

DNA energy constraints shape biological evolutionary trajectories

Piero Fariselli^{1*}, Cristian Taccioli^{2*}, Luca Pagani^{3,4}, and Amos Maritan⁵

¹Dipartimento di Scienze Biomediche, Università di Torino, Via Santena 19, Torino, 10126, IT. ²Dipartimento di Medicina Animale, Produzione e Salute, Università di Padova, Viale dell'Università 16, Legnaro (Padova), 35020, IT. ³Dipartimento di Biologia, Università di Padova, Via Ugo Bassi 58/B, Padova, 35131, IT. ⁴Estonian Biocentre, Institute of Genomics, University of Tartu, Riia 23b, Tartu, 51010, EE. ⁵Dipartimento di Fisica e Astronomia, G. Galilei, INFN, Università di Padova, Via Marzolo 8, Padova, 35131, IT.

*These authors contributed equally to this work.

Abstract

Most living systems rely on double-stranded DNA (dsDNA) to store their genetic information and perpetrate themselves. Thus, the biological information contained within a dsDNA molecule, in terms of a linear sequence of nucleotides, has been considered the main target of the evolution. However, in this information-centred perspective, certain DNA sequence symmetries are difficult to explain. Here we show that these patterns 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. Our predictions are valid for both prokaryotes and eukaryotes, and also inform the interpretation of observed codon biases and context-dependent mutation patterns in human populations. 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 biological selective processes act as soft drivers, which only under extraordinary circumstances overtake the overall entropy content of the genome.

Introduction

The biological information contained within the dsDNA sequence, has been considered the main target of selective pressures and neutral drift¹⁻³. However, in this information-centred perspective, certain emerging traits of the genetic code, such as symmetries between nucleotides abundance⁴⁻⁷, codon preferences^{8,9}, and context-dependent mutation pattern¹⁰, 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 abundant¹¹ (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 DNA¹²⁻¹⁴. In 1968 Chargaff extended his original observation into the Chargaff's second parity rule¹⁵⁻¹⁷, 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 molecule, 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 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 Table 1S.

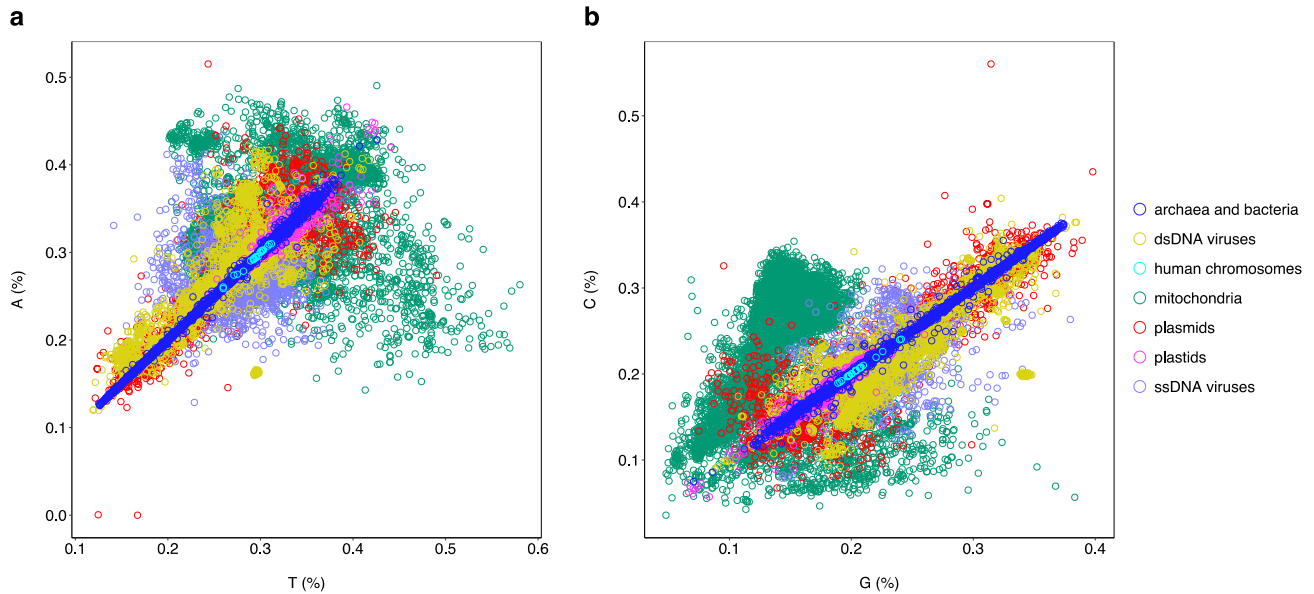


Figure 1 | Percentage of Adenine (A) versus Thymine (T) and Cytosine (C) versus guanine (G) computed for different organism genomes. **a**, Percentage scatterplot of Adenine versus Thymine. Archaea, bacteria and human chromosomes have the highest Pearson correlation values ($R^2=0.99$). Plastids, plasmids and double-stranded DNA (dsDNA) viruses have a Pearson correlation R^2 greater than 0.6. Mitochondria and single-stranded DNA (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 ($R^2=0.99$). Plastids, plasmids and double-stranded DNA (dsDNA) viruses have a Pearson correlation R^2 greater than 0.7. Mitochondria and single-stranded DNA (ssDNA) viruses do not show a significant correlation ($R^2 < 0.04$).

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 "exceptional symmetry"⁶. As an example of this, here we consider the complement pairs ("C-pairs"). In this case for example (see Figure 2), the occurrences of the nucleotide $5' \text{TTACG}^3$ 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 $3' \text{AATGC}^5$ in the same strand may differ significantly. Notice that the direction $5' \rightarrow 3'$ is conserved between RC-pairs (see Figure 1a), whereas it is inverted in C-pairs (see Figure 1b).

Generalized Chargaff's Theory

After fifty 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^{7,20,21}, stem-loops²², tandem duplications²³, duplication followed by inversions²⁴, inverted transpositions^{25,26}, and non-uniform substitutions²⁷. Notably, these explanations, although very promising, do not have predictive power, i.e. from them it is difficult to deduce testable predictions.

The causes underlying the emergence of Chargaff's second parity rule can be grouped into three main models: 1) "Physical", where the dsDNA structural properties (or energy constraints) alone are responsible for the survival of a given k-mer over another; 2) "Biological", where the presence of biological mechanisms and/or selective pressures alone (e.g. duplications, duplications followed by inversions) actively duplicate DNA sequences or convert a given k-mer into its reverse complement; 3) "Hybrid" (biological and physical), where the biological machinery and energy constraints are both involved in the abundance of a particular k-mer within a dsDNA sequence.

We tested the validity of the physical model, model 1, which is the simplest of all three since it does not rely on organism specificities or *ad hoc* biological mechanisms. This model takes into account the energy constraints implied in the DNA double helix structure, and uses the maximum entropy approach²⁸. Models 2 (biological) and 3 (hybrid) were used as putative non-null alternatives. Testing the physical model is equivalent to finding DNA sequence arrangements corresponding to the minimum free-energy²⁹. According to this principle and considering the set of all probability distributions, $P(X)$, of single strand sequence, X , of a dsDNA, the distribution maximizing the entropy is the least biased, satisfies the energy constraints, and thus makes no unwarranted additional assumptions. By using the probability distribution with the highest entropy we are choosing the model that needs the smallest amount of information to be explained.

We define X a sequence on a double helix strand (i.e. the plus strand) that interacts with its complement strand \bar{X} (the minus strand) with an energy $H(X, \bar{X})$. According to the Newton's third law, the underlying dsDNA physical nucleotides interactions are symmetric, implying 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 of the maximum entropy principle, one gets that $P(X) = P(\bar{X})$ (eq.(2), see methods). 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, $P(X) \neq P(X^c)$.

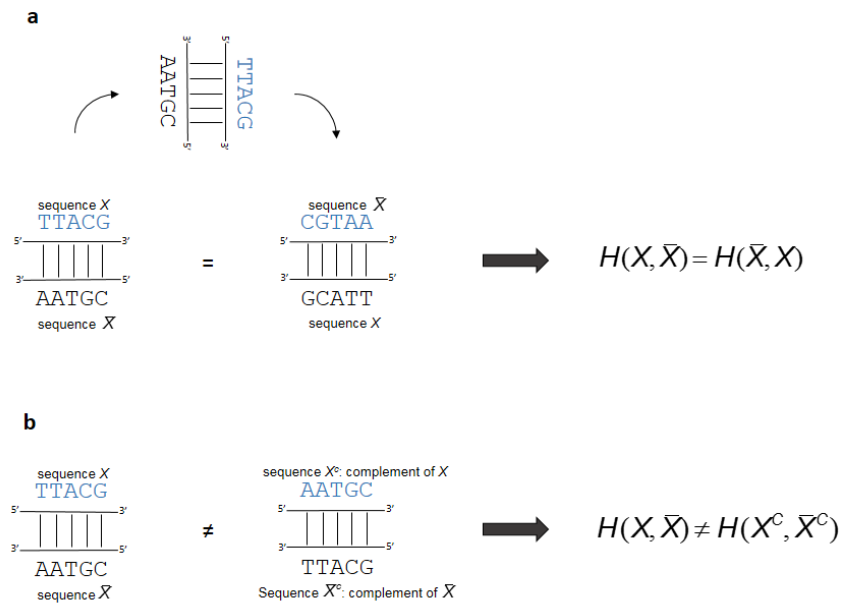


Figure 2 | Energy symmetry of dsDNA. Scheme describing the interaction energy in RC- versus C-pairs. a, The interaction energy between the sequence X (5'...TTACG...3') on plus strand and the sequence (3' ...AATGC...5') \bar{X} on the minus strand is the same of its reverse complement sequence \bar{X} (5' ...CGTAA...3') on the plus strand when interacting with the sequence X (3'...GCATT...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'...AATGC...3'), of the original sequence X on the plus strand, with its reverse complement \bar{X}^c (3'...TTACG...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 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' -> 3' is conserved between the sequences TTACG and GCATT (the same is true for AATGC and CGTAA), whereas in C-pairs (**b**) the direction of the same sequences are inverted. In other words, the sequences in (**a**) are the same, while in (**b**) are specular.

The main prediction of the equality $P(X) = P(\bar{X})$ and of the energy symmetry (equation (2), see methods), is that the maximum length of a RC-pairs we can expect to find 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 in principle span the whole dsDNA genome. Therefore, our framework represents a generalization of the preliminary observations of an “exceptional symmetry”⁶ to a more universal principle that here we call “Generalized Chargaff’s Theory” (GCT) and which is built only upon a physical approach. In this context, Chargaff’s second parity rule is just a special case of GCT corresponding to $k=1$ (k-mer of length equal to 1). On the other hand, the inequality $P(X) \neq P(X^c)$ predicts that the frequencies of a k-mer and its complement are not correlated.

When the maximum entropy is applied to ssDNA (see supplementary materials), 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. This second prediction of the model is consistent with the ssDNA virus data (Figure 1).

A third prediction of our physical formulation of GCT (equation (3), see methods) 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-pair may be constrained by the duplication mechanism. 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 analysing a set of dsDNA genomes of several species, where the data show that the energy interactions span very distant sequence positions (see Supplementary Information and Figures 1S, 2S).

A fourth prediction of GCT concerns the empirical observation that RC-pairs are extremely frequent in dsDNA genomes (see Figure 3), while any other kind of pairs, such as C-pairs, do not deviate from random expectations. Indeed, 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 (4) and (5), see methods) and tested against experimental observations. Figure 2a shows an almost perfect, k -independent match between the predicted and observed DNA sequences in the Escherichia coli genome, taken as representative genome for prokaryotes.

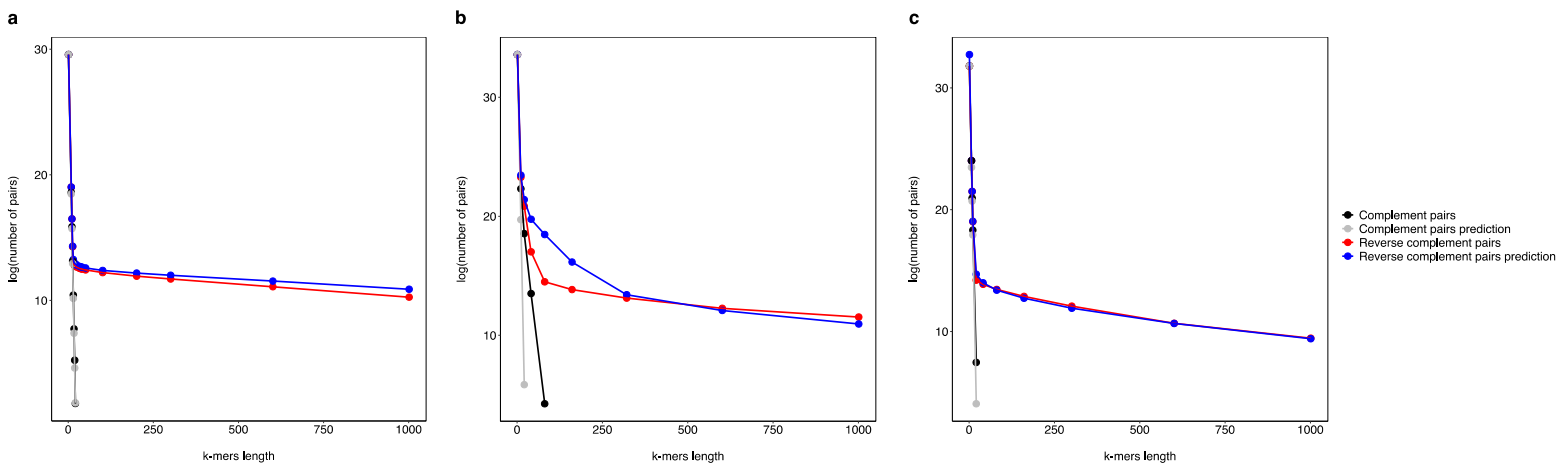


Figure 3 | Predictions of the number of RC- and C-pairs in prokaryotic and eukaryotic genomes. The logarithm of the number of RC- and C-pairs are plotted as a function of the k -mer length. The plot shows that RC-pairs are much more frequent compared to C-pairs. Moreover, the predicted and observed number of RC-pairs (blue and red dots respectively) decreases sub-linearly whereas the predicted and observed number of C-pairs (grey and black dots respectively) decreases exponentially (linearly in the graph). **a)** The predicted number of RC- and C-pairs perfectly match the observed data in Escherichia coli K12 genome. **b)** Human chromosome 22. The predicted number of RC and C-pairs are 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 a 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 (see Figure 2b), 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 2c).

This finding suggests the actual presence of two distinct portions within eukaryotic genomes: i) a stable core genome, similar to the prokaryotic DNA, which is at the GCT equilibrium, ii) regions originated from recent rearrangements events, that are either still evolving towards the equilibrium or are kept away from it by biological selective pressures. We conclude that, while for prokaryotic genomes the physical model 1 is sufficient to explain the mostly uniform distribution of observed RC-pairs, for eukaryotes, an integrated biological and physical model (“Hybrid” model 3) seems more appropriate. 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. In particular, duplications followed by inversions²⁵⁻²⁶, 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 human population

In virtue of GCT we can also observe how biological events (such as duplications contained in the repeated tracts of the human 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 biological selective 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 4, 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.

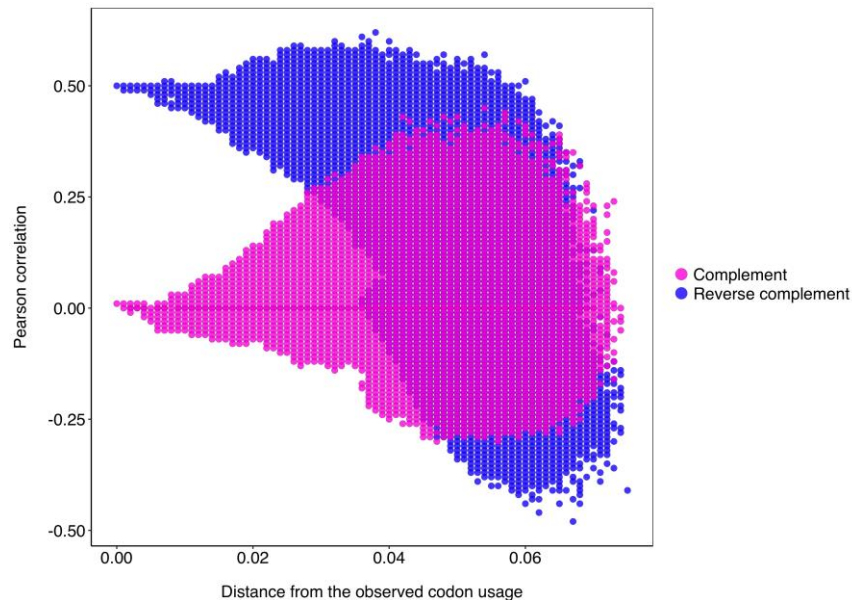


Figure 4 | Pearson correlation of the codon frequency as a function of the distance between the observed human codon usage and random permutations. The correlation is computed for the pairs of different codon frequency that are the reverse complement (RC-pairs) and simple complement (C-pairs). We generated 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 fingerprinting moving away from the true codon usage.

Remembering that in the human genome the abundance of trimers is not uniform and follows the GCT (see Supplementary Information Figure 3S), 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 Project³⁰. As recently reported¹⁰, 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, may be 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 for further details). This is what we observe from empirical data (Figure 4Sb).

Overall, our results show that processes that increase the entropy of a double-stranded DNA molecule are favoured, 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.

Methods

Maximum entropy and DNA energy constraint. Assuming that the dsDNA is at the equilibrium, and that the DNA sequence is sufficiently long, we can estimate the probability of a sequence $P(X)$ using a maximum entropy approach considering that the average energy of the double helix interaction is fixed (see supplementary materials for more details). This means to maximize the following functional (with respect to $P(X)$):

$$F = -\sum_X P(X) \ln P(X) - \lambda \sum_X P(X) - \beta \sum_X \hat{H}(X) P(X) \quad (1)$$

where for compactness we set $\hat{H}(X) = H(X, \bar{X})$. The solution of this variational problem leads to the fact that the probability of a sequence X is the same as its reverse complement sequence \bar{X} , i.e.

$$P(X) = \frac{1}{Z} e^{-\beta \hat{H}(X)} = \frac{1}{Z} e^{-\beta \hat{H}(\bar{X})} = P(\bar{X}) \quad (2)$$

because of the energy symmetry $\hat{H}(X) = H(X, \bar{X}) = H(\bar{X}, X) = \hat{H}(\bar{X})$

From the result equation (2) it is possible to show (see supplementary materials) that the expected number of k -mer w , $n(w)$, is the same of its reverse complement irrespectively of its length, or more formally:

$$\langle n(w) \rangle = \langle n(\bar{w}) \rangle \quad (3)$$

The original Chargaff's second parity rule, is just the case of k -mer of length 1 ($w \in \{A, C, G, T\}$). With this model, we can make some predictions about the expected number of RC- and C-pairs in a dsDNA long enough sequence. In particular, under some reasonable assumptions (see supplementary materials), for the expected number of C-pairs of length k , $n(w, w^c, k)$, we have:

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

where $n(k) = N - k + 1$ is the number of possible k -mer in the dsDNA sequence of length N . Similarly, the expected number of RC-pairs is:

$$\langle n(w, \bar{w}, k) \rangle \simeq \frac{1}{2} \left(\sum_{i=1}^{n(k)} \langle n(w_i) \rangle - n(k) \right) \quad (5)$$

where $\langle n(w_i) \rangle$ is the expected number of k -mer equals to the one starting at the i^{th} position of the DNA sequence and of length k . In this case we exploit equation (3) to evaluate the number of reverse complement k -mers (see supplementary materials).

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 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 possibility 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 measure the Euclidean distance $D(V,R)$ between the original V and the new random shuffled R as:

$$D(V, R) = \sqrt{\sum_i (V_i - R_i)^2} \quad (6)$$

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

References

1. Nei, M. & Zhang, J. Evolution: molecular origin of species. *Science* 282: 1428-1429 (1998).
2. Kimura, M. Evolutionary Rate at the Molecular Level. *Nature* 217:624-6 (1968)
3. King, J. L. & Jukes T. H. Non-Darwinian Evolution. *Science* 164:788-97 (1969)
4. Baisnee, P. F., Hampson, S., Baldi, P. Why are complementary DNA strands symmetric? *Bioinformatics* 18, 1021-1033 (2002).
5. Shporer, S., Chor, B., Rosset, S., Horn, D. Inversion symmetry of DNA k-mer counts: validity and deviations. *BMC Genomics* 17, 696-708 (2016).
6. Afreixo, V., Rodrigues, J. M., Bastos, C. A., Tavares, A. H., Silva, R. M. Exceptional Symmetry by Genomic Word: A Statistical Analysis. *Interdiscip Sci.* 9, 14-23 (2017).
7. Sobottka M., Hart, A. G. A model capturing novel strand symmetries in bacterial DNA. *Biochem. Biophys. Res. Commun.* 410, 823-828 (2011).
8. Plotkin, J. B., Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* 12, 32-42, (2011).
9. Athey, J., Alexaki, A., Osipova, E., Rostovtsev, A., Santana-Quintero, L. V., Katneni, U., Simonyan, V., Kimchi-Sarfaty, C. A new and updated resource for codon usage tables. *BMC Bioinformatics*, 18, 391 (2017).
10. Harris, K., Pritchard, JK. Rapid evolution of the human mutation spectrum. *Elife*, 6, e24284 (2017).
11. Zamenhof, S., Shettels, L.B., Chargaff, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Nature* 165, 756-757 (1950).
12. Watson, J. D., Crick, F. H. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 171, 737-738 (1953).
13. Wilkins M. H. F., Stokes A. R. & Wilson H. R. Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids. *Nature* 171, 738-740 (1953).
14. Franklin R. E. & Gosling R. G. Molecular configuration in sodium thymonucleate. *Nature* 171,740-741 (1953).
15. Rudner, R., Karkas, J. D., Chargaff, E. Separation of *B. subtilis* DNA into complementary strands, I. Biological properties. *Proc. Natl. Acad. Sci. U.S.A.* 60, 630-635 (1968).
16. Karkas, J. D., 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.* 60, 915-920 (1968).

17. Rudner, R., Karkas, J. D., Chargaff, E. Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proc. Natl. Acad. Sci. U.S.A. 60, 921-922 (1968).
18. Mitchell, D., Bridge, R. A test of Chargaff's second rule. Biochem. Biophys. Res. Commun. 340, 90-94 (2006).
19. Nikolaou, C., Almirantis, Y. Deviations from Chargaff's second parity rule in organellar DNA: Insights into the evolution of organellar genomes. Gene 381, 34-41 (2006).
20. Hart, A., Martínez, S., Olmos, F. A Gibbs Approach to Chargaff's Second Parity Rule. Journal of Statistical Physics 146, 408-422 (2012).
21. Cristadoro, G., Degli Esposti, M & Altmann, G. A. The common origin of symmetry and structure in genetic sequences. Scientific Reports 8, 15817 (2018).
22. Forsdyke D. R., Bell, S. J. Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. Appl. Bioinformatics 3, 3-8 (2004).
23. Siddharth, J., Netanel, R., Bruck, J. Attaining the 2nd Chargaff Rule by Tandem Duplications. Parallel and Distributed Systems Group Technical Reports, 138. (2018)
24. Okamura K., Wei J., Sherer, S. W. Evolutionary implications of inversions that have caused intra-strand parity in DNA. BMC Genomics 8, 160 (2007).
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. 103, 17828-17833 (2006).
26. Albrecht-Buehler, G. Inversions and inverted transpositions as the basis for an almost universal 'format' of genome sequences. Genomics 90, 297-305 (2007).
27. Lobry, J. R. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. 40, 326-330 (1995).
28. Jaynes, E. T. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70, 939-952 (1982).
29. Callen, H. B. Thermodynamics and an Introduction to Thermostatistics, Wiley, New York, 27-131 (1985).
30. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature, 526, 68-74 (2015).

Supplementary Information

1.	MATHEMATICAL METHODS	11
1.1	Maximum Entropy Principle and the Constraint of the DNA Double Helix Interactions.....	11
1.2	Chargaff's Second Parity Rule from Maximum Entropy Principle.....	12
1.3	Symmetry Predictions for k-mers of Arbitrary Length Using the Maximum Entropy Principle..	13
1.4	Maximum Entropy Principle in single-stranded DNA (ssDNA).....	15
2.	TABLE OF NUCLEOTIDE RATIOS COMPUTED FOR DIFFERENT ORGANISMS	15
3.	K-MERS VS REVERSE COMPLEMENT SEQUENCES ANALYSIS THROUGH LASTZ	15
4.	CODON TRIMERS FREQUENCY BARPLOT	18
5.	ANALYSIS OF CONTEXT-DEPENDENT MUTATIONS PATTERNS IN HUMAN POPULATION	18
6.	REFERENCES	20

1. MATHEMATICAL METHODS

1.1 Maximum Entropy Principle and the Constraint of the DNA Double Helix Interactions

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 $X^r = a_N a_{N-1} \dots a_1$, and 2) the complement $X^c = a_1^c a_2^c \dots a_N^c$, where a_i^c represent the complement bases, such as $A^c = T$ and $C^c = G$. Finally, the reverse complement of a sequence X is defined as $\bar{X} = (X^r)^c = (X^c)^r = a_N^c a_{N-1}^c \dots a_1^c = \bar{a}_N \bar{a}_{N-1} \dots \bar{a}_1$. The last equality comes from the fact that the reverse complement of a single nucleotide is just its complement. When we consider a dsDNA (double-stranded DNA), the two paired DNA X and \bar{X} interact with an energy $\hat{H}(X) = H(X, \bar{X})$, which 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 $\hat{H}(X) = H(X, \bar{X}) \neq H(X^c, \bar{X}^c) = \hat{H}(X^c)$. We then can define an average energy 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. 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 entropy¹ S :

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

with respect to $P(X)$ with the constraints given by eq. (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 the maximization of 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:

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

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

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

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 DNA sequence in a genome, for example, is equal to the probability of its reverse complement.

1.2 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 GCT (Generalized Chargaff's Theory): in a long-enough duplex DNA, the occurrence of a k -mer and those of its reverse complement, are almost equal. 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:

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

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:

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

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 done over all the possible duplex DNA sequences. However, for a single duplex DNA of finite length N , corresponding to the analysed 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 DNA², is the most probable outcome that we can predict by chance. Of course, this does not mean that it must be always the case, but only that the Chargaff's rule is the most probable solution that we should expect when no other relevant things happen such as new constraints, besides to eq. (1).

1.3 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 long 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:

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

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 $\langle n(\bar{w}) \rangle \simeq \langle n(w) \rangle$, from eq.(11). To find an analytical solution for $\langle n(w, \bar{w}, k) \rangle$, we observe that the expected number of \bar{w}_i in the DNA given that, at position i , there is the reverse complement k-mer w_i , is given by:

$$\sum_{j=1}^{n(k)} P(X_j = \bar{w}_i | X_i = w_i) \simeq \langle n(w_i) \rangle - 1 \quad (17)$$

where the minus one in the above equation is due to the fact that the expectation is conditioned to the fact that the i^{th} DNA site is occupied by w_i . This approximation leads to a simple formula for the number of RC-pairs given in equation (15), that is:

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

1.4 Maximum Entropy Principle in single-stranded DNA (ssDNA)

In the case of ssDNA, the interaction energy depends only in the single chain $\hat{H}(X)$, so that the average energy is:

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

Notice $H(X) \neq H(\bar{X})$. Similarly to what computed above we obtain:

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

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 $H(X)$ and the “temperature” factor β . However, we can predict that there is not a reason to expect that ssDNA genomes follow GCT.

2. TABLE OF NUCLEOTIDE RATIOS COMPUTED FOR DIFFERENT ORGANISMS

We extended and replicated previous observations^{3,4} of Chargaff’s second parity rule for 19566 dsDNA and 1397 ssDNA (single-stranded DNA) genomes including human chromosomes, bacteria, archaea, plastids, mitochondria, plasmids and viruses. In agreement with previous studies^{3,4}, we found strong Pearson correlations between the abundance of adenine versus thymine and cytosine versus guanine for human chromosomes, bacteria, archaea, plastids, plasmids and dsDNA viruses, and poor Pearson correlations for mitochondrial DNA (mtDNA) and ssDNA virus genomes (Table 1S).

Table 1S – Frequency of nucleotides in different groups of organism genomes.

Organisms	Genome size average	Number of Organisms	R ² (A vs T)	R ² (C v G)
<i>Bacteria</i>	3,584,631	1843	0.99	0.99
<i>Archaea</i>	2,010,000	170	0.99	0.99
<i>Plastids</i>	154,187	3004	0.82	0.88
<i>Bacteria plasmids</i>	55,960	2480	0.80	0.84
<i>Archaea plasmids</i>	122410	104	0.71	0.89
<i>dsDNA viruses</i>	44937	2911	0.63	0.73
<i>Mitochondria</i>	16566	9054	0.04	0.04
<i>ssDNA viruses</i>	2657	1397	0.02	0.10

The R² is the Pearson correlation coefficient between the abundance of adenine versus thymine and cytosine versus guanine

3. K-MERS VS REVERSE COMPLEMENT SEQUENCES ANALYSIS

GCT does not pose any spatial constraints to the relative positions RC-pairs sequences. This is shown in Figure 1S and provides further support to our theoretical model. The genomic data were plotted using LASTZ⁵, a program for aligning entire genomes. Thus, we aligned k-mers versus their reverse complement sequences in different bacteria genomes and we found that the RC-pairs are equally distributed along the bacteria circular dsDNA. When eukaryote chromosomes (which are more competent for events involving recombination) were studied using LASTZ, we found interesting deviations from the expectations (left side of Figure 2S). However, when repeated regions (transposons, tandem repeats, low complexity regions, etc.) were removed (right side of Figure 2S), these deviations disappeared and eukaryotic regions resembled prokaryotic genomes. This may seem to point, in eukaryotes, to a hybrid biological/physical model, indicating that it is possible to define two portions in these genomes: one that similarly to the prokaryote at the Chargaff’s equilibrium, whereas a second portion consists of repeated

regions that originated from recent duplication events and that are either still evolving towards the equilibrium or are kept as such by biological selective pressures.

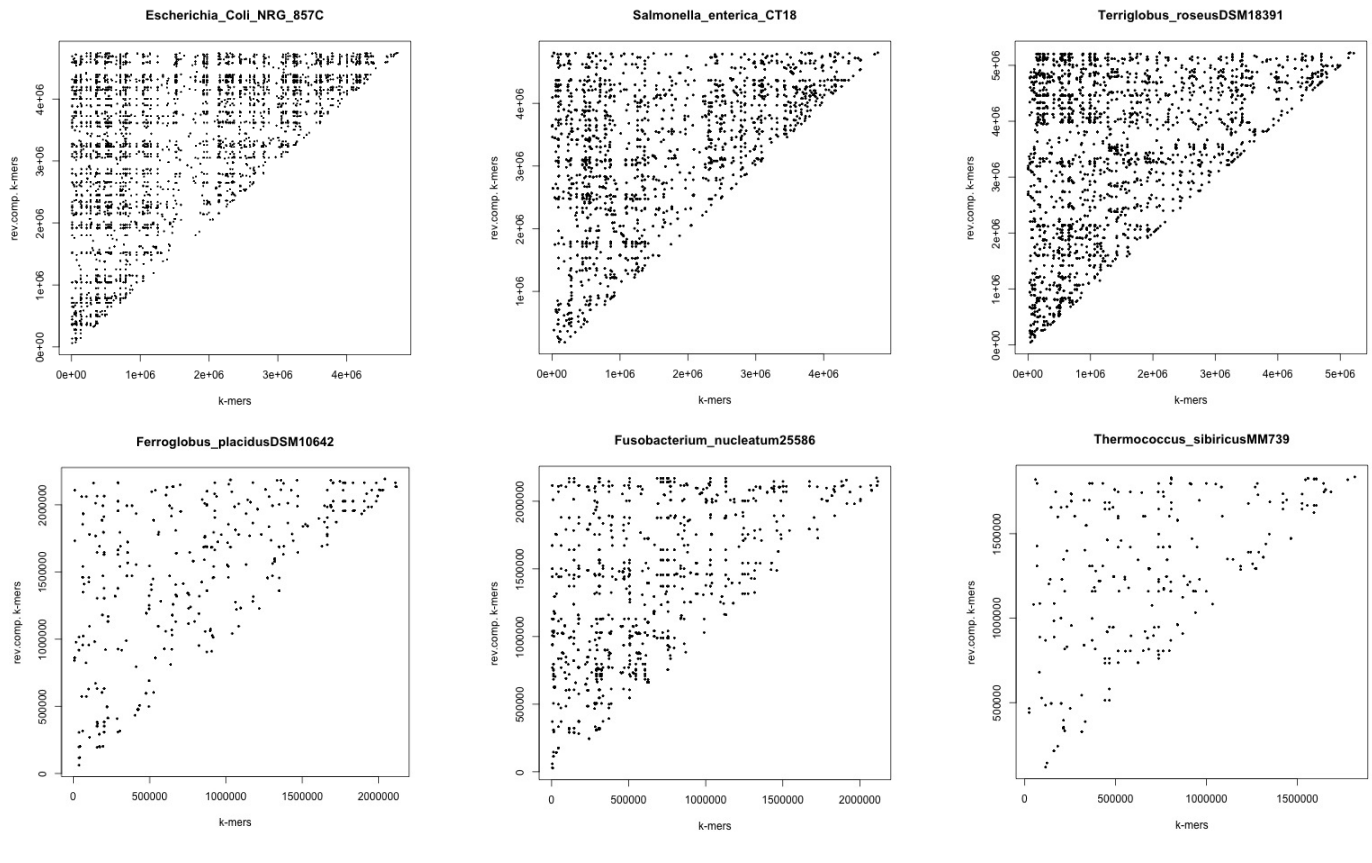


Figure 1S | Scatterplot of bacteria k-mers vs their reverse complement sequences. These plots show six bacteria k-mers against their reverse complement sequences. Each point in the x-axis represents the position of the first base of a specific k-mer, whereas each value within the y-axis is the first base of the reversed complement sequence. The six plots show that the RC-pairs are uniformly distributed all along different bacteria genomes, which is in agreement with our maximum entropy based prediction.

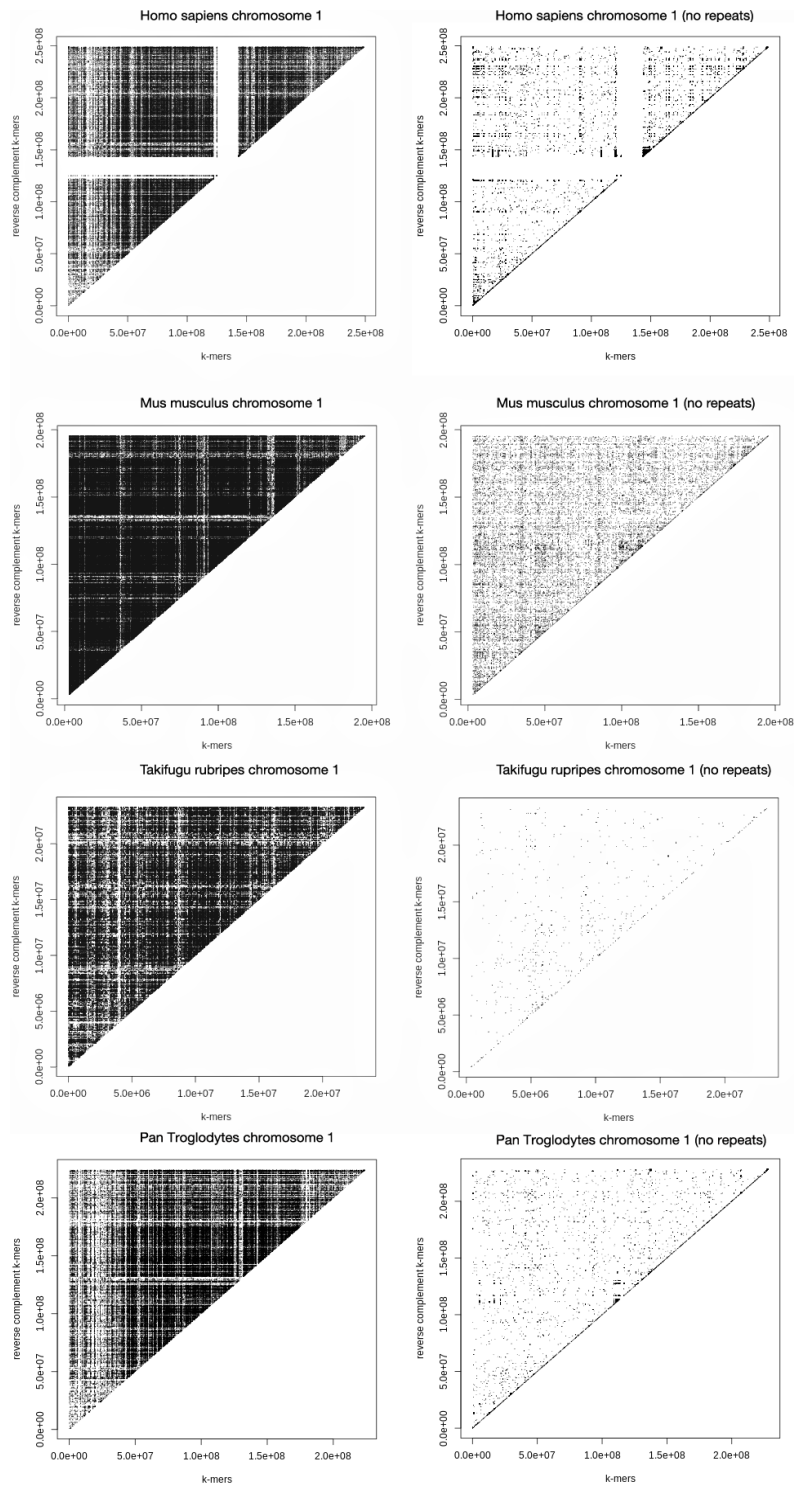


Figure 2S | Scatterplot of k-mers against their reverse complement in six eukaryotic chromosomes with and without repeated regions. The presence of a point of coordinates (x, y) shows the occurrence of a RC-pair where x represents the position of the first base of a specific k-mer, whereas the y is the first base of its reversed complement sequence along a single strand of eukaryotic genomes. RC-pairs tend to be enriched in specific regions (genomes on the left side of the figure) of the eukaryotic genomes, consistent with being the result of a sequence rearrangement or recombination (region not at equilibrium). However, when the repeated regions were removed from eukaryotic genomes (right side of the figure) the points are uniformly distributed (region at equilibrium) like in prokaryotic dsDNA.

4. CODON TRIMERS FREQUENCY BARPLOT

The human coding regions are constrained by the fact that trimer codons are translated into amino acids following a species-specific code. Thus, we investigated whether, even in a given codon bias system, the GCT is still detectable in human codon usage. A positive correlation was found when analysing the frequency of the codons and the frequency of the corresponding reverse-complement codons (Pearson correlation=0.5 with a p-value < 10⁻⁴) indicating that, still, there is a significant CGT trace in the human codon usage. The frequency trimers at whole genome level perfectly follows the GCT as shown in Figure 3S.

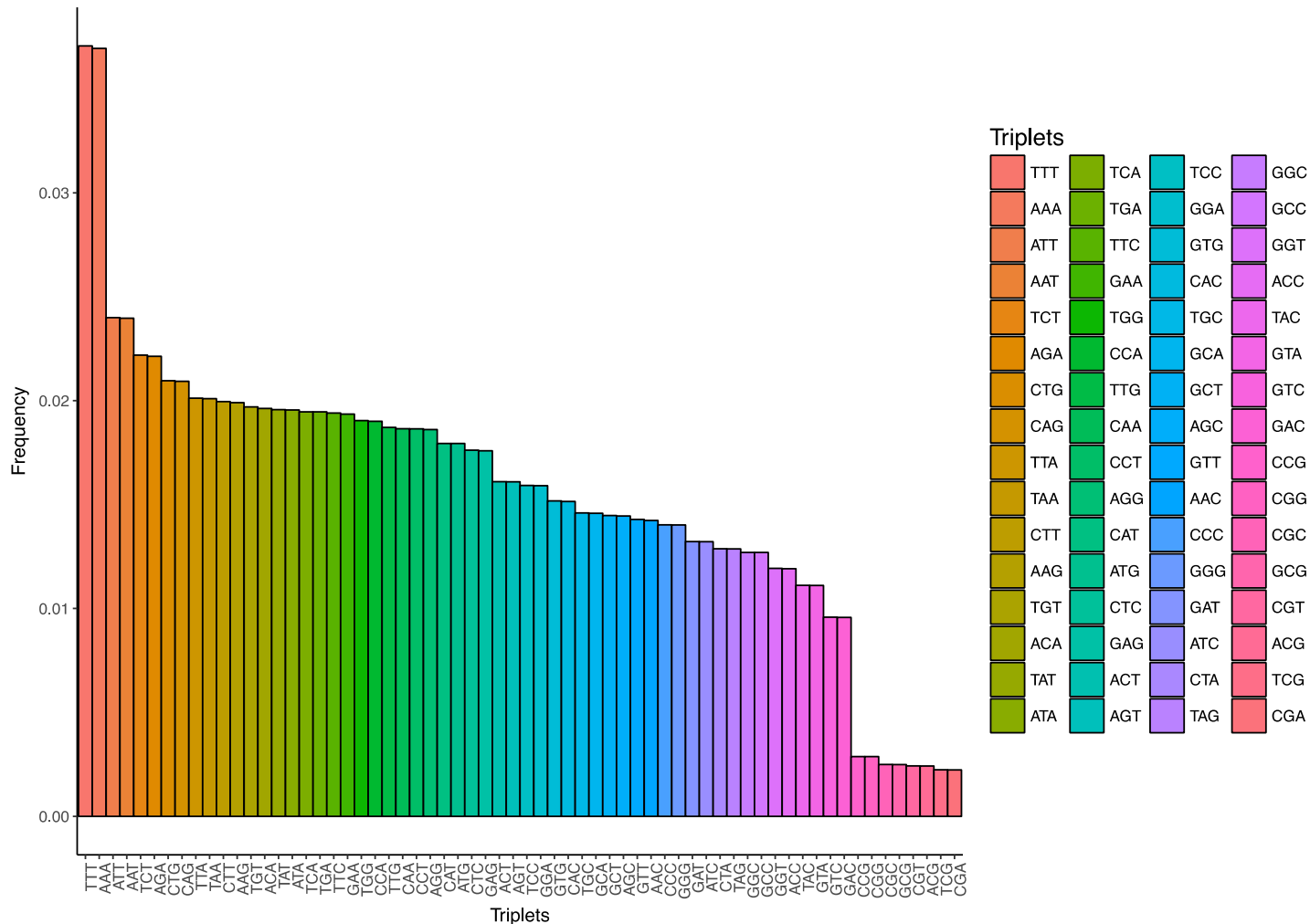


Figure 3S | Trimer as a function of their frequency (y-axis) in the human genome. The RC-pairs sit nearby as predicted by the Generalized Chargaff Theory. For sake of clarity the trimers are highlighted in colour on the right side

5. ANALYSIS OF CONTEXT-DEPENDENT MUTATIONS PATTERNS IN HUMAN POPULATION

Unbalanced distribution of context dependent mutations has been reported for the human genome when looking at trimers (considering the mutating nucleotide, the one immediately upstream and the one immediately downstream)⁶ and at longer k-mers⁷. Here we speculate that such unbalances may be partly explained by GCT and maximum entropy, given that each new

mutation, regardless of its biological impact, may be under selective constraints for its contribution towards the global energy equilibrium. To investigate this possibility, we studied 100 European (CEU) whole genome sequences from the 1000 Genome Project⁸ and focussed on NAN → NBN mutations, where A is always the ancestral (Consensus ancestral sequence from the 1000 Genomes Project, B is the derived allele and N are any two nucleotides defining the specific sequence context of a given A/B mutation. We described each of the 192 possible NAN → NBN mutation types by using the following measures: i) abundance along the genome of the mutation event (the number of times a given mutation is observed to have happened, along the genome, even just in a single individual) and ii) average population frequency of the mutations (the average fraction of individuals carrying a given mutation, calculated along the whole genome). The first measure recapitulates the properties of the human genome in terms of sequence content and, hence, the opportunity for a given NAN trimer to mutate to NBN (the more the occurrences of NAN along the ancestral genome sequence, the higher the chances of observing a NAN → NBN mutation event). The second measure is independent from the properties of the ancestral genome sequence and is instead a measure of the evolutionary success (or spread across the population) of a given mutation event (under the simplifying assumption of no recurrent mutations, and identity by descent of all carriers of a given mutation at a given genomic site). By scattering these two measures (**Figure 4Sa**) we noticed that all 192 NAN → NBN mutation types fell into three distinct categories: i) Low overall occurrences, intermediate average population frequencies (Class 1); ii) High overall occurrences, high average population frequencies (Class 2); iii) High overall occurrences, low average population frequencies (Class 3). Somewhat expectedly from Figure 4Sa, the number of occurrences of a given NAN → NBN substitution is directly proportional to the one of its NBN → NAN reverse ($R^2=0.89$, $p=10^{-16}$), while its average population frequency is inversely proportional to the average population frequency of the reverse substitution ($R^2=-0.75$, $p=10^{-16}$). We also noticed that Class 1 mutations include only transversions (purine ↔ pyrimidine substitutions), while Class 2 and Class 3 mutations include only transitions (purine-purine or pyrimidine-pyrimidine substitutions). In addition, Class 2 only include G → A and C → T, while Class 3 contains A → G and T → C mutations, suggesting a role of the directionality of these mutation types within the broader genomic “ecosystem”. We then inspected the relationship between the average population frequencies of a mutation and the one of its RC-pair (pairs involved in the GCT, here defined, for a given NAN → NBN, as mutations substituting the reverse complement of NAN into the reverse complement of NBN) and discovered a remarkably good correlation ($R^2=0.91$, $p=10^{-16}$, **Figure 4Sb**). Considering the average frequency of a given mutation type within the European population as a measure of its evolutionary success (or of its age), we then concluded that the evolutionary trajectory of any NAN → NBN mutation is paired with the trajectory of mutations substituting the reverse complement of NAN into the reverse complement of NBN, once more confirming predictions based on GCT and maximum entropy equilibrium.

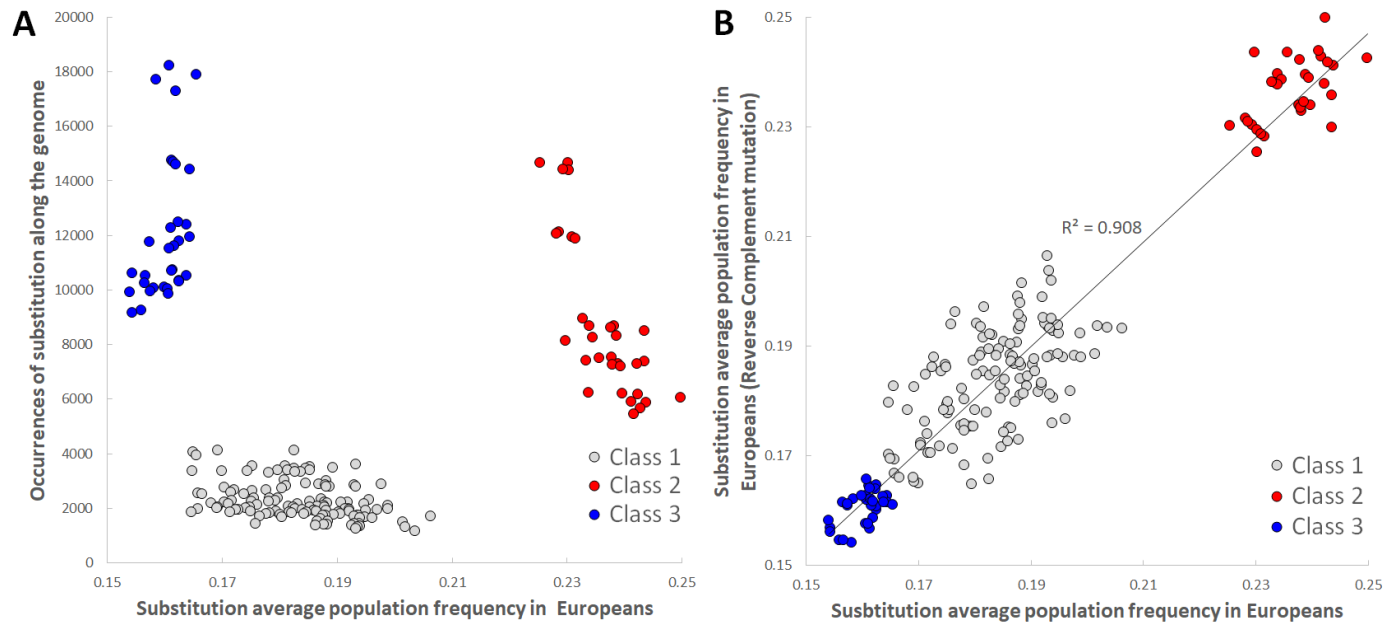


Figure 4S. Context dependent substitutions within the human genome (1000 Genomes Project European samples). **a**, The average population frequency of a given NAN->NBN substitution (x-axis) and the total occurrences of that substitution along the genome (y-axis) separate all the possible trimers into three distinct categories (Class 1, Class 2, and Class 3). **b**, The average population frequency of all the events of a given NAN->NBN type (x-axis), calculated on 1000 Genomes Europeans (CEU) correlates ($R^2=0.91$, $p=10^{-16}$) with the average population frequency of the reverse complement mutation type (y-axis: the mutation substituting the RC of NAN into the RC of NBN).

6. REFERENCES

1. Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.*106, 620-630 (1957).
2. Afreixo, V., Rodrigues, J. M., Bastos, C. A., Tavares, A. H., Silva, R. M. Exceptional Symmetry by Genomic Word: A Statistical Analysis. *Interdiscip Sci.* 9, 14-23 (2017).
3. Mitchell, D., Bridge, R. A test of Chargaff's second rule. *Biochem. Biophys. Res. Commun.* 340, 90-94 (2006).
4. Nikolaou, C., Almirantis, Y. Deviations from Chargaff's second parity rule in organellar DNA: Insights into the evolution of organellar genomes. *Gene* 381, 34-41 (2006).
5. Harris, R.S Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania University (2007)
6. Harris, K. and Pritchard, J. K. Rapid evolution of the human mutation spectrum. *eLife*, 6:e24284 (2017).
7. Aggarwala, V., Voight B.F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48:349-345 (2016)
8. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526, 68-74 (2015).