Multifractal information production of the human genome

C. Beck and A. Provata

1 Queen Mary University of London, School of Mathematical Sciences - Mile End Road, London E1 4NS, UK, EU
2 Institute of Physical Chemistry, National Center for Scientific Research “Demokritos” - 15310 Athens, Greece, EU

received 12 November 2010; accepted in final form 13 July 2011; published online 10 August 2011

PACS 89.75.Fb – Structures and organisation in complex systems
PACS 05.45.Df – Fractals
PACS 87.15.A- – Theory, modeling, and computer simulation

Abstract – We determine the Rényi entropies $K_q$ of symbol sequences generated by human chromosomes. These exhibit non-trivial behaviour as a function of the scanning parameter $q$. In the thermodynamic formalism, there are phase-transition–like phenomena close to the $q = 1$ region. We develop a theoretical model for this based on the superposition of two multifractal sets, which can be associated with the different statistical properties of coding and non-coding DNA sequences. This model is in good agreement with the human chromosome data.

DNA symbol sequences exhibit a very complicated dynamical structure. There are long-range correlations [1–11] which are particularly strong for the non-coding sequences (DNA sequences which do not code for the production of proteins, such as introns, intragenic regions, repeats etc.), whereas the coding sequences demonstrate characteristics similar to random-like processes [1–4]. The way in which coding and non-coding sequences alternate in the DNA of many organisms is described by a multifractal [12–16]. Various approaches have been suggested to map DNA sequences onto the dynamics of an associated dynamical system, such as correlated random walks [2,3], or to provide a suitable measure representation by formally mapping DNA sequences onto points of the unit interval [13,17]. The associated measures, investigated in detail by Yu et al. for a large variety of organisms [12], exhibit a non-trivial spectrum of Rényi dimensions. Multifractal properties of DNA walk profiles and associated complex phenomena related to the presence of jumps in the strand asymmetry were also recently investigated by wavelet analysis in [18]. A variety of complexity measures can be defined to characterize the complex structure of DNA. For example, in [19,20] generalized Hurst exponents were determined for DNA walks. These contain information on all higher moments of the displacement process.

In this paper we directly apply the known symbolic dynamics techniques of the thermodynamic formalism of dynamical systems [21–23] to DNA symbol sequences.

For our data analysis we will concentrate mostly on the Homo sapiens chromosome 10, as a working example. In addition, we present results on Homo sapiens chromosomes 14 and 20, the animal Mus musculus chromosome 9, and the plant Arabidopsis thaliana chromosome II, purely for comparative reasons. For all DNA the symbol space contains 4 different symbols A, G, T, C denoting the four nucleotides (Adenine, Guanine, Thymine and Cytosine). The sequence of base pair symbols obtained by moving along the DNA string, in a dynamical systems setting, can be regarded as a shift of (correlated) symbols. We are interested in the average information production produced by this shift, and in the set of all higher-order correlations of the symbols. This can be measured by various quantities which weight the rare and frequent symbol sequences in a different way. In dynamical systems theory, for a system with a generating partition, one defines the dynamical Rényi entropies as

$$K_q = \lim_{N \to \infty} \frac{1}{N} \frac{1}{1-q} \ln \sum_{i_1, \ldots, i_N} p(i_1, i_2, \ldots, i_N)^q, \; q \neq 1,$$

$$K_1 = \lim_{N \to \infty} \frac{1}{N} \sum_{i_1, \ldots, i_N} p(i_1, i_2, \ldots, i_N) \ln p(i_1, i_2, \ldots, i_N).$$

(1)

Here $p(i_1, i_2, \ldots, i_N)$ denotes the probability of the symbol sequence $i_1, i_2, \ldots, i_N$. $N$ denotes the length of the sequence and $q$ is a parameter taking real values. The above sum is taken over all allowed symbol sequences $i_1, i_2, \ldots, i_N$, i.e. over all sequences with $p(i_1, \ldots, i_N) \neq 0$. $K_1$ is the Kolmogorov-Sinai entropy, a very important
invariant in dynamical system theory. $K_q$ is the topological entropy, which counts the growth rate of allowed symbol sequences for $N \to \infty$. A much more complete characterisation is via the set of all $K_q$ with $q \in (-\infty, \infty)$. These quite generally measure the information production of the dynamical system under consideration. From this set one can proceed to the spectrum of dynamical crowding indices by Legendre transformation (see, e.g., [21,24] for details).

For the standard Bernoulli shift of $J$ different symbols, the symbols are statistically independent and occur with equal probability $p = 1/J$. We thus obtain $p_i = J^{-N}$ and $K_q = \ln J$, independent of $q$. If there are non-trivial correlations, and non-uniform probabilities, as is the case for DNA sequences, then the spectrum of $K_q$ becomes non-trivial. As an example, fig. 1 depicts the multifractal $K_q$ spectrum obtained for the *Homo sapiens* chromosome 10 (solid black line) and various other chromosomes as described in the caption. All spectra were numerically evaluated by taking into account all symbol sequences up to length $N = 8$. This length is adequate for representing the asymptotic spectrum, which is already reached for values $N \geq 6$, as was also reported in refs. [12,15].

Our goal is to compare the information production of symbolic genomic sequences with those generated by simple examples of chaotic maps. A simple example of a dynamical system with a non-trivial $K_q$ spectrum is the asymmetric tent map (fig. 2(a)), given on the unit interval $[0,1]$ by

$$f(x) = \begin{cases} x/w, & \text{for } 0 \leq x \leq w, \\ 1-x, & \text{for } w \leq x \leq 1. \end{cases}$$

The generating partition for this map corresponds to the two intervals $I_1 = [0,w]$ and $I_2 = [w,1]$. We may write the symbol “1” if an iterate $x_n$ of $f$ is in $I_1$ and “2” if it is in $I_2$. The Rényi entropies for this simple model system are given by

$$K_q = \frac{1}{1-q} \ln(w^q + (1-w)^q), \quad q \neq 1,$$

$$K_1 = w \ln w + (1-w) \ln(1-w).$$

The above chaotic dynamical system generates symbol sequences consisting of just two different symbols. An obvious generalisation is to $J$ different symbols, where the corresponding piecewise linear map has $J/2$ maxima (fig. 2(b)). In this case the $K_q$ are given by

$$K_q = \frac{1}{1-q} \ln \left(\sum_{i=1}^J w_i^q \right), \quad q \neq 1,$$

$$K_1 = \sum_{i=1}^J w_i \ln w_i$$

with $w_1 + w_2 + \ldots + w_J = 1$. The parameters $w_j$, $j = 1,\ldots,J$, correspond to the 1-point probabilities of the occurrences of the symbols $j$.

For human chromosome 10, the observed values of 1-point symbol probabilities are $w_1 = w_A = 0.291921, w_2 = w_C = 0.207966, w_3 = w_G = 0.207859$ and $w_4 = w_T = 0.292219$ [15]. The entropies $K_q$ of the human genome can be fitted neither by the above simple model with $J = 2$, which in the multifractal language corresponds to a two-scale Cantor set with a multiplicative measure, nor using $J = 4$, which corresponds to a four-scale Cantor set, choosing the same 1-point probabilities as observed in human chromosome 10. This is shown in fig. 1: the $q$-dependence of the chromosomes data is much more pronounced than that of the corresponding asymmetric chaotic map that shifts 4 symbols. We thus need a more sophisticated approach to reproduce the observed multifractal information production of genomic sequences.

The idea developed in the sequel is to take into account the different dynamical properties of the coding and noncoding strings which constitute the chromosomes. The symbol sequence probabilities are, in general, different for
each of those regions, and are denoted by $p^c(i_1, \ldots, i_N)$ and $p^{nc}(i_1, \ldots, i_N)$, respectively. In the following, inspired by the multifractal formalism, we consider sequences of size $N$ as part of longer sequences and we write $N = -\log \epsilon$, where $\epsilon$ is the partition “box size”. The limit $N \to \infty$ corresponds to “box size” $\epsilon \to 0$, and the $K_q$ are then identical (up to a multiplicative factor) to the Rényi dimensions $D_q$ of a multifractal that encodes the dynamical properties.

When the dynamical partition function

$$Z(q) := \sum_{i_1, \ldots, i_N} p(i_1, \ldots, i_N)^q \sim \epsilon^{(q-1)K_q}$$

(5)

is evaluated, there are contributions from both types of strings. We thus have

$$Z(q) \approx N_c \sum p^c(i_1, \ldots, i_N)^q + N_{nc} \sum p^{nc}(i_1, \ldots, i_N)^q$$

$$\sim N_c \epsilon^{(q-1)K_q^c} + N_{nc} \epsilon^{(q-1)K_q^{nc}},$$

(6)

where the numbers $N_c$, $N_{nc}$ determine how many strings are in the coding and non-coding region, respectively. If $N_c, N_{nc}$ are independent of $\epsilon$, then the Rényi entropies of the entire system are given by the term that dominates the partition function for $\epsilon \to 0$, i.e.

$$K_q = \begin{cases} \min(K_q^c, K_q^{nc}), & \text{for } q > 1, \\ \max(K_q^c, K_q^{nc}), & \text{for } q < 1. \end{cases}$$

(7)

In the thermodynamic formalism of dynamical systems, this means that the free energy $(q - 1)K_q$ exhibits a phase transition (non-analytic behaviour) at the critical value $q_{\text{critical}} = 1$ (see also [23] for other systems exhibiting phase transitions in the Rényi entropies). Clearly such a behaviour can only be seen if one uses other entropy measures than the usual KS entropy (corresponding to $q = 1$) for the investigation of the information production of the human genome. This once again illustrates the importance to study the entire multifractal spectrum $K_q$.

The above simple phase transition model of $K_q$ agrees well with the genome data, see fig. 3. Figure 3(a) shows two approximations of the human chromosome data via two different multifractal sets. For the modeling, multifractal sets with $J = 4$ different symbols were taken into account, since the genome consists of 4 nucleotides. For simplicity only one effective scale $v$ was introduced into each of the two sets, leading to

$$K_q = \frac{1}{1-q} \ln(v^q + 3v^{q'})$$

(8)

$$K_1 = v \ln v + 3v' \ln v',$$

where $v' = (1 - v)/3$. The first one approximates well the human chromosome 10 data when $q \to \infty$ with $v = v_+ = 0.447$ but fails in the region $q \to -\infty$, see fig. 3(a) (red circles). The second multifractal set approximates the data in the opposite region, with $v = v_- = 0.126$, see figure 3(a) (blue squares). Here our notation $v_\pm$ indicates in which $q$-region the corresponding parameter $v$ is relevant.

In fig. 3(b) the red dashed line is a composite of the two multifractal sets, based on forming the maximum, respectively the minimum, according to eq. (7). This approximates the data well in the entire $q$-region. In fig. 3 the values of the limit entropies $K_{\pm \infty}$ were fitted to give the best coincidence with the data. Note that the region $q \to -\infty$ is dominated by very rare symbol sequences and the region $q \to +\infty$ by the most frequent ones. Also, it should be clear that finite-size effects demonstrated in the genomic data make a sharp phase transition unobservable since, as in our numerical analysis, only symbol sequences of finite size are investigated. Our hypothesis in the following is to associate the blue curve (squares) in fig. 3 with the non-coding sequences and the red curve (circles) with the coding ones.

In the thermodynamic formalism of dynamical systems, the role of the free energy is played by the function $\tau_q = (q - 1)K_q$ rather than $K_q$ itself [21]. It is therefore useful to analyze this function in somewhat more detail. $\tau_q$ is shown in fig. 4, with the solid black line representing
the human chromosome 10 and the red triangles originating from the composite model. Again we see evidence for the presence of a critical value $q_{\text{critical}}$ with phase-transition-like behaviour. An abrupt change of slope is clearly observable in the area $0 \leq q \leq 4$, though, of course, the precise value of the critical $q$-value cannot be located due to finite-size effects. Our composite model predicts that $\tau_q$ is a continuous but non-differentiable function of $q$ at $q_{\text{critical}} = 1$, which in the thermodynamic analogy corresponds to a 1st-order phase transition. The relevant transition area is designated in fig. 4 by two perpendicular dashed lines.

At this point let us mention that the usefulness of our phase transition model based on Rényi entropies is certainly not restricted to biological systems and that similar techniques as the ones described in this letter can be applied to other complex systems, in particular if the underlying dynamical structure resembles that of a multiplicative random cascade model with different scales [25–27].

So far our composite multifractal model shows a phase transition at $q = 1$, since by construction the two Cantor sets were joint at the $q = 1$ scale, see eq. (7). On the other hand, it is known that the numbers $N_c$ and $N_{nc}$ can depend on $\epsilon$ in a significant way. Long-range correlations are demonstrated in the non-coding, while short-range ones are displayed by the coding sequences [1–3,15]. The structure of (mostly) non-coding sequences, as interwoven with coding sequences, forms a (multi-)fractal as well. This means the above numbers $N_c$ and $N_{nc}$ scale with $\epsilon$ and thus the critical value $q_{\text{critical}}$ can shift to different values. This is clearly observed in the present data, both in the $K_q$ spectrum (see fig. 3) and in the $\tau_q$ one (see fig. 4). Figure 4 indicates that the critical point is slightly displaced to a value $q_{\text{critical}} \approx 2 > 1$. As we shall see below, this behaviour can be understood from the domination of the long-range correlated non-coding sequences, $N_{nc} \gg N_c$, which are known to cover approximately 97% of the human genome.

Mathematically, if we assume that the coding sequences scale as

$$N_c \sim \epsilon^{-d_c},$$

and the non-coding ones as

$$N_{nc} \sim \epsilon^{-d_{nc}},$$

then the critical point $q_{\text{critical}}$ is determined by the relative dominance of the two exponents in eq. (6), i.e. by the condition

$$(q_{\text{critical}} - 1)K_q^{(c)} - d_c = (q_{\text{critical}} - 1)K_q^{(nc)} - d_{nc},$$

which, depending on the numbers $d_c$ and $d_{nc}$, can shift the critical value away from 1. Solving for $q_{\text{critical}}$ we obtain

$$q_{\text{critical}} = 1 + \frac{d_{nc} - d_c}{K_q^{(nc)} - K_q^{(c)}}.$$

At $q \approx 2$ we see from fig. 3 that $K_q^{(nc)}$ (blue squares) is bigger than $K_q^{(c)}$ (red circles). Hence eq. (12) implies that $d_{nc} > d_c$. This, on the other hand, implies

$$N_{nc} \sim \epsilon^{-d_{nc}} \gg N_c \sim \epsilon^{-d_c},$$

consistent with the fact that (for higher eucaryotes) the size $N_{nc}$ of non-coding sequences dominates over the size $N_c$ of coding ones.

Although we have mainly presented our calculation for human chromosome 10, the other chromosomes show similar behaviour, as can be expected from the observed similarity of the multifractal spectra shown in fig. 1. The fitted $v_{\pm}$ values that enter into eq. (8) in this case are $v_+ = 0.448$, $v_- = 0.1295$ for human chromosome 14 and $v_+ = 0.448$, $v_- = 0.136$ for human chromosome 20. Similar values were obtained for the mouse chromosome 9, $v_+ = 0.435$, $v_- = 0.125$. The plant Arabidopsis thaliana is a test organism in biology whose genome is known to be relatively dense in coding, in comparison with other plants and higher eucaryotes in general. This fact is visible in its Rényi entropy spectrum in fig. 1, which is closer to that of a single asymmetric tent map shifting 4 symbols. Overall, the $q$-dependence of the $K_q$ spectrum of this plant is less pronounced, consistent with the lower non-coding percentage of this chromosome. Still our method can be applied in a similar way and the computed $v_{\pm}$ values for this case are $v_+ = 0.430$, $v_- = 0.155$.

To conclude, we have shown that the information production of genomic sequences, if regarded as a shift of the four symbols A, C, G, T, is very complex and can only be fully understood by considering the entire spectrum of Rényi entropies $K_q$. The multifractal structure can be approximated to a great extent by a superposition of two processes, one describing the system for $q > q_{\text{critical}}$ and one for $q < q_{\text{critical}}$, corresponding roughly to coding and non-coding DNA characteristics.
REFERENCES

[1] Li W. and Kaneko K., Europhys. Lett., 17 (1992) 655.
[2] Peng C. K., Buldyrev S. V., Goldberger A. L., Havlin S., Sciortino F., Simons M. I. and Stanley H. E., Nature, 356 (1992) 168.
[3] Provata A. and Almirantis Y., Physica A, 247 (1997) 482; Almirantis Y. and Provata A., J. Stat. Phys., 97 (1999) 233.
[4] Mantegna R. N., Buldyrev S. V., Goldberger A. L., Havlin S., Peng C. K., Simons M. and Stanley H. E., Phys. Rev. Lett., 73 (1994) 3169.
[5] Voss R. F., Phys. Rev. Lett., 68 (1992) 3805.
[6] Ebeling W. and Nicolis G., Chaos, Solitons Fractals, 2 (1992) 635.
[7] Arneodo A., d’Aubenton-Carafa Y., Audit B., Bachy E., Muzy J. F. and Therme S., Eur. Phys. J. B, 1 (1998) 259.
[8] Arneodo A., d’Aubenton-Carafa Y., Audit B., Bachy E., Muzy J. F. and Therme S., Physica A, 249 (1998) 439.
[9] Usatenko O. V., Yampol’skii V. A., Kechedzhy K. E. and Mel’nyk S. S., Phys. Rev. E, 68 (2003) 061107.
[10] Afreixo V., Ferreira P. J. S. G. and Santos D., Phys. Rev. E, 70 (2007) 031910.
[11] Carpena P., Bernaola-Galvan P., Coronado A. V., Hackenberg M. and Oliver J. L., Phys. Rev. E, 75 (2007) 032903.
[12] Yu Z.-G., Anh V. and Lau K.-S., Phys. Rev. E, 64 (2001) 031903.
[13] Gutierrez J. M., Rodriguez M. A. and Abramson G., Physica A, 300 (2001) 271.
[14] Su Z.-Y., Wu T. and Wang S.-Y., Chaos, Solitons Fractals, 40 (2009) 1750.
[15] Provata A. and Katsalouplis P., Phys. Rev. E, 81 (2010) 026102.
[16] Zhou L.-Q., Yu Z.-G., Deng J.-Q., Anh V. and Long S.-C., J. Theor. Biol., 232 (2005) 559.
[17] Hao B.-L., Physica A, 282 (2000) 225; Hao B.-L., Lee H. C. and Zhang S.-Y., Chaos, Solitons Fractals, 11 (2000) 825.
[18] Nicolay S., Brodie de Brodie E. B., Touchon M., Audit B., d’Aubenton-Carafa Y., Therme S. and Arneodo A., Phys. Rev. E, 75 (2007) 032902.
[19] Kantelhardt J. W., Zschiegner S. A., Koscielny-Bunde E., Bunde A., Havlin S. and Stanley H. E., Physica A, 316 (2002) 87.
[20] Beck C. and Schlogl F., Thermodynamics of Chaotic Systems: An Introduction (Cambridge University Press, Cambridge) 1995.
[21] Tel T., Z. Naturforsch., 43a (1988) 1154.
[22] Csordás P. and Szépfalusy P., Phys. Rev. A, 39 (1989) 4767.
[23] Takayasu H., Fractals in the Physical Sciences (Manchester University Press, Manchester) 1990.
[24] Mandelbrot B. B., Fisher A. and Calvet L., Cowles Found. Discuss. Pap., No. 1164 (1997).
[25] Arneodo A., Bacry E. and Muzy J. F., J. Math. Phys., 39 (1998) 4142.
[26] Boggache M. I., Eichner J. F. and Bunde A., Phys. Rev. Lett., 99 (2007) 240601.

58002-p5