PLOS ONE

OPEN ACCESS

Citation: Bogachev MI, Kayumov AR, Bunde A (2014) Universal Internucleotide Statistics in Full Genomes: A Footprint of the DNA Structure and Packaging? PLoS ONE 9(12): e112534. doi:10.1371/journal.pone.0112534

Editor: Enrique Hernandez-Lemus, National Institute of Genomic Medicine, Mexico

Received: July 19, 2014

Accepted: October 7, 2014

Published: December 1, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the manuscript. The source data used for the analysis are publicly available from the GenBank database at ncbi.nlm.nih.gov/genome and are sufficient to reproduce the results. The accession numbers of the genomic sequences we used in our analysis are as follows: 1. Archaea 1.1. Optimal living T below 50˚C: NC_002607.1, NC_003552.1, NC_003901.1, NC_007355.1, NC_007426.1, NC_007681.1, NC_007796.1, NC_007955.1, NC_009135.1, NC_009634.1, NC_009635.1, NC_010364.1, NC_013967.1, NC_014222.1, NC_015574.1, NC_015948.1 1.2. Optimal living T between 50˚C and 80˚C: NC_000909.1, NC_002578.1, NC_002689.2, NC_002754.1, NC_005877.1, NC_013156.1, NC_013407.1, NC_013769.1, NC_013887.1, NC_014122.1, NC_015216.1, NC_015865.1 1.3. Optimal living T above 80˚C: NC_000868.1, NC_000917.1, NC_000961.1, NC_003106.2, NC_003364.1, NC_003413.1, NC_003551.1, NC_006624.1, NC_008701.1, NC_009073.1, NC_009135.1, NC_009376.1, NC_011529.1, NC_012804.1, NC_012883.1, NC_013741.1, NC_014804.1, NC_015320.1, NC_015474.1, NC_015680.1 2. Bacteria 2.1. Very low GC content: NC_005945.1, NC_002951.2, NC_003030.1, NC_004116.1, NC_004193.1, NC_007168.1, NC_008599.1, NC_008600.1, NC_009004.1, NC_009135.1, NC_009441.1, NC_011658.1, NC_012563.1, NC_012632.1 2.2. Low GC content: NC_000964.3, NC_002737.1, NC_004668.1, NC_006840.2, NC_009614.1, NC_009848.1, NC_010161.1, NC_010296.1, NC_010554.1, NC_010842.1, NC_013768.1, NC_014125.1,

RESEARCH ARTICLE

Universal Internucleotide Statistics in Full Genomes: A Footprint of the DNA Structure and Packaging?

Mikhail I. Bogachev[1]*, Airat R. Kayumov[2], Armin Bunde[3]

1. Radio Systems Department & Biomedical Engineering Research Center, Saint Petersburg Electrotechnical University, Saint Petersburg, Russia, **2.** Department of Genetics & Institute of Fundamental Medicine and Biology, Kazan (Volga Region) Federal University, Kazan, Tatarstan, Russia, **3.** Institut für Theoretische Physik, Justus-Liebig-Universität Giessen, Giessen, Hessen, Germany

*Mikhail.Bogachev@physik.uni-giessen.de

Abstract

Uncovering the fundamental laws that govern the complex DNA structural organization remains challenging and is largely based upon reconstructions from the primary nucleotide sequences. Here we investigate the distributions of the internucleotide intervals and their persistence properties in complete genomes of various organisms from *Archaea* and *Bacteria* to *H. Sapiens* aiming to reveal the manifestation of the universal DNA architecture. We find that in *all* considered organisms the internucleotide interval distributions exhibit the same q -exponential form. While in prokaryotes a single q-exponential function makes the best fit, in eukaryotes the PDF contains additionally a second q -exponential, which in the human genome makes a perfect approximation over nearly 10 decades. We suggest that this functional form is a footprint of the heterogeneous DNA structure, where the first q-exponential reflects the universal helical pitch that appears both in pro- and eukaryotic DNA, while the second q -exponential is a specific marker of the large-scale eukaryotic DNA organization.

Introduction

Understanding the complex structure of DNA as the carrier of genetic information is one of the major challenges of modern science. The architectural complexity of eukaryotic DNA is associated with its extensive functional versatility including accomodation and realization of genetic information as well as controlling the cell activity and its adaptation to various conditions. It is known that the human genome is about three orders of magnitude larger than the typical

NC_014498.1, NC_014922.1, NC_015660.1, NC_014248.1, NC_015213.1 2.3. Intermediate GC content: NC_010473.1, NC_002935.2, NC_003450.3, NC_004088.1, NC_004757.1, NC_005363.1, NC_006576.1, NC_007384.1, NC_007606.1, NC_009792.1, NC_010067.1, NC_010741.1, NC_012214.1, NC_012491.1, NC_012578.1, NC_013016.1, NC_013209.1, NC_014551.1, NC_015152.1, NC_015566.1, NC_015634.1, NC_015663.1 2.4. High GC content: NC_002755.2, NC_002929.2, NC_003062.2, NC_004129.6, NC_006932.1, NC_007406.1, NC_007761.1, NC_008095.1, NC_009656.1, NC_011283.1, NC_012483.1, NC_012490.1, NC_012560.1, NC_012803.1, NC_013722.1, NC_014550.1, NC_014618.1, NC_014638.1, NC_015757.1 3. A. queenslandica: NW_003546242.1 4. S. kowalevskii: NW_003105101.1 5. A. gambiae: NC_004818.2, NT_078266.2, NT_078268.4 6. O. latipes: NC_019859.1, NC_019860.1, NC_019861.1, NC_019862.1, NC_019863.1, NC_019864.1, NC_019865.1, NC_019866.1, NC_019867.1, NC_019868.1, NC_019869.1, NC_019870.1, NC_019871.1, NC_019872.1, NC_019873.1, NC_019874.1, NC_019875.1, NC_019876.1, NC_019877.1, NC_019878.1, NC_019879.1, NC_019880.1, NC_019881.1, NC_019882.1 7. G. gallus: NC_006088.3, NC_006089.3, NC_006090.3, NC_006091.3, NC_006092.3, NC_006093.3, NC_006094.3, NC_006095.3, NC_006096.3, NC_006097.3, NC_006098.3, NC_006099.3, NC_006100.3, NC_006101.3, NC_006102.3, NC_006103.3, NC_006104.3, NC_006105.3, NC_006106.3, NC_006107.3, NC_006108.3, NC_006109.3, NC_006110.3, NC_006111.3, NC_006112.2, NC_006113.3, NC_006114.3, NC_006115.3, NC_006119.2, NC_006126.3, NC_006127.3 8. F. catus: NC_018723.1, NC_018724.1, NC_018725.1, NC_018726.1, NC_018727.1, NC_018728.1, NC_018729.1, NC_018730.1, NC_018731.1, NC_018732.1, NC_018733.1, NC_018734.1, NC_018735.1, NC_018736.1, NC_018737.1, NC_018738.1, NC_018739.1, NC_018740.1, NC_018741.1 9. P. troglodytes: NC_006468.3, NC_006469.3, NC_006470.3, NC_006490.3, NC_006471.3, NC_006472.3, NC_006473.3, NC_006474.3, NC_006475.3, NC_006476.3, NC_006477.3, NC_006478.3, NC_006479.3, NC_006480.3, NC_006481.3, NC_006482.3, NC_006483.3, NC_006484.3, NC_006485.3, NC_006486.3, NC_006487.3, NC_006488.2, NC_006489.3, NC_006491.3, NC_006492.3 10. H. sapiens: NT_008705.16, NT_009237.18, NT_009759.16, NT_024524.14, NT_026437.12, NT_037852.6, NT_010393.16, NT_024972.8, NT_010859.14, NT_011255.14, NT_077402.2, NT_011387.8, NT_113952.1, NT_028395.3, NT_022221.13, NT_022517.18, NT_037622.5, NT_006576.16, NT_007592.15, NT_007819.17, NT_023736.17, NT_008413.18.

Funding: The financial support of this work was provided by the Deutsche Forschungsgemeinschaft (Justus-Liebig-Universitaet Giessen, project No.

microbial genome, while its coding part is only about 1.5 orders of magnitude larger. This indicates that the increase of the genome size during evolution is mainly caused by the accumulation of the noncoding DNA. While the functionality of the noncoding DNA is still under debate, there is a major agreement about its architectural role in the formation of the complex eukaryotic DNA structure. This suggests that the genomic evolution is not limited to the introduction of new genes but also includes considerable complication of the DNA spatial structure. Understanding the DNA structural evolution is especially important since it significantly contributes to the control of the gene expression, DNA replication, recombination and repair mechanisms [1,2].

The DNA consists of two complementary polynucleotide chains which at small scales form a double helix with a helical pitch of about 10–11 base pairs (bp) [1,2] that is universal for all cellular life. At larger scales, the DNA structure varies considerably between different domains of life, the simple *prokaryotes* exemplified by *Archaea* and *Bacteria* and the *eukaryotes* characterized by a cell that contains a nucleus (a large group of organisms ranging from yeast, fungi and plants to animals including *H. Sapiens*). While in *Bacteria* the DNA is located in a relatively free manner in the cytoplasm, with random attachments to the cell membrane and without any characteristic structural scales, in *Archaea* the DNA is additionally wrapped around the histones. In contrast, in the *eukaryotes* the DNA structure is more complex and, additionally to these two structural levels, constitutes several other packaging levels with larger characteristic scales. For an extensive review on the eukaryotic DNA structure, we refer to [2].

The primary structure of DNA is determined by a sequence that consists of four nucleotides, namely adenosine (A), cytosine (C), guanosine (G) and thymidine (T). The second polynucleotide chain can be normally reconstructed from the first one due to their complementarity, provided that A is opposed to T and G is opposed to C, and thus statistical analysis can be performed on a single sequence. The two types of base pairs have considerably different bonding energies characterized by the bond enthalpies -11.8 for A:T and -23.8 kcal/mol for G:C, respectively [3]. Nevertheless the occurrence of either G:C or A:T in the primary sequence leads to the identical tertiary architecture of the base pair and thus their alteration does not perturb the double helix structure [1].

It has been revealed earlier that the DNA sequences exhibit long-range correlations (LRC) with rather monofractal properties [4–9]. It has been established that there are two scaling regimes in the DNA sequences that are separated by a prolonged crossover typically between 100 bp and 1 kbp. Below the crossover the correlations are characterized by Hurst exponents H close to 0.5 for the prokaryotes and close to 0.6 for the eukaryotes. Above the crossover, H is around 0.8 in both domains. Grossberg *et al.* first proposed that the long-range correlations in the DNA primary sequences are related to its three-dimensional structure [10]. Next the LRC have been associated the formation of the DNA bending profile [11], and the two separate scaling regimes were attributed to the different hierarchical levels of the DNA structure [12,13]. The relationship

BU 534/24-1, AB), by the Ministry of Education and Science of the Russian Federation (St. Petersburg Electrotechnical University, assignment No. 2014/187, MB) and by the subsidy of the Russian Government to support the Program of competitive development of the leading academic centers (Kazan Federal University, AK). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

between the LRC in the DNA primary sequence and its elastic bend stiffness related to the DNA loops formation has been further investigated [14,15].

The original approach to the LRC in DNA is based on the ''DNA walk'', which increases by one when a pyrimidine (C or T) is observed and decreases by one when a purine (A or G) is observed in a DNA sequence [4]. Here we follow a more direct approach to the DNA primary structure based on the persistence properties of the intervals between the same nucleotides (A-A, C-C, G-G, T-T) in the DNA sequence. The central quantities here are the distribution of the intervals and their autocorrelation function that are signatures of both linear and nonlinear correlations in the analysed DNA sequences. In random sequences, the probability density function (PDF) of the intervals is a simple exponential $P(l)=(1/L)\exp(-l/L)$, where L is the average interval length, and the intervals are uncorrelated. In LRC sequences similar nucleotides tend to follow each other, and thus short (long) intervals are more likely to be followed by short (long) intervals. When there are purely linear correlations in the sequence, one expects that the PDF follows a stretched exponential $\ln[P(l)] \propto -(l/L)^{\gamma}$, with exponent $\gamma=2-2H$. The intervals are also LRC and their autocorrelation function (ACF) decays by a power law $C(s) \propto s^{-\gamma}$ with the same correlation exponent γ [16–18]. In the presence of nonlinear correlations, the PDF gets even broader and decays by a power-law $P(l) \sim (l/L)^{-\delta}$, where the exponent δ changes with the strength of the LRC in the data [19,20]. In some complex systems with nonlinear LRC like returns in financial markets exceeding a certain threshold also q-exponential PDFs of the intervals have been observed [21]. In both cases, the ACF of the intervals decays by a power law [16,18,19].

Materials and Methods

We assessed the complete genomes of organisms at different evolutionary positions ranging from *Archaea* (48 species) and *Bacteria* (72 species) to various eukaryotes including *H. Sapiens* from the NCBI GenBank [22]. From each genomic sequence we obtained the series of consecutive intervals between the same nucleotides (A-A, C-C, G-G, T-T). The procedure of the assessment of the four internucleotide interval sequences from the DNA primary sequence is shown in Fig. 1. We focused on the two major quantities characterizing the intervals series, the probability distribution functions (PDFs) and the autocorrelation functions (ACFs).

To obtain the distribution, we first counted the number of occurrences of intervals of a certain size of l nucleotides, which constituted the histrogram $H(l)$. To obtain the probability density function (PDF), we next divided $H(l)$ by the total number of fragments in the genome. By definition, the PDF is normalized, $\sum_{l_{min}}^{l_{max}} P(l)=1$, where l_{min} and l_{max} are the shortest and the longest intervals observed in the studied sequence. Since large fragments occur very rarely, the statistics becomes poor for large l, and the functional form of the PDF can no longer be observed visually.

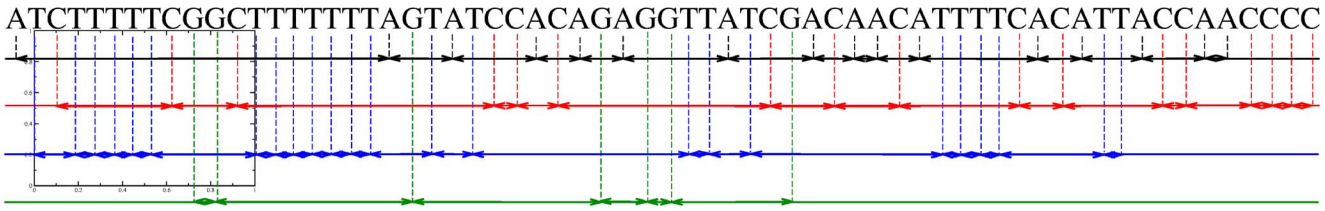

Figure 1. The procedure of the assessment of the four internucleotide interval sequences from the DNA primary sequence.

doi:10.1371/journal.pone.0112534.g001

In order to improve the statistics gradually with increasing size l, we have chosen logarithmic binning, which is widely used in statistical physics in order to determine the behavior of the tails of the distribution. In the logarithmic binning, one counts the number of fragments with sizes between a_i and a_{i+1}, $a_i \propto b^i$, where i is the number of bin. We have tuned the parameter b to achieve the best visualization of our results, in particular $b = 1.22$ for the PDFs obtained from the single DNA sequences and $b = 1.4$ for the total PDFs over several DNA sequences. The respective l value was associated with the center of the bin in log-scale, i.e., at the geometric average. For example, when the bin included values from $l = 5$ to $l = 7$, we averaged $P(5)$, $P(6)$ and $P(7)$, with the associated $l = \sqrt[3]{5 \cdot 6 \cdot 7}$, respectively. Once a bin which contained less than three occurrences was found, the analysis was stopped. For obtaining the functional form of $P(l)$ over several decades, we typically used a double-logarithmic presentation.

Finally, in order to eliminate the changes in the average size of fragments, and to concentrate solely on the shape of the distributions, we used the average interval L as a characteristic scale for the size distributions. We rescaled the PDFs by dividing the sizes l by their mean value L for every particular genome. To keep the normalization, we also multiplied the PDFs by L, and thus finally obtained $LP(l)$ as a function of l/L.

Next we calculated the linear two-point autocorrelation function (ACF) $C_x(s)$ of the interval series

$$C_l(s) = \frac{1}{\sigma_l^2 (N-s)} \sum_{i=1}^{N-s} (l_i - L)(l_{i+s} - L), \qquad (1)$$

where s is the scale parameter and N is the total number of intervals in the sequence, and applied a similar logarithmic binning procedure to it.

Results

Internucleotide interval distributions

Figure 2 shows the PDFs of the inter-nucleotide intervals in the DNA of *Archaea* (48 species), the believed predecessor of all other forms of life, *Bacteria* (72 species) and *H. Sapiens* (22 chromosomes) calculated from the complete genomes obtained from the NCBI GenBank [22]. The PDFs are provided in scaled form,

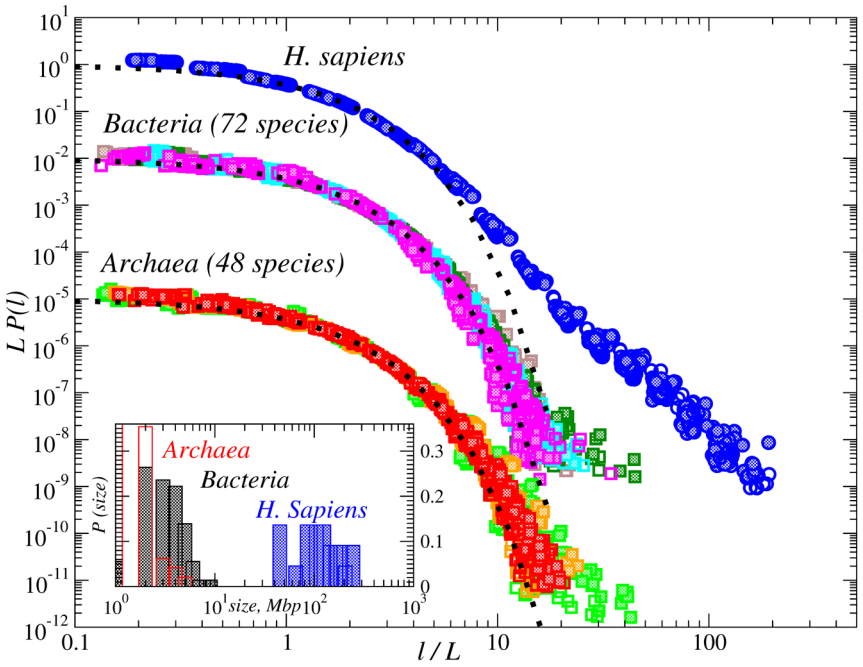

Figure 2. PDFs of the inter-nucleotide intervals A-A, T-T (open symbols); G-G, C-C (full symbols) in the DNA sequences from *Archaea* (◊) (48 species), *Bacteria* (□) (72 species) and *H. Sapiens* (○) (22 chromosomes). For comparison, dotted lines show corresponding exponential PDFs. The inset shows the size distribution plots of the DNA sequences considered.

doi:10.1371/journal.pone.0112534.g002

i.e., the intervals l are given in units of the average interval L (around 4), and the PDFs are multiplied by the same L value to keep the normalization. For comparison, the figure shows also simple exponential $P(l) = 1/L \exp(-l/L)$ distributions (by dotted lines. At scales $l/L > 1$ the empirical PDFs are significantly broader than the simple exponential distributions rejecting the hypothesis of the random positioning of nucleotides. In archaeal and bacterial genomes, the simple exponential appears close to the lower bound for the empirical PDFs, while the upper bound resembles a power law at the tail of the distribution. Additionally pronounced scattering at the tail of the PDFs can be observed and thus a particular functional form can hardly be determined from the analysis of individual sequences. This scattering could be attributed both to the heterogeneity of the considered *Archaea* and *Bacteria* as well as to the finite size effects taking into account the limited size of the considered DNA sequences (see the inset in Fig. 2). In contrast, in the *H. Sapiens* genome the scattering is much less pronounced, and the empirical PDFs exhibit a specific two-compound shape with clearly significant deviations from a single exponential distribution.

In order to determine the PDFs more accurately, instead of considering the individual histograms $H(l)$ and PDFs $P(l)$, we next calculate the total histograms $H(l)$ of inter-nucleotide intervals over *all Archaea* and *Bacteria* genomes as well as the total histogram over all 22 chromosomes in the *H. Sapiens* genome. The results are shown in Fig. 3 in the units of the respective average intervals L. The

figure shows that the total PDF in *Bacteria* can be well approximated by a single q-exponential distribution

$$P(l) = \frac{1}{L}\frac{A}{[1+(q-1)\beta(l/L)]^{1/(q-1)}} \quad (2)$$

with $q = 1.1 \pm 0.01$, $\beta = 1.5$ and $A = 1.5$ over 8 orders of magnitude limited by the genome size. The q-exponential distribution is a special case of the generalized Pareto distribution

$$GP(l) \sim \frac{1}{\sigma}\left[1+\frac{\xi(l/L)}{\sigma}\right]^{-(1/\xi)-1}, \quad (3)$$

where $\sigma = 1/[\beta(2-q)]$ and $\xi = (q-1)/(2-q)$. The q-exponential distribution extremizes, under simple constraints, the nonadditive entropy which is the generalization of the Boltzmann-Gibbs entropy [26]. In the limit $q \to 1$ the q-exponential distribution reduces to a simple exponential. At small arguments $l/L < 1$ the q-exponential it behaves as $A - \beta(l/L)$ for all q values. A similar functional form but with $q = 1.11 \pm 0.02$ can be observed in the total PDF for the *Archaea*. It is also notable that in prokatyores the PDFs for the stronger bonded nucleotides are slightly broader than for the weakly bonded ones.

In the human genome, the description by a single q-exponential is valid only at small and intermediate scales $l/L < 20$, while at large scales $l/L > 40$ another q-exponential *with the same* $q = 1.11$ as in *Archaea*, but now with $\beta = 1.5$ and $A = 10^{-5}$ makes a perfect fit. If we add both q-exponentials, we obtain an excellent fit over nearly 10 orders of magnitude limited by the human chromosome size (see the size distribution plot in the inset of Fig. 3). It is remarkable that in the *H. Sapiens* genome there is a perfect collapse for the internucleotide interval distributions for each of the nucleotides A, C, G or T.

Given $L \cong 4$ the first q-exponential is valid until approximately 100 bp that is close to the characteristic scale of the DNA wrapping cycle (146 bp in *H. Sapiens*) [2,27] and thus might correspond to the first two DNA structural levels, namely the double helix and the DNA wrapping around the histones (nucleosome formation) [2,12,13]. The second q-exponential likely reflects the eukaryotic DNA organization at larger scales. It is interesting that, despite of the apparent otherness of different DNA structural levels, the q value is identical, that suggests striking similarities in their optimization principles.

Figure 4 shows (in addition to the results of Fig. 3) the scaled PDFs of internucleotide intervals for seven more eukaryotic species at various evolutionary levels such as *A. queenslandica* (sponge native to the Great Barrier Reef), *S. kowalevskii* (marine acorn worm), *A. gambiae* (malaria mosquitoes), *O. latipes* (Japanese rice fish), *G. gallus* (chicken), *F. catus* (domestic cat) and *P. troglodytes* (common chimpanzee). The functional fits in the figure are identical to those in Fig. 3. The inset shows the average intervals L for different nucleotides. The figure shows that, despite of the discrepancy in the average interval L (shown in the inset of Fig. 4), after rescaling the PDFs in *all* higher eukaryotes obey a similar

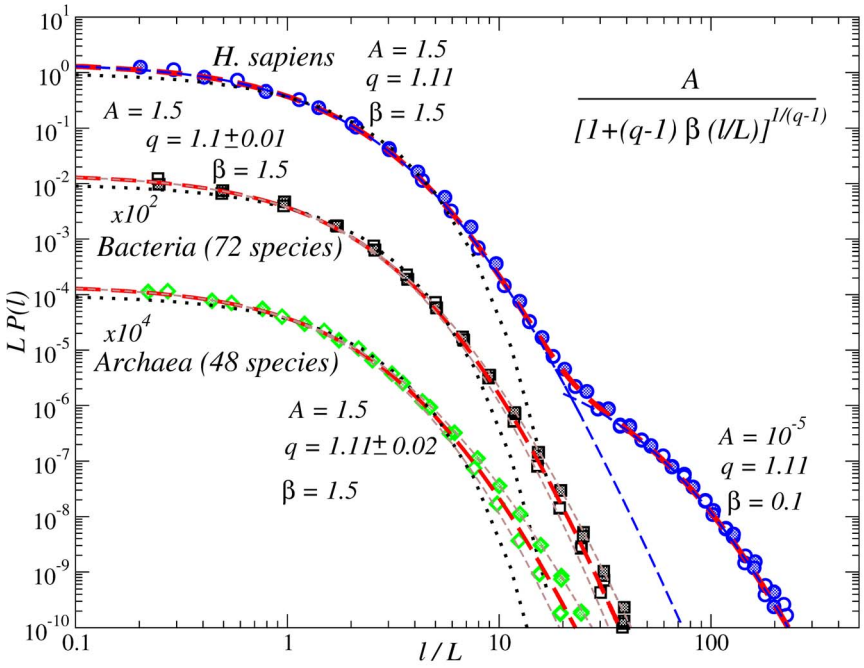

Figure 3. PDFs of the inter-nucleotide intervals A-A, T-T (open symbols); G-G, C-C (full symbols) in the DNA sequences from *H. Sapiens* and *Bacteria* full genomes (in scaled form). Dashed lines show the best fits by a q-exponential distribution $A/[1+(q-1)\beta(l/L)]^{1/(q-1)}$. While in *Bacteria* the approximation by a single q-exponential with $q=1.1$ and $\beta=1.5$ is possible, in *H. Sapiens* a sum of two q-exponentials with $q=1.11$ and $\beta=1.5$ and 0.1 makes the best fit. To avoid overlapping, the PDFs for *Bacteria* are shifted downwards by two decades. For comparison, dotted lines show corresponding exponential PDFs.

doi:10.1371/journal.pone.0112534.g003

functional form as in *H. Sapiens*. We like to emphasize that for all 10 studied examples the first part of the distribution up to about $l/L<20$ is identical supporting the hypothesis that the observed q-exponential is a footprint of the universal DNA double helix structure that is independent of the evolutionary position of the organism.

Figure 4 additionally indicates that some deviations in the large-scale part of the distribution can be observed for several organisms at intermediate evolutionary positions such as *A. queenslandica*, *S. kowalevskii* and *O. latipes*. The figure shows that the PDFs still follow the universal q-exponential for $l/L<20$. At larger scales there are moderate deviations from the second q-exponential (universal for higher eukaryotes). Besides their evolutionary positions, we like to note that these organisms are all water living with environmentally dependent body temperature and thus the deviations from the universal PDF could be a reflection of their adaptation to the living environment associated with more pronounced thermodynamical constraints.

The inset in Fig. 4 shows that the same water living organisms are characterized by the largest average interval for G and C, i.e., their genomes has the lowest fraction of G:C base pairs and the highest fraction of A:T base pairs among considered species, respectively. Due to the differences in their bonding energies,

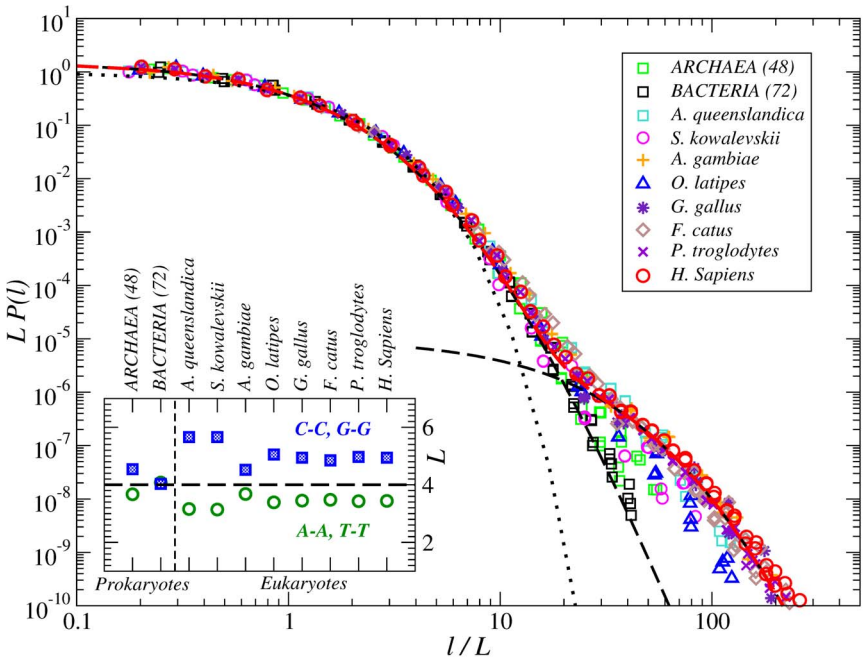

Figure 4. PDFs of the internucleotide intervals in the DNA from full genomes of ten different organisms at different evolutionary positions from *Archaea* and *Bacteria* to *H. Sapiens*. The thin dashed line shows an approximation by a single q-exponential, while the thick dashed line shows an approximation by a sum of two q-exponentials. For comparison, the dotted line shows the corresponding exponential PDF. The inset shows the evolution of the average interval L separately for strongly (G:C) and weakly (A:T) bonded nucleotides.

doi:10.1371/journal.pone.0112534.g004

the relative fractions of G:C and A:T base pairs (GC-/AT-content) are closely associated with the optimal environmental temperature that is limited by the DNA thermostability at the high end and by the energy required for the DNA unwinding during replication and the low end. On the other hand, broader interval distributions for larger average intervals is a typical sign of multifractality [18,19]. However, in multifractal data this effect is commonly observed over a wide scale range, contrast to the observations in Fig. 4. It is also known that in eukaryotic genomes GC-rich and AT-rich regions exhibit alterations that are associated with the DNA replication domains and thus the average G:C or A:T intervals over the whole genome are often not representative for most of its particular fragmetns.

To further evaluate the effects of the G:C/A:T content and of the environmental temperature, we considered prokaryotes with less complex genome structure. Figure 5 shows the PDFs of the inter-nucleotide intervals in the DNA from full genomes of *Bacteria* classified into four groups according to the fraction of G and C in their genomes and *Archaea* classified into three groups according to their optimal living temperature, from normal environment to extermophiles. The figure shows that in *Bacteria* the distributions for larger average intervals L appear slightly broader. Additionally for some groups in both *Bacteria* and *Archaea* the PDFs for G and C appear typically broader than for A and T that could be

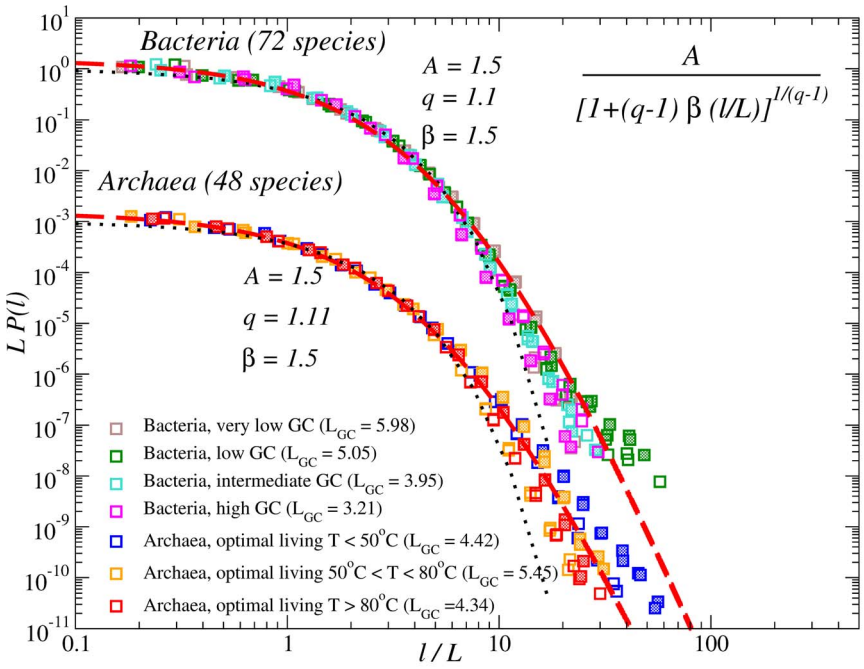

Figure 5. PDFs of the internucleotide intervals in the DNA from full genomes of *Bacteria* classified into four groups according to the fraction of G and C in their genomes (GC-content) and *Archaea* classified into three groups according to their optimal living temperature, from normal environment to extermophiles. The dashed lines show approximations by a single q-exponential with $q=1.1$ for *Bacteria* and $q=1.11$ for *Archaea*. Open symbols denote distributions for A and T, while full symbols denote distributions for G and C.

doi:10.1371/journal.pone.0112534.g005

attributed to more pronounced thermodynamical constraints on the location of the stronger bonded nucleotides. In *Archaea* the dependence of the PDFs on the environmental temperature for different species can be also clearly observed and is especially pronounced when comparing the normal ($T<50^{\circ}$C) and the extermophile ($T>80^{\circ}$C) groups, despite of their average intervals between G:C base pairs L_{GC} being nearly identical. Our results indicate that moderate deviations from the universal PDFs observed under peculiar conditions like environmentally dependent body temperature in a wide range and/or extreme living temperatures cannot be fully described by the average interval L_{GC} as a single parameter within a framework of a simple multifractal model.

To check whether the observed functional forms are related to the coding/noncoding DNA fragments, we also considered the PDFs for the transcribed DNA sequences (which contain both the coding exons and the noncoding introns) and the complementary DNA (cDNA) sequences which contain only coding parts of DNA. Figure 6 shows the corresponding results for the same organisms as in Fig. 3. Since the sequences considered here are quite short, we used the semi-logarithmic presentation to best distinguish from exponential functional form. The figure shows explicitly that both in *Bacteria* and in the *H. Sapiens* genomes at least in the small- and medium-scale regime $l/L<20$ the fit by a single

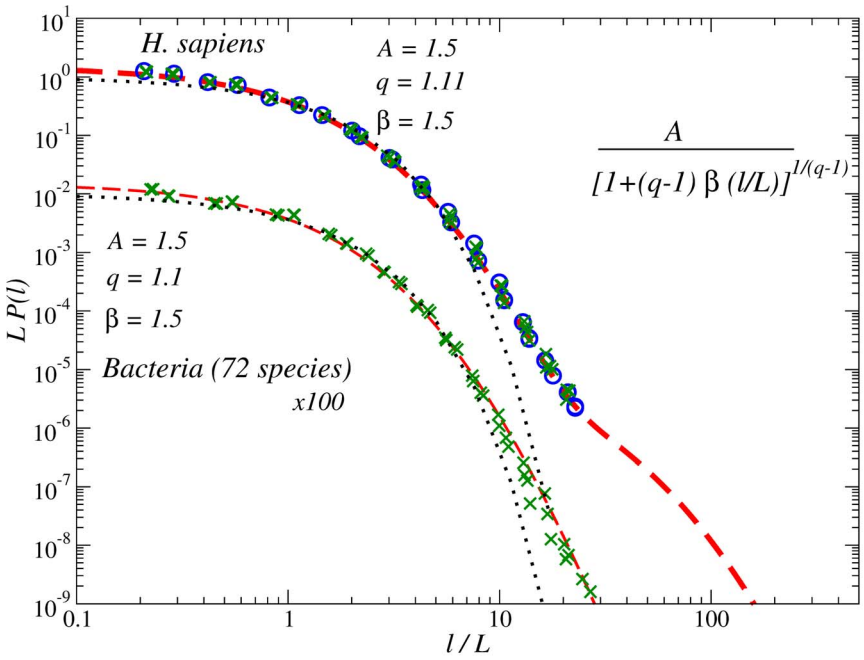

Figure 6. Similar to Fig. 3 but shows the corresponding PDFs for the transcribed DNA (∘) and for the complementary DNA (cDNA, ×). Fits are the same as in Fig. 3. To avoid overlapping, the PDFs for *Bacteria* are shifted downwards by two decades. For comparison, dotted lines show corresponding exponential PDFs.

doi:10.1371/journal.pone.0112534.g006

q-exponential is also valid for the transcribed and for the cDNA sequences. Since there are no introns (noncoding intragenic DNA) in *Bacteria* their transcribed DNA contains solely coding DNA (that corresponds to the eukaryotic cDNA). Both the transcribed and the complementary DNA sequences are relatively short and the large-scale behaviour could not be determined accurately.

In the *H. Sapiens* genome we also considered repetitive DNA that is found in multiple copies throughout the genome. Since the repetitive DNA consists of relatively short fragments, for further analysis we constructed artificial sequences by concatenating repetitive DNA fragments recongnized by the repeat-masker algorithm [23,24] and the remaining non-repetitive fragments between them, respectively. The PDFs of internucleotide intervals for such sequences are shown in Fig. 7. The figure shows that, while for the repetitive DNA the PDFs follow roughly the same functional form with double *q*-exponential shape, for the remaining non-repetitive DNA significant deviations from this behaviour can be observed. This further indicates that, while the first *q*-exponential is universal for all DNA fragments, the second *q*-exponential is entirely determined by the noncoding DNA, especially by the intergenic regions where the repetitive DNA is mainly located.

Internucleotide interval arrangement

To evaluate the inter-nucleotide interval arrangement we also studied the autocorrelation function (ACF) of the interval sequences. For comparison with

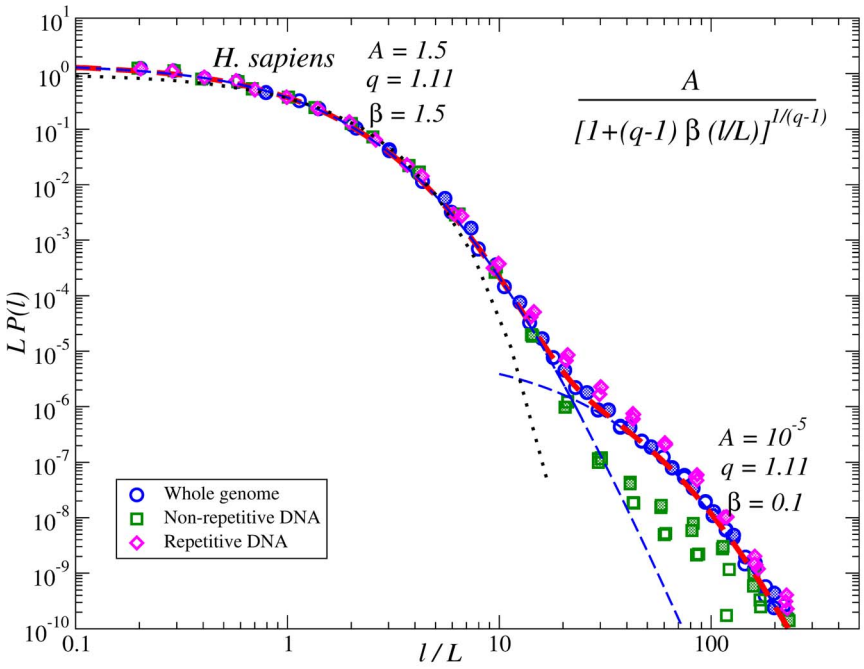

Figure 7. Similar to Fig. 6 but shows the corresponding PDFs for the human genome after elimination of the repeat-masked DNA. For comparison, the dotted line shows corresponding exponential PDF.

doi:10.1371/journal.pone.0112534.g007

earlier results based on the the DNA walk analysis, we have calculated the fluctuation functions using second-order detrended fluctuation analysis (DFA) [7] exemplified for the DNA walks of the *H. Sapiens* and *Bacteria* genomes that are shown in Fig. 8(a). Additionally to the DNA walks, we also calculated the DFA for the internucleotide interval sequences, shown in the same figure. To improve the presentation, we have divided the fluctuation functions $F(l)$ by the square root of their arguments l such that the absence of correlations corresponds to the horizontal line in the plot. The figure shows that both fluctuation functions closely reproduce the previous findings of [12,13]. The empirical exponents of the observational data at small scales are $\alpha = 0.59 \cong 0.6$ for *H.sapiens* and $\alpha = 0.53 \cong 0.5$ for *Bacteria*, while at large scales $\alpha = 0.75$ in both pro- and eukaryotes, that is consistent with the typical H values obtained earlier by wavelet-based analysis in [12,13]. In our case the crossover appears closer to 1 kbp in both cases, that can be attributed to the higher order detrending used in the DFA analysis (for more details of the effect of the detrending order on the crossover position, we refer to [25]).

Finally, the ACFs for the same internucleotide interval sequences are shown in Fig. 8(b). The figure shows that in the human genome the ACF up to about 10 kbp can be reasonably well approximated by a power-law (PL), $C(s) \propto s^{-\gamma_{\text{eff}}}$, with an effective correlation exponent $\gamma_{\text{eff}} \cong 0.3$, that is consistent with the H values around 0.8. Our results however indicate that the same correlation exponent γ_{eff} also characterizes the ACF of the intervals at the smaller scales below 200 bp. Notable

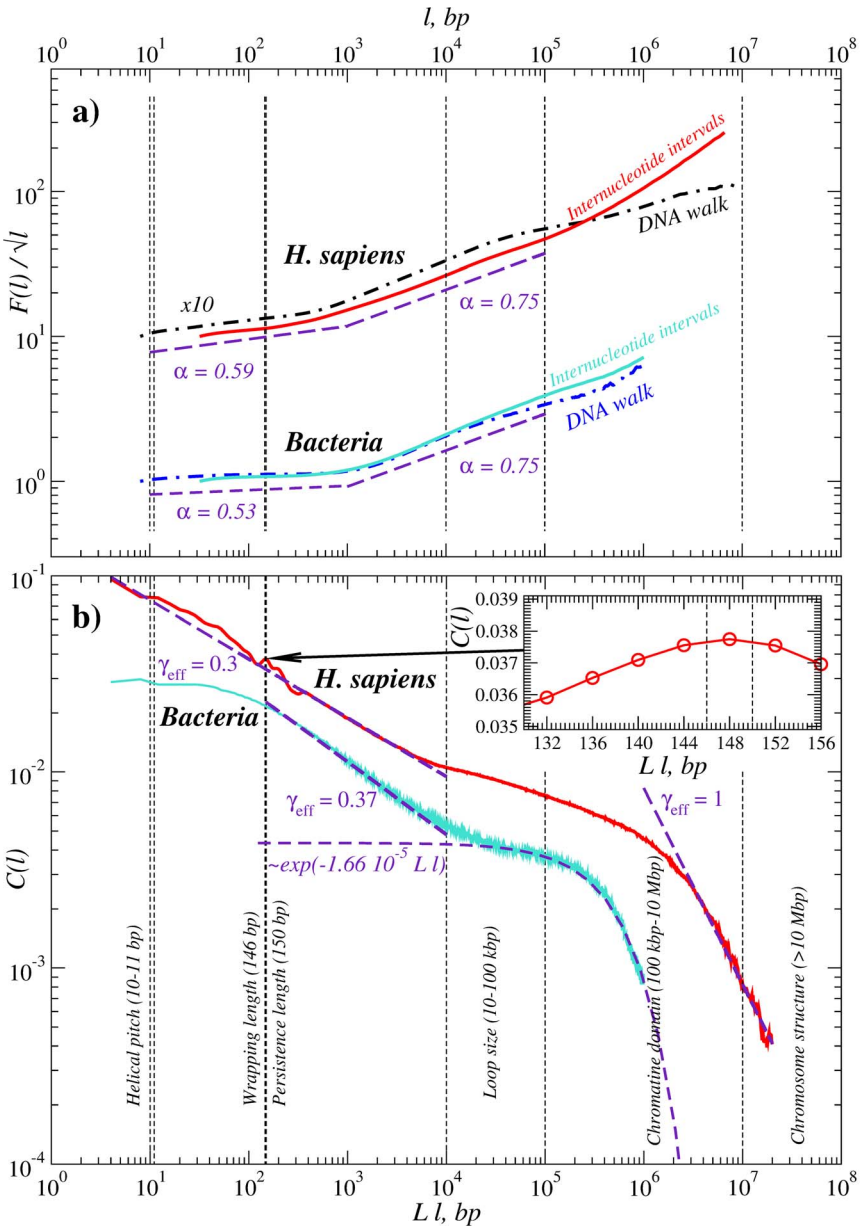

Figure 8. Comparison of the internucleotide interval statistics with the DNA walk analysis: (a) DFA fluctuation functions of the internucleotide intervals (full lines) and of the DNA walks (dashdot lines) and (b) the ACFs of the inter-nucleotide intervals (full lines), all provided for the DNA sequences of *H. Sapiens* (upper curves in each panel) and *Bacteria* (lower curves in each panel) full genomes. For the internucleotide interval sequences, the arguments are multiplied by the average interval $L \cong 4$ to provide all results in the same units of base pairs. Vertical dashed lines refer to the approximate boundaries of characteristic scaling regimes for different hierarchical levels of eukaryotic DNA packaging structure (exemplified for *H. Sapiens*, following [2]).

doi:10.1371/journal.pone.0112534.g008

violations from this behavior could be attributed to the characteristic scales of the DNA packaging structure like the helical pitch at 10–11 bp and the wrapping length at 146 bp that shows up as a spike in the ACF (see the inset in Fig. 8(b)).

In contrast, in *Bacteria* the power law regime can be observed only between 200 bp and 10 kbp with a slightly larger correlation exponent γ_{eff} about 0.37, that is consistent with the H values about 0.8 observed in earlier studies by DNA walk analysis [12,13]. Below 100 bp the ACF demonstrates a crossover that ends up with a plateau below approx. 40 bp, with a single notable deviation around 10–11 bp corresponding to the helical pitch, that is the only characteristic scale in the prokaryotic DNA packaging.

Despite the considerable discrepancy of the shape of the ACF obtained for the internucleotide intervals in the bacterial and in the human genome, the persistence in the arrangement of nucleotides at scales below 10 kbp both in *Bacteria* and *H. Sapiens* could be interpreted in the framework of the same simple model, where long-range correlations and random ''white'' noise are superimposed in the interval sequence. It has been shown recently that in the simulated LRC data, the lag 1 ACF is given by $C(1)=(2-\gamma)(1-\gamma)/2$ [28], yielding $C(1)\cong 0.6$ for $\gamma=0.3$ and $C(1)\cong 0.5$ for $\gamma=0.37$. However, the observed $C(1)$ values in Fig. 8 are considerably lower, in particular $C(1)\cong 0.1$ for the human DNA and $C(1)\cong 0.03$ for the bacterial DNA. This indicates that the interval sequences, while exhibiting LRC, also contain additive random ''white'' noise, which is much more pronounced in bacterial than in the human genome. For an analytical treatment of the superimposed LRC and random noise, we refer to [28].

At large scales above 10 kbp in *Bacteria* the ACF follows a clear exponential decay over the next two decades corresponding to the rather randomized organization of the DNA-membrane attachments to the cell membrane [29]. In contrast, in the *H. Sapiens* genome in the same scale range there is a crossover to a regime with even more pronounced LRC indicating persistence in the large-scale structural organization. This crossover is consistent with the borderline between the chromatin ''compaction'' and ''looping'' regimes. The breakdown of LRC in the human DNA occurs well above 1 Mbp that is two orders of magnitude higher than in *Bacteria*. The rapid decay of the ACF with $\gamma_{\text{eff}}\cong 1$ well above 1 Mbp suggests rather uncorrelated arrangements at these scales. For a more detailed overview of the eukaryotic DNA structural levels, we refer to Fig. 1 in [2], p.49.

Discussion

The observed universality in the functional form of the internucleotide interval distributions indicates the universality of the DNA compaction in the eukaryotic cell nucleus while prokaryotic (bacterial and archaeal) DNA is spreaded over the cell space in a relatively random manner. It has been recently revealed that the DNA structural organization and packaging contributes significantly to its robustness against UV radiation, chemical agents, electromagnetic fields and other external stress factors [30–33]. The observed striking universality in the internucleotide interval distributions could be attributed to the similarity in the cell stress pattern perceived by all eukaryotic species from their living environment that leads to stunningly similar structural optimization. The

universal q-exponential fit as a typical background distribution could also be used when investigating deviations from the typical pattern in eukaryotic DNA as indicators of truncations caused by mutation, insertions of mobile genetic elements and viruses, or specific adaptation to environmental conditions.

In the interval arrangement, the difference in the additive "white" noise level in *Bacteria* and in *H. Sapiens* can explain in addition to the discrepancy between the ACFs of their internucleotide interval sequences, also the differences in the earlier reported wavelet-based fluctuation functions of the respective DNA walks [12,13]. Since in *Bacteria* the noise is very pronounced, it completely overwhelmes the LRC at small scales resulting in vanishing DNA walk correlations with H close to 0.5 [12,13]. In contrast, in the *H. Sapiens* genome the noise is less pronounced, and thus the superimposed LRC is not completely hidden at very small scales as in *Bacteria*, but gives rise to an effective Hurst exponent $H \cong 0.6$. At scales between 100 bp and 1 kbp there is a prolonged crossover to the regime where LRC effects dominate the noise effects and the effective correlation exponent obtained here is consistent with the Hurst exponents obtained earlier by wavelet-based analysis [12,13]. It is known that in the presence of additive noise fluctuation analysis methods create artificial crossovers like this. Taking into account that the DNA structure at very small scales well below 100 bp is rather similar in pro- and eukaryotes, we suggest that the amount of "white" noise in the arrangement of nucleotides is the only difference, and the respective crossover appears to be an artifact of the analysis technique.

Finally, we like to note that the q-exponential description of distribution functions has been found useful in many other complex systems [21,34–37]. For an extensive review, we refer to [35]. Recently the broad occurrence of Pareto-tailed distributions in a number of economic, social and biological systems has been attributed to similar optimization principles in a common entropy maximization framework [38]. Also very recently q-exponentials have been observed in the distributions of interevent times in seismic data [39]. In another recent application in financial markets, also the interoccurrence times between daily losses exceeding a certain threshold have been analyzed [21]. It has been found that the distribution of the interoccurrence times also follows a q-exponential, where the parameter q only depends on the average interoccurrence time R and not on the respective asset (stocks, commodities, or currency exchange rates). We have found that for $R = 4$, which corresponds to $L = 4$ in the DNA example, the same q-exponential with the same parameters A, β and q describes the distribution of the interoccurrence times over the first four decades of magnitude. Moreover, the validity of exactly the same approximations has been also confirmed for the financial data with minute time resolution [40]. Due to the limited statistics of the financial data, the behaviour at larger scales cannot be studied. We consider an accidental coincidence as unlikely. The coincidence may indicate that for the DNA structure as well as for the dynamics of the financial markets similar optimization strategies hold.

Conclusions

In summary, we have investigated the DNA structure using the statistics of internucleotide intervals. The advantage of this approach is that it does not require generation of secondary synthetic sequences like DNA walks. We have shown explicitly that the distribution of the internucleotide intervals exhibits a remarkably universal q-exponential form over nearly five orders of magnitude independently of the evolutionary position of the organism that reflects the universality of the small and intermediate scale DNA structure in all organisms. Differences in the distributions can be observed at large internucleotide intervals, where a second q-exponential *with the same q value* is added for eukaryotes only and thus appears to be a specific marker of the universal large-scale eukaryotic DNA structure. We suggest that this striking universality in the statistical laws governing the DNA primary sequences is a footprint of the DNA tertiary architecture that has undergone similar evolutionary structural optimization due to the similarities in the cell stress pattern perceived by various eukaryotic species from the living environment. We have also shown that the persistence in the arrangement of the intervals is consistent with the hierarchy of the DNA structural organization and reflects the heterogeneity of the optimization patterns at different scales. Moreover, since the direct analysis of the interval sequences is capable of tracking complex models like superposition of LRC and random noise where the interpretation of the DNA walk analysis is complicated a better understanding of the LRC effects in the DNA can be achieved. Another advantage of the interval approach is that it can be easily extended to the analysis of intervals between di- or trinucleotides as well as various combinations of nucleotides.

Finally, we like to emphasize striking similarities between interval distributions in the DNA sequences and in the financial markets, that might indicate similar optimization principles in these very different complex systems.

Acknowledgments

We thank Jan W. Kantelhardt and Josef Ludescher for very valuable discussions.

Author Contributions

Conceived and designed the experiments: MB AK AB. Performed the experiments: MB AK AB. Analyzed the data: MB AK AB. Contributed reagents/materials/analysis tools: MB AK AB. Wrote the paper: MB AK AB.

References

1. **Watson J, Baker TA, Bell SP** (2014) Molecular Biology of the Gene. NY: Benjamin-Cummings Publishing Company, 2014.
2. **Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton-Carafac Y et al.** (2011) Multi-scale coding of genomic information: From DNA sequence to genome structure and function. Physics Reports 498: 45–188.

3. **Guerra CF, Bickelhaupt FM, Snijders JG, Baerends EJ** (2000) Hydrogen Bonding in DNA Base Pairs: Reconciliation of Theory and Experiment. J Am Chem Soc 122: 4117–4128.
4. **Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F et al.** (1992) Long-range correlations in nucleotide sequences. Nature 356: 168–170.
5. **Li W, Kaneko K** (1992) Long-Range Correlation and Partial 1/*f* Spectrum in a Noncoding DNA Sequence. Europhys Lett 17: 655–660.
6. **Chatzidimitriou-Dreismann CA, Larhammar D** (1993) Long-range correlations in DNA. Nature 361: 212–213.
7. **Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE et al.** (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49: 1685–1689.
8. **Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME et al.** (1995) Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys Rev E 51: 5084–5091.
9. **Arneodo A, Bacry E, Graves PV, Muzy JF** (1995) Characterizing Long-Range Correlations in DNA Sequences from Wavelet Analysis. Phys Rev Lett 74: 3293–3296.
10. **Grossberg A, Rabin Y, Havlin S, Neer A** (1993) Crumpled globule model of the three-dimensional structure of DNA. Europhys Lett 23: 373–378.
11. Goodsell DS, Dickerson RE. Bending and curvature calculations in B-DNA. Nucleic Acids Res 22: 54975503.
12. **Audit B, Thermes C, Vaillant C, d′Aubenton-Carafa Y, Muzy JF et al.** (2001) Long-Range Correlations in Genomic DNA: A Signature of the Nucleosomal Structure. Phys Rev Lett 86: 2471–2474.
13. **Audit B, Vaillant C, Arneodo A, d′Aubenton-Carafa Y, Thermes C** (2002) Long-range Correlations between DNA Bending Sites: Relation to the Structure and Dynamics of Nucleosomes. J Mol Biol 316: 903–918.
14. **Vaillant C, Audit B, Thermes C, Arneodo A** (2003) Influence of the sequence on elastic properties of long DNA chains. Phys Rev E 67: 032901.
15. **Vaillant C, Audit B, Arneodo A** (2005) Thermodynamics of DNA Loops with Long-Range Correlated Structural Disorder. Phys Rev Lett 95: 068101.
16. **Bunde A, Eichner JF, Kantelhardt JW, Havlin S** (2005) Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. Phys Rev Lett 94: 048701.
17. **Altmann EG, Kantz H** (2005) Recurrence time analysis, long-term correlations, and extreme events. Phys Rev E 71: 056106.
18. **Bogachev MI, Eichner JF, Bunde A** (2008) On the Occurence of Extreme Events in Long-term Correlated and Multifractal Data Sets. Pure Appl Geophys 165: 1195–1207.
19. **Bogachev MI, Eichner JF, Bunde A** (2007) Effect of nonlinear correlations on the statistics of return intervals in multifractal records. Phys Rev Lett 99: 240601.
20. **Bogachev MI, Kireenkov IS, Nifontov EM, Bunde A** (2009) Statistics of return intervals between long heartbeat intervals and their usability for online prediction of disorders. New J Phys 11: 063036.
21. **Ludescher J, Tsallis C, Bunde A** (2011) Universal behaviour of interoccurrence times between losses in financial markets: An analytical description. EPL 95: 68002.
22. **Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ et al.** (2013) GenBank. Nucleic Acids Res 41: D36–D42.
23. **Smit AF** (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9: 657–663.
24. **Jurka J** (2000) Repbase update: A database and an electronic journal of repetitive elements. Trends Genet 16: 418–420.
25. **Kantelhardt JW, Koscielny-Bunde E, Rego HHA, Havlin S, Bunde A** (2001) Detecting long-range correlations with detrended fluctuation analysis. Physica A 295: 441–454.
26. **Tsallis C** (1988) Possible generalization of Boltzmann-Gibbs statistics. J Stat Phys 52: 479–487.

27. **Felsenfeld G, Groudine M** (2003) Controlling the double helix. Nature 421: 448–453.

28. **Lennartz S, Bunde A** (2009) Eliminating finite-size effects and detecting the amount of white noise in short records with long-term memory. Phys Rev E 79: 066101.

29. **Toro E, Shapiro L** (2010) Bacterial Chromosome Organization and Segregation. CSH Perspectives in Biology 2: 000349.

30. **Pandita TK, Hittelman WN** (1995) Evidence of a chromatin basis for increased mutagen sensitivity associated with multiple primary malignancies of the head and neck. Int J Cancer 61: 738–743.

31. **Pandita TK, Richardson C** (2009) Chromatin remodeling finds its place in the DNA double-strand break response. Nucleic Acids Res 37: 1363–1377.

32. **Kong MG, Kroesen G, Morfill G, Nosenko T, Shimizu T et al.** (2009) Plasma medicine: an introductory review. New J Phys 11: 115012.

33. **Hunt CR, Ramnarain D, Horikoshi N, Iyengar P, Pandita RK et al.** (2013) Histone Modifications and DNA Double-Strand Break Repair after Exposure to Ionizing Radiations. Radiation Research 179: 383–392.

34. **Malacarne LC, Mendes RS, Lenzi EK** (2001) q-exponential distribution in urban agglomeration. Phys Rev E 65: 017106.

35. **Tsallis C** (2009) Introduction to nonextensive statistical mechanics: Approaching a complex world. NY: Springer,2009.

36. **Andrade JS, da Silva GFT, Moreira AA, Nobre FD, Curado EMF** (2010) Thermostatistics of Overdamped Motion of Interacting Particles. Phys Rev Lett 105: 260601.

37. **Nobre FD, Rego-Monteiro MA, Tsallis C** (2011) Nonlinear relativistic and quantum equations with a common type of solution. Phys Rev Lett 106: 140601.

38. **Peterson J, Dixit PD, Dill KA** (2013) A maximum entropy framework for nonexponential distributions. Proc Nat Acad Sci U S A 110: 20380–20385.

39. **Antonopoulos CG, Michas G, Vallianatos F, Bountis T** (2014) Evidence of q-exponential statistics in Greek seismicity. Physica A 409: 71–77.

40. **Ludescher J, Bunde A** (2014) Universal behavior of the interoccurrence times between losses in financial markets: Independence of the time resolution. *Preprint*.