SCIENTIFIC REPORTS

OPEN

Superstatistical model of bacterial DNA architecture

Mikhail I. Bogachev[1,2], Oleg A. Markelov[1], Airat R. Kayumov[2] & Armin Bunde[3]

Received: 08 September 2016
Accepted: 18 January 2017
Published: 22 February 2017

Understanding the physical principles that govern the complex DNA structural organization as well as its mechanical and thermodynamical properties is essential for the advancement in both life sciences and genetic engineering. Recently we have discovered that the complex DNA organization is explicitly reflected in the arrangement of nucleotides depicted by the universal power law tailed internucleotide interval distribution that is valid for complete genomes of various prokaryotic and eukaryotic organisms. Here we suggest a superstatistical model that represents a long DNA molecule by a series of consecutive ~150 bp DNA segments with the alternation of the local nucleotide composition between segments exhibiting long-range correlations. We show that the superstatistical model and the corresponding DNA generation algorithm explicitly reproduce the laws governing the empirical nucleotide arrangement properties of the DNA sequences for various global GC contents and optimal living temperatures. Finally, we discuss the relevance of our model in terms of the DNA mechanical properties. As an outlook, we focus on finding the DNA sequences that encode a given protein while simultaneously reproducing the nucleotide arrangement laws observed from empirical genomes, that may be of interest in the optimization of genetic engineering of long DNA molecules.

The ability of DNA to adopt different structural conformations is a key instrument of many biological processes such as interaction with proteins, replication and transcription. Therefore a better understanding of the DNA mechanical and thermodynamical properties and their impact on its conformational abilities is essential to reveal many regulatory mechanisms at molecular scale and their reflection in the biological systems performance at macroscopic scale. Considerable progress in both theoretical and experimental biophysics in recent decades led to the design and experimental verification of sophisticated mathematical models capable of describing various DNA structural conformations and their physical properties[1,2].

The DNA consists of two complementary polynucleotide chains which form a double helix with a helical pitch of about 10–11 base pairs (bp) that is universal for all kingdoms of life[3]. 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[4]. For an extensive review on the DNA structural organization, we refer to[5].

Due to its extremely compact packaging, DNA is very efficient as a carrier of genetic information, that is represented by the sequence of nucleotides in its primary structure. Recent success in genetic engineering resulted in the record-breaking amount of information that can be written on synthetic DNA (up to ~200 Mb of data has been reported as of 2016), that is comparable with the entire human genome size, with synthetic DNA patches of thousands to dozens of thousands of base pairs (bp) being already a laboratory routine. One of the key challenges in genetic engineering of long DNA segments is the reproduction of its physical properties such that the synthetic molecule exhibits similar conformational abilities like the DNA in the living cell. The most straightforward approach is to follow the laws that govern the architecture of the host/model organisms and organize the synthetic DNA in a similar way, that in turn requires a better understanding of these laws.

While as a carrier of genetic information the DNA is often treated simply as a sequence of consecutively arranged nucleotides, in fact its molecular structure is much more complex including multiple packaging levels

[1]Biomedical Engineering Research Centre, St. Petersburg Electrotechnical University, St. Petersburg, 197376, Russia. [2]Molecular Genetics of Microorganisms Lab, Institute of Fundamental Medicine and Biology, Kazan (Volga Region) Federal University, Kazan, Tatarstan, 420008, Russia. [3]Institut für Theoretische Physik, Justus-Liebig-Universität Giessen, 35392 Giessen, Germany. Correspondence and requests for materials should be addressed to M.I.B. (email: Mikhail.Bogachev@physik.uni-giessen.de)

and exhibiting largely heterogeneous properties. From the pure information theory point of view the DNA content is drastically redundant. Even in *Bacteria* where about 97% of DNA encode functional proteins, 20 possible amino acids are being translated from 64 possible trinucleotides that, providing the necessity for the stop codons, already indicates approximately three-fold redundancy. In higher eukaryotes, only about 3% of the DNA encode functional proteins, while up to one half of the remaining 97% non-coding DNA consists of repetitive DNA sequences. The discovery of pronounced long-range correlations (LRC) in DNA sequences in the early 90 s[6–10] only added to this redundancy. LRC in the primary DNA sequence is a signature of clustering of certain nucleotides or nucleotide complexes. In recent years these clustering effects has been utilized in a number of alignment-free bioinformatic approaches via the analysis of the probability distributions of different oligonucleotides in genomes or oligopeptides in corresponding proteomes[11–19].

Clustering of nucleotides and nucleotide complexes have been attributed to the structural complexity of DNA represented by the famous fractal globule model[20] that has been recently supported by sequence-based reconstructions of three-dimensional whole-genome architecture[21–25].

Contrast to the multi-level hierarchical eukaryotic DNA architecture, there is a single regular structural level in bacterial DNA, the double helix[3]. At larger scales, it is localized in a relatively free manner in the cytoplasm, with random attachments to the cell membrane and packed in supercoils of varying size. In terms of its mechanical properties, double stranded DNA has a single specific structural scale characterized by its persistence length of about 50 nm, that corresponds to approximately 150 bp[26–28]. The DNA segments displaced by more than persistence length can be considered as no longer mechanically coupled in terms of their tangential directions, and thus structural conformations that are separated by more than 150 bp at a first glance could be treated as independent.

In a recent publication[29] we have reported a non-trivial universal power law tailed distribution and power law correlations in the internucleotide interval sequences from 130 complete genomes of various organisms from *Archaea* and *Bacteria* to *H. Sapiens* that are valid over many orders of magnitude limited only by the respective genome/chromosome size. This universality is somewhat striking, since it holds for both coding and non-coding, repetitive and non-repetitive DNA segments, and exhibits only moderate variations depending on the GC content of the genomes and on the optimal living temperature of the studied organisms, and thus could be likely attributed to the universal structural and/or conformational properties of DNA that hold independently of the carried genetic information.

The observed universal power law tailed distribution is also known as the q-exponential distribution, a subclass of generalized Pareto distributions. In the literature, this class of distributions has been also associated with the maximization of generalized entropy[30] that had been originally introduced in the early 1960 s by Renyi[31] and found numerous applications in the analysis of fluctuations in various dynamical systems (see, e.g. ref. 32 and references therein; for the criticism of the concept, see ref. 33 and references therein). In the limit $q \to 1$ the q-exponential distribution reduces to a simple exponential. In recent years, the same functional form of the distribution have been numerously observed in the dynamical and structural characteristics of several very different complex systems (see, e.g. refs 29, 35–38 and references therein).

One of the prominent concepts that leads to the observed class of distributions is superstatistics. The superstatistical concept in its original version considers a macroscopic system that consists of microscopic cells exhibiting fluctuations of an intensive quantity, usually denoted as β, such as the inverse temperature or dissipation energy[39,40]. A local equilibrium is supposed in each of these microscopic cells, while it is achieved at very different β values. In this setting, according to the law of total probability, the macroscopic energy distribution is then given by

$$P(E) = \int_0^\infty P(\beta)P(E|\beta)\mathrm{d}\beta = \int_0^\infty P(\beta)\frac{1}{Z(\beta)}e^{-\beta E}\mathrm{d}\beta, \tag{1}$$

where $P(\beta)$ is the distribution of β over all local cells in the macroscopic system and $Z(\beta)$ is a normalization factor for $e^{-\beta E}$ at specified β[41]. Under a very general assumption that β is additively driven by multiple random factors and each factor is approximately Gaussian distributed, the macroscopic system is described by q-exponential distributions[40,42].

In this paper, we take the advantage of the superstatistical approach to suggest a possible explanation of the non-trivial universality in the internucleotide interval distributions. We show explicitly that both the superstatistical model and the corresponding DNA generation algorithm accurately reproduce the laws governing the empirical nucleotide arrangement properties of the bacterial DNA sequences for various GC contents and optimal living temperatures. We also discuss the relevance of our model in terms of DNA mechanical properties. Finally, as an outlook, we focus on finding DNA sequences that encode a given protein while simultaneously reproducing the nucleotide arrangement laws observed from empirical genomes, that may be of interest for the optimization of reverse translation algorithms in the genetic engineering of long DNA molecules.

Results and Discussion

Fluctuation analysis. We focus on the same datasets as previously in ref. 29 and start with bacterial DNA that exhibits the simplest structural organization. We split 72 complete bacterial genomes into four groups according to their global GC content determined as the fraction of strongly bonded base pairs (G:C) in the studied genome. For each DNA sequence, we next perform the fluctuation analysis using the widespread detrended fluctuation analysis (DFA) method (for more details on the DFA, we refer to the Methods section at the end of this paper). Since we are mostly interested in the quantities that determine the mechanical properties of DNA and its structural organization, we exchange the nucleotide base pairs by their bond enthalpies, −11.8 for A:T and −23.8 kcal/mol for G:C, respectively[4].

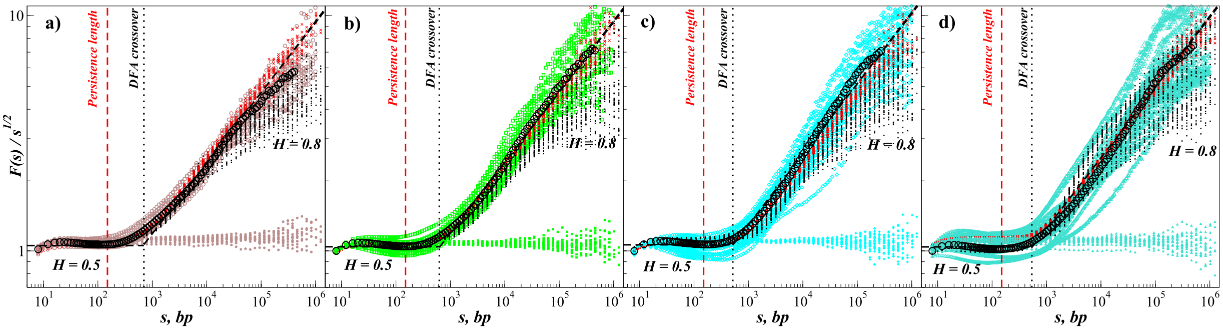

Figure 1. DFA2 fluctuation functions $F(s)$ for the sequence of base pair bond enthalpies for bacterial genomes with (**a**) very low, (**b**) low, (**c**) intermediate and (**d**) high GC content divided by $\sqrt{s}$ such that a horizontal plateau corresponds to the absence of correlations. Black circles indicate average $F(s)$ within each group, while dashed lines show the approximate model with effectively vanishing correlations ($H = 0.5$) below and pronounced long-range correlations ($H = 0.8$) above the crossover. Small coloured bubbles show $F(s)$ for randomly shuffled DNA sequences, while black bubbles show $F(s)$ for the reconstructed DNA after randomized back-and-forth translation. Corresponding model approximations are given by red $\star$. Vertical dotted lines show the position of the crossover, while the red dashed lines indicate the DNA persistence length.

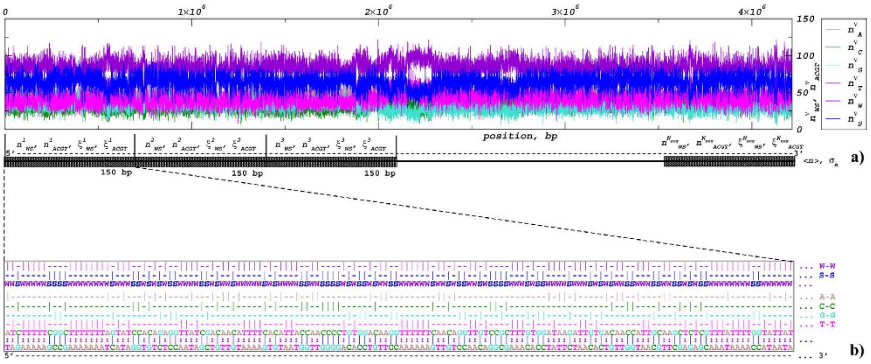

Figure 2. A schematic representation of the superstatistical model and the internucleotide interval sequences assessment procedure exemplified for the *Bacillus subtillis* genomic DNA. (**a**) The DNA sequence is being split into N_{seg} 150 bp non-overlapping segments ν characterized by the local numbers $n^{\nu}_{W,S}$ and $n^{\nu}_{A,C,G,T}$ (or their relative fractions $\xi^{\nu}_{W,S}$ and $\xi^{\nu}_{A,C,G,T}$, respectively) of weakly 'W' or strongly 'S' bonded base pairs as well as individual nucleotides 'A', 'C', 'G' and 'T', where ν runs from 1 to N_{seg}. The upper plot shows the variations of $n^{\nu}_{W,S}$ and $n^{\nu}_{A,C,G,T}$ on a genome-size scale. (**b**) An example of 150 bp DNA segment taken from the 5′ end of the *B.subtillis* chromosome is shown in the lower part of the panel, with weakly bonded base pairs connected by dots, while strongly bonded base pairs connected by full vertical lines. Extraction of intervals l between consecutive positions of similar nucleotides (A-A, C-C, G-G and T-T) is exemplified for the upstream (5′ to 3′) nucleotide chain. Consecutive occurrences of similar nucleotides (e.g., 'AA') is considered as a single internucleotide interval $l = 1$, one nucleotide between similar nucleotides (e.g., 'ABA') leads to double interval, where 'B' is any nucleotide except 'A', and so on. Intervals between consecutive positions of weakly 'W' or strongly 'S' bonded base pairs are obtained in a similar way from an auxiliary sequence, where either 'A' or 'T' are exchanged by 'W', and either 'G' or 'C' are exchanged by 'S'.

Our results depicted in Fig. 1 indicate that in all cases $F(s)$ contains two characteristic regimes, with effectively vanishing correlations below ($H = 0.5$) and pronounced long-range correlations above ($H = 0.8$) the crossover, that is also consistent with the results obtained in earlier studies for other DNA numerical representations[43,44]. Taking into account that the second-order DFA used in this study typically shows a characteristic crossover at about 3 … 5 persistence lengths $s_\times$ (see, e.g. ref. 45), the actual position of $s_\times$ at ~150 bp is consistent with the current consensus on the DNA persistence length under physiological conditions widely used in biophysical models (see, e.g. refs [26–28]). The figure also shows that, as expected, after random shuffling of the nucleotides the correlations vanish at all scales.

Superstatistical model. Following the above observations, we suggest a simple representation of DNA by a superstatistical model, that is schematically illustrated in Fig. 2. First, we split the DNA sequence into $N_{seg} = [N/150[$ non-overlapping 150 bp segments $\nu = 1, \ldots, N_{seg}$, where N is the genome size and] …] is the integer operator. We focus on the long-range inter-segment variations of the numbers $n^{\nu}_{A,C,G,T}$ and $n^{\nu}_{W,S}$ of the respective nucleotides or base pairs per each segment in the analyzed genomes (see Fig. 2a). We also determine the intervals between consecutive positions of similar nucleotides as well as strongly- or weakly bonded base pairs (see Fig. 2b). Due to the effective absence of correlations at short scales (see Fig. 1), we assume that the nucleotides are arranged randomly within each segment, and thus the internucleotide intervals l are distributed exponentially $P(l) = 1/\langle l\rangle \exp(-l/\langle l\rangle)$, where $\langle l\rangle$ is the local average interval for each given segment. However, since the numbers of respective nucleotides or base pairs per segment n exhibit pronounced segment-to-segment variations along the DNA sequence, this also leads to the variations of the local average interval $\langle l\rangle$ that is inversely proportional to n. In this model, for known $P(n)$, the marginal distribution is given by

$$P(l) = \int_0^\infty P(\beta)\frac{1}{Z(\beta)}e^{-\beta l}\mathrm{d}\beta = \int_0^\infty P\left(\frac{n}{150}\right)\frac{n}{150}e^{-nl/150}\mathrm{d}n, \tag{2}$$

where $\beta = n/150$ is the fraction of specific nucleotides or base pairs in each of the local fragments that plays the role of the local intensity parameter and is bounded between 0 and 1.

Next we focus on the shape of the distributions $P(n)$. Figure 3 exemplifies the distributions $P(n)$ for 12 representative genomes of well-known free living bacteria with different GC contents which are also widely used in molecular biology and genetic engineering. The figure shows that the shapes are very close to Gaussian family distributions, and thus are fully determined by their averages $\langle n\rangle$ and standard deviations σ_n. While a very close approximation could be provided by a Gaussian distribution, it is not physically supported, since it allows for negative values with non-zero probability. Besides a trivial solution such as an effective description by Gaussian truncated at $0 \leq n \leq 150$, similar asymptotic behaviour can be observed in several distributions with a non-negative support such as binomial, Γ- or χ^2-distributions. For large average values $\langle n\rangle$ the discrepancies between these distributions and Gaussian distributions with corresponding averages $\langle n\rangle$ and standard deviations σ_n are vanishing. The figure also shows that random shuffling of the nucleotides in the studied DNA sequences leads to considerable narrowing of the distributions, while their shapes remain close to Gaussian.

Figure 4a,b shows the parameters $\langle n\rangle$ and σ_n for the entire set of the studied genomes. Remarkably, σ_n does not depend on the average per segment number over the entire genome $\langle n\rangle$ for both strongly and weakly bonded base pairs (see Fig. 4a), in contrast to the simple assumptions like the binomial or χ^2-distributions. It is also interesting that the purine-pyrimidine alternation exhibits different properties for strongly and weakly bonded nucleotides (see Fig. 4b). While for G and C σ_n does not change significantly with the changes in their averages per segment $\langle n\rangle$, for A and T σ_n increases with increasing their averages per segment $\langle n\rangle$, like in the χ^2 model. Figure 4c,d shows the coefficients of variation $\rho = \sigma_n/\langle n\rangle$ as functions of the relative fraction $\xi = \langle n\rangle/150$ of given base pairs and individual nucleotides, respectively. The figure shows that in both cases for strongly bonded base pairs and corresponding nucleotides there is approximately algebraical decay $\rho \propto 1/\xi$. The figure also shows that random shuffling of the DNA sequence leads, as expected, to the binomially distributed numbers of strongly- and weakly bonded base pairs as well as individual nucleotides in 150 bp segments.

DNA simulation algorithm. Next we suggest an simulation algorithm which generates a DNA sequence with random genetic code where the local number of nucleotides within 150 bp segments exhibits similar distributions and long-range correlation properties, like in the empirical bacterial genomes.

In the first step, we generate a long-range correlated dataset x_S^ν with $H = 0.8$ with zero mean and unit variance consisting of N_{seg} numbers. To determine the number of strongly bonded base pairs n_S^ν in each segment ν, we next multiply x_S^ν by the standard deviation $\sigma_n = 9$ that corresponds to the average standard deviation observed from empirical genomes and add the mean $\langle n_s\rangle$ equal to the average number of strongly bonded base pairs per segment of the simulated genome.

In the second step, a simulated DNA sequence is created consisting of N_{seg} consecutive segments, with ν^{th} segment containing $n_S^\nu = x_S^\nu * \sigma_n + < n_S >$ strongly bonded base pairs 'S' that are randomly allocated within the segment. The remaining $n_W^\nu = 150 - n_S^\nu$ positions are filled with weakly bonded base pairs 'W'.

In the third step, the positions of strongly- and weakly bonded base pairs are filled by either purines (A or G) or pyrimidines (C or T) in the primary polypeptide chain by exchanging 'W' by either 'A' or 'T', and 'S' by either 'G' or 'C'. For that, two other long-range correlated dataset with $H = 0.8$ consisting of N_{seg} numbers each that are independent of the first dataset are created. One of these datasets gives the probabilities p_A^ν that a purine 'A' replaces 'W' for each segment ν. Accordingly, the number of 'A' in segment ν is given by $n_A^\nu = p_A^\nu n_W^\nu$, and the number of 'T' is given by $n_A^\nu = p_T^\nu n_W^\nu = (1 - p_A^\nu)n_W^\nu$. We found that the mean value $\langle p_A\rangle = 0.5$ and standard deviation $\sigma_{p_A} = 0.06$ leads to the best agreement with the empirical data, independently of both local and global GC contents.

In contrast, for strongly bonded base pairs, 'S' is exchanged by purine 'G' with probabilities taken from the third long-range correlated dataset with mean $< p_G^\nu > = 0.5$, but now with standard deviation $\sigma_{p_G} = \rho/k$, where $\rho = \sigma_n/\langle n\rangle$ is the coefficient of variation of the average number of strongly bonded base pairs per segment that is adjusted according to the global GC content of the studied genome, and $k = 1.75$ is the empirically adjusted correction coefficient. After that the remaining 'S' are exchanged by 'C'. In all cases, particular positions of either purines or pyrimidines within segments are chosen randomly, due to vanishing short-range correlations.

We have simulated 72 datasets each corresponding to a single bacterial genome in terms of its size and global GC content. Figure 4 indicates that the simulated DNA follows the regression lines that indicate the representative σ_n and ρ for the given $\langle n\rangle$ and ξ, respectively. Indeed, a more specific adjustment to the particular host/model organism genome properties is possible by taking its specific σ_n instead of the universal $\sigma_n = 9$ and fitting a specific empirical coefficient k for the purine-pyrimidine alternation procedure from empirical genome data.

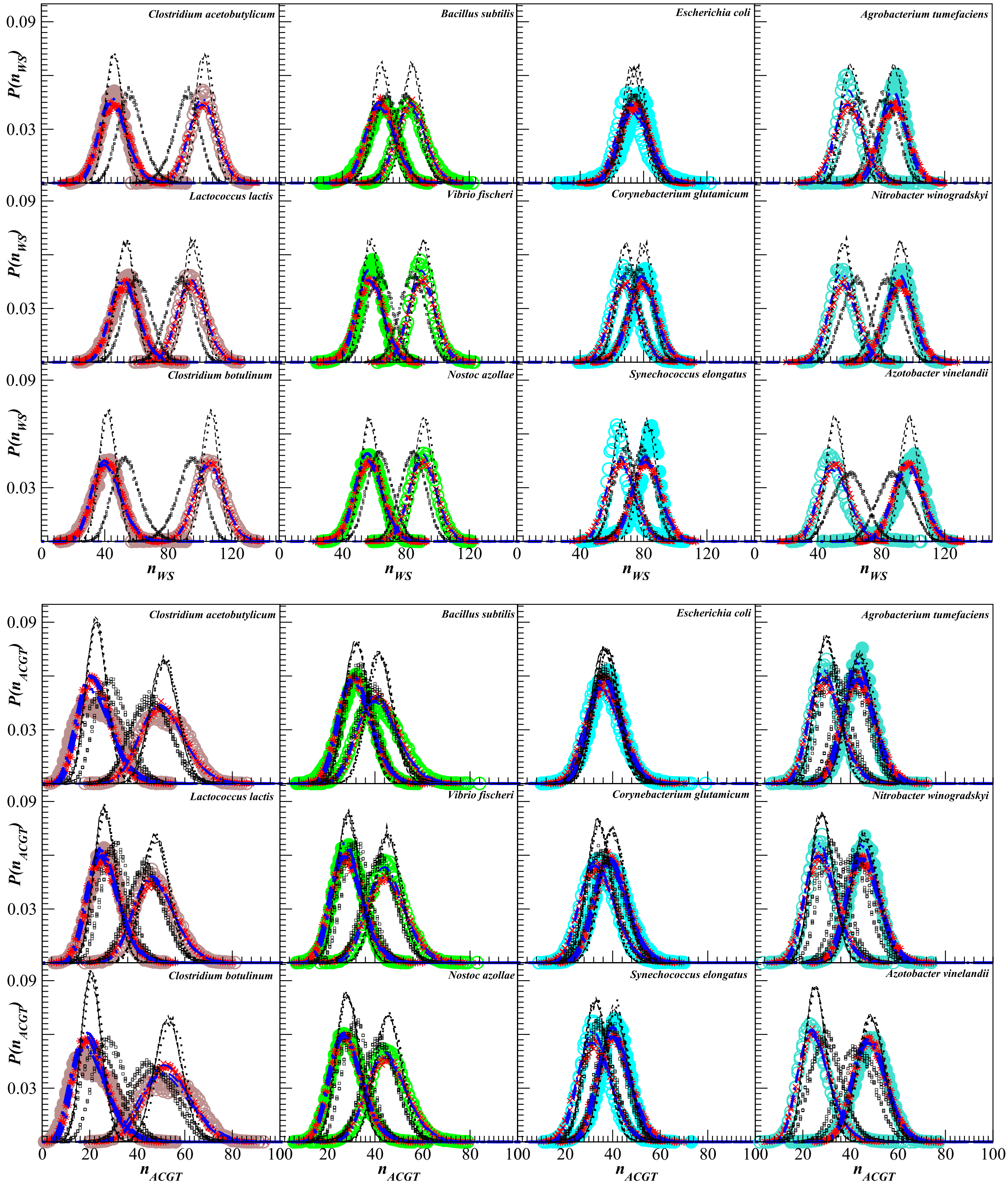

Figure 3. Distribution of the numbers of strongly (●) and weakly (○) bonded base pairs (upper panels) as well as individual nucleotides (lower panels) in the local 150 bp DNA segments for genomes of 12 representative organisms, 3 from each of the very low, low, intermediate and high GC conteof the local numbersnt groups (with GC content increasing from left to right). Corresponding model approximations are given by red ⋆ for strongly and by red × for weakly bonded base pairs. Blue dash-dotted lines show model approximations by Γ-distributions. Black dotted and dashed lines show the same distributions for the randomly shuffled DNA sequences and for the randomly generated nucleotide sequences with the same GC content like in the corresponding empirical sequences, respectively. Small black □ symbols show the same distributions for DNA reconstructed from the corresponding proteome after a randomized back-and-forth translation test.

Nevertheless, Fig. 3 shows that the distributions of the numbers of given nucleotides or base pairs in local 150 bp segments can be well reproduced by the superstatistical model and by the DNA simulation algorithm

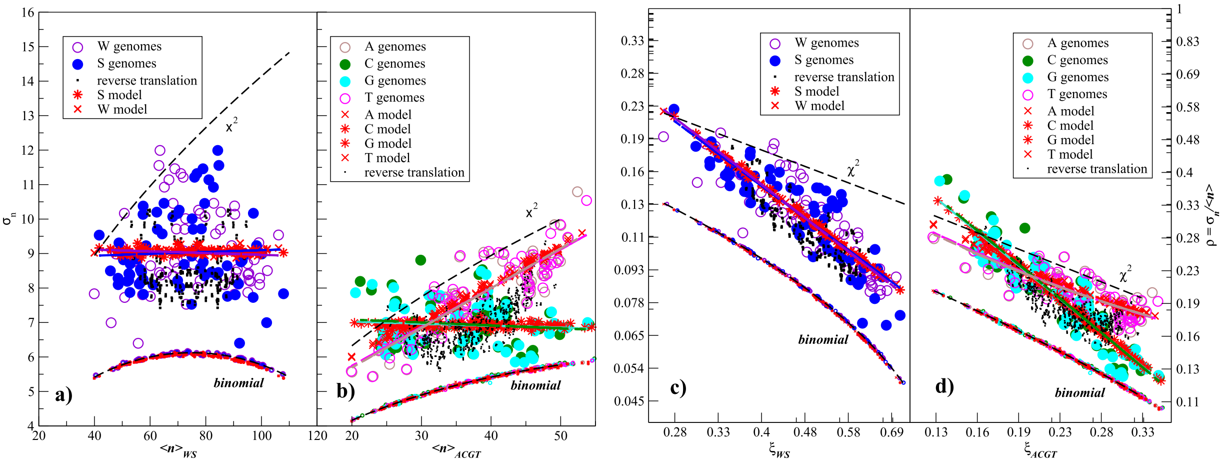

Figure 4. Averages $\langle n\rangle$ and standard deviations σ_n of the local numbers (**a**) of strongly ● or weakly ○ bonded base pairs as well as (**b**) of individual nucleotides A, C, G and T in local 150 bp DNA segments from 72 bacterial genomes. Small coloured symbols show the same quantities for the shuffled DNA sequences as well as for the randomly generated nucleotide sequences with the same GC content like in the empirical data, both following the binomial distributions. Linear regressions are given by full lines, while theoretical combinations of $\langle n\rangle$ and σ_n for the binomial and χ^2-distributions are given by dashed lines. Corresponding data for the simulated DNA sequences are shown by red * and red × for strongly and weakly bonded nucleotides, respectively. Small black □ symbols show the same data for the DNA sequences obtained by a reverse translation of the corresponding proteomes, as a result of the back-and-forth translation test. (**c**,**d**) The same data expressed as a relative measure, given by the coefficients of variation $\rho = \sigma_n/\langle n\rangle$ as a function of the relative fraction of the given base pairs or nucleotides ξ_{WS} or ξ_{ACGT}, respectively.

despite of using not exact, but typical σ_n values for the given GC content. To emphasize the reproducibility of our results for a larger set of bacterial genomes, more data are shown in the Supplementary information (available) as Figs S1 and S2 for very low, Figs S5 and S6 for low, Figs S9 and S10 for intermediate as well as Figs S13 and S14 for high overall GC content.

Internucleotide interval distributions.

Since neither binomial nor χ^2 distributions fit the combinations of $\langle n\rangle$ and σ_n observed in the empirical data, we next follow a more generalized model with a positive support that could represent $P(n)$, and thus also the relative quantity $P(\beta)$ for all considered cases, that is the Γ-distribution

$$P(\beta) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\beta^{\alpha-1}\exp(-\lambda\beta), \tag{3}$$

where $\beta = n/150$ plays the role of the local intensity parameter, $\Gamma(\alpha)$ is the Γ-function, α is the shape parameter and λ is the rate parameter. The average of the Γ-distribution is given by $\langle n\rangle = \alpha/\lambda$ and its variance equals $\sigma_n^2 = \alpha/\lambda^2$. Accordingly, the shape parameter can be expressed as $\alpha = 1/(\sigma_n/\langle n\rangle)^2 = 1/\rho^2$ such that it depends only on the coefficient of variation $\rho = \sigma_n/\langle n\rangle$, and the rate parameter $\lambda = \alpha/\langle n\rangle$ adjusts for the average number of a given nucleotide per 150 bp segment $\langle n\rangle$. Figure 3 shows that Γ-distributions with above parameters provide reasonable quality approximations despite of using not exact, but typical σ_n values for the given GC content.

For the entire DNA molecule, we have the superposition of multiple DNA segments with exponential internucleotide interval distributions characterized by local averages $\langle l\rangle$ that are inversely proportional to the local numbers of nucleotides n, that in turn are proportional to the local "intensity parameter" β, and thus the overall interval distribution for the entire genome can be determined in the framework of the above superstatistical concept. Following Eq. 1 with Γ-distributed $P(\beta)$ one easily obtains

$$P(l) \propto \int_0^{\infty} \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\beta^{\alpha-1}\exp(-\lambda\beta)\exp(-\beta l)\,\mathrm{d}\beta = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\frac{\Gamma(\alpha)}{(l+\lambda)^{\alpha}} = \left(\frac{\lambda}{l+\lambda}\right)^{\alpha} \tag{4}$$

that is asymptotically equivalent to the q-exponential distribution $P(l) \propto (1 + (q-1)(\alpha/\lambda)l)^{-1/(q-1)} = (1 + (q-1)\langle n\rangle l)^{-1/(q-1)}$, where $q = 1/(1+\alpha) = 1 + (\sigma_n/\langle n\rangle)^2 = 1 + \rho^2$ determines the slope of the asymptotic power law behaviour and $\langle n\rangle$ adjusts the position of the crossover, in agreement with our previous empirical findings[29]. Accordingly, the observational power law tailed distribution can be obtained directly from the law of total probability and also can be expressed in a simpler form than the q-exponential. Furthermore, it seems likely that the reason for the wide occurrence of q-exponentials in complex systems is simply the superposition of some locally independent patterns with different local concentrations, that can be explained by similar superstatistical models.

Figure 5 shows that the internucleotide interval distributions are also quite well reproduced by both the superstatistical model as well as by the corresponding distributions in the simulated DNA. The fitted distributions exhibit the same values of α and λ parameters as predicted by the above analytical treatment. In particular, for

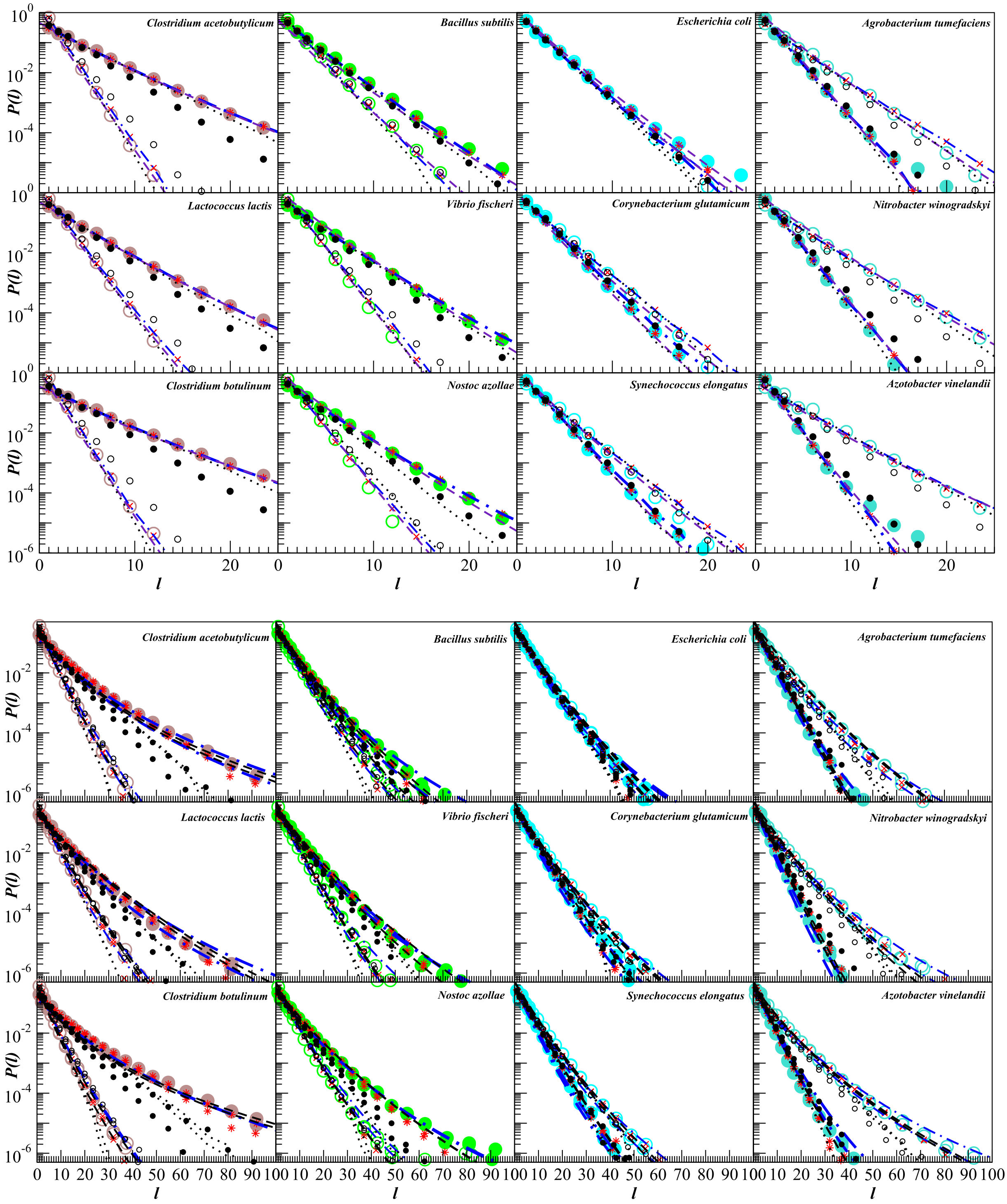

Figure 5. Distribution of the intervals between strongly (coloured ●) and weakly (coloured ○) bonded base pairs (upper panels) as well as between individual nucleotides (lower panels) for genomes of 12 representative organisms, 3 from each of the very low, low, intermediate and high GC content groups (with GC content increasing from left to right). Corresponding model approximations are given by red * for strongly and by red × for weakly bonded base pairs. Blue dash-dotted lines show results of numerical integration according to Eq. 2, while black dashed lines show corresponding approximations by q-exponential distributions. Dotted lines show the same distributions for the randomly shuffled DNA sequences. Small black ● symbols show the same distributions after the randomized back-and-forth translation test.

the distributions of intervals between strongly or weakly bonded base pairs, α decreases from $\alpha \approx 200$ to $\alpha \approx 20$ with increasing the coefficient of variation ρ from 0.07 to 0.23, that corresponds to decreasing ξ from $\xi \approx 0.7$ to $\xi \approx 0.3$ (see also outliers in Fig. 4). For the distributions of intervals between G and C nucleotides, α increases from $\alpha \approx 6$ to $\alpha \approx 50$ for ξ increasing from $\xi \approx 0.13$ to $\xi \approx 0.35$, while for the distributions of intervals between G and C nucleotides, α increases from $\alpha \approx 10$ to $\alpha \approx 30$ for ξ increasing from $\xi \approx 0.13$ to $\xi \approx 0.35$, due to their more narrow range of the coefficients of variation ρ (see also Fig. 4). The figure also shows that random shuffling of the DNA sequence, as expected, leads to the reduction of the internucleotide interval distributions to simple exponentials that can be observed as straight lines in the semi-logarithmic plots. To emphasize the reproducibility of our results for a larger set of bacterial genomes, more internucleotide interval distributions and their model predictions are shown in the Supplementary information (available) as Figs S3 and S4 for very low, Figs S7 and S8 for low, Figs S11 and S12 for intermediate as well as Figs S15 and S16 for high overall GC content.

Back-and-forth translation test. Next we simulated translations of each studied genome to obtain corresponding proteomes, and then a reverse translation to obtain a DNA sequence that encodes the given proteome. While the first procedure is straightforward and unambiguous, the second one is less trivial due to the inherent redundancy of the genetic code. We used a reverse translation procedure that randomly selects one of the triplets corresponding to a given amino acid residue. Taking into account that well above 90% of the bacterial genome contains coding DNA, there are only minor differences between the sizes of the original and the back-and-forth translated datasets, and thus corresponding finite size effects should be comparable. Despite that, as one can clearly observe from the above figures, there are significant discrepancies between the distributions of the numbers of given nucleotides in local 150 bp segments as well as between the distributions of the internucleotide intervals for the original and back-and-forth translated DNA. Deviations are especially pronounced when the GC content is significantly different from 50%. This indicates that the random choice of triplets corresponding to a given amino acid disrupts the variations of the local average numbers $\langle n \rangle$ and fractions ξ of the strongly- and weakly bonded base pairs as well as individual nucleotides and, especially, the long-range correlations in the genome, leading to the significant changes in the fluctuation functions (see Fig. 1) as well as to the corruption of the internucleotide interval distributions (see Fig. 5). Moreover, the randomized back-and-forth translation test leads to the distributions that are considerably further away from the empirical data than similar distributions for the randomly shuffled DNA sequences. This is easy to explain, since random shuffling preserves the total number and thus also the fractions of nucleotides, while back-and-forth translation does not. More sophisticated reverse translation algorithms are often adjusted to fit the global properties of the given host/model organism such as its GC content or the fraction of each codon in the genome, which only partially resolves the problem, since the long-range correlations are nevertheless corrupted.

The effects of optimal living temperature and extreme living conditions. In a recent study[29] we have observed moderate discrepancies between the internucleotide interval distributions in organisms with different optimal living temperatures that might be a sign of their adaptation to the external conditions. In terms of genome statistics, an easily observed sign of adaptation to non-typical thermodynamical constraints are specific laws governing the variation of both global and local GC contents. We next verify the validity of the suggested model for a cohort of extremophile organisms belonging to the *Archaea* kingdom. Remarkably, among three groups consisting of genomes from organisms with optimal living temperatures below 50 °C, between 50 °C and 80 °C and above 80 °C, the first and the last groups have comparable average GC contents[29].

Supplementary Figure S17 (available) shows the fluctuation functions *F(s)* for the studied archaeal genomes. The figure shows that there are significant variations of the crossover position both in the normal and high living temperature groups. While it has been shown that DNA persistence length decreases with increasing temperature[28], there might be additional effects arising from the adaptation to other extreme living factors such as high salinity and high pressure.

In the Supplementary information (available) similar distributions as in Fig. 3 for extremophile *Archaea* can be observed. In particular, Figs S18 and S19, Figs S22 and S23 as well as Figs S26 and S27 (available) show the corresponding distributions for the organisms with optimal living temperatures below 50 °C, between 50 °C and 80 °C and above 80 °C, respectively. The figures show that in several observed cases there are considerable deviations between the empirical and the model distributions, indicating that direct extrapolation of the laws governing the local arrangement of nucleotides in free living *Bacteria* does not always work for other microbial organisms such as extremophile *Archaea*. Distribution asymmetry, skewness and other shape modifications such as flattening of the distribution indicated by a plateau in its central part can be observed in several examples from different groups of extremophile genomes.

Supplementary Figs S20 and S21, Figs S24 and S25 as well as Figs S28 and S29 (available) show the intervals between strongly and weakly bonded base pairs as well as between individual nucleotides in the genomes of extremophile *Archaea* with optimal living temperatures below 50 °C, between 50 °C and 80 °C and above 80 °C, respectively. The figures show that, while the results of numerical integration according to Eq. 2 leading to the power law tailed approximations according to Eq. 4 agree with the empirical internucleotide interval distributions to a certain extent, the DNA simulation algorithm in several cases fails to reproduce the asymptotic behaviour of the distributions indicating that the nucleotide arrangement differs from that one observed in *Bacteria*, and only an effective phenomenological description is possible.

Hierarchical model, or superstatistics of superstatistics. Contrast to bacterial DNA, eukaryotic genomes exhibit complex multi-level hierarchical structural organization. In a recent publication[29], we reported that in the genomes of higher eukaryotes including humans internucleotide intervals exhibit a complex two-compound distribution that could be well approximated by two additive q-exponentials with $q \approx 1.11$, that is

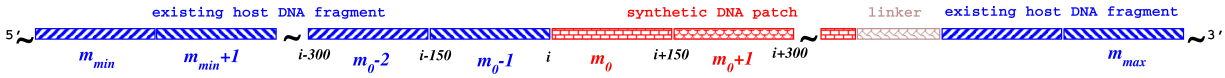

Figure 6. A schematic representation of the synthetic DNA patch insertion into existing host DNA. Existing part of host DNA upstream of the synthetic DNA patch is split into 150 bp segments $m_0 - 1$, $m_0 - 2$, …, m_{min} and is used to calculate the optimized local concentrations ξ_{WS} and ξ_{ACGT} in the synthetic DNA patch segments m_0, $m_0 + 1$ and so on, according to Eq. 5. Further downstream the synthetic patch is connected with the remaining part of the host DNA (with or without an optional linker) in the way that the local concentrations ξ_{WS} and ξ_{ACGT} in following DNA fragments m until m_{max} also follow Eq. 5.

asymptotically equivalent to the power law tailed distributions of form (4) with $\alpha \approx 9$. Supplementary Figure S30 (available) shows that such a distribution can be to a certain extent reproduced in a simulated DNA sequence that is organized as a hierarchical cascade. The first level, like in the bacterial DNA model, contains 150 bp segments, that for eukaryotes is also close to another important characteristic scale, the 146 bp cycle of DNA wrapping around a histone[5]. Next, 150 segments are organized in a single supersegment, and an additional long-range correlated variability parameter with the same $H = 0.8$ is added for these larger 150^2 bp supersegments, that are now of the common size of chromatin loops. Finally, a third-level variability parameter also with $H = 0.8$ is added for 150^3 bp supersupersegments that correspond to the typical size of isochores, large fragments of eukaryotic genomes that are characterized by stable GC content. By tuning the weights for these large-scale variability parameters, one can obtain a distribution that exhibits similar shape like in the empirical genomes of higher eukaryotes.

Indeed, the approximate reproduction of the distribution of internucleotide intervals cannot guarantee that the simulated DNA sequence would correspond to the same spatial structure as the empirical DNA exhibits. The eukaryotic DNA architecture has seven known characteristic scales, which should be implemented explicitly in a proper model setting to reproduce its spatial organization. However, we believe that similar principles based on superstatistical approach could be applied to characterize also the large-scale organization of eukaryotic DNA, and the internucleotide interval distributions could be useful as a testbed to verify the implementation of the respective nucleosome packaging and chromatin positioning models.

Predicting DNA mechanical properties. The suggested model could be also used for the statistical prediction of some mechanical properties of long DNA strands. It is well known that particular composition of nucleotides determines the local DNA curvature[46], bending energy and several other important mechanical properties. In particular, at short scales the local sub-elastic chain (LSEC)[47,48] and at long scales the worm-like chain (WLC) models[49,50] accurately predict the local bending energy, that is proportional to the linear and squared local cumulative bending angles for given DNA segments, respectively[51]. The suggested superstatistical model can be easily extended to describe the corresponding DNA curvature or bending energy distributions by considering the local cumulative bending angle Θ as the local intensity parameter β and to obtain the respective macroscopic curvature or energy distributions from Eq. 1. In the absence of particular nucleotide composition, representative statistics for a DNA fragment of given size and GC content could be obtained by using the representative model parameters that have been used above to simulate the DNA sequences.

Adahptation of synthetic DNA patches to the host organism architecture: An outlook towards potential implications for genetic engineering. In most practical scenarios, a synthetic genetic construction has to be inserted at a specific position i into the DNA of the host organism (see Fig. 6). The simplest case is when only one of the synthetic DNA strands carries a gene, e.g., in the direction from 5′ to 3′. Our goal is to suggest the synthetic insertion that would appear like a natural extrapolation of the preceding DNA of the host/model organism. For that, we first analyze the preceding host DNA by splitting it into 150 bp non-overlapping segments at positions $i - 150$, $i - 300$ and so on and calculating the local fractions of strongly and weakly based pairs ξ_{WS} as well as the local fractions of individual nucleotides $\xi_{A,C,G,T}$ in each segment. Next we use our suggested model to predict the local fractions of both strongly/weakly bonded based pairs as well as individual nucleotides by the optimal extrapolation of the same quantities from the preceding host DNA onwards to the synthetic fragment following ref. 52

$$\xi_m = \sum_{j=m_{min}}^{m_{max}} a_j \xi_{m-j}, \quad a_j = (H - 1/2)(j + 1)^{H-3/2}, \tag{5}$$

where m is the current segment number that changes between m_{min} and m_{max} spanning over both the preceding piece of host DNA and the synthetic DNA patch. The calculation starts from ξ_{m_0} that corresponds to the first segment m_0 in the synthetic DNA patch, while the data used for calculation considers also the fractions of strongly/weakly bonded base pairs in the preceding segments of the host DNA by taking into account also $j < m_0$. The prediction is continued iteratively for 150 bp segments of the synthetic DNA patch up to its completion. After the local fractions of both strongly/weakly based pairs as well as of individual nucleotides are obtained for each 150 bp segment, they can be used as target quantities for the reverse translation optimization. During the reverse translation procedure, the current fractions of strongly/weakly bonded base pairs are calculated from already back-translated codons. In each ambiguous case, when a single amino acid residue can be represented by several codons, one should always choose that codon that leads to the local fraction of strongly/weakly bonded base pairs characterized by the minimum square displacement from the prediction provided by the model for the current

150 bp segment. Next, the same procedure is repeated for the purine-pyrimidine alternation, separately for choosing between G and C in strongly and between A and T in weakly bonded base pairs, in each case taking the respective model parameters, this way resolving the remaining ambiguity.

Finally, similar optimization strategy can be employed to find the best ending position of the synthetic DNA patch including, if necessary, an optimal size and content of the linker between the synthetic patch and the forward piece of host DNA. If the synthetic DNA patch has to encode functional genes on each of its strands, an optimization problem with two equations for the directions from 5′ to 3′ and from 3′ to 5′ similar to Eq. 5 arises, and a best fit can be found iteratively.

Conclusion

To summarize, we have suggested a superstatistical model that is capable of reproducing the statistical laws that govern the arrangement of nucleotides in the primary sequence of bacterial DNA. The model considers the entire DNA molecule as a concatenation of non-overlapping segments, with each segment characterized by its own number (or fraction) of strongly/weakly bonded base pairs as well as individual nucleotides. The size of the segments is determined by the DNA persistence length that is approximately 150 bp under physiological conditions. We assume that within each segment the nucleotides are allocated randomly, corresponding to the local equilibrium scenario, while the numbers (or fractions) of different nucleotides in consecutive segments alternate in a long-range correlated manner with $H = 0.8$. Based on this model, we have also suggested a DNA simulation procedure that explicitly reproduces the typical distributions of internucleotide intervals for bacterial genomes with very different GC contents. We have also shown that the predictions by the model can to a certain extent reproduce similar properties in other microbial genomes, while some deviations can be observed in several extremophile *Archaea*, that might be attributed to their adaptation to the extreme living conditions such as high temperatures, pressure and salinity, that also affects the DNA persistence length. We also show that the suggested superstatistical model can be organized in a hierarchical cascade, that could lead to a reasonable reproduction of the internucleotide interval distributions in higher eukaryotic genomes, and thus might be also useful as a testbed for the respective nucleosome packaging and chromatin positioning models. Finally, as an outlook, we suggest how the proposed model could be utilized to facilitate the adaptation of the synthetic genetic constructions by choosing the most appropriate codons that would lead to the best reproduction of the empirical laws governing the nucleotide arrangement at particular positions in the genomes of the respective host/model organisms, that is an important issue for genetic engineering of long DNA patches.

Methods

Detrended fluctuation analysis.

The detrended fluctuation analysis (DFA) method was originally suggested by Peng *et al.* and generalized by Kantelhardt *et al.* and Hu *et al.*[8,45,53]. It is based on the analysis of the profile, or the cumulative sum $Y_j = \sum_{i=1}^{j} y_i$ of the raw series of numbers y_i. The profile is split into N_s non-overlapping segments of size s. In each segment k one determines the best polynomial fit $y_k(j)$ and obtains the variance $F_k^2(s) = (1/s)\sum_{j=1}^{s}(Y_{[(k-1)s+j]} - y_k(j))^2$ between the local trend and the profile in each segment k. Finally, one obtains the fluctuation function $F(s)$ by averaging over all segments

$$F(s) \equiv \left\{ \frac{1}{N_s} \sum_{k=1}^{N_s} F_k^2(s) \right\}^{1/2}. \tag{6}$$

For long-range correlated data $F(s)$ scales with s as $F(s) \sim s^H$, where H is the Hurst exponent directly related to the autocorrelation function $C(s)$ by $C(s) \sim s^{-\gamma}$, where $\gamma = 2 - 2H$. When the correlations are vanishing, $C(s) = 0$ for $s > 0$, and $H = 1/2$.

Fluctuation analysis deals with sequences of numbers, not nucleotides, and thus in general the results depend on the way how the studied DNA sequence has been replaced by a numeric sequence. Early studies focused mainly on the replacement of a given nucleotide (A, C, G or T) or their combination by one, while the others by zero[54], or on the "DNA walks" which increase by one when a pyrimidine (C or T) is observed and decrease by one when a purine (A or G) is observed in a DNA sequence[7]. Since we are mostly interested in the properties that determine the mechanical and thermodynamical properties of DNA, we exchange the nucleotide base pairs by their bond enthalpies, −11.8 for A:T and −23.8 kcal/mol for G:C, respectively[4].

Interval distributions.

Distributions of intervals between similar items in a series are known to explicitly reflect the persistence properties of data sequences. If the items of interest are allocated randomly, intervals between them follow a simple exponential distribution $P(l) = (1/\langle l \rangle)\exp(-l/\langle l \rangle)$, where $\langle l \rangle$ is the average interval, and are uncorrelated. In linearly long-range correlated data, one expects the asymptotic PDF of the intervals to follow a stretched exponential $\ln [P(l)] \propto -(l/\langle l \rangle)^{\gamma}$, with exponent $\gamma = 2 - 2H$[55,56], where H is the Hurst exponent characterizing the LRC. In the presence of nonlinear correlations, the PDF gets even broader and decays asymptotically by a power-law $P(l) \sim (l/\langle l \rangle)^{-\delta}$, where the exponent δ decreases when the LRC gets more pronounced[57,58]. Related analytical results on gap sizes and cluster sizes for widespread models including random walks, Levy flights as well as for systems exhibiting phase transitions have been obtained[59–61]. In recent years, interval distributions have been shown to efficiently reflect structural and dynamical features of complex systems in physics, biology, geoscience, climate, finance and many other applications[35–37,55–58,62]. Besides already mentioned application to the primary structure of DNA[29], very recently an approach to the prediction of structural, localization and functional properties of unknown proteins based on their explicit reflection in the distributions of intervals between similar amino acid residues has been suggested[63].

References

1. Bustamante, C., Bryant, Z. & Smith, S. B. Ten years of tension: single-molecule DNA mechanics. *Nature* **421(6921),** 423–427 (2003).
2. Bryant, Z., Oberstrass, F. C. & Basu, A. Recent developments in single-molecule DNA mechanics. *Curr. Opin. Str. Biol.* **22(3),** 304–312 (2012).
3. Watson, J., Baker, T. A. & Bell, S. P. *Molecular Biology of the Gene* (7th Edition). (NY, Benjamin-Cummings Publishing Company, 2014)
4. Guerra, C. F., Bickelhaupt, F. M., Snijders, J. G. & Baerends, E. J. Hydrogen Bonding in DNA Base Pairs: Reconciliation of Theory and Experiment. *J Am. Chem. Soc.* **122,** 4117–4128 (2000).
5. Arneodo, A. *et al.* Multi-scale coding of genomic information: From DNA sequence to genome structure and function. *Physics Reports* **498,** 45–188 (2011).
6. Li, W. & Kaneko, K. Long-Range Correlation and Partial $1/f^{\alpha}$ Spectrum in a Noncoding DNA Sequence. *Europhys. Lett.* **17,** 655–660 (1992).
7. Peng, C.-K., Buldyrev, S. V., Havlin, S. *et al.* Long-range correlations in nucleotide sequences. *Nature* **356,** 168–170 (1992).
8. Peng, C.-K., Buldyrev, S. V., Havlin, S. *et al.* Mosaic organization of DNA nucleotides. *Phys. Rev. E* **49,** 1685–1689 (1994).
9. Buldyrev, S. V. *et al.* Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. E* **51,** 5084–5091 (1995).
10. Arneodo, A., Bacry, E., Graves, P. V. & Muzy, J. F. Characterizing long-range correlations in DNA Sequences from wavelet analysis. *Phys. Rev. Lett.* **74,** 3293–3296 (1995).
11. Sandberg, R. *et al.* Capturing whole-genome characteristics in short wequences using a naive-bayessian classifier. *Genome Res.* **11,** 1404–1409 (2001).
12. Hao, B. & Ji, Q. Prokaryote phylogeny without sequence alignment: From avoidance signature to composition distance. *J. Bioinf. Comp. Biol.* **2,** 1–19 (2004).
13. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive bayessian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *J. Bioinf. Comp. Biol.* **2,** 1–19 (2004).
14. Liu, Z., DeSantis, T. Z., Andersen, G. L. & Knight, R. Accurate taxonomy assignments from 16S RNA sequences produced by highly parallel pyrosequences. *Nucl. Acids Res.* **36,** e120 (2008).
15. Chor, B., Horn, D., Goldman, N., Levy, Y. & Massingham, T. Genomic DNA k-mer spectra: models and modalities. *Genome Biol.* **10,** R108 (2009).
16. Kuksa, P. & Pavlovic, V. Efficient alignment-free DNA barcode analysis. *BMC Bioinformatics* **10,** 59 (2009).
17. DeSantis, T. Z. *et al.* Simrank: Rapid and sensitive general-purpose k-mer search tool. *BMC Ecology* **11,** 11 (2011).
18. LaRosa, M., Fiannaca, A., Rizzo, R. & Urso, A. Alignment-free analysis of barcode sequences by means of compression-based methods. *BMC Bioinformatics* **14,** S4 (2013).
19. LaRosa, M., Fiannaca, A., Rizzo, R. & Urso, A. Probabilistic approach modeling for the analysis and classification of genomic sequences. *BMC Bioinformatics* **16,** 52 (2015).
20. Grosberg, A., Rabin, Y., Havlin, S. & Neer, A. Crumpled globule model of the three-dimensional structure of DNA. *Europhys. Lett.* **23,** 373–378 (1993).
21. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
22. McNally, J. G. & Mazza, D. Fractal geometry in the nucleus. *EMBO J.* **29,** 2–3 (2010).
23. Mirny, L. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* **19,** 37–51 (2011).
24. Schram, R. D., Barkema, G. T. & Schiessel, H. On the stability of fractal globules. *J. Chem. Phys.* **138,** 224901 (2013).
25. Tamm, M. V., Nazarov, L. I., Gavrilov, A. A. & Chertovich, A. V. Anomalous diffusion in fractal globules. *Phys. Rev. Lett.* **114,** 178102 (2015).
26. Bednar, J. *et al.* Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. *J Mol. Biol.* **254,** 579–594 (1995).
27. Vologodskaia, M. & Vologodskii, A. Contribution of the intrinsic curvature to measured DNA persistence length. *J Mol. Biol.* **317,** 205–2013 (2002).
28. Geggier, S., Kotlyar, A. & Vologodskii, A. Temperature dependence of DNA persistence length. *Nucl. Acids Res.* **39(4),** 1419–1426 (2011).
29. Bogachev, M. I., Kayumov, A. R. & Bunde, A. Universal internucleotide statistics in full genomes: A footprint of the DNA structure and packaging? *PLoS One* **9,** e0112534 (2014).
30. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **52,** 479–487 (1988).
31. Renyi, A. On measures of entropy and information. *Proc. 4th Berkeley Symp. Math. Stat. Prob.* **1,** 547–561 (1961).
32. Grassberger, P. & Procaccia, I. Dimensions and entropies of strange attractors from a fluctuating dynamics approach. *Physica D* **13,** 34–54 (1984).
33. Nauenberg, M. Critique of q-entropy for thermal statistics. *Phys. Rev. E* **67,** 036114 (2003).
34. Presse, S. Nonadditive entropy maximization is inconsistent with Bayessian updating. *Phys. Rev. E* **90,** 052149 (2014).
35. Ludescher, J., Tsallis, C. & Bunde, A. Universal behaviour of interoccurrence times between losses in financial markets: An analytical description. *EPL* **95(6),** 68002 (2011).
36. Ludescher, J. & Bunde, A. Universal behavior of the interoccurrence times between losses in financial markets: Independence of the time resolution. *Phys. Rev. E* **90(6),** 062809 (2014).
37. Tsallis, C. Inter-occurrence times and universal laws in finance, earthquakes and genomes. *Chaos, Solitions and Fractals* **88,** 254–266 (2016).
38. Tamazian, A., Nguyen, V. D., Markelov, O. A. & Bogachev, M. I. Universal model for collective access patterns in the Internet traffic dynamics: A superstatistical approach. *EPL* **115,** 10008 (2016).
39. Beck, C. Dynamical foundations of nonextensive statistical mechanics. *Phys. Rev. Lett.* **87,** 180601 (2001).
40. Beck, C. & Cohen, E. G. D. Superstatistics. *Physica A* **322,** 267–275 (2003).
41. Naimark, O. B. Structural-scaling transitions and localized distortion modes in the DNA double helix. *Phys. Mesomech.* **10(1),** 33–45 (2007).
42. Touchette, H. & Beck, C. Asymptotics of superstatistics. *Phys. Rev. E* **71(1),** 016131 (2005).
43. Audit, B. *et al.* Long-Range Correlations in Genomic DNA: A Signature of the Nucleosomal Structure. *Phys. Rev. Lett.* **86,** 2471–2474 (2001).
44. Audit, B. *et al.* Long-range Correlations between DNA Bending Sites: Relation to the Structure and Dynamics of Nucleosomes. *J Mol. Biol.* **316,** 903–920 (2002).
45. Kantelhardt, J. W., Koscielny-Bunde, E., Rego, H. H. A., Havlin, S. & Bunde, A. Detecting long-range correlations with detrended fluctuation analysis. *Physica A* **295,** 441–454 (2001).
46. Goodsell, D. S. & Dickerson, R. E. Bending and curvature calculations in B-DNA. *Nucl. Acids. Res.* **22(24),** 5497 (1994).
47. Wiggins, P. A., *et al.* High flexibility of DNA on short length scales probed by atomic force microscopy. *Nature Nanotechnol.* **1(2),** 137–141 (2006).
48. Mazur A. K. & Maaloum, M. DNA flexibility on short length scales probed by atomic force microscopy. *Phys. Rev. Lett.* **112(6),** 068104 (2014).

49. Bresler, S. E. & Frenkel, Y. I. On the character of brownian motion of long organic chains and on the elastic properties of the rubber. *J Exp. Theor. Phys.* **9,** 1094–1106 (1939).
50. Kratky, O. & Porod, G. Röntgenuntersuchung gelöster Fadenmoleküle. *Rec. Trav. Chim. Pays-Bas*. **68,** 1106–1123 (1949).
51. Vologodskii, A. & Frank-Kamenetskii, M. D. Strong bending of the DNA double helix. *Nucl. Acids Res.* **41(14),** 6785–6792 (2013).
52. Mandelbrot, B. B. Gaussian *Self-Affinity and Fractals*. (NY, Springer, 2002)
53. Hu, K., Ivanov, P. C., Chen, Z., Carpena, P. & Stanley, H. E. Effect of trends on detrended fluctuation analysis. *Phys. Rev. E* **64,** 011114 (2001).
54. Voss, R. F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* **68,** 3805 (1992).
55. Bunde, A., Eichner, J. F., Kantelhardt, J. W. & Havlin, S. Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Phys. Rev. Lett.* **94,** 048701 (2005).
56. Altmann, E. G. & Kantz, H. Recurrence time analysis, long-term correlations, and extreme events. *Phys. Rev. E* **71,** 056106 (2005).
57. Bogachev, M. I., Eichner, J. F. & Bunde, A. Effect of nonlinear correlations on the statistics of return intervals in multifractal records. *Phys. Rev. Lett.* **99,** 240601 (2007).
58. Bogachev, M. I. & Bunde, A. On the predictability of extreme events in records with linear and nonlinear long-range memory: Efficiency and noise robustness. *Physica A* **390,** 2240 (2009).
59. Schehr, G. & Majumdar, S. N. Universal order statistics of random walks. *Phys. Rev. Lett.* **108,** 040601 (2012).
60. Majumdar, S. N., Mounaix, P. & Schehr, G. Exact statistics of the gap and time interval between the first two maxima of random walks and Lévy flights. *Phys. Rev. Lett.* **111,** 070601 (2013).
61. Bar, A., Majumdar, S. N., Schehr, G. & Mukamel, D. Exact extreme-value statistics at mixed-order transitions. *Phys. Rev. E* **93,** 052130 (2016).
62. Bogachev, M. I. & Bunde, A. Universality in the precipitation and river runoff. *EPL* **97,** 48011 (2012).
63. Bogachev, M. I., Kayumov, A. R., Markelov, O. A. & Bunde, A. Statistical prediction of protein structural, localization and functional properties by the analysis of its fragment mass distributions after proteolytic cleavage. *Sci. Rep*. **6,** 22286 (2016).

Acknowledgements

We like to acknowledge the support of this work by the Ministry of Education and Science of the Russian Federation within the state assignment of St. Petersburg Electrotechnical University, by the Russian Science Foundation (project No. 15-14-00046) as well as by the subsidy of the Russian Government to support the Program of competitive development of Kazan Federal University.

Author Contributions

M.B. and A.K. conceived the experiments, M.B. and O.M. conducted the experiment(s), M.B., A.K. and A.B. analyzed the results. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Bogachev, M. I. *et al.* Superstatistical model of bacterial DNA architecture. *Sci. Rep.* **7**, 43034; doi: 10.1038/srep43034 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SCIENTIFIC REPORTS

OPEN

Corrigendum: Superstatistical model of bacterial DNA architecture

Mikhail I. Bogachev, Oleg A. Markelov, Airat R. Kayumov & Armin Bunde

Scientific Reports 7:43034; doi: 10.1038/srep43034; published online 22 February 2017; updated 22 December 2017

This Article contains errors. An improper normalization factor was inadvertently applied, resulting in an incorrect form of Eq. (4). In the Introduction section,

"In this setting, according to the law of total probability, the macroscopic energy distribution is then given by

$$P(E)=\int_0^\infty P(\beta)P(E|\beta)\,\mathrm{d}\beta=\int_0^\infty P(\beta)\frac{1}{Z(\beta)}e^{-\beta E}\,\mathrm{d}\beta$$

where P(β) is the distribution of β over all local cells in the macroscopic system and Z(β) is a normalization factor for e$^{-\beta E}$ at specified β[41]."

should read:

"In this setting, according to the law of total probability, the macroscopic energy distribution is then given by

$$P(E)=\int_0^\infty W(\beta)P(E|\beta)\,\mathrm{d}\beta=\int_0^\infty W(\beta)\frac{1}{Z(\beta)}e^{-\beta E}\,\mathrm{d}\beta$$

where W(β) is the distribution of β over all local cells in the macroscopic system and Z(β) is a normalization factor for e$^{-\beta E}$ at specified β[41]."

Therefore in the Results and Discussion section under the subheading 'Superstatistical model',

"Due to the effective absence of correlations at short scales (see Fig. 1), we assume that the nucleotides are arranged randomly within each segment, and thus the internucleotide intervals l are distributed exponentially $P(l)=1/\langle l\rangle\exp(-l/\langle l\rangle)$, where $\langle l\rangle$ is the local average interval for each given segment. However, since the numbers of respective nucleotides or base pairs per segment n exhibit pronounced segment-to-segment variations along the DNA sequence, this also leads to the variations of the local average interval $\langle l\rangle$ that is inversely proportional to n. In this model, for known $P(n)$, the marginal distribution is given by

$$P(l)=\int_0^\infty P(\beta)\frac{1}{Z(\beta)}e^{-\beta l}\,\mathrm{d}\beta=\int_0^\infty P\left(\frac{n}{150}\right)\frac{n}{150}e^{-nl/150}\,\mathrm{d}n,$$

where $\beta=n/150$ is the fraction of specific nucleotides or base pairs in each of the local fragments that plays the role of the local intensity parameter and is bounded between 0 and 1.

Next we focus on the shape of the distributions $P(n)$."

should read:

"Due to the effective absence of correlations at short scales (see Fig. 1), we assume that in each of the N_{seg} segments of length 150 bp we have n_ν randomly located sites, where n_ν could stand for the number of S,W,A,C,G or T in segment ν. Then in the ν-th segment, the mean interval length $\langle l_\nu\rangle=150/n_\nu$ and the probability $P_\nu(l)$ of finding an interval of length l is given by

$$P_\nu(l) = (1/\langle l_\nu\rangle)\ (1 - 1/\langle l_\nu\rangle)^{l-1} \cong (1/\langle l_\nu\rangle)\ \exp\ (-l/\langle l_\nu\rangle) = (n_\nu/150)\ \exp(-l\, n_\nu/150).$$

Accordingly, the probability $P(l)$ of finding an interval of length l in all N_{seg}segments is given by

$$P(l) = \frac{\sum_{\nu=1}^{N_{\text{seg}}} n_\nu P_\nu(l)}{\sum_{\nu=1}^{N_{\text{seg}}} n_\nu}.$$

By introducing $\beta_\nu = n_\nu/150$ that specifies the local fraction of the considered sites in segment ν, $P(l)$ becomes

$$P(l) = \frac{\sum_{\nu=1}^{N_{\text{seg}}} \beta_\nu^2\ \exp(-l\beta)}{\sum_{\nu=1}^{N_{\text{seg}}} \beta_\nu}.$$

By introducing the probability density $W(\beta)$ the sum over the segments can be replaced by an integral,

$$P(l) = \frac{\int_0^1 W(\beta)\,\beta\ \exp(-l\beta)\,\mathrm{d}\beta}{\int_0^1 W(\beta)\beta\,\mathrm{d}\beta}$$

By substituting $P(\beta) = W(\beta)\beta/\int_0^1 W(\beta)\beta\,\mathrm{d}\beta$ we finally obtain

$$P(l) = \int_0^1 P(\beta)P(l|\beta)\,\mathrm{d}\beta \tag{2}$$

in agreement with the law of total probability (1).

Next we focus on the shape of the distributions $P(\beta)$. For simplicity, we study the related distributions of the numbers n of sites in each fragment $P(n) = P(150\cdot\beta)/150$. "

In addition, under the subheading "Internucleotide interval distributions",

"Since neither binomial nor χ^2 distributions fit the combinations of $\langle n\rangle$ and σ_n observed in the empirical data, we next follow a more generalized model with a positive support that could represent $P(n)$, and thus also the relative quantity $P(\beta)$ for all considered cases, that is the Γ-distribution

$$P(\beta) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\beta^{\alpha-1}\ \exp(-\lambda\beta) \tag{3}$$

where $\beta = n/150$ plays the role of the local intensity parameter, $\Gamma(\alpha)$ is the Γ-function, α is the shape parameter and λ is the rate parameter."

should read:

"Since neither binomial nor χ^2 distributions fit the combinations of $\langle n\rangle$ and σ_n observed in the empirical data, we next follow a more generalized model with a positive support that could represent $P(n)$ for all considered cases, that is the Γ-distribution

$$P(n) = \frac{\lambda^\alpha}{\Gamma(\alpha)}n^{\alpha-1}\ \exp(-\lambda n) \tag{3}$$

where $\beta = n/150$ plays the role of the local intensity parameter, $\Gamma(\alpha)$ is the Γ-function, α is the shape parameter and λ is the rate parameter."

Furthermore,

"Following Eq. (1) with Γ-distributed $P(\beta)$ one easily obtains

$$P(l) \propto \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)}\beta^{\alpha-1}\ \exp(-\lambda\beta)\ \exp(-\beta l)\,\mathrm{d}\beta = \frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha)}{(l+\lambda)^\alpha} = \left(\frac{\lambda}{l+\lambda}\right)^\alpha \tag{4}$$

that is asymptotically equivalent to the q-exponential distribution.

$P(l) \propto (1 + (q-1)\ (\alpha/\lambda)l)^{-1/(q-1)} = (1 + (q-1)\ \langle n\rangle l)^{-1/(q-1)}$, where $q = 1/(1+\alpha) = 1 + (\sigma_n/\langle n\rangle)^2 = 1 + \rho^2$ determines the slope of the asymptotic power law behaviour and $\langle n\rangle$ adjusts the position of the crossover, in agreement with our previous empirical findings[29]."

should read:

"Accordingly, for Γ-distributed $P(n)$ the local fractions $W(\beta)$ are also Γ-distributed as $W(\beta)= [\lambda_0{}^{\alpha}/\Gamma(\alpha)]\,\beta^{\alpha-1}\ \exp\ (-\lambda_0\beta)$, with $\langle\beta\rangle=\alpha/\lambda_0$ and $\sigma_\beta^2 = \alpha/\lambda_0^2$, where the shape parameter α is the same as for $P(n)$, while the rate parameter $\lambda_0=150\cdot\lambda$. Then $P(\beta)$ is also Γ-distributed with the shape parameter $(\alpha+1)$, and thus the marginal distribution of intervals follows

$$\begin{aligned} P(l) &= \int_0^\infty \frac{\lambda_0{}^{\alpha}}{\Gamma(\alpha)}\beta^{\alpha-1}\ \exp(-\lambda_0\beta)\,\beta^2\ \exp(-\beta l)\,d\beta \\ &= \frac{\lambda_0{}^{\alpha}}{\Gamma(\alpha)}\frac{\Gamma(\alpha)\ (\alpha+1)\,\alpha}{(l+\lambda_0)^{\alpha+2}} \\ &= \frac{\lambda_0{}^{\alpha}(\alpha+1)\,\alpha}{(l+\lambda_0)^{\alpha+2}}, \end{aligned} \tag{4}$$

that is equivalent to the q-exponential distribution $P(\tau) = C\ [1 + b(q-1)l]^{-1/(q-1)}$ with $q = 1 + 1/(\alpha+2)$, $b = (\alpha+2)/\lambda_0$ and $C = (\alpha+1)\alpha/\lambda_0^2$, in agreement with our previous empirical findings[29].

This discrepancy, while providing different asymptotic behavior, can hardly be observed for real bacterial DNA sequences. The reason is that $\alpha = 1/\rho^2$ is inversely proportional to the squared coefficient of variation $\rho = \sigma_\beta/\langle\beta\rangle$ of the fractions of given sites β. Since under normal conditions long DNA segments with nearly no strongly bonded base pairs cannot exist over long time due to their instability, and long DNA segments with nearly solely strongly bonded base pairs would require irrelevantly high energy consumption for unwinding that precede replication and transcription thereby lacking its function, we always remain at $\rho \ll 1$. This leads to $\alpha \gg 1$ thus making the decay rates of α and $\alpha+2$ hardly distinguishable in observational plots. Our tests indicate that within the available range of $P(l)$ up to the maximum internucleotide interval l that can be observed in studied bacterial DNA sequences this correction does not change significantly the PDF shapes and affects neither the validity of the figures of $P(l)$ in the original version of this Article nor the conclusions that have been drawn specifically for the bacterial DNA. However, this correction may appear important when modeling considerably longer (e.g. eukaryotic) DNA sequences as well as other complex systems where similar laws can be observed with $\rho \cong 1$,with exponentially distributed intensity parameter β as a prominent example.