epl draft

Multifractal information production of the human genome

Christian Beck[1] and Astero Provata[1,2]

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

Abstract. - We determine the Renyi entropies K_q of symbol sequences generated by human chromosomes. These exhibit nontrivial 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) 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 [12–15]. 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 [13,16]. The associated measures, investigated in detail by Yu et al. for a large variety of organisms [12], exhibit a non-trivial spectrum of Renyi dimensions.

In this paper we directly apply the known symbolic dynamics techniques of the thermodynamic formalism of dynamical systems [17–19] to DNA symbol sequences. For our data analysis we will concentrate mostly on the human genome (chromosome 10) as a working example. For DNA the symbol space contains 4 different symbols A,G,T,C denoting the four nucleotides (Adenine, Guanine, Thymine and Cytosine). Translations along the DNA string 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 Renyi entropies as

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

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

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_0 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. [17, 20] 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 $p(i_1, \ldots, i_N) = p^N = J^{-N}$ 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 nontrivial. As an example, in
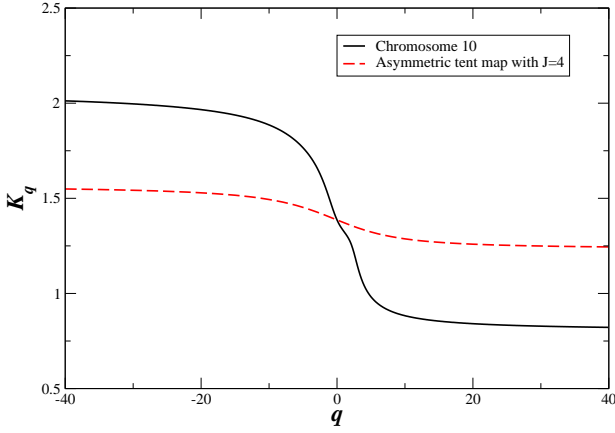
Christian Beck1 and Astero Provata1,2

Fig. 1: (Colour online) The spectrum of Renyi entropies K_q for human chromosome 10 (black solid line). The dashed line shows the corresponding spectrum of an asymmetric tent map with 4 symbols and the same 1-point probabilities as chromosome 10.

fig.1 the solid black line shows the multifractal K_q spectrum obtained for the human chromosome 10. The spectrum was 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 references [12, 15].

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

$$f(x) = \begin{cases} \frac{x}{w} & \text{for } 0 \leq x \leq w \\ \frac{1-x}{1-w} & \text{for } w \leq x \leq 1 \end{cases} \qquad (2)$$

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

$$K_q = \frac{1}{1-q} \ln(w^q + (1-w)^q), \ q \neq 1 \qquad (3)$$
$$K_1 = w \ln w + (1-w) \ln(1-w)$$

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. 2b). In this case the K_q are given by

$$K_q = \frac{1}{1-q} \ln(w_1^q + w_2^q + \ldots + w_J^q), \ q \neq 1 \qquad (4)$$
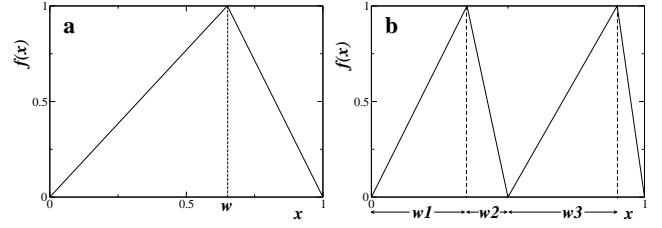$$K_1 = \sum_{i=1}^{J} w_i \ln w_i$$

Fig. 2: (Colour online) Example of an asymmetric tent map with (a) 1 maximum (shift of 2 symbols) and (b) 2 maxima (shift of 4 symbols).

with $w_1 + w_2 + \ldots + w_J = 1$. The parameters w_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 neither be fitted 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 4-scale Cantor set, choosing the same 1-point probabilities as observed. 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 the human genome.

The idea developed in the sequel is to take into account the different dynamical properties of the coding and non-coding 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 ϵ is the partition 'box size'. The limit $N \rightarrow \infty$ corresponds to 'box size' $\epsilon \rightarrow 0$, and the K_q are then identical (up to a multiplicative factor) to the D_q of a multifractal that encodes the dynamical properties.

When the dynamical partition function

$$Z(q) := \sum_{i_1, \ldots, i_N} p(i_1, \ldots, i_N)^q \sim \epsilon^{(q-1)K_q} \qquad (5)$$

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

$$\begin{aligned} Z(q) &\approx N_c \sum p^{(c)}(i_1, \ldots, i_N)^q + N_{nc} \sum p^{(nc)}(i_1, \ldots i_N)^q \\ &\sim N_c \epsilon^{(q-1)K_q^{(c)}} + N_{nc} \epsilon^{(q-1)K_q^{(nc)}}, \end{aligned} \qquad (6)$$

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 ϵ, then the Renyi entropies of

the entire system are given by the term that dominates the partition function for $\epsilon \to 0$, i.e

$$K_q = \begin{cases} \min(K_q^{(c)}, K_q^{(nc)}) & \text{for } q > 1 \\ \max(K_q^{(c)}, K_q^{(nc)}) & \text{for } q < 1. \end{cases} \quad (7)$$

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 $q_{critical} = 1$ (see also [19] for other systems exhibiting phase transitions in the Renyi 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 3a shows two approximations of the human chromosome data via two different multifractal sets. For the modelling multifractal sets with $J = 4$ different symbols were taken into account, since the genome consists of 4 nucleotides. For simplicity only one effective scale w_1 was introduced into each of the two sets, leading to

$$K_q = \frac{1}{1-q} \ln(w_1^q + 3w_2^q), \; q \neq 1 \quad (8)$$
$$K_1 = w_1 \ln w_1 + 3w_2 \ln w_2$$

where $w_2 = (1 - w_1)/3$. The first one approximates well the chromosome 10 data when $q \to \infty$ with $w_1 = 0.447$ but fails in the region $q \to -\infty$, see fig. 3a (red circles). The second multifractal set approximates the data in the opposite region, with $w_1 = 0.126$, see fig. 3a (blue squares). In fig 3b 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 [17]. It is therefore useful to analyze this function in somewhat more detail. τ_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 $q_{critical}$ with phase-transition-
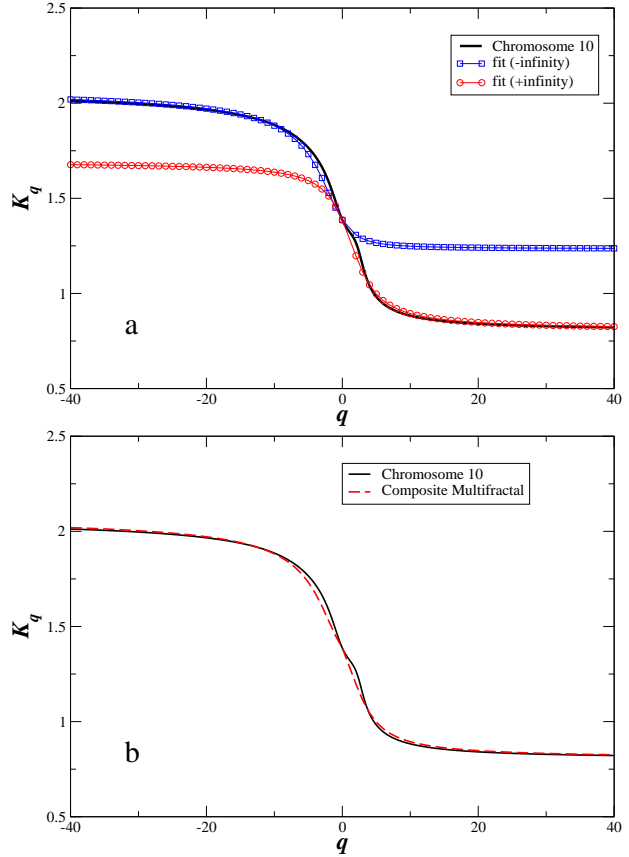

Fig. 3: (Colour online) a) Separate approximations to the K_q spectrum for $q \to -\infty$ (blue squares) and $q \to +\infty$ (red circles). b) Composite multifractal spectrum (red triangles) and multifractal spectrum of Chromosome 10, organism Homo Sapiens (solid black line).

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 model predicts that τ_q is a continuous but non-differentiable function of q at $q_{critical} = 1$, which in the thermodynamic analogy corresponds to a 1st-order phase transition. The relevant transition area is designated in fig. 4 by two perpedicular dashed lines.

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 ϵ in a significant way. Long range correlations are demonstrated in the noncoding, while short range ones are displayed by the coding sequences [1–3,15]. The structure of (mostly) noncoding sequences, as intervowen with coding sequences, forms a (multi-)fractal as well. This means the above numbers N_c and N_{nc} scale with ϵ and thus the critical value $q_{critical}$ 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 τ_q one (see fig. 4). Fig. 4 indicates that the critical point is slightly displaced
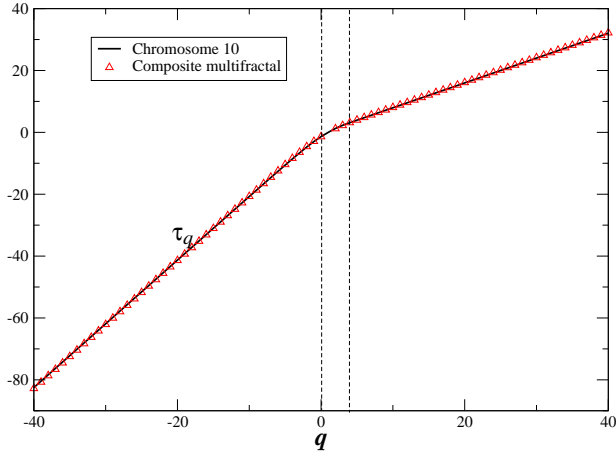
Fig. 4: (Colour online) The τ_q of Chromosome 10, organism Homo Sapiens (solid black line) and of the composite multi-fractal spectrum (red triangles)

to a value $q_{critical} \approx 2 > 1$. As we shall see below, this behaviour can be understood from the domination of the long range correlated noncoding sequences, $N_{nc} >> N_c$, which are known to cover approximately 97% of the human genome.

Mathematically, if we assume that the coding sequences scale as

$$N_c \sim \epsilon^{-d_c} \qquad (9)$$

and the non-coding ones as

$$N_{nc} \sim \epsilon^{-d_{nc}}, \qquad (10)$$

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

$$(q_{critical} - 1)K_q^{(c)} - d_c = (q_{critical} - 1)K_q^{(nc)} - d_{nc}, \quad (11)$$

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

$$q_{critical} = 1 + \frac{d_{nc} - d_c}{K_q^{(nc)} - K_q^{(c)}}. \qquad (12)$$

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

$$N_{nc} \sim \epsilon^{-d_{nc}} >> N_c \sim \epsilon^{-d_c}, \qquad (13)$$

consistent with the fact that the number N_{nc} of non-coding sequences dominates over the number N_c of coding ones.

To conclude, we have shown that the information production of the human genome, 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 Renyi entropies K_q. The multifractal structure can be approximated to a great extent by a superposition of two processes, one describing the system for $q > q_{critical}$ and one for $q < q_{critical}$, corresponding roughly to coding and non-coding DNA characteristics.

REFERENCES

[1] W. Li and K. Kaneko, *Europhys. Lett.*, **17** (1992) 655.
[2] C. K. Peng, S. V. Buyldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. I. Simons and H. E. Stanley, *Nature*, **356** (1992) 168.
[3] A. Provata and Y. Almirantis,, *Physica A*, **247** (1997) 482; Y. Almirantis and A. Provata, *J. Stat. Phys.*, **97** (1999) 233.
[4] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons and H. E. Stanley, *Phys. Rev. Lett.*, **73** (1994) 3169.
[5] R. F. Voss, *Phys. Rev. Lett.*, **68** (1992) 3805.
[6] W. Ebeling and G. Nicolis, *Chaos, Solitons & Fractals*, **2** (1992) 635.
[7] A. Arneodo, Y. d'Aubenton-Carafa, B. Audit, E. Bacry, J. F. Muzy and C. Thermes, *European Physical Journal B*, **1** (1998) 259-263.
[8] A. Arneodo, Y. d'Aubenton-Carafa, B. Audit, E. Bacry, J. F. Muzy and C. Thermes , *Physica A*, **249** (1998) 439-448.
[9] O. V. Usatenko, V. A. Yampol'skii, K. E. Kechedzhy and S. S. Mel'nyk, *Phys. Rev. E*, **68** (2003) 061107.
[10] V. Afreixo, P; J. S. G. Ferreira and D. Santos, *Phys. Rev. E*, **70** (2007) 031910.
[11] P. Carpena, P. Bernaola-Galvan, A. V. Coronado, M. Hackenberg and J.L Oliver, *Phys. Rev. E*, **75** (2007) 032903.
[12] Z.-G. Yu, V. Anh and K.-S. Lau, *Phys. Rev. E*, **64** (2001) 031903.
[13] J.M. Gutierrez, M.A. Rodriguez and G. Abramson, *Physica A*, **300** (2001) 271-284.
[14] Z.-Y. Su, T. Wu and S.-Y. Wang,, *Chaos, Solitons & Fractals*, **40** (2009) 1750-1765.
[15] A. Provata and P. Katsaloulis,, *Phys. Rev. E*, **81** (2010) 026102.
[16] B.-L. Hao, *Physica A*, **282** (2000) 225-246; B.-L. Hao, H.C. Lee and S.-Y. Zhang, *Chaos, Solitons & Fractals*, **11** (2000) 825-836.
[17] C. Beck and F. Schlogl, *Thermodynamics of Chaotic Systems: An Introduction* (Cambridge University Press, Cambridge) 1995.
[18] T. Tél, *Z. Naturforsch.*, **43a** (1988) 1154.
[19] P. Csordás and P. Szépfalusy, *Phys. Rev. A*, **39** (1989) 4767.
[20] H. Takayasu, *Fractals in the Physical Sciences* (Manchester University Press, Manchester) 1990.