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

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

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} .”

should read:

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

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

where $W(\beta)$ 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} .”

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

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

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

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

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

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

By substituting $P(\beta) = W(\beta) \beta \int_0^1 W(\beta) \beta d\beta$ we finally obtain

$$P(l) = \int_0^1 P(\beta) P(l|\beta) d\beta \quad (2)$$

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

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

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

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

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

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

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 findings²⁹.”

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

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

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 findings²⁹.

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.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017