A \( \kappa \)-statistical analysis of the Y-chromosome

N. T. C. M. Souza\textsuperscript{1}, D. H. A. L. Anselmo\textsuperscript{1}, R. Silva\textsuperscript{1,3}, M. S. Vasconcelos\textsuperscript{2} and V. D. Mello\textsuperscript{3}

\textsuperscript{1} Departamento de Física Teórica e Experimental, Universidade Federal do Rio Grande do Norte
Natal-RN, 59072-970, Brazil
\textsuperscript{2} Escola de Ciência e Tecnologia, Universidade Federal do Rio Grande do Norte - 59072-970, Natal-RN, Brazil
\textsuperscript{3} Universidade do Estado do Rio Grande do Norte, Departamento de Física - Mossoró-RN, 59610-210, Brazil

received 3 September 2014; accepted in final form 17 October 2014
published online 10 November 2014

PACS 89.70.Cf – Entropy and other measures of information
PACS 87.15.-v – Biomolecules: structure and physical properties
PACS 87.14.gk – Nucleic acids: DNA

Abstract – An analysis of the coding sequence for the Y-chromosome (\textit{Homo sapiens}) has been performed, embedded in the formalism of \( \kappa \)-statistics, which naturally encompasses long-range correlations. In this formalism, the entropy has been written as a function of \( \kappa \) (called the deformation—or entropic—parameter). The \( \kappa \)-entropy has been linked directly to dimensional parameters defined for the DNA chain associated with the chromosome Y. Our analysis indicates that there are certain regions of chromosome Y which exhibit linearity between \( S_\kappa \) entropy and sample sizes for some particular values of \( \kappa \), implying that, on these regions, the information contained on the DNA increases monotonically linearly with the sample size, and also depicts an internal order.

Introduction. – The DNA is formed by two linear polymer chains consisting of smaller units called nucleotides. Composing each nucleotide, one finds the known nitrogen bases: Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). It is through these bases that the two linear chains are connected to each other by hydrogen bonds through the nucleotides, these in turn are connected in pairs according to the rules of Chargaff \cite{1}. In the \textit{Homo sapiens} species, the DNA is grouped in 23 pairs of chromosomes, located in the nucleus of cells and mitochondria. Being a “soft” molecule, DNA is highly subject to the change of its properties (conductivity, structure, etc.), due to some external influence \cite{2}.

From a statistical point of view, many studies on the statistic of genomic sequences have shown degrees of complexity in the primary structure of DNA. An important statistical property of the DNA molecule is related to observation of long-range correlations, which has been investigated using the so-called DNA walk model \cite{3} (see also \cite{4,5} for an analysis based on the \( 1/f \) spectrum). Some frameworks have been proposed in order to investigate the statistics of genomic sequences, as well as the DNA molecule, \textit{e.g.} Levy-walk models \cite{6}, linguistic methods \cite{7}, wavelet approaches \cite{8}, multifractal information and Rényi entropy \cite{9}, network approach \cite{10}, among others (see \cite{11,12} and references therein).

On the other hand, the Tsallis framework has been a useful tool in the study of systems which present long-range correlations between the constituent parts \cite{13}. In particular, there are theoretical efforts to study the DNA molecule through the Tsallis approach, \textit{e.g.} the size distributions of non-coding DNA (including introns and intergenic regions) in human chromosomes have been studied by using the \( q \)-exponential distribution that emerges from the Tsallis framework \cite{14,15}. Another effort has also used the Tsallis statistics to show the behavior of the electronic specific heat at low temperature by considering a quasi-periodic model for the DNA molecules, as well as parts of the real genomic DNA sequence \cite{16,17}. More recently, using entropy analysis and phase plane concepts, the DNA information has been successfully investigated in the non-extensive framework \cite{18}. Here, it is worth asking: i) Is there another framework, beyond the Tsallis one, useful to estimate the role of long-range correlation in the DNA molecule? ii) What is the entropic effect on the geometric organisation of the chromosome Y?

To address such issues, we use the generalized statistics of Kaniadakis \cite{19, 20, 21}, which is characterized by an additive entropy\textsuperscript{4} and a power-law behaviour, by featuring the stationary distribution of random variables of the system.

\textsuperscript{4}In contrast, the Tsallis entropy is non-additive.
We investigate the $\kappa$-entropic effect on the geometry of the chromosome $Y$, as well as the role of the long-range correlations in the DNA molecule, which naturally emerge in the context of generalized statistics [13,19,20]. Specifically, we will use Kaniadakis’ definition of an adapted entropy together with the concept of block entropy to describe the relationships between a set of $\kappa$ parameters and the linear dimensions of the chromosome $Y$. Here, we follow the approach proposed by the author in [22], who demonstrated analytically that the static structures of deterministic Cantor sets with fractal dimension $d_f$, calculated in the framework of Tsallis’ statistics, are characterized by a non-extensive $q$-exponent, i.e., $q = 1/(d_f - d)$. In contrast, we will approach the problem within the ambit of the generalized statistics of Kaniadakis [19–21].

This paper is organized as follows: a discussion about the Kaniadakis framework is made in the next section. Later on, we give a brief description of the Y-chromosome, and relate its properties that are important to the present study. Finally, in the conclusions section we summarize our main findings.

Kaniadakis framework. – Through a kinetic foundation, in a non-linear phase, a system of particles can be governed by a principle called the Kinetic Interaction Principle (KIP), as proposed by Kaniadakis [19]. According to the KIP, it is always possible to obtain a stationary statistical distribution consistent with the system constraints, which “imposes” a specific form for the (generalized) entropic distribution. By following this direction, and admitting the momenta and its distribution function, respectively. Several physical features of the $\kappa$-distribution have also been theoretically investigated as, for instance, the self-consistent relativistic statistical theory [20,21], non-linear kinetics [23], the $H$-theorem from a generalization of the chaos molecular hypothesis [24,25], the $\kappa$-Weibull distribution as one model for extreme-event return intervals in finite-size systems [26], and the reexamination of the blackbody radiation in the context of $\kappa$-framework [27]. Moreover, another investigation has shown that the $\kappa$-statistics is able to describe the entropy of a Cantor set [28].

$\kappa$-Description of DNA: chromosome $Y$. – The Y-chromosome is composed of approximately 50 million base pairs in size, containing more than 400 genes that represent no more than 2% of all pairs mapped by the human genome project [29]. This chromosome contains information that is associated with, among other features, male infertility due to spermatogonial failure, growth control and sex determination.

Although there is no rule of iteration that could lead us to that sequence construction, however, and in the context of entropy of block, we can apply the method of block scanning and associate a block entropy $S_{\kappa}(s, L)$ to chromosome $Y$, by describing the way each nucleotide base in the DNA sequence presents itself in relation to the rest of the chain.

The scanning of the coding sequence of chromosome $Y$ is made by using the block-scanning method [28]. Here, we analyze the DNA sequence using blocks of size $s$, where $s = 1, 2, 4$. The choice of these values will become clear later. Then we use the functional $S_{\kappa}$, as expressed in

\[
S_{\kappa} = -\int d^3 p f \ln_{\kappa} f,
\]

which fully recovers standard Boltzmann-Gibbs entropy, $S_{\kappa = 0}(f) = -\int f \ln f d^3 p$. Here, $p$ and $f$ represent the momenta and its distribution function, respectively. Several physical features of the $\kappa$-distribution have also been theoretically investigated as, for instance, the self-consistent relativistic statistical theory [20,21], non-linear kinetics [23], the $H$-theorem from a generalization of the chaos molecular hypothesis [24,25], the $\kappa$-Weibull distribution as one model for extreme-event return intervals in finite-size systems [26], and the reexamination of the blackbody radiation in the context of $\kappa$-framework [27]. Moreover, another investigation has shown that the $\kappa$-statistics is able to describe the entropy of a Cantor set [28].

$\kappa$-Description of DNA: chromosome $Y$. – The Y-chromosome is composed of approximately 50 million base pairs in size, containing more than 400 genes that represent no more than 2% of all pairs mapped by the human genome project [29]. This chromosome contains information that is associated with, among other features, male infertility due to spermatogonial failure, growth control and sex determination.

Although there is no rule of iteration that could lead us to that sequence construction, however, and in the context of entropy of block, we can apply the method of block scanning and associate a block entropy $S_{\kappa}(s, L)$ to chromosome $Y$, by describing the way each nucleotide base in the DNA sequence presents itself in relation to the rest of the chain.

The scanning of the coding sequence of chromosome $Y$ is made by using the block-scanning method [28]. Here, we analyze the DNA sequence using blocks of size $s$, where $s = 1, 2, 4$. The choice of these values will become clear later. Then we use the functional $S_{\kappa}$, as expressed in

\[
S_{\kappa} = -\int d^3 p f \ln_{\kappa} f,
\]

which fully recovers standard Boltzmann-Gibbs entropy, $S_{\kappa = 0}(f) = -\int f \ln f d^3 p$. Here, $p$ and $f$ represent the momenta and its distribution function, respectively. Several physical features of the $\kappa$-distribution have also been theoretically investigated as, for instance, the self-consistent relativistic statistical theory [20,21], non-linear kinetics [23], the $H$-theorem from a generalization of the chaos molecular hypothesis [24,25], the $\kappa$-Weibull distribution as one model for extreme-event return intervals in finite-size systems [26], and the reexamination of the blackbody radiation in the context of $\kappa$-framework [27]. Moreover, another investigation has shown that the $\kappa$-statistics is able to describe the entropy of a Cantor set [28].

It is important to mention here two properties of $S_{\kappa}$: when we make a composition of two independent subsystems, i.e., $p_{ij} = p_i \otimes p_j$ with $\otimes$ being the so-called $\kappa$-product, the $\kappa$-entropy is in general additive and extensive (for details, see [21]).

From a mathematical point of view, the $\kappa$-framework is based on the $\kappa$-exponential and the $\kappa$-logarithm functions defined as

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

with

\[
\ln_{\kappa}(\exp_{\kappa}(x)) = \exp_{\kappa}(\ln_{\kappa}(x)) = x.
\]

In the case $\kappa \to 0$, these expressions reduce to the usual exponential and logarithmic functions. For a continuous distribution function $f$, the $\kappa$-entropy associated with the $\kappa$-framework is given by

\[
S_{\kappa} = -\int d^3 p f \ln_{\kappa} f,
\]

which fully recovers standard Boltzmann-Gibbs entropy, $S_{\kappa = 0}(f) = -\int f \ln f d^3 p$. Here, $p$ and $f$ represent the momenta and its distribution function, respectively. Several physical features of the $\kappa$-distribution have also been theoretically investigated as, for instance, the self-consistent relativistic statistical theory [20,21], non-linear kinetics [23], the $H$-theorem from a generalization of the chaos molecular hypothesis [24,25], the $\kappa$-Weibull distribution as one model for extreme-event return intervals in finite-size systems [26], and the reexamination of the blackbody radiation in the context of $\kappa$-framework [27]. Moreover, another investigation has shown that the $\kappa$-statistics is able to describe the entropy of a Cantor set [28].

$\kappa$-Description of DNA: chromosome $Y$. – The Y-chromosome is composed of approximately 50 million base pairs in size, containing more than 400 genes that represent no more than 2% of all pairs mapped by the human genome project [29]. This chromosome contains information that is associated with, among other features, male infertility due to spermatogonial failure, growth control and sex determination.
A $\kappa$-statistical analysis of the Y-chromosome

and we calculate computationally the entropy for each coding segment of DNA, as a function of the size $L$ of the nucleotide sequence present in chromosome Y. Therefore, the $\kappa$-entropy can be written as

$$S_\kappa(s, L) = - \sum_i \left[ p_i(s, L) \right]^{1 + \kappa} - \left[ p_i(s, L) \right]^{1 - \kappa} / 2\kappa,$$

where $s$ is the block size used to sweep the sequence of nucleotides, and $L$ is the length of the chain, measured in units of number of nitrogenous bases (nB). The sum in (8) will depend on the value assumed for $s$. When $s = 1$, the DNA is scanned in order to consider a single base at a time. In this case, the index $i$ will run from 1 to 4, which corresponds to the existing four nitrogenous bases on the chromosome under study: A, T, G and C. When $s = 2$, the bases are taken two by two (here, blocks of type XY and YX are taken as the same). Thus, there are ten different possibilities to combine the bases, namely, AA, AT, AG, AC, TT, TG, TC, GG, GC, CC. The previous schema remains when we consider $s = 4$, however, we observe that the number of blocks to take into account inside the string will be 35: AAAT, AAAG, AAAC, AATT, AATG, AATC, AAGG, AAGC, ACC, ACCC, ATTT, TTTC, TTGC, TTTG, TTGG, TGGC, CGCT, CCC, CCC, CGTC, GGTC, GCC, GCC, GCC, GCC, GGG, GGG, GGG, GGG, GGG, GGG, GGG, GGG, GGG, GGG.

Because we are performing a box-counting analysis, we have considered that blocks as ATGC and GCAT are taken as equal within our scheme, and therefore are counted just once. Although from a functional point of view these arrangements could be considered distinct from each other (see paragraph below), we will consider them statistically identical. Energetically speaking, however, our assumption is physically reasonable, since the energy of the block remains the same in both orders [2].

In this work, we applied a filter to NCBI database for chromosome Y, and discarded each of the explicitly non-coding sequences. Also, inside the gene itself there are some sub-sequences of nucleotides which contain the instructions to produce proteins, while other sub-sequences do not follow this rule. The sub-sequences which have the instructions are also called coding sequences. However, even in the remaining data, some of the nucleotides do not play a role of heredity.

A few words must be said regarding the so-called junk DNA [30]. This nomenclature derives from a paradox (“C-value paradox”), which states that more complex organisms should have longer encoding sequences, but actually the opposite occurs [31]. For example, lungfish DNA is around 30 times larger than the human DNA, and even some flowers have also a much larger genomic encoding than ours [32]. So, in this context, some author argue that a large part of a DNA sequence is composed by non-coding (junk) DNA. Although there exists some controversy about this nomenclature, and the meaning of a functional element (see, e.g., [33,34]), in our numerical analysis we explicitly discarded the (apparently) non-coding parts. Since the role of the so-called junk DNA is still unclear, we decided to restrict our scanning to the “coding” sets.

Results and discussion. – From now on we will use eq. (3), along with the concept of block entropy, to associate to the coding sequence of human DNA (more specifically to that which concerns the Y-chromosome) an entropy which describes statistically the DNA arrangement, as well as measures long-range statistical correlations between the $n$-tuples sets of genomic bases. In fig. 1 we depict our building blocks of information, namely the probabilities of finding the four given bases, when scanning the Y-chromosome with a block of size $s = 1$. A similar analysis is shown in fig. 2, however considering the coupling of the ten different base pairs. With this result, together with the definition of block of information, we are able to determine the $S_\kappa$-entropy. In both figures, the probability saturates. It is noticeable that for $s = 2$, the probability of occurrence of pairs made of equal bases...
**Fig. 3:** (Colour on-line) (a) $\kappa$-block entropy as a function of the chain length for $\kappa = 3.4$ Here, the box size is $s = 2$, so we consider 10 different pairs of bases. Inset: strong oscillations occur for low values of the chain length. (b) Same as (a), but for several values of the entropic parameter $\kappa$.

(TT, AA, CC, GG) is lower than those of different ones, with the exception of GC pairs, which is caused obviously by the lower occurrence of these base pairs, for the Y-chromosome, that can be inferred from fig. 1. Now we analyze the $\kappa$-entropy associated with these probabilities. The behavior of $S_\kappa(s = 2, L)$ with $L$ is marked, initially (fig. 3(a), see also the insert) by strong oscillations. This is expected, because when we proceed to scan the coding chain, any significant changes in the counting of a block carries a drastic change in the probability of one being faced with it and, consequently, the entropy can oscillate strongly at this stage. Thereafter, the entropy presents a (mostly linear) increase, and then saturates. In fig. 3(b) we consider several values of the entropic parameter $\kappa$. In the region where $S_\kappa(s, L)$ increases more linearly, we have observed that the ratio of increasing is directly linked to the entropic parameter $\kappa$. Depending on the value of $\kappa$, the generalized entropy in this region may suffer a rapid increase, can grow slowly and even grow linearly with $L$.

We realized that all values of $\kappa$ make $S_\kappa(s, L)$ linear with respect to $L$ (extensive), at least on some of the coded region under consideration. When we increase the box size ($s = 4$), the behavior of the $\kappa$-entropy changes accordingly, as one can see in fig. 4. Here, we decided to make an analysis of the $\kappa$-entropy for different values of the entropic parameter $\kappa$, so it can be compared to fig. 3(b). Also, in this case the $s = 4$ block size allows a richer set of 4-tuples bases. From this figure, it is also very clear that there are regions of system length which allow a linear increase of the entropy, independently of the value of the entropic parameter, and of the box size $s$. This region belongs to the interval $L \in [0.25, 0.52]$ (in units of $10^7$ nB).

**Conclusions.** – An analysis of the Y-chromosome has been performed by using the $\kappa$-statistical formalism adapted to the block concept. DNA molecules actually display a multi-fractal pattern [35], what could lead us to consider that DNA information is characterized by the statistical concept of disorder. Indeed, the $\kappa$-statistics is suitable to analyze the long-range correlations of the DNA molecules. However, in the ambit of $\kappa$-entropy, we can always find a suitable $\kappa$ that makes the system linear with $L$ and, consequently, extensive in this sense. This means that, at least for some portions, the Y-chromosome behaves non-fractally. By non-fractal we mean a structure which presents a linear behavior, and therefore is extensive, in contrast to fractals, which feature a power-law mathematical behavior in all scales. Indeed, in a similar analysis made in [28], applied to the Cantor set, it was determined that there is *only one* value of $\kappa$ which makes the entropy linear. Here, on the opposite, there is a full range of linearity for all the values of $\kappa$. Additionally, we have found that the values of $\kappa$ that make $S_\kappa$ linear with $L$ are contained in a region where $\kappa$-entropy shows no maximum since they are outside the limit $|\kappa| \leq 1$. Another aspect that we have noticed in this analysis, and which refers to description of the coding sequence in a physical framework concerning the thermodynamics, it is the emerging possibility of describing the DNA inside a $\kappa$-theory of ensembles. How to do this? We know that DNA is subject to various intra- and extra-structural forces which stabilize it. One of these interactions occurs within the intra-structural ambit: the base pairs that are stacked exert forces on each other. This type of interaction is known in the literature...
as stacking interaction. The energies associated with each pair of interaction are well known. Thereby, we could make the connection with the thermodynamics through a deformed partition function $Z_{\kappa}$, associated, in turn, to a statistical weight of the form $\frac{\exp(-\beta\epsilon_{i})}{\kappa}$ and hence, we could obtain all the $\kappa$-thermodynamics of the system, such as the specific heat $c_{\kappa}$.

***

We acknowledge the financial support received from the Brazilian Founding Agency (CAPES). RS is very grateful to CNPq for the grants under which this work was carried out.

REFERENCES

[1] Zameroff S., Brawermann F. and Chargaff E., Biochim. Biophys. Acta, 9 (1952) 402.
[2] Yakushevich L. V., Nonlinear Physics of DNA (Wiley-VHC) 2004.
[3] Peng C.-K., Buldyrev S. V., Goldberger A. L., Havlin S., Sciontin F., Simons M. and Stanley H. E., Nature, 356 (1992) 168.
[4] Li W. and Kaneko K., Europhys. Lett., 17 (1992) 655.
[5] Voss R. F., Europhys. Lett., 68 (1992) 3805.
[6] Buldyrev S. V., Goldberger A. L., Havlin S., Peng C.-K., Simons M. and Stanley H. E., Phys. Rev. E, 47 (1993) 4514.
[7] Mantegna R. N., Buldyrev S. V., Goldberger A. L., Havlin S., Peng C.-K., Simons M. and Stanley H. E., Phys. Rev. E, 52 (1995) 2039.
[8] Arneodo A., Bacry E., Graves P. V. and Muzy J. F., Phys. Rev. Lett., 74 (1995) 3293.
[9] Beck C. and Provata A., EPL, 95 (2011) 58002.
[10] Provata A. and Beck C., Phys. Rev. E, 86 (2012) 046101.
[11] Nakajima T., Biology, 113 (2013) 67.
[12] Provata A. and Almirantis Y., Fractals, 8 (2000) 15.
[13] Gell-Mann M. and Tsallis C., Noneextensive Entropy: Interdisciplinary Applications (Oxford University Press, New York) 2004.
[14] Oikonomou Th. and Provata A., Eur. Phys. J. B, 50 (2006) 259.
[15] Oikonomou Th., Provata A. and Tirnakli U., Physica A, 387 (2008) 2653.
[16] Moreira D. A., Albuquerque E. L., Da Silva L. R. and Galvao D. S., Physica A, 387 (2008) 5477.
[17] Albuquerque E. L., Fulco U. L., Freire V. N., Caetano E. W. S., Lyra M. L. and Moura F. A. B. F., Phys. Rep., 535 (2014) 139.
[18] Machado J. A. T., Costa A. C. and Queihas M. D., Nonlinear Anal.: Real World Appl., 12 (2011) 3135.
[19] Kaniadakis G., Physica A, 296 (2001) 405.
[20] Kaniadakis G., Phys. Rev. E, 66 (2002) 056125.
[21] Kaniadakis G., Phys. Rev. E, 72 (2005) 036108.
[22] Provata A., Physica A, 381 (2007) 148.
[23] Kaniadakis G., Phys. Lett. A, 288 (2001) 283.
[24] Silva R., Phys. Lett. A, 352 (2006) 17.
[25] Silva R., Eur. Phys. J. B, 54 (2006) 499.
[26] Hristopulos D. T., Petarakis M. P. and Kaniadakis G., Phys. Rev. E, 89 (2014) 052142.
[27] Ourabah K. and Tribeche M., Phys. Rev. E, 89 (2014) 062130.
[28] Souza N. T. C. M., Anselmo D. H. A. L., Mello V. D. and Silva R., Phys. Lett. A, 378 (2014) 1691.
[29] Retrieved February 2nd, 2014, http://www.ncbi.nlm.nih.gov/projects/mapview/.
[30] Eddy S. R., Curr. Biol., 22 (2012) R898.
[31] Thomas C. A. Jr., Annu. Rev. Genet., 5 (1971) 237.
[32] Gregory T. R., Nat. Rev. Genet., 5 (1971) 237.
[33] Doolittle W. F., Proc. Natl. Acad. Sci. U.S.A., 110 (2013) 5294.
[34] Pennisi E., Science, 37 (2012) 1159.
[35] Rifaat R. and Kissner W., in Proceedings of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering, Vol. 2 (IEEE) 1999, p. 801.