Analysis of human DNA through power-law statistics

M. O. Costa,[1,*] R. Silva,[1,2,†] D. H. A. L. Anselmo,[2,‡] and J. R. P. Silva[1,§]
[1]*Departamento de Física, Universidade do Estado do Rio Grande do Norte, Mossoró, 59610-210, Brazil*
[2]*Universidade Federal do Rio Grande do Norte, Departamento de Física, Natal-RN, 59072-970, Brazil*

(Received 21 September 2018; published 11 February 2019)

We report an analysis of *Homo sapiens* DNA through the formalism of κ statistics, which encompasses power-law correlations and provides an optimization principle that permits us to model distinct physical systems; i.e., the power-law distribution of the length of DNA bases is calculated from a general model which follows arguments similar to those proposed in Maxwell's deduction of statistical distributions. The viability of the model is tested using a data set from a catalog of proteins collected from the Ensembl Project. The results indicate that the short-range correlations, always present in coding DNA sequences, are appropriately captured through the Kaniadakis power-law distribution, adequately describing the cumulative length distribution of DNA bases, in contrast with the case of the traditional exponential statistical model.

DOI: 10.1103/PhysRevE.99.022112

I. INTRODUCTION

The DNA molecule in several eukaryotic organisms has been widely studied from a statistical physics standpoint [1–9]. In this connection, various analyses, including random-walk simulations [10–12], wavelet transforms [13,14], and 1D Ising models [15], have successfully demonstrated that these statistical frameworks had enabled us to face the growing DNA sequences data. All these efforts have led to the conclusion that the DNA is associated with an aggregation phenomenon, resulting in a fractal cluster with power-law correlations in space or time. Furthermore, some statistical properties (e.g., long-range and short-range correlations, among others) have also been widely discussed in the context of many living organisms (see, e.g., [16] and references therein). Specifically, some approach on the length distribution of both coding and noncoding sequences of many living organisms, including human DNA, has been investigated in connection with the long-range and short-range correlations [17–21].

On the other hand, statistical frameworks based on the so-called generalized entropies have been used in order to investigate several complex systems [22,23]. From the DNA molecule standpoint, Refs. [24,25] have used the Tsallis statistics to describe both coding and noncoding human DNA sections. The behavior of the electronic specific heat at low temperature, by considering a quasiperiodic model of the DNA molecules, as well as parts of the real genomic DNA sequence have also been discussed through nonadditive statistics [26,27]. Thus the Tsallis approach, which has used a power-law distribution as an efficient tool in order to capture the so-called long-range correlations (LRCs) present in DNA molecules, has also been used as a useful framework in this subject (see, e.g., [28] and the references therein for this connection).

Within these entropic generalizations, there are other frameworks which have also been considered as a consistent approach in order to face complex systems [29]. Indeed, by considering connections between the Tsallis statistics and the DNA sequence, we are going to propose a statistical framework in order to address statistical issues associated with the DNA which is based on the concept of another generalized entropy as well as the power-law distribution. Like Tsallis entropy, the so-called Kaniadakis entropy depends on a free parameter (the κ parameter) and provides a power-law distribution rather than the exponential one [30–32]. Recently, this statistical framework has also been used to investigate some complex systems [33], and an application of the Kaniadakis framework has been proposed in connection with the DNA molecule. This study addresses the introduction of the κ-entropic effect on the geometry of the Y chromosome and the role of the correlations in the DNA molecule in order to encompass the concept of block entropy. The relationships between a set of κ-entropic parameters and the linear dimensions of the Y chromosome were calculated [8]. Here, however, we follow a route based on the introduction of the short-range correlations (SRCs) among nucleotides of the human DNA sequence through an approach which provides a distribution of length of the nucleotides. By taking into account a statistical model which follows a universal optimization principle, we use similar statistical arguments addressed in the calculation of the power-law distribution of molecular velocities [34]; these correlations among nucleotides should be captured addressing the distributions of length measured in base pairs (bp). In order to test the viability of the model, we use data of proteins compiled by the Ensembl Project [36]. It is possible to show clearly that the distributions of length follow a power-law instead of an exponential distribution. Therefore the SRCs should be characterized by a power-law distribution being more regulable with all the human chromosomes.

*marconeoliveira@df.ufpe.br
†raimundosilva@fisica.ufrn.br
‡Corresponding author: doryh@fisica.ufrn.br
§joseronaldo@uern.br

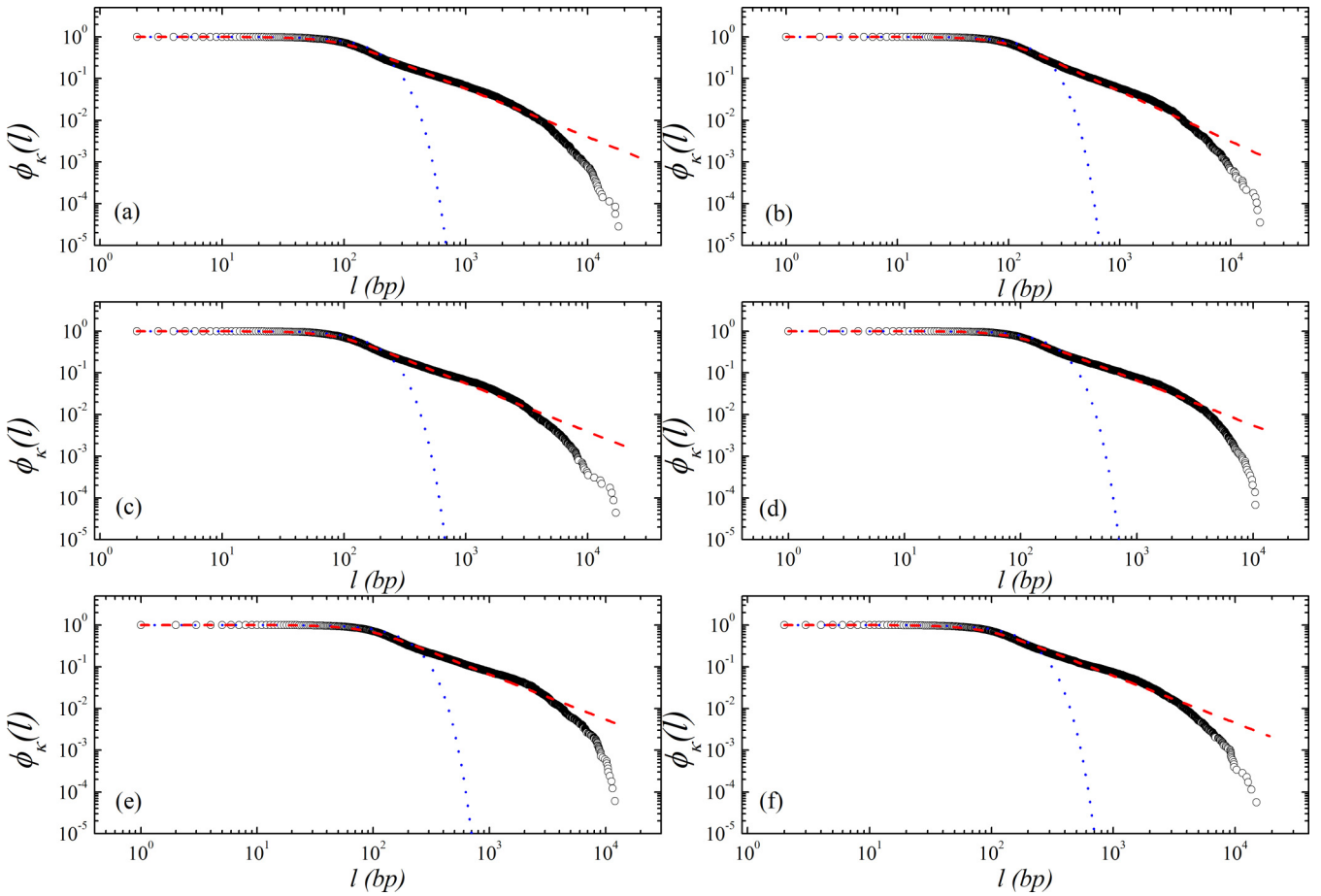

FIG. 1. Cumulative distribution of sizes for proteins. The fit of the distribution using the Kaniadakis formalism is the red dashed line, and the Gaussian distribution is represented by a blue dotted line. Panel (a) corresponds to the chromosome 01, panel (b) corresponds to chromosome 02, and so on.

This paper is organized as follows. In the next section, we will propose the analytical model which captures the statistical correlations among the length distributions of DNA, measured in base pairs (bp). In Sec. III, by using a catalog of proteins collected from the Ensembl Project, we numerically test the viability of the model showing that the SRCs are well described through the κ power-law distribution instead of an exponential one. The main conclusions are presented in Sec. IV.

II. κ DISTRIBUTIONS OF LENGTH OF DNA BASES: THE MODEL

We assume that in a volume V, each "protein" has a length belonging to the interval $[\vec{l}, \vec{l}+d\vec{l}]$; therefore in our approach we consider the probability that l_i lies on the interval $[l_i, l_i + dl_i]$, $i = x, y, z$. We also assume that the distribution can be decomposed onto its Cartesian components, and the distribution for each component is independent of the other ones. Thus, we can write the length distribution as

$$F(l)d^3l = f(l_x)f(l_y)f(l_z)dl_x dl_y dl_z \tag{1}$$

with $\vec{l} = l_x\hat{i} + l_y\hat{j} + l_z\hat{k}$ and $l = \sqrt{l_x^2 + l_y^2 + l_z^2}$.

Consequently, by using the generalized formalisms of Kaniadakis, we can write

$$F(l)d^3l = \exp_\kappa\{\ln_\kappa[f(l_x)] + \ln_\kappa[f(l_y)] + \ln_\kappa[f(l_z)]\}dl_x dl_y dl_z. \tag{2}$$

Here, κ is the nonadditive parameter, and the generalized exponential and logarithm functions $\exp_\kappa$ and $\ln_\kappa$ are given by

$$\exp_\kappa(x) = [\sqrt{1+(\kappa x)^2} + \kappa x]^{\frac{1}{\kappa}},$$
$$\ln_\kappa(x) = \frac{x^\kappa - x^{-\kappa}}{2\kappa}, \tag{3}$$

for the Kaniadakis formalism [30–32]. The standard expression (1) is recovered in the limit $\kappa \to 0$, and for this limit, the generalized logarithm and exponential functions are also reduced to their standard forms.

In order to determine the distribution functions F and f, one must first apply the generalized logarithm on Eq. (2) and derive it with respect to l_i:

$$\frac{\partial \ln_\kappa[F(l)]}{\partial l_i} = \frac{\partial \ln_\kappa[f(l_i)]}{\partial l_i}; \tag{4}$$

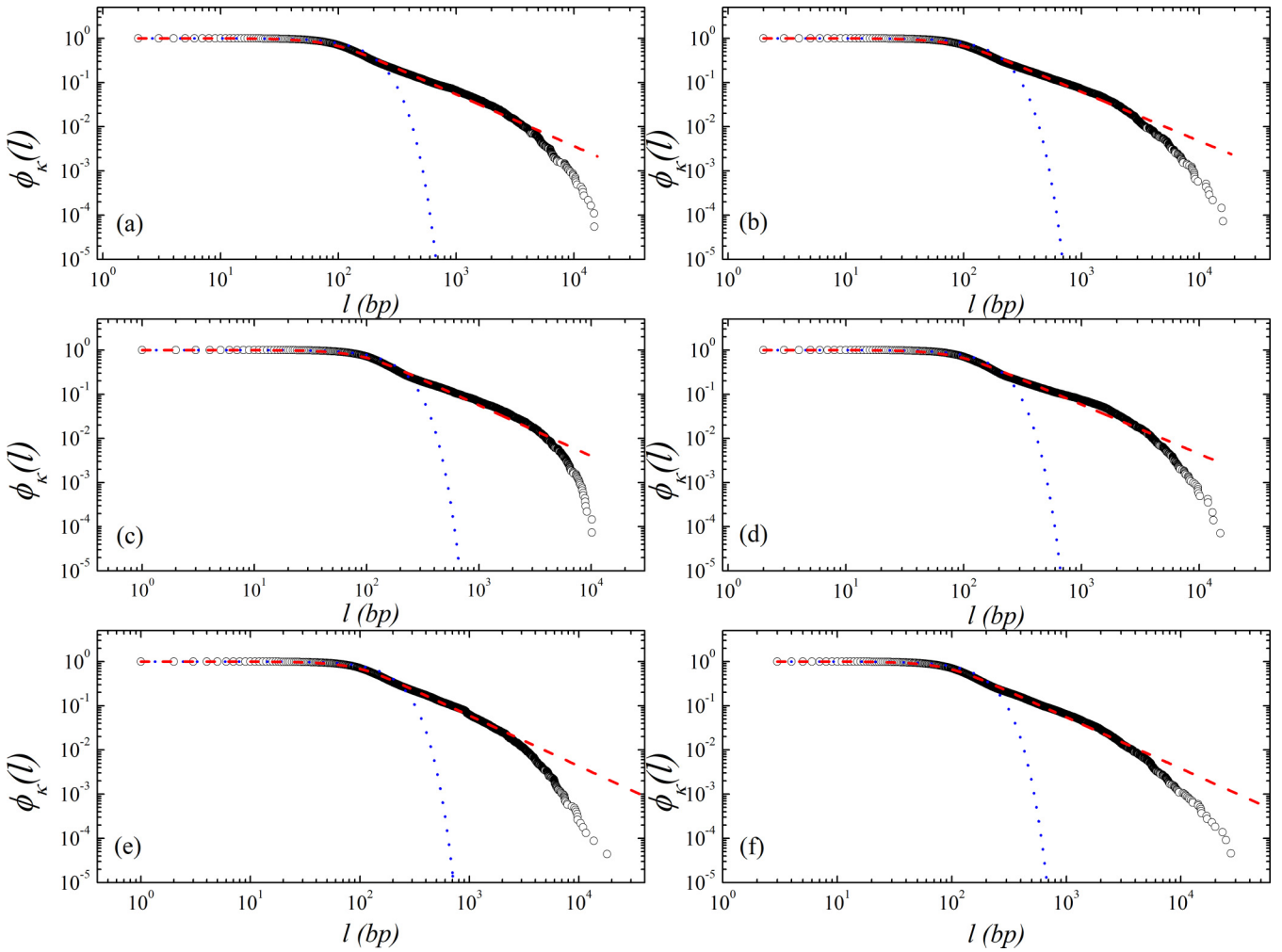

FIG. 2. Same as Fig. 1, but now for chromosomes 07 to 12.

as a consequence we obtain the relation

$$\frac{\partial \ln_\kappa[F(l)]}{\partial F(l)}\frac{dF(l)}{dl}\frac{1}{l} = \frac{1}{l_i}\frac{\partial \ln_\kappa[f(l_i)]}{\partial l_i}. \tag{5}$$

We notice that the left side of Eq. (5) is a constant, independent of whichever index is used. So we can write

$$\Phi_\kappa(l) = \frac{1}{l}\frac{\partial \ln_\kappa[F(l)]}{\partial F(l)}\frac{dF(l)}{dl}. \tag{6}$$

In this manner, we can match both Eqs. (6) and (5),

$$\Phi_\kappa(l) = \frac{1}{l_i}\frac{\partial \ln_\kappa[f(l_i)]}{\partial l_i}. \tag{7}$$

Equation (7) can be satisfied if all its members are equal to the same constant, and this equality does not depend on any components of the vector $\vec{l}$.

Now we use the Kaniadakis formalism to obtain the distributions F and f. We can make $\Phi_\kappa(l) = -\frac{2}{\sigma_\kappa^2}$, where the minus sign provides the correct normalization, and the factor $\frac{2}{\sigma_\kappa^2}$ is also introduced for mathematical convenience. Also, the parameter σ_κ is the width of the distribution in the formalism of Kaniadakis:

$$\frac{1}{l_i}\frac{d \ln_\kappa[f(l_i)]}{dl_i} = -\frac{2}{\sigma_\kappa^2}, \quad f(l_i) = \exp_\kappa\left(-\frac{l_i^2}{\sigma_\kappa^2}\right). \tag{8}$$

Hence, it is straightforward to show that the distribution $F(l)$ is given by

$$F(l) = \exp_\kappa\left(-\frac{l^2}{\sigma_\kappa^2}\right). \tag{9}$$

The probability is found in the same manner as before, on the interval $[l, l+dl]$:

$$F(l) = \int f(l) d^3l, \tag{10}$$

where $d^3l = l^2\sin(\theta)d\theta d\varphi dl$. We can write Eq. (10), for a given l, as

$$F_\kappa(l) = \iint \left\{\sqrt{1+\left[\kappa\left(\frac{l^2}{\sigma_\kappa^2}\right)\right]^2} - \kappa\frac{l^2}{\sigma_\kappa^2}\right\}^{\frac{1}{\kappa}} l^2\sin(\theta)d\theta d\varphi. \tag{11}$$

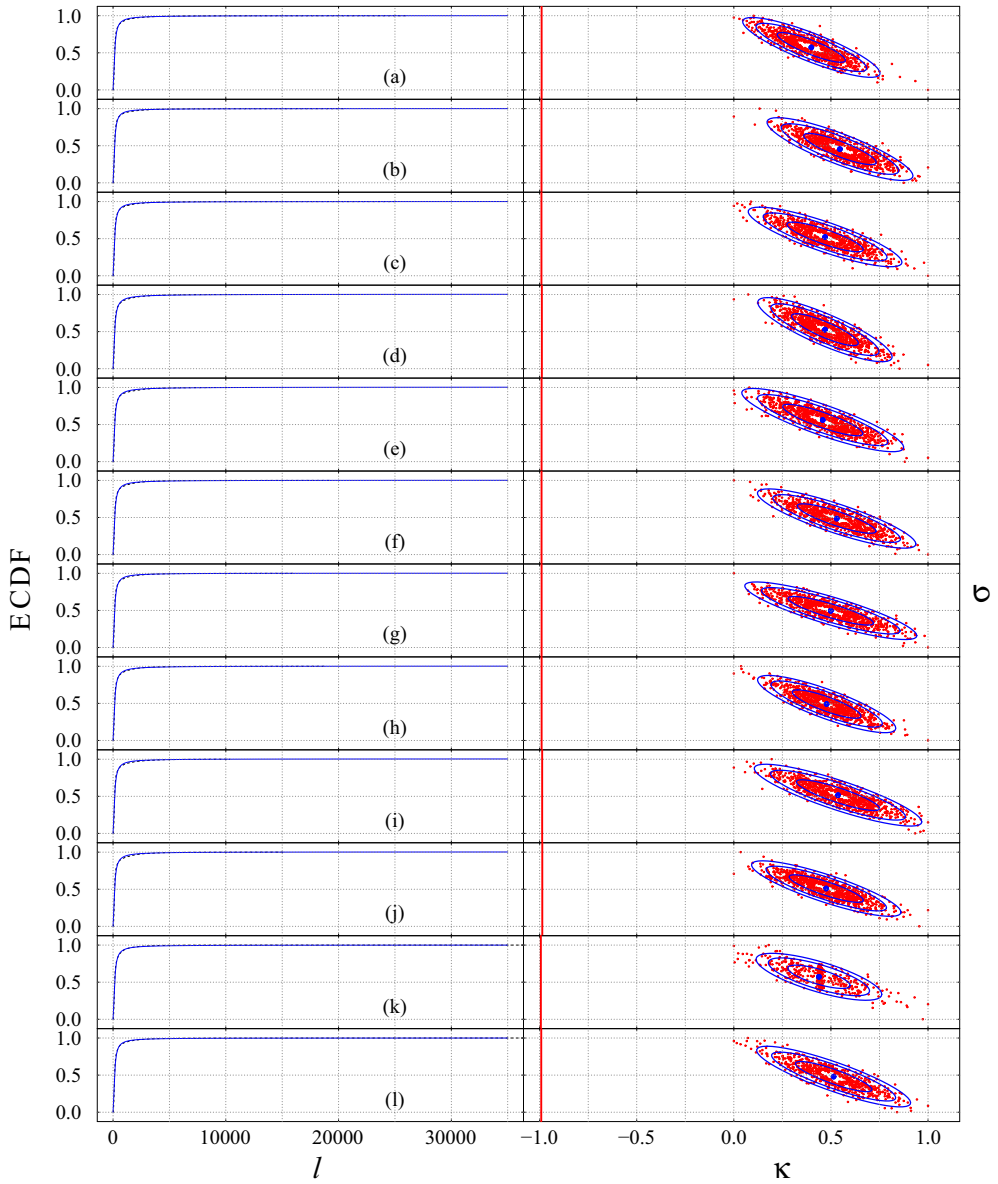

FIG. 3. Left panels: Empirical cumulative distribution functions for chromosome data sample are represented by a dashed black line and their respective best-fit curves are the continuous blue line. The best-fit parameters κ and σ are indicated. Right panels: The 68%, 95%, and 99% confidence ellipses in addition the 1000 κ-σ values computed in bootstrap replications. The axes κ and σ were scaled from 0 to 1, and the red line corresponds to $\kappa = 0$ on that scale. Panels (a), (b), (c), . . ., (l) correspond to chromosomes 01, 02, . . ., 12, respectively.

The result of the integration gives us

$$F_\kappa(l) = 4\pi l^2 \exp_\kappa\left(-\frac{l^2}{\sigma_\kappa^2}\right). \tag{12}$$

Now, we assume that the distribution (12) belongs to the same universality class of some distributions investigated previously in the same nonadditive context [34,35]. This assumption is based on the universal optimization which arises among these different complex systems, and it will be confirmed through several statistical analyses performed in the next section. So the length distribution of molecules in the Kaniadakis formalism is then given by

$$\phi_\kappa(l) = l \exp_\kappa\left(-\frac{l^2}{\sigma_\kappa^2}\right). \tag{13}$$

In the next section, we are going to test the statistical viability of the distribution of the length of proteins using the catalog of proteins collected from the Ensembl Project [36].

III. NUMERICAL RESULTS AND ANALYSIS

We made our numerical investigation by using a catalog of proteins collected from the Ensembl Project [36]. In our analysis, we considered the coding bases (exons). On these biological databases, the size of a sequence of proteins is given in terms of the number of base pairs (bp). However, there are statistical fluctuations in the distribution of sizes of proteins; thus we decided to analyze the cumulative distributions, so the suppression of fluctuations was made possible, and we compared it with the distribution functions of probabilities in

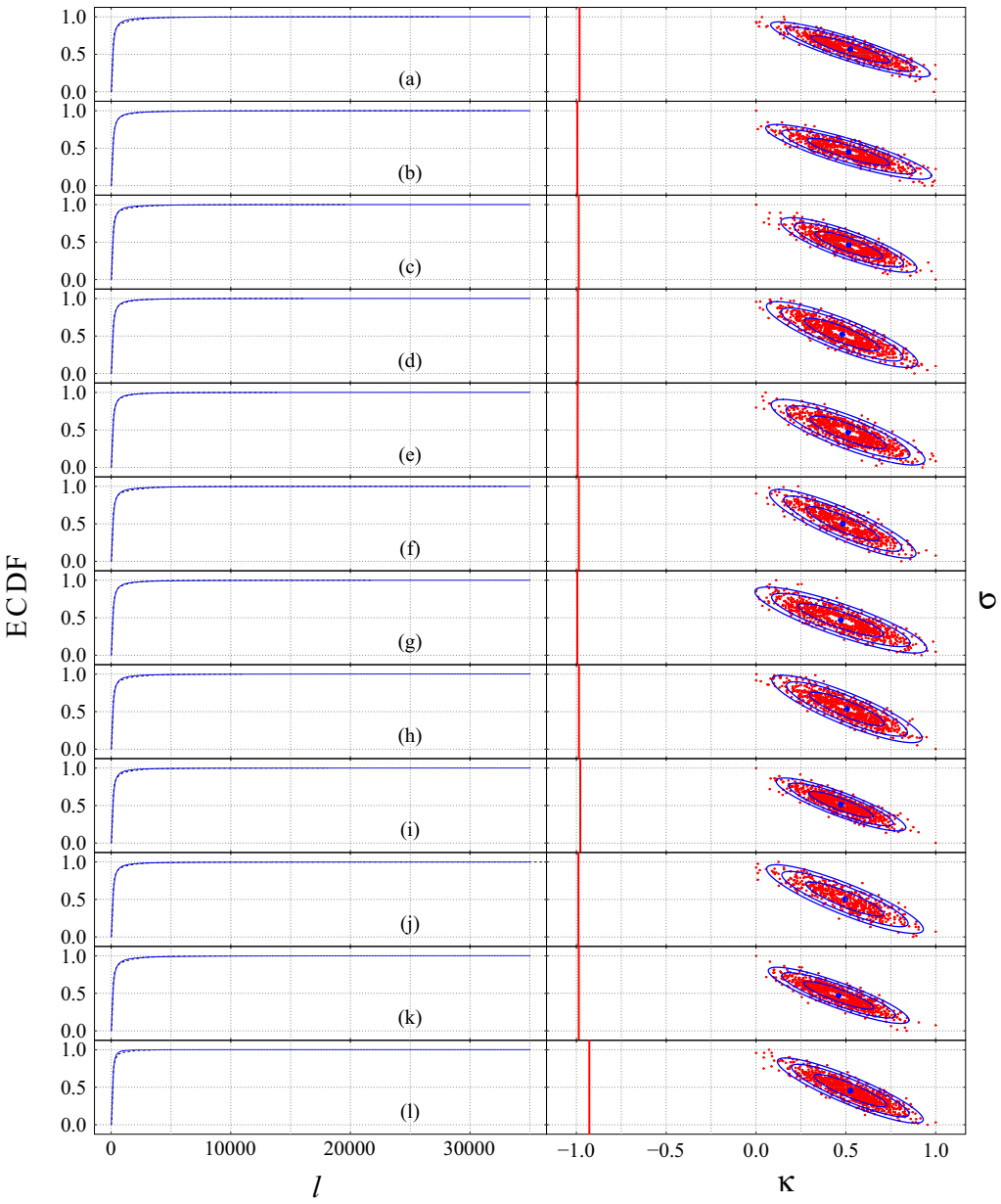

FIG. 4. The same as Fig. 3, except that panels (a), (b), (c), . . ., (j) correspond to chromosomes 13, 14, . . ., 22, respectively. Panels (k) and (l) correspond to chromosomes X and Y, respectively.

the κ Maxwellian, given by

$$\phi_\kappa(x) = A_\kappa \int_0^x x \left\{ \sqrt{1 + \left[\kappa \left(\frac{x}{\sigma_\kappa} \right)^2 \right]^2} - \kappa \left(\frac{x}{\sigma_\kappa} \right)^2 \right\}^{\frac{1}{\kappa}} dx. \tag{14}$$

As seen before, A_κ are normalization factors, and the deviation (widths) of the distributions are given by σ_κ. By solving Eq. (14) analytically the resulting normalized κ distribution is given by

$$\phi_\kappa(x) = 1 - \left[\exp_\kappa \left(-\frac{x^2}{\sigma_\kappa^2} \right) \left(\sqrt{1 + \kappa^2 \frac{x^4}{\sigma_\kappa^4}} + \kappa^2 \frac{x^2}{\sigma_\kappa^2} \right) \right]. \tag{15}$$

In Figs. 1 and 2, we show the cumulative distributions for human chromosomes 01 to 12. The qualitative and quantitative behaviors of the remaining chromosomes are quite analogous, and we have left the detailed analysis to be depicted in Figs. 3 and 4. The distribution function (13) was used to fit the protein catalog from the database, in order to obtain the best values for $\phi_\kappa(l)$ and, consequently, the best-fitting values for κ and σ_κ for the cumulative κ distribution. In Figs. 1 and 2, the κ-Maxwellian distributions are displayed in red dashed lines. For comparison, we decided to plot the Gaussian distribution (in dotted blue), which was obtained by taking $\kappa \to 0$, in expression (13). Clearly, the κ-Maxwellian distribution fits very well the whole range of base lengths (the statistical variations were also calculated). In contrast, the usual Gaussian distribution has a poor fitting behavior, mainly in the region where the curvature of the cumulative distribution changes.

TABLE I. Main characteristics and best-fit parameters of the data set of chromosomes analyzed. The columns are identification of the chromosome (1); sample size (2); first (3), second (4), and third (5) quartiles of the data set; best-fit parameters and their respective 95% confidence intervals (6 and 7); residual standard error and the achieved convergence tolerance of the fit (8 and 9).

Crm$^{(1)}$	$N_{bp}^{(2)}$	$Q1^{(3)}$	$Q2^{(4)}$	$Q3^{(5)}$	$\kappa^{(6)}$	$\sigma^{(7)}$	RMSE$^{(8)}$	$\delta^{(9)}$
1	35504	96	144	246	$0.636^{+0.002}_{-0.002}$	$112.20^{+0.872}_{-0.782}$	0.014	0.002
2	28236	90	138	239	$0.626^{+0.002}_{-0.002}$	$109.80^{+0.716}_{-0.644}$	0.013	0.006
3	22819	92	141	244	$0.636^{+0.002}_{-0.002}$	$109.30^{+0.851}_{-0.801}$	0.014	0.008
4	14750	94	142	258	$0.656^{+0.002}_{-0.002}$	$107.20^{+0.911}_{-1.021}$	0.017	0.006
5	16632	94	142	257	$0.656^{+0.002}_{-0.002}$	$107.40^{+0.965}_{-0.867}$	0.017	0.002
6	17663	95	144	258	$0.643^{+0.002}_{-0.002}$	$111.50^{+0.902}_{-0.910}$	0.015	0.004
7	18265	92	140	244	$0.633^{+0.002}_{-0.002}$	$110.40^{+0.920}_{-0.857}$	0.015	0.009
8	13785	92	141	256	$0.649^{+0.002}_{-0.002}$	$107.40^{+0.866}_{-0.893}$	0.015	0.007
9	13584	94	141	239	$0.640^{+0.003}_{-0.003}$	$109.40^{+1.208}_{-1.032}$	0.019	0.002
10	14041	93	139	233	$0.646^{+0.003}_{-0.002}$	$105.50^{+1.032}_{-1.233}$	0.002	0.002
11	22584	94	144	268	$0.638^{+0.001}_{-0.001}$	$114.00^{+0.613}_{-0.588}$	0.011	0.009
12	21774	92	140	243	$0.634^{+0.002}_{-0.002}$	$109.80^{+0.716}_{-0.726}$	0.013	0.006
13	6275	92	137	240	$0.646^{+0.004}_{-0.004}$	$106.70^{+1.706}_{-1.501}$	0.025	0.003
14	12615	89	144	274	$0.640^{+0.001}_{-0.001}$	$111.80^{+0.583}_{-0.501}$	0.009	0.008
15	14484	91	138	234	$0.631^{+0.002}_{-0.002}$	$109.00^{+1.038}_{-0.992}$	0.018	0.002
16	18182	92	142	247	$0.631^{+0.002}_{-0.002}$	$112.20^{+0.791}_{-0.727}$	0.013	0.007
17	23728	91	141	249	$0.631^{+0.001}_{-0.001}$	$111.60^{+0.645}_{-0.679}$	0.012	0.005
18	6423	94	144	259	$0.643^{+0.002}_{-0.003}$	$111.50^{+1.191}_{-1.084}$	0.018	0.003
19	24261	87	139	263	$0.641^{+0.001}_{-0.001}$	$107.80^{+0.468}_{-0.548}$	0.009	0.005
20	8537	92	140	243	$0.635^{+0.003}_{-0.003}$	$109.80^{+1.283}_{-1.125}$	0.020	0.004
21	4019	92	144	242	$0.618^{+0.004}_{-0.004}$	$115.34^{+1.271}_{-1.399}$	0.020	0.007
22	8425	92	142	252	$0.626^{+0.002}_{-0.003}$	$113.60^{+1.035}_{-0.970}$	0.016	0.002
X	12558	92	138	249	$0.659^{+0.002}_{-0.003}$	$103.70^{+1.164}_{-1.106}$	0.019	0.009
Y	1779	89	120	190	$0.525^{+0.012}_{-0.013}$	$117.80^{+2.706}_{-2.502}$	0.035	0.000

This behavior was observed in all chromosomes. Indeed, in Fig. 2 the curves related to the chromosomes 07 to 12 can be observed. Here the κ distribution is also a remarkably good fitting function for the experimental data, in opposition to the Gaussian, blue dotted, curves.

Now we proceed to make a statistically detailed analysis. The best-fit parameters κ and σ were estimated by using the method described in Costa *et al.* [9] (see also Silva *et al.* [37]) which comprises three steps. The first one consists of computing the empirical cumulative distribution function (ECDF), defined as

$$F_e(l) = \frac{1}{n}, \quad l \in \{l_1, l_2, \ldots, l_n\}, \tag{16}$$

where a cumulative probability of an l value that repeats k times is given by k/n [38,39]. In the second step the theoretical cumulative distribution function, $\phi_\kappa(x)$ [Eq. (15)], is fitted to the ECDF using a Gauss-Newton algorithm to minimize the sum of squares of the residuals. The last step is to estimate the 95% confidence intervals for the best-fit parameters κ and σ by using bootstrap resampling. Table I summarizes the main characteristics of the samples for each chromosome and the respective estimated best-fit parameters. In the table, the first column presents the chromosome data set, Crm. The second column gives the length of the sample, given in number of base pairs, N_{bp}. The columns indicate by $Q1$, $Q2$, and $Q3$ the first, second, and third quartiles of the data set. The best-fit parameters, κ and σ, are given in the sixth and seventh columns, respectively. The eighth and ninth columns display the residual standard error, RMSE, and the achieved convergence tolerance, δ, in each fit.

The left panels of Figs. 3 and 4 show the best-fit curves (in continuous blue) and corresponding ECDFs for the analyzed chromosomes data samples. The panels to the left of Figs. 3 and 4 present the 68%, 95%, and 99% confidence ellipses on the κ-σ computed for 1000 bootstrap replications of κ and σ parameters for each chromosome. It should be noted that the axes were scaled to range between 0 (minimum value of κ or σ) and 1 (maximum value of κ or σ). In that scale, the line in red corresponds to $\kappa = 0$.

IV. CONCLUSIONS

Although several statistical approaches have been used in order to study the DNA molecules, mainly, taking into accounts different statistical tools to investigate the correlations [10–15], we have proposed a model based on the power-law statistics which captures the SRCs. The model was based

primarily on an extension of Maxwell's deduction of the molecular distribution. Indeed, the model captured the statistical correlations, where SRCs were associated with power laws. Moreover, it was also shown that the exponential functions were ruled out in order to capture the SRCs. The core of our approach is related to the distributions of length measured in base pairs (bp), being mathematically characterized by a power law rather than exponential behavior. The viability of the model was tested considering the data set of proteins compiled by the Ensembl Project [36].

Specifically, the empirical cumulative distribution function (ECDF), based on the method developed in Ref. [37], was adapted to calculate the distribution length of proteins. By using the *Homo sapiens* DNA, we showed that the phenomena of SRCs are statistically consistent with the power-law distribution of the model (κ-Maxwellian distributions), rather than the exponential length distribution which fails for all the curves, mainly when the inclination changes. The SRCs present on *Homo sapiens* DNA can be statistically tested through the power-law distribution with the value of $\kappa \sim 0.6$ for almost all chromosomes, except the Y with $\kappa \sim 0.5$. The exponential distribution, on the other hand, should be statistically ruled out in order to capture the SRCs (it is visually shown in Figs. 1 and 2 and considering $\kappa = 0$ in Fig. 3). In particular, the values of the entropic parameter κ found here are within the same range found in Ref. [8]. It is worth mentioning that a similar model, however being based on the Tsallis framework, has been proposed in order to investigate the SRCs presenting in human DNA [9]. From the statistical standpoint, at least in the context of this analysis (the cumulative length distribution of human DNA), it is not possible to know which model is better. We, however, believe that a Bayesian analysis (which can compare models) could be done, in order to answer this question.

Finally, it is worth emphasizing that in order to obtain a general approach based on the model proposed in this paper, we need to include the noncoding DNA, which exhibits long-range correlations (see, for instance, [21]). This issue will be investigated in future work.

ACKNOWLEDGMENTS

The authors are very grateful to the Brazilian agency CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - No. 065.815.684-57) for financial support. R.S. thanks CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - No. 303613/2015-7) for financial support.

[1] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[2] H. E. Stanley *et al.*, Physica A **273**, 1 (1999).
[3] A. K. Mohanty and A. V. S. S. Narayana Rao, Phys. Rev. Lett. **84**, 1832 (2000).
[4] W. Li and D. Holste, Phys. Rev. E **71**, 041910 (2005).
[5] C. Vaillant, B. Audit, and A. Arneodo, Phys. Rev. Lett. **95**, 068101 (2005).
[6] M. V. Koroteev and J. Miller, Phys. Rev. E **84**, 061919 (2011).
[7] A. Carbone, Sci. Rep. **3**, 2721 (2013).
[8] N. T. C. M. Souza, D. H. A. L. Anselmo, R. Silva, M. S. Vasconcelos, and V. D. Mello, Europhys. Lett. **108**, 38004 (2014).
[9] M. O. Costa, J. R. P. Silva, R. Silva, D. H. A. L. Anselmo, and J. S. Alcaniz (unpublished).
[10] C.-K. Peng *et al.*, Nature (London) **356**, 168 (1992).
[11] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[12] W. Li, Comput. Chem. **21**, 257 (1997).
[13] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).
[14] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J. F. Muzy, and A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001).
[15] A. Colliva, R. Pellegrini, A. Testori, and M. Caselle, Phys. Rev. E **91**, 052703 (2015).
[16] A. Provata and Y. Almirantis, J. Stat. Phys. **106**, 23 (2002).
[17] Y. Almirantis and A. Provata, J. Stat. Phys. **97**, 233 (1999).
[18] A. Provata and Y. Almirantis, Fractals **8**, 15 (2000).
[19] P. Katsaloulis, T. Theoharis, and A. Provata, Physica A **316**, 380 (2002).
[20] P. Katsaloulis, T. Theoharis, W. M. Zheng, B. L. Hao, T. Bountis, Y. Almirantis, and A. Provata, Physica A **366**, 308 (2006).
[21] A. Provata and Th. Oikonomou, Phys. Rev. E **75**, 056102 (2007).
[22] M. Gell-Mann and C. Tsallis, (eds.), *Nonextensive Entropy: Interdisciplinary Applications* (Oxford University Press, New York, 2004).
[23] G. Kaniadakis, Eur. Phys. J. B **70**, 3 (2009).
[24] Th. Oikonomou and A. Provata, Eur. Phys. J. B **50**, 259 (2006).
[25] Th. Oikonomou, A. Provata, and U. Tirnakli, Physica A **387**, 2653 (2008).
[26] D. A. Moreira, E. L. Albuquerque, L. R. da Silva, and D. S. Galvão, Physica A **387**, 5477 (2008).
[27] E. L. Albuquerque *et al.*, Phys. Rep. **535**, 139 (2014).
[28] M. I. Bogachev, A. R. Kayumov, and A. Bunde, PloS One **9**, e112534 (2014).
[29] C. Beck, Contemp. Phys. **50**, 495 (2009); J. Naudts, *Generalised Thermostatistics* (Springer-Verlag, London, 2011); R. Hanel, S. Thurner, and M. Gell-Mann, Proc. Natl. Acad. Sci. USA **108**, 6390 (2011); **109**, 19151 (2012); **111**, 6905 (2014); U. Tirnakli and E. Borges, Sci. Rep. **6**, 23644 (2016).
[30] G. Kaniadakis, Physica A **296**, 405 (2001).
[31] G. Kaniadakis, Phys. Rev. E **66**, 056125 (2002).
[32] G. Kaniadakis, Phys. Rev. E **72**, 036108 (2005).
[33] F. Clementi, M. Gallegati, and G. Kaniadakis, Eur. Phys. J. B **57**, 187 (2007); F. Clementi, T. di Matteo, M. Gallegati, and G. Kaniadakis, Physica A **387**, 3201 (2008); F. Clementi, M. Gallegati, and G. Kaniadakis, J. Stat. Mech. (2009) P02037; (2012) P12006; B. Trivellato, Entropy **15**, 3471 (2013); O. J. Tapiero, Physica A **392**, 3051 (2013); M. L. Bertotti and G. Modenese, Eur. Phys. J. B **85**, 261 (2012).
[34] J. C. Carvalho, R. Silva, J. D. do Nascimento, and J. R. De Medeiros, Europhys. Lett. **84**, 59001 (2008); J. C. Carvalho,

J. D. do Nascimento, R. Silva, and J. R. De Medeiros, Astrophys. J. **696**, L48 (2009); E. P. Bento, J. R. P. Silva, and R. Silva, Physica A **392**666 (2013).
[35] R. Silva, Phys. Lett. A **352**, 17 (2006).
[36] The Ensembl Project, http://www.ensembl.org/info/about/index.html.
[37] J. R. P. Silva, M. M. F. Nepomuceno, B. B. Soares, and D. B. Freitas, Astrophys. J. **777**, 20 (2013).
[38] G. J. Kerns, *Introduction to Probability and Statistics Using R.* (Lulu. com, 2010).
[39] Y.-P. Qin, G. Z. Xie, X.-T. Zheng, and S.-M. Wu, Astron. Astrophys. Suppl. Ser. **132**, 301 (1998).