


DNA sequence symmetries from randomness: the origin of the Chargaff's second parity rule

Piero Fariselli , Cristian Taccioli, Luca Pagani and Amos Maritan

Corresponding author: Piero Fariselli, Dipartimento di Scienze Biomediche, Università di Torino, Via Santena 19, Torino 10126, Italy. Tel.: +39 011 6705871; E-mail: piero.fariselli@unito.it

Authors Piero Fariselli and Cristian Taccioli contributed equally to this work.

Abstract

Most living organisms rely on double-stranded DNA (dsDNA) to store their genetic information and perpetuate themselves. This biological information has been considered as the main target of evolution. However, here we show that symmetries and patterns in the dsDNA sequence can emerge from the physical peculiarities of the dsDNA molecule itself and the maximum entropy principle alone, rather than from biological or environmental evolutionary pressure. The randomness justifies the human codon biases and context-dependent mutation patterns in human populations. Thus, the DNA 'exceptional symmetries,' emerged from the randomness, have to be taken into account when looking for the DNA encoded information. Our results suggest that the double helix energy constraints and, more generally, the physical properties of the dsDNA are the hard drivers of the overall DNA sequence architecture, whereas the selective biological processes act as soft drivers, which only under extraordinary circumstances overtake the overall entropy content of the genome.

Key words: genome evolution; DNA symmetries; DNA k-mers; sequence analysis; Chargaff's second parity rule; codon usage

Introduction

The biological information contained within a dsDNA genome, in terms of a linear sequence of nucleotides, has been traditionally considered as the main target of selective pressures and neutral drift [1–3]. However, in this information-centered perspective, certain emerging traits of the genetic code, such as symmetries between nucleotides abundance [4–7], codon preferences [8,9] and context-dependent mutation pattern [10], are difficult to explain. In 1950, Erwin Chargaff made the important observation that the four nucleotides composing a double helix of DNA (adenine, A; cytosine, C; guanine, G and thymine, T) are

symmetrically abundant [11] (number of A = number of T and number of C = number of G). This symmetry, named Chargaff's first parity rule, played a crucial role in the discovery, in 1953, of the double helix structure of DNA [12–14]. In 1968, Chargaff extended his original observation into the Chargaff's second parity rule [15–17], which states that the same sets of identities found for a double helix DNA also hold on every single strand of the same molecule. In other words, in every single strand of a dsDNA genome, the number of adenines is almost equal to the number of thymines and the number of guanines is almost equal to the number of cytosines. This rule does not hold for

Piero Fariselli is full professor at the Department of Medical Sciences of the University of Turin, Italy. His research interests include bioinformatics, machine learning, software development and modeling of biological systems.

Cristian Taccioli is associate professor at the MAPS Department at the University of Padova, Italy. His main interests focus on genomics, cancer genomics and development of computational genomics software.

Luca Pagani is assistant professor at the Department of Biology of the University of Padova, Italy. His main research topics are: molecular anthropology, human population genomics, ancient DNA, computational genomics, natural selection and human biodemography.

Amos Maritan is full professor at the Department of Physics of the University of Padova, Italy. His main research interests are in the statistical mechanics of out-of-equilibrium systems, with interdisciplinary applications ranging from the physics of biopolymers to ecology and biogeography.

Submitted: 4 December 2019; Received (in revised form): 27 February 2020

© The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

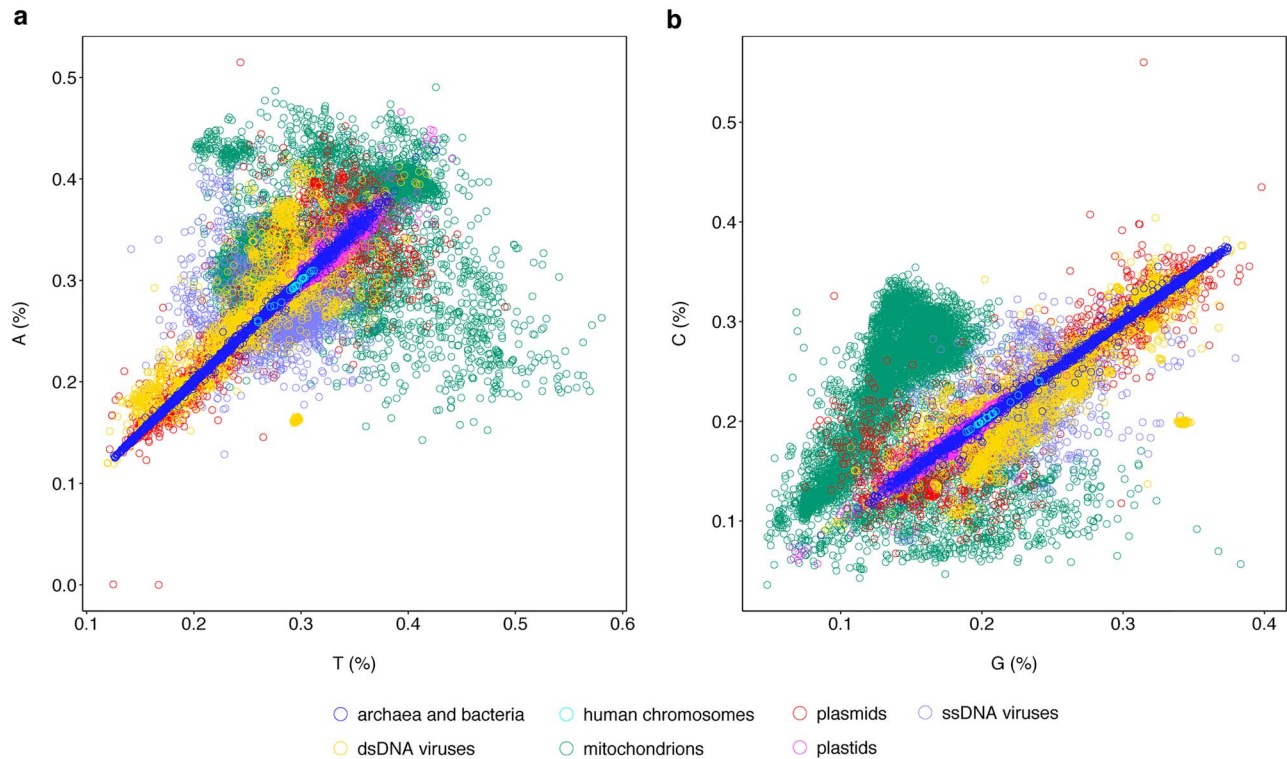


Figure 1. Percentage of adenine (A) versus thymine (T) and cytosine (C) versus guanine (G) computed for different organism genomes. The frequencies are computed on the reference strand for the dsDNA genomes and for the reported strand in case of ssDNA genomes. (a) Percentage scatterplot of adenine versus thymine. Archaea, bacteria and human chromosomes have the highest Pearson correlation values ($R^2 = 0.99$). Plastids, plasmids and dsDNA viruses have a Pearson correlation R^2 greater than 0.6. Mitochondria and ssDNA viruses do not show a significant correlation ($R^2 < 0.04$). (b) A similar graph was obtained plotting the percentage of cytosine (C) versus the percentage of guanine (G) using the same set of organism genomes. Archaea, bacteria and human chromosomes have the highest Pearson correlation values ($R^2 = 0.99$).

single-stranded DNA (ssDNA), and it has been found to be globally valid for all the dsDNA genomes with the exception of mitochondria [18,19]. An updated confirmation of these previous observations is reported in Figure 1 and in Supplementary Table 1S available online at <https://academic.oup.com/bib>, based on all reference genomes downloaded from the NCBI repository (<ftp://ftp.ncbi.nlm.nih.gov/genomes/>).

Chargaff's second parity rule has previously been extended to all the possible k-mers up to 10 bases [4,5] within a dsDNA molecule and holds only for k-mers and their reverse complements (here named 'RC-pairs'), but not for any alternative permutation of the reverse complement k-mers. This symmetry, which holds for all the dsDNA genomes, has been recently named as 'exceptional symmetry' [6]. As an example of this, here we consider the complement pairs ('C-pairs'). In this case for example (Figure 2), the occurrences of the nucleotide sequence $5' \text{TTACG} 3'$ and its reverse complement sequence $5' \text{CGTAA} 3'$ (are RC-pairs) in a single strand of a dsDNA genome are almost the same. Conversely, the frequencies of the C-pairs $5' \text{TTACG} 3'$ and $5' \text{AATGC} 3'$ in the same strand may differ significantly. Notice that the direction $5' \rightarrow 3'$ is conserved between RC-pairs (Figure 1a), whereas it is inverted in C-pairs (Figure 1b).

After 50 years from the discovery of Chargaff's second parity rule, there is not a generally accepted justification for its emergence, although several explanations have been proposed based on different models and hypothesis, such as statistical [5, 6, 7, 20, 21], stem-loops [22], tandem duplications [23], duplication followed by inversions [24], inverted transpositions [25, 26] and non-uniform substitutions [27].

All these explanations share a bottom-up approach and use the relations found in the data to build a model. Some models proposed statistical distributions of the data and showed that simple Markov models could not explain the symmetries found in dsDNA genomes [4,5,7]. Among the most relevant achievement is the fact that the maximum k-mer length for which the extended Chargaff's second parity rule is significantly detectable, is a logarithmic function of the dsDNA sequence length [4,5]. In particular, Shporer and co-workers [5], found a very precise estimation of the slope for the maximum length for the k-mer (k_{\max}) as a function of the genome size (L), which is $k_{\max} \cong 0.73 \ln(L)$.

Other approaches assume that the sequence symmetries might have originated by biological mechanisms, such as stem-loops (as in the secondary structure of the RNA) [22], which account for local symmetries. Alternative models include inversions (by cutting a dsDNA and connecting it again but inverting the strands) [24], or tandem duplications before inversion rearrangements [23]. In this particular case, the authors have shown that a computational model, based on sequence generation that creates reverse complement tandem duplications, could satisfy Chargaff's second parity rule 'even when the duplication lengths are very small when compared to the length of sequences' [23]. In line with these recent models, Albrecht-Buehler [25, 26] was among the first in proposing the hypothesis of inversions followed by transpositions. He also showed that the random flipping of the bases between the two strands leads to the Chargaff second parity rule. Unfortunately, this illuminating idea does not extend to reverse complement pairs longer than one nucleotide.

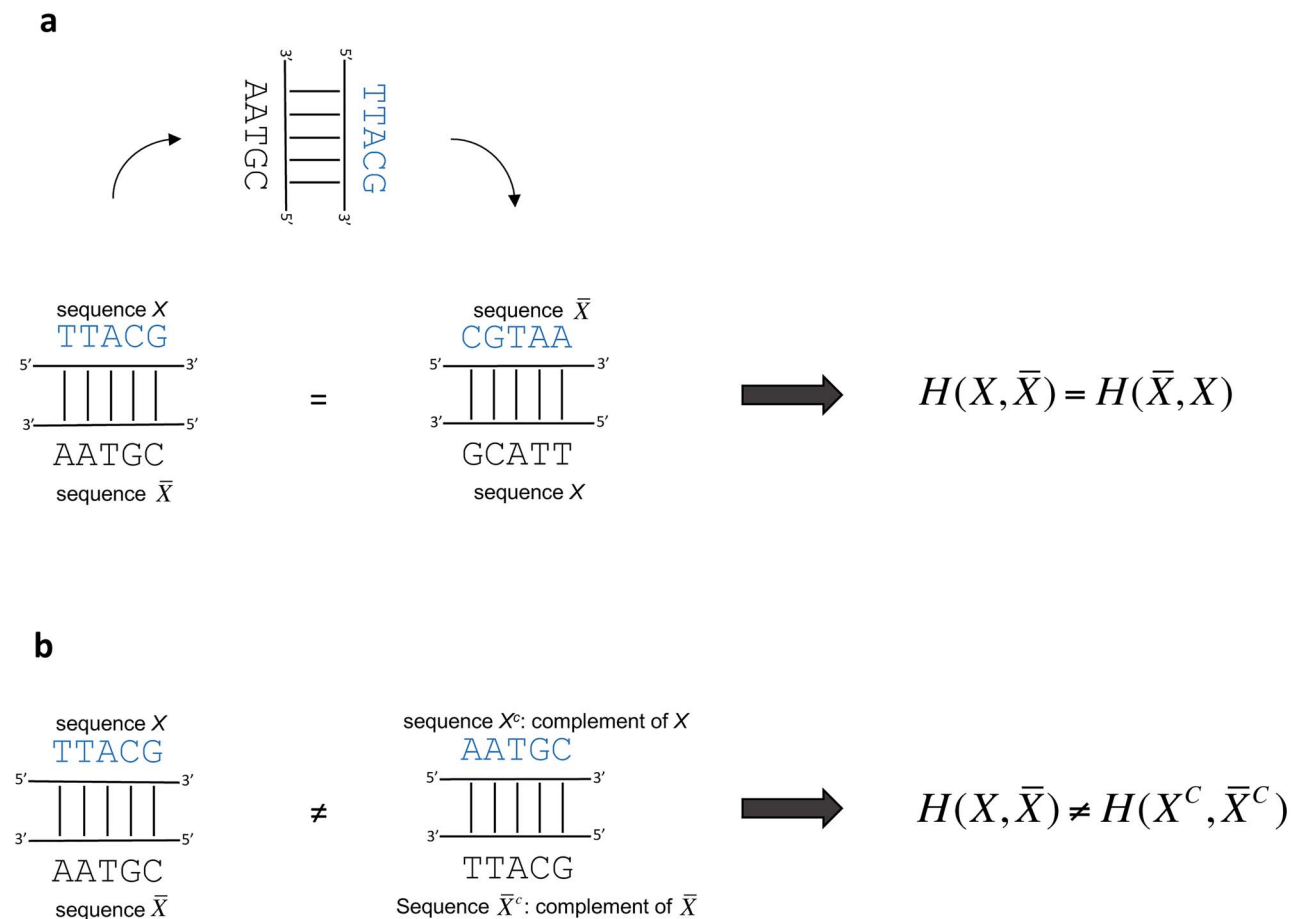


Figure 2. Energy symmetry of dsDNA. Scheme describing the interaction energy in RC- versus C-pairs. (a) The interaction energy between the sequence X ($5' \dots \text{TTACG} \dots 3'$) on plus strand and the sequence ($3' \dots \text{AATGC} \dots 5'$) \bar{X} on the minus strand is the same of its reverse complement sequence \bar{X} ($5' \dots \text{CGTAA} \dots 3'$) on the plus strand when interacting with the sequence X ($3' \dots \text{GCATT} \dots 5'$) on the minus strand. Thus, in RC-pairs, this implies that $H(X, \bar{X}) = H(\bar{X}, X)$. (b) The energy equality does not hold when taking into account the interaction energy between the complement, X^c ($5' \dots \text{AATGC} \dots 3'$), of the original sequence X on the plus strand, with its reverse complement \bar{X}^c ($3' \dots \text{TTACG} \dots 5'$) on the minus strand, that is $H(X, \bar{X}) \neq H(X^c, \bar{X}^c)$. Thus, this energy symmetry leads to the prediction that the probability of finding a specific k-mer on a strand is the same of finding its reverse complement k-mer on the same strand when we apply the energy constraint to a maximum entropy approach. In particular, this leads to Chargaff's second parity rule because of the sequences X and \bar{X} contain the same number of A-T and C-G. Note that in RC-pairs (a) that direction $5' \rightarrow 3'$ is conserved between the sequences TTACG and GCATT (the same is true for AATGC and CGTAA), whereas in C-pairs (b), the directions of the same sequences are inverted. In other words, the sequences in (a) are the same, while in (b) are specular.

Notably, these explanations, although very promising, do not have predictive power, i.e. from them it is difficult to deduce testable predictions. Here, we propose a deductive approach: we start with a minimal hypothesis (maximum randomness with the average energy constraint), and from that, we derive several predictions.

We hypothesize that the leading force shaping the DNA sequence in the genomes is the entropy and that the major cause of all these symmetries is the randomness. However, randomness does not imply uniformity and equality. As an example, the 'random' process of blindly throwing stones in a rugged landscape generates a higher probability of finding the stones in the valleys than on top of the hills.

Materials and methods

Genomic data

We accessed the NCBI database (<https://www.ncbi.nlm.nih.gov/genome/>) and downloaded all the reference and representative genomes for archaea, bacteria, dsDNA and ssDNA viruses,

mitochondria, plasmids, plastids and the following eukaryotes: *Homo sapiens*, *Pan troglodytes*, *Mus musculus* and *Takifugu rubripes*. We also downloaded freely accessible human population data from the 1000 Genomes Project database (<https://www.internationalgenome.org>).

Maximum entropy principle and the constraint of the DNA double helix interactions

We assume that the dsDNA of most of the current organisms achieved the equilibrium, and we define $P(X)$ as the probability of seeing the genome X . We postulate that and one significant contribution to $P(X)$, besides the evolutionary pressure is the thermodynamic stability of the DNA. Thus, we estimate the probability of a sequence $P(X)$ using a maximum entropy approach considering that the 'average energy' of the interactions is constant (canonical ensemble). For the sake of compactness, we introduce the following notation. We define a DNA sequence of length N the string $X = a_1 \dots a_N$, where a_i is one of the possible four nucleotides {A, C, G, T}. We define two functions of a DNA sequence X : (1) the reverse sequence

$X' = a_N a_{N-1} \dots a_1$, and (2) the complement $X^c = a_1^c a_2^c \dots a_N^c$, where a_i^c represents the complement bases, such as $A^c = T$ and $C^c = G$. Finally, the reverse complement of a sequence X is defined as $\bar{X} = (X')^c = (X^c)' = a_N^c a_{N-1}^c \dots a_1^c = \bar{a}_N \bar{a}_{N-1} \dots \bar{a}_1$. The last equality comes from the fact that the reverse complement of a single nucleotide is just its complement. When we consider a double-stranded DNA (dsDNA), the two paired DNA X and \bar{X} interact with an energy $\hat{H}(X) = H(X, \bar{X})$. $H(X, \bar{X})$ includes both the intra-chain (single-strand) interactions $H_S(X)$ (and $H_S(\bar{X})$), and the inter-chain (double strand) interactions $H_D(X, \bar{X})$, resulting in $H(X, \bar{X}) = H_D(X, \bar{X}) + H_S(X) + H_S(\bar{X})$. Although its complete form is not computable, we can recognize that it is symmetric under the exchange of X with \bar{X} , i.e. $\hat{H}(X) = H(X, \bar{X}) = H(\bar{X}, X) = \hat{H}(\bar{X})$, since the force exerted by X on \bar{X} is the same as the force that \bar{X} exerts on X , and its strength depends on the specific DNA sequence X . On the other hand, with the exception of specific cases, in general, $\hat{H}(X) = H(X, \bar{X}) \neq H(X^c, \bar{X}^c) = \hat{H}(X^c)$. To obtain the inequality is sufficient to consider the directionality of the covalent bonds between each nucleotide in the single strands. Formally, the average energy is expressed as:

$$\langle E \rangle = \sum_X \hat{H}(X) P(X) = \sum_X H(X, \bar{X}) P(X) \quad (1)$$

where $P(X)$ is the probability of occurrence of the DNA strand X , and the sum is over all possible DNA sequences (all possible genomes of this size). Our purpose is to find the most probable and less informative distribution among the ones satisfying the constraint in Eq. (1). This is achieved by maximizing the information entropy [28] S :

$$S = - \sum_X P(X) \ln(P(X)) \quad (2)$$

with respect to $P(X)$ with the constraints given by equation (1) and by the normalization condition $\sum_X P(X) = 1$. By introducing the appropriate Lagrange multipliers, λ and β , the function to maximize is:

$$F = - \sum_X P(X) \ln(P(X)) - \lambda \sum_X P(X) - \beta \sum_X \hat{H}(X) P(X) \quad (3)$$

After maximizing F , the probability can be written as:

$$P(X) = \frac{e^{-\beta \hat{H}(X)}}{Z} = \frac{e^{-\beta H(X, \bar{X})}}{Z} \quad (4)$$

Where the constant Z is the partition function:

$$Z = \sum_X \exp \left\{ -\beta \hat{H}(X) \right\} \quad (5)$$

The Lagrange multiplier β is related to E in eq. (1) by the equation:

$$E = - \frac{\partial \ln Z}{\partial \beta} \quad (6)$$

From the symmetry of the interaction energy ($\hat{H}(X) = H(X, \bar{X}) = H(\bar{X}, X) = \hat{H}(\bar{X})$), it follows that:

$$P(X) = \frac{1}{Z} e^{-\beta \hat{H}(X)} = \frac{1}{Z} e^{-\beta \hat{H}(\bar{X})} = P(\bar{X}) \quad (7)$$

This indicates that the probability of the existence of a genome is equal to the probability of its reverse complement. The energy constraint forces the two strands to have the same probability of occurring within a given genome.

Chargaff's second parity rule from maximum entropy principle

We will now show that the symmetry of the energy and the double helix interaction jointly with the maximum entropy principle is the origin of generalized Chargaff's theory (GCT), newly introduced here, that we enunciate as: in a long-enough duplex DNA, the occurrences of a k -mer and that of its reverse complement, are almost equal. Our assumption is that the counts obtained over a single strand of a dsDNA sequence (but long enough, theoretically infinite) are the same as of the summation over all possible sequences (weighted by their probabilities).

GCT can be made more formal, as follows: the expected number of a DNA segment (k -mer) of length k , $w = (a_1 \dots a_k)$, indicated as $\langle n(w) \rangle$, and the expected number of its reverse complement $\langle n(\bar{w}) \rangle$, $\bar{w} = (\bar{a}_k, \bar{a}_{k-1}, \dots, \bar{a}_1)$ are equal. Let N be the total length of the duplex DNA. By definition of the expectation value, we have that:

$$\langle n(w) \rangle = \langle n(a_1, \dots, a_k) \rangle = \sum_{i=1}^{N-k+1} P(X_i = w) = \sum_{i=1}^{N-k+1} P(a_1 \dots a_{i+k-1}) \quad (8)$$

where $X_i = (a_i \dots a_{i+k-1})$ indicates the segment of the DNA sequence X from position i to position $i+k-1$. The equality of $\langle n(w) \rangle$ and $\langle n(\bar{w}) \rangle$ is a consequence of the equality of the probabilities $P(X_i = w) = P(X_{N-i-k+2} = \bar{w})$. This follows by computing the average of the occurrence of the event $X_i = w$ with the probability distribution (7) for a generic position i in the DNA sequence as:

$$P(X_i = w) = \frac{1}{Z} \sum_{X'} \delta(X'_i, w) e^{-\beta \hat{H}(X')} \quad (9)$$

where $\delta(X'_i, w) = 1(0)$ if $X'_i = w$ ($X'_i \neq w$). Since we sum over all possible sequences X' , for a generic function $f(X_i)$, we have that the identity $\sum_{X'} f(X') = \sum_{X'} f(\bar{X})$ holds. Thus, we have:

$$\begin{aligned} P(X_i = w) &= \frac{1}{Z} \sum_{X'} \delta(X'_i, w) e^{-\beta \hat{H}(X')} = \frac{1}{Z} \sum_{X'} \delta(X'_{N-i-k+2}, \bar{w}) e^{-\beta \hat{H}(X')} \\ &= P(X_{N-i-k+2} = \bar{w}) \end{aligned} \quad (10)$$

Using equation (8), we finally obtain the desired result:

$$\langle n(w) \rangle = \langle n(\bar{w}) \rangle \quad (11)$$

which is exact if the average is computed over all the possible duplex DNA sequences. However, for a single duplex DNA of finite length N , corresponding to the analyzed cases, equation (11) is only approximate and becomes exact in the $N \rightarrow \infty$ limit. This result proves the GCT. Furthermore, this also furnishes the explanation of the original Chargaff's second parity rule, such as:

$$\langle n(a) \rangle = \langle n(\bar{a}) \rangle \quad \forall a \in \{A, C, G, T\} \quad (12)$$

According to our derivation, the 'exceptional symmetry' found in the same strand of natural duplex DNA [6] is the most

probable outcome that we can predict by chance. Of course, this does not mean that it must always be the case, but only that the Chargaff's second parity rule is the most probable solution that we should expect when no other relevant things happen such as new constraints, besides to equation (1).

GCT also predicts that equation (11) is valid also for any kind of marginal probabilities, even the one obtained using 'gapped', k-mers, such as k-mers that contain wild characters. For examples, the expected number of occurrences of the sequence $5' \text{tcNtNNNGa}3'$ should be equal to the expected number of $5' \text{tcNNNaNGa}3'$, where 'N' stands for any possible nucleotides.

It is worth noticing that our theory deals with a whole dsDNA sequence and it is not suitable to explain local frequency variation of specific DNA fragments.

Symmetry predictions for k-mers of arbitrary length using the maximum entropy principle

From the application of the maximum entropy principle, we predict that at equilibrium, in a long-enough dsDNA, the numbers of k-mers and their reverse complements tend to be the same. However, other kinds of sequence symmetries, such as k-mers and their simple complement k-mers, are predicted not to hold. Here, we study two completely different cases: (i) the C-pairs of k-mers, $w = (a_1, a_2, \dots, a_k)$ and their complements $w^c = (\bar{a}_1, \bar{a}_2, \dots, \bar{a}_k)$, (ii) RC-pairs of k-mers, w and their reverse complements, $\bar{w} = (\bar{a}_k, \bar{a}_{k-1}, \dots, \bar{a}_1)$. We can compute the conditional probability to find the k-mer \bar{w} at position j given that there is a k-mer w at position i and the conditional probability to find the k-mer w^c at position j given that there is a k-mer w at position i :

$$P(X_j = \bar{w} | X_i = w) = \frac{\sum_X P(X) \delta(X_i, w) \delta(X_j, \bar{w})}{\sum_X P(X) \delta(X_i, w)} \quad (13)$$

and

$$P(X_j = w^c | X_i = w) = \frac{\sum_X P(X) \delta(X_i, w) \delta(X_j, w^c)}{\sum_X P(X) \delta(X_i, w)} \quad (14)$$

By summing over i and j in the equations (13) and (14), we obtain the expected number of C-pairs ($\langle n(w, w^c, k) \rangle$) and RC-pairs ($\langle n(w, \bar{w}, k) \rangle$):

$$\begin{aligned} \langle n(w, \bar{w}, k) \rangle &\equiv \frac{1}{2} \sum_{i,j=1}^{n(k)} P(X_j = \bar{w}_i | X_i = w_i); \\ \langle n(w, w^c, k) \rangle &\equiv \frac{1}{2} \sum_{i,j=1}^{n(k)} P(X_j = w^c_i | X_i = w_i) \end{aligned} \quad (15)$$

where $n(k)$ is the number of possible k-mers of size k , $n(k) = N - k + 1$. We can estimate the two sums introduced above by evaluating the expected value of the pairs of k-mers along a naturally duplex genome sequence of length N . Concerning the evaluation of $\langle n(w, w^c, k) \rangle$, the maximum entropy solution does not pose constraints to the C-pairs, thus assuming that the four bases occur in a completely uncorrelated way, the probability $P(X_j = w^c | X_i = w)$ simplifies as $P(X_j = w^c | X_i = w) = P(X_j = w^c) \simeq \prod_a f_a^{n(a)}$, where f_a is the frequency on the base a in the DNA strand and $n(a)$ is the number of time the base a appears in the k-mer w^c . With this assumption, the probability decreases exponentially with the length k . If we assume equal probability for all the bases, the probability of the k-mer becomes $P(w^c) \simeq 4^{-k}$, and the

expectation can be estimated as:

$$\langle n(w, w^c, k) \rangle \simeq \frac{1}{2} n(k)^2 4^{-k} \quad (16)$$

On the contrary, in naturally duplex DNA, the maximum entropy solution implies that the number RC-pairs has the same chance to appear since the double helix constraint imposes that $\langle n(\bar{w}) \rangle \simeq \langle n(w) \rangle$, form equation (11). To find an analytical solution for $\langle n(w, \bar{w}, k) \rangle$, we observe that the frequencies of the k-mers and their reverse complements are predicted to be the same for long sequences ($n(k) = N - k + 1 \gg 1$). Thus, we can assume that the square of the difference of the frequencies ($f(w) = n(w)/n(k)$) goes to zero for long DNA sequences (large $n(k)$), as:

$$\frac{1}{2} \sum_w (f(w) - f(\bar{w}))^2 \simeq \frac{1}{n(k)} \quad (17)$$

From equation (17) and the definition of the frequency, we obtain a simple formula for the number of RC-pairs given in equation (15), that is:

$$\begin{aligned} \langle n(w, \bar{w}, k) \rangle &\simeq \frac{1}{2} \left(\sum_w n(w)n(\bar{w}) \right) \simeq \frac{1}{2} \left(\sum_w n(w)^2 - n(k) \right) \\ &= \frac{1}{2} \left(\sum_{i=1}^{n(k)} n(w_i) - n(k) \right) \end{aligned} \quad (18)$$

Maximum entropy principle in single-stranded DNA

In the case of ssDNA, the interaction energy depends only in the single chain $H_S(X)$, so that the average energy is:

$$\langle E \rangle = \sum_X H_S(X) P(X) \quad (19)$$

Notice that in general $H_S(X) \neq H_S(\bar{X})$. Similarly, to what computed above we obtain:

$$P(X) = \frac{e^{-\beta H_S(X)}}{Z} \neq \frac{e^{-\beta H_S(\bar{X})}}{Z} = P(\bar{X}) \quad (20)$$

Since in ssDNA $P(X) \neq P(\bar{X})$, there are no constraints applicable as in the dsDNA so that the compositions and sequence regularities depend on the type of single-stranded energy $H_S(X)$ and the 'temperature' factor β . However, we can predict that there is not a reason to expect that ssDNA genomes follow GCT.

Codon usage analysis

Each codon in the human genome has a reported frequency (codon usage frequency). We evaluated the Pearson correlation between the frequencies of each codon with respect to the frequencies of the corresponding reverse-complement codons (or complement codons). This is done by creating a list V , whose elements are the frequency of the alphabetically ordered codons. Then we generated two other lists V^{RC} and V^C , whose elements are the frequencies of the reverse complement codons and complement codons of V , respectively. Then, we computed the Pearson correlation between V and V^{RC} , and between V and V^C . Furthermore, we generated 10^7 different random codon list R , by shuffling the original frequencies in V (by shuffling the elements V_i of V). However, in order to keep the same amino acid

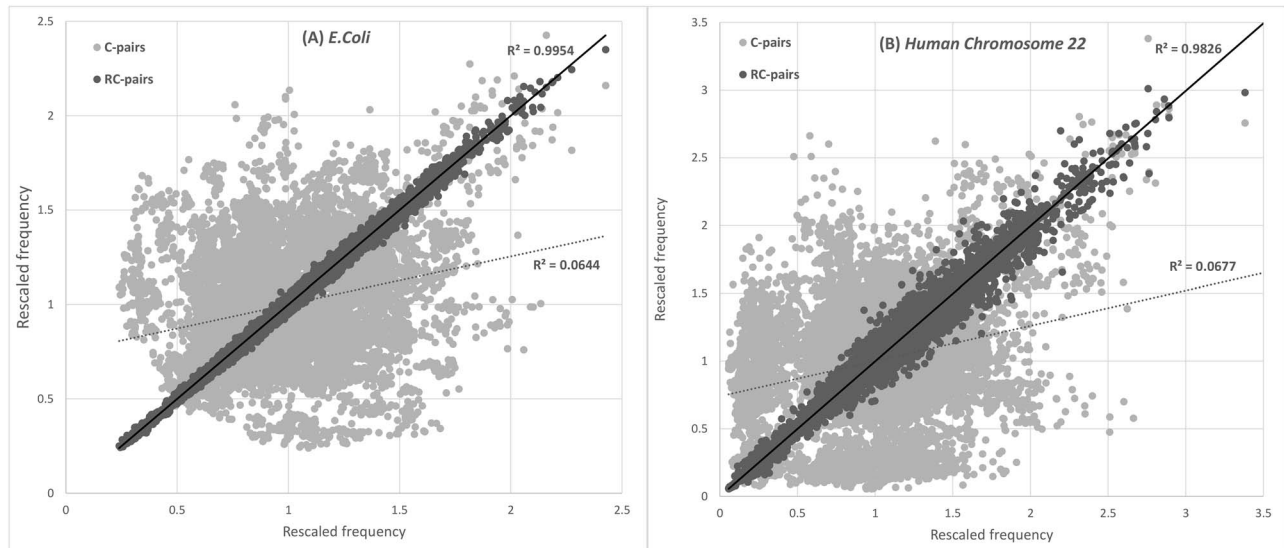


Figure 3. Plot of C-pair and RC-pair rescaled frequencies in gapped k-mers. (A) *Escherichia coli* genome, (B) human chromosome 22. The x-axis indicates the k-mer rescaled frequency and the y-axis the corresponding values for the complement (C-) or reverse complement (RC-) pairs. In the figure, we report k-mers of the form $a*b$, $ab*c$, $ab*cd$ and $abc*de$, where a , b , c , d and e are the specific nucleotides, while $*$ represents a gap. For each type of k-mer, we evaluated gaps of size 5, 10, 20, 50, 100, 500, 1000, 5000 and 10 000 bases. To show the different k-mers in the same plot, we rescaled each k-mer frequency, by dividing it for the corresponding uniform distribution $1/N$ (N is the number of possible gapped pairs, which is 16, 64, 256 and 1024, for the four types of gapped k-mer used). The best fit with the R^2 value is also reported.

frequency, and avoiding improbable amino acid frequencies in the genome, the shuffling is performed only inside the groups of the codons that code for the same residue (or stop codons). The number of possibilities is still extremely high, due to the degeneracy of the code. For each new random codon R , we computed as before, the Pearson correlation between R and the corresponding reverse-complement R^{RC} and Complement R^C lists. At the same time, we measured the Euclidean distance $D(V,R)$ between the original V and the new random shuffled R as:

$$D(V,R) = \sqrt{\sum_i (V_i - R_i)^2} \quad (21)$$

In Figure 4, we plot the Pearson correlation between R and R^{RC} (or R^C) as a function of $D(V,R)$.

Results

DNA symmetries from randomness

To find the probability distribution of the dsDNA sequences, we applied the maximum entropy approach [28], taking into account the energy constraints dictated by the DNA double helix structure. This is equivalent to finding DNA sequence arrangements corresponding to the minimum free-energy [29]. According to this principle, the distribution maximizing the entropy is the least biased, among the ones satisfying the energy constraints. By using the probability distribution with the highest entropy, we are choosing the model that needs the smallest amount of information to be explained.

The double-energy interaction and its intrinsic symmetry (due to Newton's third law) shift the probability away from the uniform distribution. Given a dsDNA with the two strands X (the plus strand) and \bar{X} (the minus strand), the interaction energy $H(X,\bar{X})$ is equal to $H(\bar{X},X)$. This implies that the energy does not change if X and \bar{X} are interchanged, that is

$H(X,\bar{X}) = H(\bar{X},X)$ (Figure 2a). As a consequence, the probabilities of a sequence and its reverse complement are equal, $P(X) = P(\bar{X})$ (equation (7)). Notably, the energy symmetry does not hold when comparing the interaction energy $H(X,\bar{X})$ with that of the complement sequence X^c (its C-pair) on plus strand and its base-pairing sequence on minus strand $H(X,\bar{X}) \neq H(X^c,\bar{X}^c)$ (Figure 2b). This implies that, in general, $P(X) \neq P(X^c)$ and the directionality of the covalent bond of a single DNA strand is sufficient to break the symmetry ($P(\bar{X}) = P(X) \neq P(X^c)$).

The main prediction which is deducible from the equality $P(X)=P(\bar{X})$ (equation (7)) is that the expected numbers of RC-pairs are equally balanced. This is only technically correct for infinitely long dsDNA sequences, and we can foresee deviations when the genome size decreases (as in the case of the DNA of viruses and organelles).

The maximum entropy solution predicts that the maximum length of RC-pairs in a genome is not constrained to any specific length of a k-mer and depends only on the range of the energy interaction, which could span the whole dsDNA genome. Therefore, our framework represents a generalization of the preliminary observations of an 'exceptional symmetry' [6] to a more general principle that here we call 'Generalized Chargaff's Theory' and which is built only upon a physical approach.

In this context, Chargaff's second parity rule is deduced from the maximum entropy and represents just a special case of GCT corresponding to $k=1$ (k-mer of length equal to 1). On the other hand, inequality $P(X) \neq P(X^c)$ predicts that the frequencies of a k-mer and its complement are not correlated.

A second prediction of the model is that the Chargaff's second parity rule can be extended to 'gapped' k-mers, such as k-mers that contain 'wildcards' in different positions inside the k-mers, such as the RC-pairs ${}^5'actNNNgNa^3'$ and ${}^5'tNcNNNagt^3'$, where 'N' stands for any possible nucleotides, while it is not necessarily found for the corresponding gapped C-pairs (${}^5'actNNNgNa^3'$ and ${}^5'tgaNNNcNt^3'$). This prediction is verified in both the *Escherichia coli* genome and in the Human chromosome 22 (Figure 3).

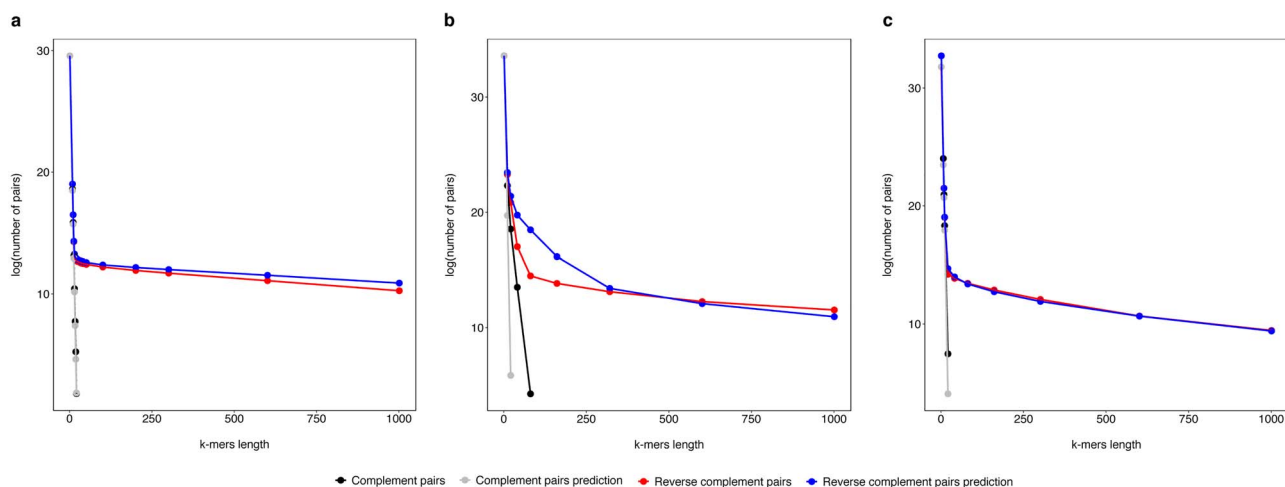


Figure 4. Predictions of the number of RC- and C-pairs in prokaryotic and eukaryotic genomes. The logarithm of the number of RC- and C-pairs is plotted as a function of the k -mer length. The plot shows that RC-pairs are much more frequent compared to C-pairs. Moreover, the predicted and observed number of RC-pairs (blue and red dots, respectively) decreases sub-linearly, whereas the predicted and observed number of C-pairs (gray and black dots, respectively) decreases exponentially (linearly in the graph). (a) The predicted number of RC- and C-pairs perfectly matches the observed data in *Escherichia coli* K12 genome. (b) Human chromosome 22. The predicted number of RC and C-pairs is in agreement with the empirical data, but not at the same degree as for *Escherichia coli* (a). (c) Human chromosome 22 with removed repeated regions. Now the match between our predictions and empirical data is of the same quality as for prokaryotes (a).

When the maximum entropy is applied to ssDNA, in the absence of any other known single-stranded energy interaction constraints, no correlation between the abundances of A and T (and between C and G, or of any other RC-pair) in ssDNA is expected (equation (20)). This third prediction of the model is consistent with the ssDNA virus data (Figure 1).

A fourth prediction of our physical formulation of GCT (equation 11) is the freedom for relative positions of RC-pairs in the dsDNA sequences, unlike alternative biological models that imply duplications with inversions and other local phenomena of DNA rearrangements, where relative positions of the two sequence of a RC-pairs may be constrained by the duplication mechanisms. According to our maximum entropy solution, the probability of finding a RC-pair in a double helix sequence is determined by the range of the interaction energy $H(X, \bar{X})$, and not necessary nearby within a dsDNA genome. This is confirmed by analyzing a set of dsDNA genomes of several species, where the data show that the energy interactions span very distant sequence positions (see Supplementary data and Figures 1S and 2S available online at <https://academic.oup.com/bib>).

A fifth prediction of GCT concerns the empirical observation that RC-pairs are extremely frequent in dsDNA genomes (Figure 4 and Supplementary Figures 5S–10S available online at <https://academic.oup.com/bib>), while any other kind of pairs, such as C-pairs, are expected to decrease exponentially with their length. This prediction is very important in light of the statistical limit derived for the maximal k -mer length by Shporer and coworkers [5]. According to their derivation, there is a limit for the k -mer length in a dsDNA after which significant deviation can be found in the Chargaff's second parity rule [4,5]. This limit, which is proportional to the natural logarithm of the sequence length, assigns maximum k -mer lengths of 6, 8 and 10 for genome sizes of about 1, 10 and 200 M bases, respectively [5].

Here, we extend beyond the k -mer statistics, without contradicting it, by predicting to find very long and more frequent RC-pairs in a genome (we may expect violations for short-sized dsDNAs). A prediction of the expected number of RC-pairs and C-pairs, as a function of k (length of a k -mer), can be analytically

derived (equations 16 and 18) and tested against experimental observations. Figure 4a shows an almost perfect, k -independent match between the predicted and observed DNA sequences in the *Escherichia coli* genome, taken as a representative genome for prokaryotes (results for other genomes are reported as supplementary materials).

When an eukaryotic genome is considered, here represented by the human chromosome 22 (the smallest in terms of nucleotides abundance), we find that the predictions of our equations (equations 4 and 5, see methods) are still in good agreement with the empirical data (Figure 4b), but not at the same degree as for prokaryotes. However, when repeated regions (transposons, tandem repeats, low complexity regions, etc.) are removed, we found an optimal improvement between predictions and observations (Figure 4c).

This finding suggests the actual presence of two distinct portions within eukaryotic genomes: (a) a stable core genome, similar to the prokaryotic DNA, which is at the GCT equilibrium, (b) regions originated from recent rearrangements events, that are either still evolving towards the equilibrium or are kept away from it by biological selective pressures. In evolutionary terms, we can imagine that, among all the chromosomal rearrangements affecting a genome, the ones that maintain or facilitate the emergence of RC-pairs are those that help the dsDNA genome to reach the equilibrium (maximum entropy or minimum free energy) and are hence positively selected in light of their energy balance rather than just for their biological information content.

Among the previously hypothesized mechanisms, the duplications followed by inversions [25–26], which create reverse complement sequences and led to GCT, are the most probable outcome in terms of genomic thermodynamic equilibrium. In this energetic view, the observed difference in abundance of various k -mers can be interpreted as a result of the difference of free energy of the relative nucleotide sequences.

Codon usage and evolutionary patterns in the human population

In virtue of GCT, we can also observe how biological events (such as duplications contained in the repeated tracts of the human

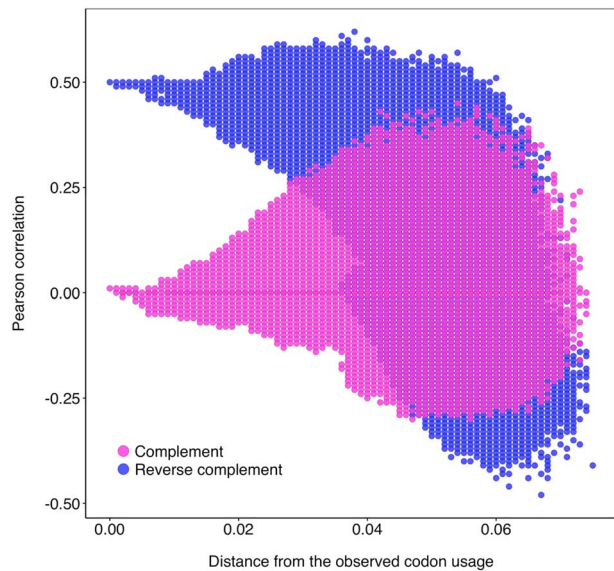


Figure 5. Pearson correlation of the codon frequency as a function of the distance between the observed human codon usage and the random permutations. The correlation is computed for the pairs of different codon frequency that are the reverse complement (RC-pairs) and simple complement (C-pairs). We generated 10^7 random samples by shuffling the codon frequencies inside the group coding for the same amino acid, to keep constant the natural amino acid abundance. The R^2 between the computed Pearson (y-axis) and the distance (x-axis) for the RC-pairs is 0.65 ($R^2 = 0$ for the C-pairs). This indicates that there is a tendency to lose the Chargaff's second parity rule fingerprinting moving away from the true codon usage.

genome) might generate detectable deviations from the maximum entropy solution. We focused on coding regions, which sequences are constrained by the fact that, under certain reading frames, tri-nucleotides are translated into amino acids following a species-specific code. Here, we investigate whether, even in a given codon bias system, the GCT is still detectable as an entropic tendency. Through the human codon usage, we found a significantly positive correlation between the frequency of the codons and the frequency of the corresponding reverse-complement codons (Pearson correlation = 0.5 with a P -value $< 10^{-4}$). On the contrary, there is no correlation between the codon frequency and those of their complement codons (Pearson correlation $R^2 = 0.0$ with a P -value = 1). This indicates that, despite selective biological pressures, there is a significant GCT trace in the human codon bias. We went further on this path, and we generated 10^7 random permutations of the observed codon usage frequencies to evaluate the stability of the GCT signal in the human codon usage. The 10^7 random codon frequencies preserve the amino acid abundance to maintain the same protein composition for each randomly generated codon usage (see 'Methods'). We then evaluate the Pearson correlation of the codon frequency as a function of the distance between the 'true' human codon usage and the corresponding random permutations. The results, reported in Figure 5, show that the more the simulated codon usage is distant from the true one, the less (on average) the RC-pairs correlate. Conversely, this does not happen for the C-pairs. The result confirms that the GCT signal in the human codon bias is robust. In broader terms, assuming 'perfect GCT compliance' as the energy equilibrium, we could tentatively see the energy needed to deviate from the equilibrium while keeping a certain level of GCT 'unbalance' (i.e. and R^2 of 0.5) within the adopted codon usage, as an upper limit

for the evolutionary 'energy gain' represented by the usage of alternative codon usage.

Chargaff's signature can also be found in pieces of the genomes, such as coding sequences (CDS) and long non-coding RNA regions (lncRNA). In these cases, Chargaff's scores are 0.96 and 0.98 for CDS lncRNA, respectively (see Supplementary data available online at <https://academic.oup.com/bib>). The corresponding Chargaff's score for the whole chromosomes is close to 1.

Remembering that in the human genome the abundance of trimers is not uniform and follows the GCT (see Supplementary Figure 3S available online at <https://academic.oup.com/bib>), another validation of the role played by the dsDNA energy constraints is provided by the observation of context-dependent mutations within the human genomes of the human 1000 Genomes Project [30]. As recently reported [10], when stratifying the occurrences of a given mutation $A \rightarrow B$ within a human population by considering the 5' and 3' context nucleotides, the relative abundance of the various $NAN \rightarrow NBN$ trimers ($k=3$) is not uniformly distributed (N can be any of the four DNA nucleotides for two fixed nucleotides A and B). We can assume that, among other causes, GCT as a major driver toward the equilibrium of the genome, maybe one of the reasons for the observed non-uniformity of the mutation $NAN \rightarrow NBN$. If this is the case, we expect that at population level, the evolutionary success of a given mutation type, approximated by its average population frequency, should match the average population frequency of the mutation type substituting the RC of NAN into the RC of NBN, hence ensuring compliance with GCT (see Supplementary Information available online at <https://academic.oup.com/bib>). This is what we observe from empirical data (see Supplementary Figure 4Sb available online at <https://academic.oup.com/bib>).

Discussion

Here, we showed that the intra-strand symmetries in the dsDNA emerge from the double-helix structure and the randomness. In this respect, we defined the principle of the GCT, which expands the unexpected 'exceptional symmetry' found in double-helix DNA as just the most likely and simplest probability distribution attainable for a duplex DNA. It should be emphasized that we obtained this result by solely imposing the base-pairing energy constraint, due by the symmetrical nature of the DNA double helix, and using the entropy maximization without specifying an energy form at a quantitative level. These two simple physical ingredients, rather than biological events, seem to be able to explain most of the observed genome-wide patterns that generalize the discovery made by Chargaff in 1968 (Chargaff's second parity rule).

It is worth noticing that a simple model that generates random sequences with the same dsDNA frequency, such as $P(A)=P(T)$ and $P(C)=P(G)$, by construction, satisfy Chargaff's second parity rule for the single bases (without explaining it). However, this process does not create the asymmetry between RC and C-pairs found in the genomes. Moreover, the numbers of RC-pairs would decrease exponentially with the k -mer length, which disagrees with both our predictions and the experimental observations.

GCT makes predictions that have been confirmed in empirical data, including dsDNA viruses and intracellular organelles, with the exception of mitochondria. Future work will be needed to further address the sequence length and biochemical peculiarities of these organelles.

Overall, our results show that processes that increase the entropy of a dsDNA molecule (such as inverted duplications and other such biological events [22,25,26]) are favored, and we speculate that exceptions to this trend may provide future opportunities to measure the energetic content of the biological information embedded in dsDNA sequences shaped by the natural selection.

Furthermore, deciphering the mechanisms that favor the long-term survival of a random DNA sequence over another will provide crucial insights to research fields focused on the understanding the basic structure and evolution of dsDNA genomes, or with designing synthetic DNA constructs.

Author contributions

P.F., C.T. and A.M. conceived the study. P.F. C.T and L.P run the computations. All the authors analyzed the data and wrote the manuscript.

Key Points

- We introduce a new paradigm where the DNA free-energy equilibrium, rather than the biological information it encodes, is the first target of evolutionary forces. In a metaphorical way, the genome is often referred to as of a book, where the ink represents the biological information it encodes. With our work, we are shifting the focus to include also the paper of which the book is made, and which constitutes the bulk of it, within the broader picture.
- Most of the intra-strand symmetries and unexpected patterns are simply due to the randomness under the double-helix constraint. This solves the puzzle of the Chargaff's second parity rule after more than 50 years of its first enunciation.
- Computational analyses of the model reveal that the 'Chargaff's second parity rule' is a strong signature in the selection of the codon bias in the human genome.
- Mutational frequency in the human population appears influenced by the underlying most probable k-mer frequency in the genome, thus related to the k-mer equilibrium distribution in the genome.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Acknowledgements

We thank T. Bobbo, A. Boattini and C. Vischioni for critical reading of our manuscript. P.F. was supported by the Italian Ministry for Education, University and Research under the programme 'Dipartimenti di Eccellenza 2018-2022', Project code D15D18000410001. C.T. was supported by the University of Padova through M.A.P.S. department under the programme 'SID2017 Project code BIRD171214'. L.P. was supported by the European Union through the 'European Regional Development Fund', Project No. 2014-2020.4.01.16-0024, MOBT53. A.M. was supported by 'Excellence Project 2017' of the Cariparo Foundation.

References

1. Nei M, Zhang J. Evolution: molecular origin of species. *Science* 1998;**282**:1428–9.
2. Kimura M. Evolutionary rate at the molecular level. *Nature* 1968;**217**:624–6.
3. King JL, Jukes TH. Non-Darwinian evolution. *Science* 1969;**164**:788–97.
4. Baisnee PF, Hampson S, Baldi P. Why are complementary DNA strands symmetric? *Bioinformatics* 2002;**18**:1021–33.
5. Shporer S, Chor B, Rosset S, et al. Inversion symmetry of DNA k-mer counts: validity and deviations. *BMC Genomics* 2016;**17**:696–708.
6. Afreixo V, Rodrigues JM, Bastos CA, et al. Exceptional symmetry by genomic word: a statistical analysis. *Interdiscip Sci* 2017;**9**:14–23.
7. Sobottka M, Hart AG. A model capturing novel strand symmetries in bacterial DNA. *Biochem Biophys Res Commun* 2011;**410**:823–8.
8. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* 2011;**12**:32–42.
9. Athey J, Alexaki A, Osipova E, et al. A new and updated resource for codon usage tables. *BMC Bioinformatics* 2017;**18**:391.
10. Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. *Elife* 2017;**6**: e24284 e24284.
11. Zamenhof S, Shettels LB, Chargaff E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Nature* 1950;**165**:756–7.
12. Watson JD, Crick FH. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 1953;**171**:737–8.
13. Wilkins MHF, Stokes AR, Wilson HR. Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids. *Nature* 1953;**171**:738–40.
14. Franklin RE, Gosling RG. Molecular configuration in sodium thymonucleate. *Nature* 1953;**171**:740–1.
15. Rudner R, Karkas JD, Chargaff E. Separation of *B. subtilis* DNA into complementary strands, I. biological properties. *Proc Natl Acad Sci U S A* 1968;**60**:630–5.
16. Karkas JD, Rudner R, Chargaff E. Separation of *B. subtilis* DNA into complementary strands II template functions and composition as determined by transcription with RNA polymerase. *Proc Natl Acad Sci U S A* 1968;**60**: 915–20.
17. Rudner R, Karkas JD, Chargaff E. Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. *Proc Natl Acad Sci U S A* 1968;**60**:921–2.
18. Mitchell D, Bridge R. A test of Chargaff's second rule. *Biochem Biophys Res Commun* 2006;**340**:90–4.
19. Nikolaou C, Almirantis Y. Deviations from Chargaff's second parity rule in organellar DNA: insights into the evolution of organellar genomes. *Gene* 2006;**381**: 34–41.
20. Hart A, Martínez S, Olmos F. A Gibbs approach to Chargaff's second parity rule. *J Stat Phys* 2012;**146**:408–22.
21. Cristadoro G, Degli Esposti M, Altmann GA. The common origin of symmetry and structure in genetic sequences. *Sci Rep* 2018;**8**:15817.
22. Forsdyke DR, Bell SJ. Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. *Appl Bioinformatics* 2004;**3**:3–8.

23. Siddharth J, Netanel R, Bruck J. Attaining the 2nd Chargaff Rule by Tandem Duplications. Parallel and Distributed Systems Group Technical Reports, **138** (2018). <https://resolver.caltech.edu/CaltechAUTHORS:20180105-092230028>
24. Okamura K, Wei J, Sherer SW. Evolutionary implications of inversions that have caused intra-strand parity in DNA. *BMC Genomics* 2007;**8**:160.
25. Albrecht-Buehler G. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. *Proc Natl Acad Sci U S A* 2006;**103**:17828–33.
26. Albrecht-Buehler G. Inversions and inverted transpositions as the basis for an almost universal 'format' of genome sequences. *Genomics* 2007;**90**:297–305.
27. Lobry JR. Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol* 1995;**40**:326–30.
28. Jaynes ET. On the rationale of maximum-entropy methods. *Proc IEEE* 1982;**70**:939–52.
29. Callen HB. *Thermodynamics and an Introduction to Thermostatistics*. New York: Wiley, 1985, 27–131.
30. The 1000 genomes project consortium. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.