-GG(RC), GG(C)-CC(R)]. For human genome the inverse proportionality between very low relative frequencies in [GC(D)-GC(RC), CG(C)-CG(R)] and [AA(D)-TT(RC), TT(C)-AA(R)] quadruplets are emphasized. All these species satisfy strand symmetry, i.e., the Chargaffś second parity rule, as seen from identical frequency values of all four dinucleotides in each quadruplet (see the text).

would be needed. Quadruplets with 1,048,576 combinations of 10 nucleotides in oligonucleotides explain why only rare $f(\mathrm{D}) = f(\mathrm{RC})$ groups exist in each strand of the human genome. For this reason, the genome quadruplet structure explains the perplexing question of why the equality of D = RC frequencies gradually disappears for longer oligonucleotides.

3.2. The natural symmetry law of DNA creation and conservation

The crucial question is how the DNA genome preserves the quadruplet structure and the total number of nucleotides specific for each species during the whole evolution. We have shown that DNA is a well-organized molecule due to its mirror symmetries within each of the 20 different constituent quadruplets (Rosandić et al., 2016). Accidental grouping of nucleotides cannot create quadruplet symmetries. On the other hand, random insertions of mono- or oligonucleotides in spite of the Watson-Crick rule would dissolve the quadruplet structure with their symmetries, i.e., Chargaff's second parity rule within the DNA molecule, so no species would be able to preserve the same total number of nucleotides in its genome.

Nature has a tendency for simple solutions. Accordingly, it follows that the DNA molecule was created under the influence of the Natural symmetry law of DNA creation and conservation. This law states that the same mono/oligonucleotide enters not only one strand but simultaneously both strands in direct-reverse 5′3↔3′5 direction and each of them is coupled with its complement pair in the opposite strand (Figs. 1 and 2). In this way, the complete "butterfly" quadruplet D-RC, C-R symmetries in both strands are created.

In accordance with the Natural symmetry law of DNA creation and conservation, an inverse proportionality of nucleotide frequencies appears also between $Q\mathrm{box}_{\mathrm{D\text{-}RC}}$ and $Q\mathrm{box}_{\mathrm{C\text{-}R}}$. If one increases, the other decreases or vice versa (Figs. 1 and 3c). Quadruplets composed of symmetrical trinucleotides have the same frequencies in both *Q*-boxes, as already discussed. In this way, the same frequencies of all members of the same quadruplet are preserved (Fig. 3). The next step in protecting the total number of nucleotides in the whole genome also involves an inverse proportionality between the frequencies of A+T rich and C+G rich quadruplet matrices. Namely, the inverse proportionality relation is also present for quadruplet frequencies between the corresponding purine-pyrimidine bases of A+T rich and C+G rich groups of trinucleotides from our trinucleotide classification (Table 1). An example of this inverse proportionality is the A+T rich quadruplet frequency *f*(ATG(D)-CAT(RC)-TAC(C)-GTA(R)) that is opposite to the corresponding C+G rich quadruplet frequency within the A↔C and T↔G amino-keto pairs *f*(CGT(D)-ACG(RC)-GCA(C)-TGC(R)) (Table 1).

This is shown by comparing chromosome 1 and 22 in evolutionary adjacent species in *H. sapiens neanderthalensis* and *H. sapiens* (Figs. 3, 6, Table S2). In all quadruplets from the A+T rich matrix of chromosome 22 in *H. sapiens neanderthalensis*, the frequencies of A and T nucleotides decrease, while in all quadruplets of C+G rich matrix the frequencies of C+G nucleotides simultaneously increase. In their chromosome 1, this difference is absent (Fig. 3). It should be stressed that in the alphabetic ordering of trinucleotides, without the organization into A+T rich and C+G rich quadruplets, these relations between trinucleotides are not recognizable.

Consequently, the inverse proportionality relation between the frequencies of trinucleotides in both *Q*-boxes, as well as between A+T rich and C+G rich quadruplets keeps the average value of trinucleotide frequencies the same within each quadruplet and within the whole genome. In this way, a kind of circular cycle for the protection of the genome is closed (Fig. 7).

One might argue that quadruplet features would be a secondary consequence to strand symmetry (Chargaff's second parity rule) rather than the cause of it. However, DNA is an autonomous system. It has no miraculous properties with which to recognize the D↔RC form of trinucleotides, e.g. ATG↔CAT, GCT↔AGC, TTT↔AAA in the same strand. We show that, according to the natural symmetry law of creation and conservation, DNA accepts mono/oligo mutations in such a way that the same nucleotides enter both strands of DNA simultaneously regardless of their locations. According to the Watson-Crick rule, mutations also "catch" their complements in opposite strands (A↔T, C↔G) and the core of the quadruplet symmetries is created. We show that in this way each quadruplet fulfills Chargaff's second parity rule and accordingly is a consequence and not the cause of quadruplet features.

The current hypothesis of strand symmetry (Albrecht-Buehler, 2006) is that the inversion/inverted transpositions have been numerous during evolution to create the strand symmetry. However, evolution is a continuous process, posing some questions: 1) How is the process of inversion/inverted transposition interrupting once the symmetries are established? 2) Why are transitional forms today practically absent? 3) Does it mean that the evolution stops after symmetries are established?

Analyzing fascinating purine-pyrimidine quadruplet mirror symmetries, we established our hypothesis of Natural symmetry law of DNA creation and conservation that in its simplicity reminds of Occamś razor: the Natural symmetry law stating that at the beginning, the insertions/deletions must simultaneously enter/exit into both strands of DNA. In conjunction with the Watson-Crick rule this automatically leads to quadruplet symmetries. Thus the Watson-Crick pairing by itself is not sufficient to lead to symmetries and to explain DNA creation. In analogy to the natural laws embedded in the structure of Universe, on the basis of our investigations we suggest that the Natural symmetry law of DNA creation and conservation is embedded in the molecule of life.

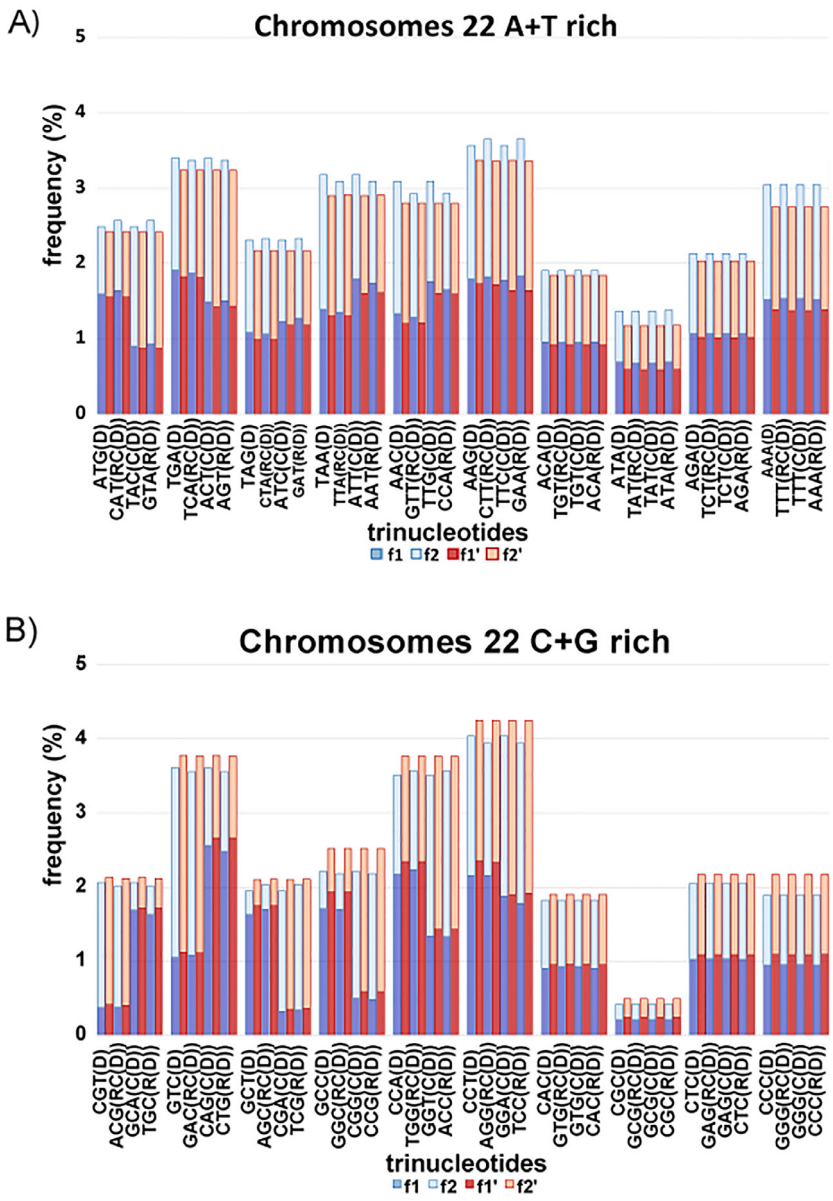


Fig. 6. Comparison of quadruplet matrices for chromosome 22 of *H. sapiens sapiens* (blue) and *H. sapiens neanderthaliensis* (red). Blue: *H. sapiens sapiens*; red: *H. sapiens neanderthaliensis*. A) A+T rich and B) C+G rich trinucleotides. It can be seen that with respect to *H. sapiens sapiens*, in all quadruplets of C+G rich matrix of *H. sapiens neanderthaliensis* the frequencies of C and G nucleotides increase while in A+T rich matrix inverse proportionally of the frequencies of A and T nucleotides decrease (supplementary Table 2). In C+G rich matrix the increase is most pronounced for trinucleotides consisting of only C and G nucleotides, while in A+T rich matrix the decrease is most pronounced for trinucleotides consisting of only A and T nucleotides. This inverse proportionality between A+T rich and C+G rich matrices preserves the total number of nucleotides in the whole genome.

3.3. Quadruplet symmetries and Chargaff's second parity rule in eukaryotes and prokaryotes

We have shown that the quadruplet DNA symmetry structure and the ensuing Chargaff's second parity rule appear in prokaryotes (bacteria and archaea) and all eukaryotes – from *S. cerevisiae*, among the evolutionary oldest living eukaryotes, all the way to hominids, including the extinct *H. sapiens neanderthalensis* and the contemporary *H. sapiens sapiens* (Rosandić et al., 2016). Exceptions are rare prokaryotes like some bacteria with extremely reduced genomes (McCutcheon and Moran, 2012; Bondarev et al., 2013; Aruni et al., 2015): *Candidatus tremblaya princeps* (138,927 bp) (Fig. 7a,b, Table S4), *Candidatus hodgkinia cicadicola* (143,795 bp) and bacteria with reduced genomes: *Filifactor alocis* (1,931,012 bp), and *Pseudovibrio* (5,916,833 bp). These genomes contain 20 quadruplets, but only one mirror symmetry in each of the two strands following the Watson-Crick trinucleotide pairing D-C↔RC-R. For example:

$$\begin{array}{l} \text{top strand} \quad \overset{\text{Q-box}_{\text{D-C}}}{\text{ATG TAC}} \leftrightarrow \overset{\text{Q-box}_{\text{RC-R}}}{\text{CAT GTA}} \\ \text{bottom strand} \quad \text{TAC ATG} \leftrightarrow \text{GTA CAT} \end{array}$$

Their quadruplets create $\text{Qbox}_{\text{D-C}}$ and $\text{Qbox}_{\text{RC-R}}$ with an identical frequency sum between both strands inside each box but not between the two boxes. They also show the structural mirror symmetry in form in each strand. Symmetries are at the core of our study and now we can say that all DNA species have some form of quadruplet symmetries. However, they do not comply with the Chargaff's second parity rule, but as explained in Fig. 8, they

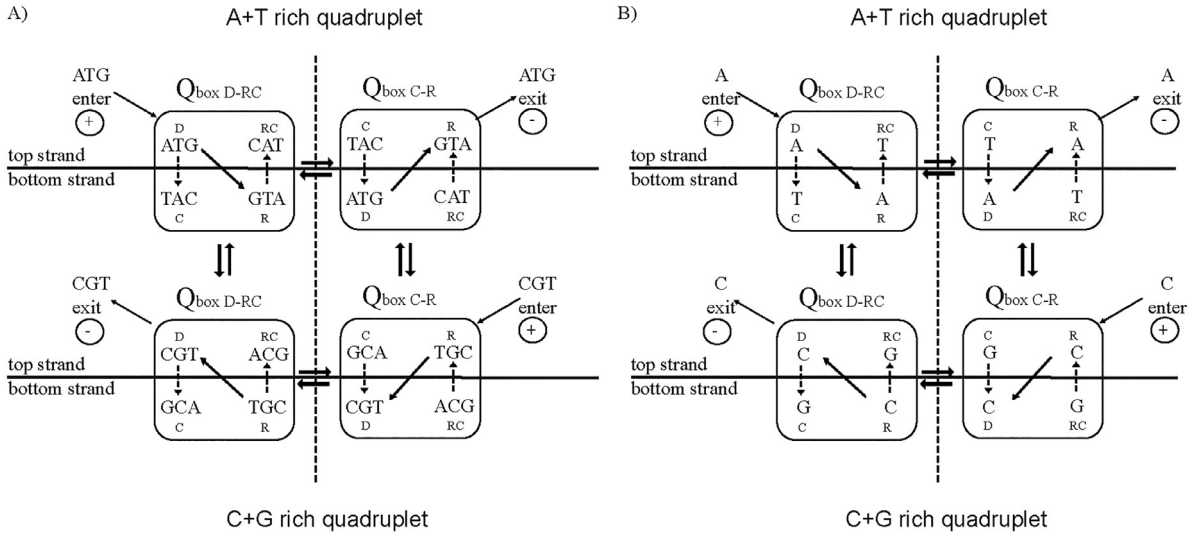


Fig. 7. The "circular cycle" of quadruplet for relative frequency protection. The entry of A) trinucleotide (ATG) or B) mononucleotide (A) into the top strand of DNA leads to the entry of all four members of the same quadruplet in the form of $Qbox_{D\text{-}RC}$ due to the natural symmetry law of the creation and conservation of DNA. Simultaneously, the same trinucleotides exit inverse proportionally from $Qbox_{C\text{-}R}$ (see real illustration in Fig. 3a,b) or vice versa. As a result, quadruplet symmetries are preserved and the frequencies of quadruplet members in both strands of DNA remain the same. Analogously, the entry of trinucleotide into, for example a A+T rich quadruplet, leads inverse proportionally to the exit of its purine-pyrimidine pair from a C+G rich quadruplet (see real illustration in Fig. 6a,b) and vice versa. In this way, a closed "circular cycle" is established and the total number of nucleotides in the genome remains unchanged.
Every insertion of a nucleotide (for example A) leads to the entry of its complement partner (T) in another strand. In relative percentages, this simultaneously leads to decrease of complementary C-G base pairs. The rule is valid for quadruplets as well as for A+T and C+G rich matrices. At the same time, the Natural law of DNA creation and conservation has the role of creating quadruplets and their symmetries, which results in Chargaffś second parity rule.

exhibit a tendency toward this equality. Despite the differences in genome size, these bacteria are all symbionts (McCutcheon and Moran, 2012; Bondarev et al., 2013; Aruni et al., 2015). However, some symbionts, even with extremely reduced genomes like *Candidatus carsonella ruddi* (162,589 bp), are characterized by quadruplet symmetries based on Chargaff's second parity rule (Fig. 7). It is shown that the same complete symmetries and Chargaff's second parity rule are present in 1309 free-living bacteria like *Escherichia coli* or *Helicobacter pylori* (see Material and Methods) and 133 archaea (Zhang and Huang, 2008; Rosandić et al., 2016). We have explained that an organism that fulfills Chargaff's second parity rule also complies with the Natural symmetry law of DNA creation and conservation.

4. Discussion

Each violation of the circular cycle of trinucleotides causes significant consequences for balance according to the rules of the Natural symmetry law of DNA creation and conservation. Changes of trinucleotide frequency ratio between trinucleotides in *Q*-boxes as well as in A+T rich and C+G rich chromosome matrices result in an increase or decrease in the total number of trinucleotides in the whole genome and create the possibility for the evolution of new species. On this basis, we argue that evolution is a consequence of accidental mutations and repositioning within DNA, but that this is carried out under strict rules of the Natural symmetry law of DNA creation and conservation in all examined eukaryotes and in prokaryotes, like free-living bacteria and archaea, and most of the symbionts.

In this framework, we show that evolution is a dynamical interplay between divergent mutational forcing and natural selection that allows the development of new species, and on the other hand, of convergent forcing in the form of symmetries, which introduce order and thus protect newly created species. Thus, we could say that the random processes of mutation increases disorder in the biological system, i.e. contributes to the increase of entropy in the biological system, while the symmetry forcing imposes an increase of order, i.e., causes a decrease of disorder (entropy) in the biological system. These two opposing tendencies may resemble rich phenomena of nonlinear dynamical systems present in nature (Bak, 1996; Deisboeck and Kresh, 2006).

In classical Darwinism, the mechanism of biological evolution consists of random mutations and natural selection. According to the present state of art, the entropy (disorder) increase during evolution only through random mutations presents a challenge. We argue that a mechanism for the explanation of this problem is the Natural symmetry law of DNA creation and conservation within genome. This automatically restricts disorder, i.e. the increase of entropy during genome evolution, while simultaneously enabling the evolution of species.

"In its full complexity the question "What is life" is multifaceted, opening many pathways and challenges (Schrödinger, 1944; Rosen, 1991; Maturana and Varela, 1980; Ganti, 2003). Here, we might ask "What is life viewed in light of DNA quadruplet symmetries?" related to the problem of increase in order (decrease in entropy) during evolution. In this study, we were surprised to find that the basic A+T rich and C+G rich quadruplet matrices of relative nucleotide frequencies are similar for all chromosomes (Fig. 4 and Table S3), for example in the S. cerevisiae, one of the simplest living eukaryotes from Kingdom Fungi, and in the chimpanzee, which is five hundred million years younger. It appears that primarily the whole matrices with their quadruplet symmetries are largely multiplied during evolution, and not only individual trinucleotides or their quadruplets."

Both A+T rich and C+G rich matrices should be considered as basic units of the quadruplet pattern. At the same time, matrix multiplication ensures symmetries and the optimal nucleotide number for all codons in coding the genome for the synthesis of 20 natural amino acids for proteinogenesis necessary for the creation of specific species. Apart from the coding part of DNA, the presence of large noncoding DNA sequences is also needed in order to create all 20 quadruplets with their symmetries. Only by such combination, the quadruplet symmetry structure of the genome is preserved.

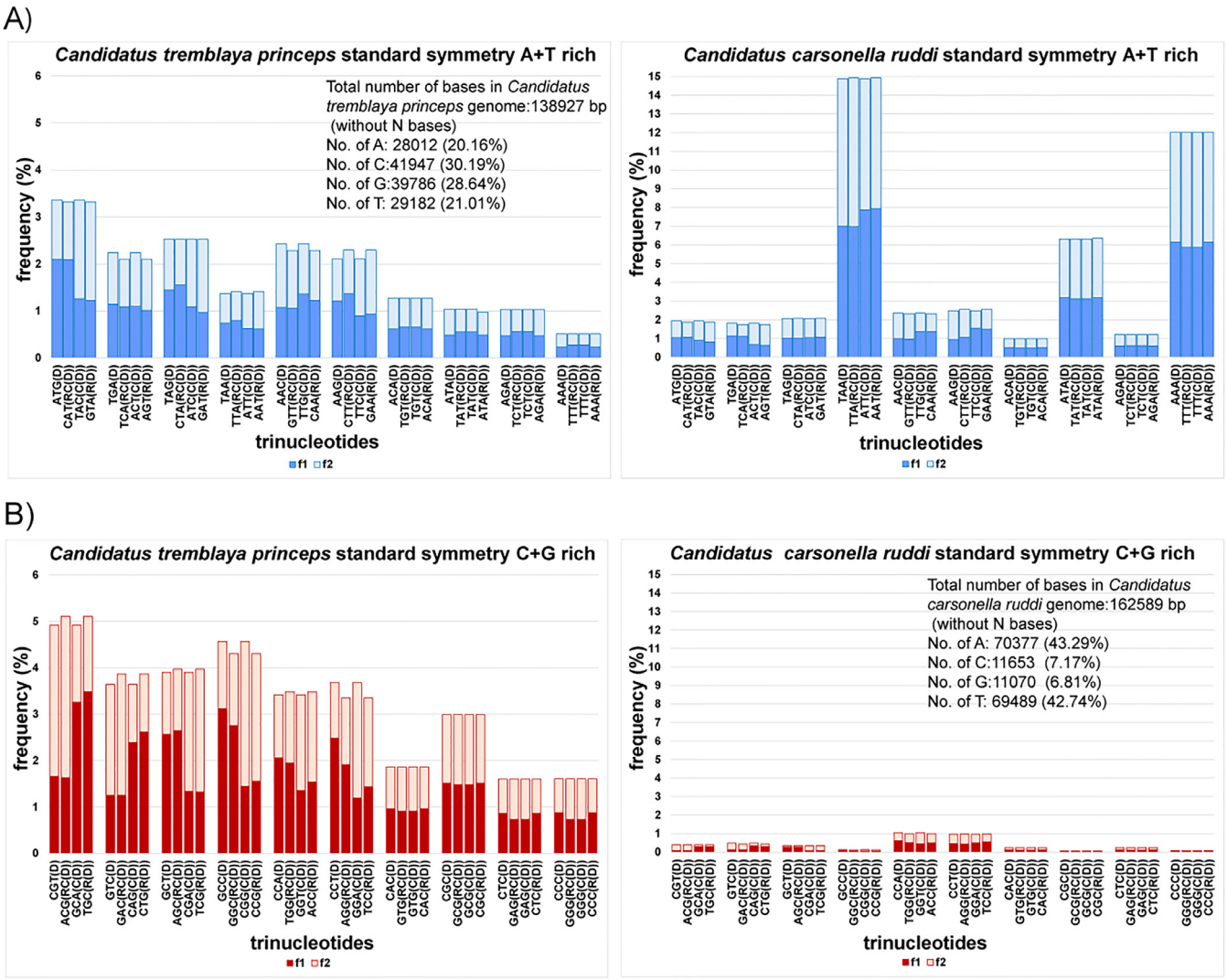


Fig. 8. Quadruplet matrices for symbiont bacteria with an extremely reduced genome: (138,927 bp) *Candidatus tremblaya princeps* and *Candidatus carsonella ruddi* (162,589 bp). A) A+T rich quadruplet matrix: dark blue top strand, light blue bottom strand; B) C+G rich quadruplet matrix: dark red top strand, light red bottom strand. *C. tremblaya princeps*: quadruplet matrices show the Watson–Crick pairing $f(D) = f(C)$ and $f(RC) = f(R)$ between the two strands, but some quadruplets (A+T rich: ATG, TAG, ACA and C+G rich: GCT, CAC, CGC) tend to equality $f(D) = f(RC)$, and $f(C) = f(R)$ and exhibit quadruplet symmetries on the principle of Chargaff's second parity rule. *C.carsonella ruddi:* All Quadruplet matrices exhibit symmetries on the principle of Chargaff's second parity rule. Evolutionary close species could have very different quadruplet matrices between two symbiont bacteria like *Candidatus tremblaya princeps* and *Candidatus carsonella ruddi*, as shown in this figure. The difference in relative percentages between quadruplets in symbionts is the consequence of adjustment of individual symbiont to its host. On the other hand, the quadruplet matrices are mutually similar between eukaryotes with a large evolutionary distance, as for example yeast and chimpanzee (compare with Fig. 4).

5. Conclusion

We argue that for all living organisms the following fundamental principle for the structure of DNA is present: quadruplets with mirror symmetries between two identical purines (A–A, G–G) and between two identical pyrimidines (T–T, C-C). As we show in some rare symbionts, there is only one quadruplet mirror symmetry within each strand of DNA. In eukaryotes and in free-living bacteria and archaea among prokaryotes, there are two quadruplet mirror symmetries: within each strand and between the two strands of DNA. Thus, the mirror symmetries between the two strands in a quadruplet are a fundamental influence of the Natural symmetry law of DNA creation and conservation of DNA. The understanding of DNA quadruplet symmetries resulted in our classification. We recognize sophisticated symmetries and incorporate them in our extraordinary "ideal" genetic code. We show that DNA quadruplet mirror symmetries solve the etiology of Chargaff's second parity rule. Symmetry forcing imposes an increase of order and explains the decrease of disorder (entropy) in the biological system while simultaneously enabling the evolution of species. We might hypothesize that because of strong DNA quadruplet symmetries, possible consequences could arise due to some uncontrollable interventions into the structure of genome. The Natural symmetry law of DNA creation and conservation can contribute to the most prominent role of the DNA molecule in the creation and evolution of life.

6. Materials and methods

In order to calculate quadruplet frequencies, we downloaded a sequence from the NCBI site for the human (ftp://ftp.ncbi.nih.gov/genomes/Homo_sapiens/ARCHIVE/ANNOTATION_RELEASE.107/Assembled_chromosomes/seq/, 6.1.2015.), chimpanzee (ftp://ftp.ncbi.nih.gov/genomes/Pan_troglodytes/ARCHIVE/ANNOTATION_RELEASE.103/Assembled_chromosomes/seq/, 27.1.2015), S.cerevisiae genome (http://www.ensembl.org/Saccharomyces_cerevisiae/Info/Index, 28.1.2015), for *Candidatus tremblaya princeps* (annotation GCF_000219195.1_ASM21919v1, ftp://ftp.ncbi.nih.

gov/genomes/refseq/bacteria/Candidatus_Tremblaya_princeps/all_assembly_versions/suppressed/, 14.10.2016) and for *Candidatus Hodgkinia cicadicola* (ftp://ftp.ncbi.nih.gov/genomes/refseq/bacteria/Candidatus_Hodgkinia_cicadicola/all_assembly_versions/suppressed/GCF_000021505.1_ASM2150v1, 14.10.2016). The sequence for *Klepsiella pneumonie, Bacillus subtilis, Rickettsia prowazekii* str. Breinl and *Mycoplasma genitalium* were downloaded from the following links: ftp://ftp.ensemblgenomes.org/pub/bacteria/release-33/fasta/bacteria_21_collection/klebsiella_pneumoniae_subsp_pneumoniae_dsm_30,104/dna/, https://www.ncbi.nlm.nih.gov/nuccore/255767013?report=genbank, http://bacteria.ensembl.org/Rickettsia_prowazekii_str_breinl/Info/Index, https://www.ncbi.nlm.nih.gov/nuccore/108885074. We got the Neanderthal genome sequence from the Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig. From files in the VCF format, we extracted nucleotide sequences. For calculating trinucleotides, we used a custom-made computational program in c# that calculates trinucleotide frequencies from the DNA sequence for nucleotides in the step of one sliding window (we neglected the trinucleotides containing the *N* base, the program "CLT__Find" is free for download at http://genom.hazu.hr/tools.html). We extracted the coding sequence for human chromosome 1 using a custom-made program also in c#, from ftp://ftp.ensembl.org/pub/release-83/fasta/homosapiens/cdna/, downloaded on March 9, 2016.

Relative frequencies of all 1309 bacteria and 133 archaea quadruplets are calculated for the whole genomes of bacteria and archaea, and in all human chromosomes. Because of very large numbers involved we donť perform randomization of genome, that would lead to similar percentages (25%) of A, T, C and G nucleotides, as can be inferred from our previous work (Rosandić et al., 2016, Fig. 4).

We performed randomization for all human chromosomes by randomly selecting segments of 100 bases and forming a combined randomized sequence of 200 000 bp and computing relative frequencies of trinucleotides from quadruplets. In such reduced randomized sequences for all chromosomes, we also obtain the strand symmetry (not included in this manuscript).

Authors' contributions

M.R. and I.V. contributed equally. M.R. conceived the theory, I.V. wrote the computer programs and performed computations, M.R. and V.P. jointly directed the research, developed methodology, contributed to investigations and wrote the paper. All authors have approved the final manuscript.

Funding

This work has been fully supported by Croatian Science Foundation under the project IP 2014 09 3626.

Declaration of Competing Interest

The authors declare no competing financial interests.

CRediT authorship contribution statement

Marija Rosandić: Conceptualization, Formal analysis, Investigation, Visualization, Methodology, Writing - original draft, Writing - review & editing. **Ines Vlahović:** Formal analysis, Investigation, Software, Visualization, Writing - review & editing. **Vladimir Paar:** Formal analysis, Investigation, Methodology, Writing - original draft, Writing - review & editing.

Acknowledgments

We thank the reviewers for their very helpful comments and suggestions.

We thank Janet Kelso from the Max Planck Institute for Evolutionary Anthropology in Leipzig for providing us with Neanderthal genome. We thank Matko Glunčić for useful discussions.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jtbi.2019.08.016.

References

Afreixo, V., Bastos, C.A.C., Garcia, S.P., Rodrigues, J.M.O.S, Pinho, A.J., Ferreira, P.J.S.G., 2013. The breakdown of the word symmetry in the human genome. J. Theor. Biol. 335, 153–159.
Afreixo, V., Rodrigues, J.M.O.S, Bastos, C.A.C., 2015. Analysis of single-strand exceptional word symmetry in the human genome: new measures. Biostatistics 16 (2), 209–221.
Afreixo, V., Rodrigues, J.M.O.S, Bastos, C.A.C., Silva, R.M., 2016. The exceptional genomic word symmetry along DNA sequences. BMC Bioinformatics 17 (1), 59.
Albrecht-Buehler, G., 2006. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. PNAS 103 (47), 17828–17833.
Albrecht-Buehler, G., 2007. The three classes of triplet profiles of natural genomes. Genomics 89 (5), 596–601.
Aruni, A.W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B.N., Fletcher, H.M., 2015. Filifactor alocis – a new emerging periodontal pathogen. Microb. Infect. 17 (7), 517–530.
Bak, P., 1996. How Nature works: the Science of Self-Organized Criticality. Copernicus, New York.
Baisnee, P.F., Hampson, S., Baldi, P., 2002. Why are complementary DNA strands symmetric? Bioinformatics 18 (8), 1021–1033.
Barbieri, M., 2012. Code biology – a new science of life. Biosemiontics 5 (3), 411–437.
Bashford, J.D., Tsohantjis, I.P., Jarvis, D., 1998. A supersymmetric model for the evolution of the genetic code. PNAS 95 (3), 987–992.
Bindi, L., et al., 2015. Natural quasicrystal with decagonal symmetry. Sci. Rep. 5 (1), 9111.
Bondarev, V., Richter, M., Romano, S., Piel, J., Schwedt, A., Schulz-Vogt, H.N., 2013. The genus Pseudovibrio contains metabolically versatile bacteria adapted for symbiosis. Environ. Microbiol. 15, 2095–2113.
Chargaff, E., 1951. Structure and function of nucleic acids as cell constituents. Fed. Proc. 10, 654–659.
Chen, L., Zhao, H., 2005. Negative correlation between compositional symmetries and local recombination rates. Bioinformatics 21 (21), 3951–3958.
Deisboeck, T., Kresh, J.Y. (Eds.), 2006. Complex Systems Science in Biomedicine. Springer.
Findley, G.L., Findley, A.M., McGlynn, S.P., 1982. Symmetry characteristics of the genetic code. PNAS 79 (22), 7061–7065.
Forsdyke, D.R., Bell, S.J., 2004. Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to easy chemical observations. Appl. Bioinform. 3, 3–8.
Forger, M., Hornos, Y.M.M., Hornos, J.E.M., 1997. Global aspects in the algebraic approach to the genetic code. Phys. Rev. E56 (6), 7078–7082.
Frenkel, Z.M., Trifonov, E.N., 2012. Origin and evolution of genes and genomes. Crucial role of triplet expansions. J. Biomol. Struct. Dyn. 30 (2), 201–210.
Frenkel, Z.M., Barzily, Z., Volkovich, Z., Trifonov, E.N., 2013. Hidden ancient repeats in DNA: mapping and quantification. Gene 528 (2), 282–287.
Ganti, T., 2003. The Principles of Life. Oxford University Press, Oxford.
Glazebrook, J.F., Wallace, R., 2012. "The frozen accident" as an evolutionary adaptation: a rate distortion theory perspective on the dynamics and symmetries of genetic coding mechanisms. Informatica 36, 53–73.
Gross, D.J., 1996. The role of symmetry in fundamental physics. PNAS 93 (25), 14256–14259.
Kong, S.G., et al., 2009. Inverse symmetry in complete genomes and whole-genome inverse duplication. PLoS ONE 4 (11), e7553.
Koonin, E.V., 2011. Are there laws of genome evolution? PLoS Comput. Biol. 7 (8), e1002173.
Kosmann-Schwarzbach, Y., 2010. The Noether theorems: Invariance and Conservation Laws in the Twentieth Century. Springer.
Kozirev, S.V., Khrennikov, A.Y., 2010. 2-Adic numbers in genetics and Rumerś symmetry. Dokl. Math. 81 (1), 128–130.
Mainzer, K., 2005. Symmetry and Complexity: The Spirit and Beauty of Nonlinear Science. World Scientific, Singapore.
Mascher, M., Schubert, I., Scholz, U., Friedel, S., 2013. Patterns of nucleotide asymmetries in plant and animal genomes. BioSystems 111 (3), 181–189.
Maturana, H., Varela, F., 1980. Autopoiesis and cognition. The realization of Living. Reidel, Dordrecht.

McCutcheon, J., Moran, N.A., 2012. Extreme genome reduction in symbiotic bacteria. Nature. Rev. Micro 10 (1), 13–26.
Michel, C.J., Pirillo, G., 2013. A permuted set of circular code coding the 20 amino acids in variant nuclear codes. J. Theor. Biol. 319, 116–121.
Mitchell, D., Bridge, R., 2006. A test of Chargaff's second rule. Biochem. Biophys. Res. Commun. 340 (1), 90–94.
Monod, J., 1978. On symmetry and function in biological systems. In: Ullmann, A. (Ed.), Selected Papers in Molecular Biology By Jacques Monod. Academic Press, Cambridge, Massachusetts, pp. 701–713.
Muller, A., 2003. The beauty of symmetry. Science 300, 749–750.
Nikolajewa, S., Friedel, M., Beyer, A., Wilhelm, T., 2005. The new classification scheme of the genetic code, its early evolution, and tRNA usage. J. Bioinformatics Compxut. Biol. 4 (02), 609–620.
Nikolau, C., Almirantis, Y., 2006. Deviation from Chargaff's second parity rule in organellar DNA. Insights into the evolution of organellar genomes. Gene 381, 34–41.
König, N.D., Wiss, G.D., 1918. Invariante variationsprobleme, Zu Göttingen, Mathphys. Klasse 235–257.
Okamura, K., Wei, J., Scherer, S.W., 2007. Evolutionary implications of inversions that have caused intra-strand parity in DNA. BMC Genomics 8 (1), 160–166.
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. 2 (3), 228–240.
Prabhu, V.V., 1993. Symmetry observations in long nucleotide sequences. Nucleic Acids Res. 21 (12), 2797–2800.
Qi, D., Cuticchia, A.J., 2001. Compositional symmetries in complete genomes. Bioinformatics 17 (6), 557–559.
Ramos, A.F., Innocentini, G.C., Forger, F.M., Hornos, J.E., 2010. Symmetry in biology: from genetic code to stochastic gene regulation. IET Syst. Biol. 4 (5), 311–329.
Rapoport, A.E., Trifonov, E.N., 2013. Compensatory nature of Chargaff's second parity rule. J. Biomol. Struct. Dyn. 31 (11), 1324–1326.
Rosandić, M., Glunčić, M., Paar, V., 2013a. Start/stop codon like trinucleotides extensions in primate alpha satellites. J. Theor. Biol. 317, 301–309.
Rosandić, M., Paar, V., Glunčić, M., 2013b. Fundamental role of start/stop regulators in whole DNA and new trinucleotide classification. Gene 531 (2), 184–190.
Rosandić, M., Paar, V., 2014. Codon sextets with leading role of serine create "ideal" symmetry classification scheme of the genetic code. Gene 543 (1), 45–52.
Rosandić, M., Vlahović, I., Glunčić, M., Paar, V., 2016. Trinucleotide's quadruplet symmetries and natural symmetry law of DNA creation ensuing Chargaff's second parity rule. J. Biomol. Struct. Dyn. 34 (7), 1383–1394.
Rosen, R., 1991. Life itself. A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life. ColumbiaUniversity Press, New York.
Rudner, R., Karkas, J.D., Chargaff, E., 1968. Separation of B. Subtilis DNA into complementary strands. III. Direct analysis. PNAS 60 (3), 921–922.
Schrödinger, E., 1944. What is Life?. Cambridge University Press, Cambridge.
Shu, J.J., 2016. New integrated symmetrical table for genetic codes. J. Biosyst. 151, 21–26.
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. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171, 737–738.
Wigner, E.P., 1969a. Problems of symmetry in old and new physics. Bull. Amer. Math. Soc. 75, 891–913.
Wigner, E.P., 1969b. In: Seeger, R.J., Cohen, R.S. (Eds.). In: Physics and the Explanation of Life, 11. Boston Studies in the Philosophy of Science, pp. 119–132.
Yamagishi, M.E.B., Herai, R.H., 2011. Chargaff's "grammar of biology": new fractal-like rules. arXiv:1112.1528 [q-bio.GN].
Zee, A., 2007. Fearful Symmetry: The Search For Beauty in Modern Physics. Princeton University Press, Princeton.
Zhang, S.H., 2015. Persistence and breakdown of strand symmetry in the human genome. J. Theor. Biol. 370, 202–204.
Zhang, S.H., Huang, Y.Z., 2008. Characteristics of oligonucleotide frequencies across genomes: conservation versus variation, strand symmetry, and evolutionary implications. Nature Precedings Available online at: http://hdl.handle.net/10101/npre.2008.2146.1.
Zhang, H., Li, P., Zhong, H.S., Zhang, S.H., 2013. Conservation vs. variation of dinucleotide frequencies across bacterial and archaeal genomes: evolutionary implications. Front. Microbiol. 4, 269.
Zhang, S.H., Wang, L., 2011. A novel common triplet profile for GC-rich prokaryotic genomes. Genomics 97, 330–331.
Zhang, S.H., Wang, L., 2012. Two common profiles exist for genomic oligonucleotide frequencies. BMC Res. Notes 5, 639.